ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Πρόβλεψη της κρίσιμης θερμοκρασίας υπεραγωγών με χρήση τεχνικών μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μιχαήλ Διαμαντόπουλος

Επιβλέπων: Κωνσταντίνος Κουσουρής, Επίκουρος
Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2019

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF APPLIED MATHEMATICAL AND PHYSICAL SCIENCES

Predicting the critical temperature of a superconductor by using machine learning techniques

DIPLOMA THESIS

Michael Diamantopoulos

Professor: Konstantinos Kousouris

Athens, September 2019

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

Πρόβλεψη της κρίσιμης θερμοκρασίας υπεραγωγών με χρήση τεχνικών μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μιχαήλ Διαμαντόπουλος

Επιβλέπων: Κωνσταντίνος Κουσουρής, Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Σεπτεμβρίου 2019.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
……	……	……
Κωνσταντίνος Κουσουρής Επίκουρος Καθηγητής Ε.Μ.Π.	Γεώργιος Τσιπολίτης Καθηγητής Ε.Μ.Π.	Αικατερίνη Τζαμαριουδάκη Ερευνήτρια Α ΕΚΕΦΕ «Δημόκριτος»

Αθήνα, Σεπτέμβριος 2019

(Υπογραφή)

……..

Μιχαήλ Διαμαντόπουλος

Διπλωματούχος Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών Ε.Μ.Π.

Περίληψη

Σκοπός της συγκεκριμένης πτυχιακής εργασίας είναι η κατασκευή ενός μοντέλου που θα είναι σε θέση να προβλέπει με όσο το δυνατόν μεγαλύτερη ακρίβεια, την κρίσιμη θερμοκρασία (T_c) ενός υλικού, δηλαδή τη θερμοκρασία κάτω από την οποία το υλικό αποκτά ιδιότητες υπεραγωγού, βασιζόμενο στη χημική του σύσταση. Για την επίτευξη του σκοπού αυτού, χρειάστηκε να δοκιμαστούν διάφορες τεχνικές και μοντέλα μηχανικής μάθησης (ελάχιστα τετράγωνα, δένδρα απόφασης, μέθοδοι ενδυνάμωσης, νευρωνικά δίκτυα κ.α.) και να αξιολογηθούν με βάση την απόδοση τους ώστε το τελικό μοντέλο παλινδρόμησης, να έχει όσο το δυνατόν μεγαλύτερη ακρίβεια στις προβλέψεις του και να είναι όσο το δυνατόν πιο γενικευμένο.

Στα πλαίσια της πτυχιακής εργασίας αυτής αναλύονται οι βασικότερες και πιο διαδεδομένες τεχνικές μηχανικής μάθησης, που εφαρμόζονται σε προβλήματα παλινδρόμησης. Έτσι μπορεί να αποτελέσει έναν οδηγό για κάποιον που καλείται να αντιμετωπίσει ένα πρόβλημα παλινδρόμησης, ώστε να έχει μια καλή εικόνα του εύρους των τεχνικών που μπορεί να χρησιμοποιήσει αλλά και της βασικής θεωρίας που κρύβεται πίσω από αυτές.

Λέξεις Κλειδιά: Παλινδρόμηση, Μηχανική Μάθηση, Πρόβλεψη, Μέθοδος Ελαχίστων Τετραγώνων, Δένδρα Απόφασης, Τεχνητά Νευρωνικά Δίκτυα, Κ-Κοντινότεροι Γείτονες, Υπεραγωγοί

Abstract

The purpose of this diploma thesis is to develop a model that will be able to predict as accurately as possible the critical temperature (T_c) of a superconductor, the temperature below which the material acquires superconducting properties, based on its chemical properties. To achieve this goal, various techniques and machine learning models (least squares, decision trees, boosting methods, and neural networks) were tested and evaluated on their performance, so that the final regression model is as accurate and generalized as possible.

In this thesis, we also, analyze the most basic and most widely used machine learning techniques applied to regression problems. So it can be a guide for someone who has to deal with a regression problem in order to get a good idea of the range of techniques he can use and the basic theory behind them.

Keywords: Regression, Machine Learning, Prediction, Least Squares, Decision Trees, Artificial Neural Networks, Boosting, KNN, Superconductors

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του επίκουρου καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, κ. Κωνσταντίνου Κουσουρή, τον οποίο θα ήθελα να ευχαριστήσω θερμά τόσο για την ανάθεση της συγκεκριμένης εργασίας όσο και για όλα όσα μου έμαθε. Παράλληλα, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στους γονείς μου που ήταν πάντα δίπλα μου και με υποστήριζαν σε όλη τη διάρκεια των σπουδών μου.

Περιεχόμενα

ΜΕΡΟΣ Ι – Εισαγωγή στη Μηχανική Μάθηση

Τη δεκαετία του 1940, δημιουργήθηκε το πρώτο χειροκίνητο σύστημα πληροφορικής και έγινε γνωστό ως ENIAC (Electronic Numerical Integrator and Computer). Μέχρι τότε, η λέξη υπολογιστής, χρησιμοποιούταν ως χαρακτηρισμός για τους ανθρώπους που είχαν πολύ ανεπτυγμένες αριθμητικές υπολογιστικές ικανότητες. Για το λόγο αυτό ο ENIAC ονομάστηκε αριθμητικό υπολογιστικό μηχάνημα έχοντας απώτερο σκοπό, να μιμείται την ανθρώπινη σκέψη και να μαθαίνει όπως ο άνθρωπος.

EIMC — Electronic Numerical Integrator and Computer | Image: www.computerhistory.org

Μια δεκαετία αργότερα το 1950, παρουσιάζεται το πρώτο ηλεκτρονικό παιχνίδι, που ισχυριζόταν, ότι μπορεί να νικήσει τους παγκόσμιους πρωταθλητές στο σκάκι. Το πρόγραμμα αυτό αποτέλεσε σημείο αναφοράς της εποχής του και βοήθησε τους παίχτες σκάκι να βελτιώσουν σε μεγάλο βαθμό τις δεξιότητες τους. Την ίδια εποχή, ο Frank Rosenblatt, εφευρίσκει τον Perceptron, που ήταν ένας αδύναμος ταξινομητής όταν χρησιμοποιούταν μόνος του, όταν όμως συνδυαζόταν και με άλλους ίδιους ταξινομητές σε ένα δίκτυο, είχε απίστευτες δυνατότητες, οι οποίες τον κατέστησαν επαναστατική ανακάλυψη για την εποχή του. Ο Perceptron, ήταν στην πραγματικότητα το πρώτο νευρωνικό δίκτυο που εφευρέθηκε.

Τα χρόνια που ακολούθησαν υπήρξε μια στασιμότητα στην έρευνα στον τομέα της μηχανικής μάθησης και των νευρωνικών δικτύων κυρίως λόγω της αδυναμίας των υπολογιστών της εποχής να πραγματοποιήσουν δύσκολες υπολογιστικά διαδικασίες ώστε να λύσουν συγκεκριμένα προβλήματα και της έλλειψης δεδομένων. Τα πράγματα όμως αλλάζουν τη δεκαετία του 1990, όταν η επιστήμη των υπολογιστών

συναντά τη στατιστική, δίνοντας έτσι χώρο για πιθανολογικές προσεγγίσεις της τεχνητής νοημοσύνης (artificial intelligence). Παράλληλα η ραγδαία εξέλιξη της τεχνολογίας και των υπολογιστών σε συνδυασμό με τον μεγάλο όγκο δεδομένων που παράγονται καθημερινά, δίνουν τη δυνατότητα στους επιστήμονες της εποχής να κατασκευάσουν έξυπνα συστήματα που έχουν την ικανότητα να αναλύουν και να μαθαίνουν μέσα από μεγάλα σύνολα δεδομένων. Αξιοσημείωτο γεγονός για την εποχή, το «Deep Blue», σύστημα της IBM, να νικάει τον παγκόσμιο πρωταθλητή στο σκάκι, Garry Kasparov.

Garry Kasparov faced off against Deep Blue, IBM's chess-playing computer, in 1997. Deep Blue was able to imagine an average of 200,000,000 positions per second. Kasparov ended up losing the match. Credit: Peter Morgan/Reuters

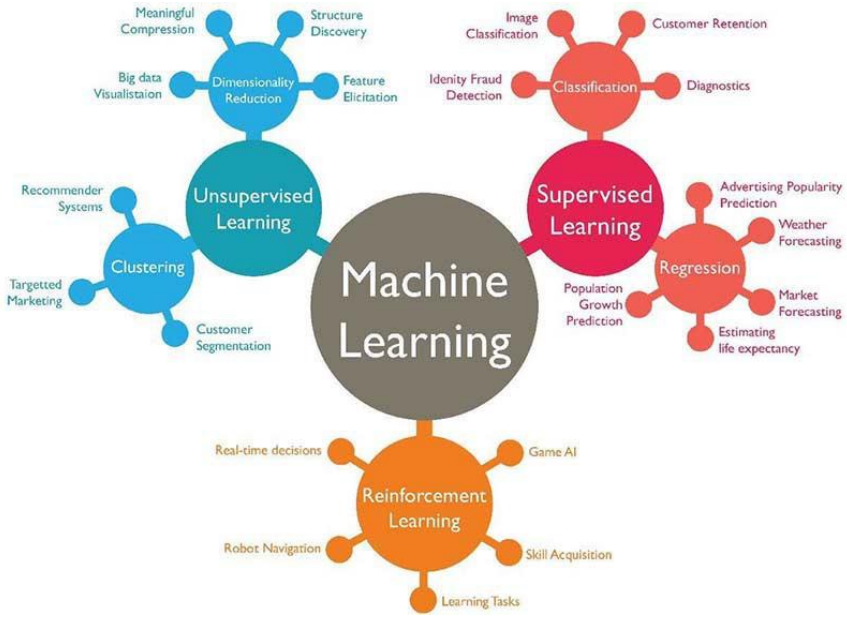
Ωστόσο εδώ προκύπτει το ερώτημα, πως ορίζει κάποιος τον όρο μηχανική μάθηση;! Σύμφωνα με τον ορισμό του Αμερικάνου πρωτοπόρου στον τομέα των ηλεκτρονικών παιχνιδιών και της τεχνητής νοημοσύνης, Arthur Samuel, οι αλγόριθμοι μηχανικής μάθησης, παρέχουν τη δυνατότητα στους υπολογιστές, να μαθαίνουν από δεδομένα, και να βελτιώνουν την απόδοση και τη ακρίβεια τους, χωρίς αυτοί να έχουν αρχικά προγραμματιστεί λεπτομερώς. Ο σκοπός της μηχανικής μάθησης, είναι να χτίζει αλγορίθμους, που παίρνουν δεδομένα (είσοδος) και χρησιμοποιούν στατιστική ανάλυση για να προβλέψουν την τιμή συγκεκριμένων μεταβλητών (έξοδος), ενώ παράλληλα μπορούν να προσαρμόζονται σε καινούρια δεδομένα ανανεώνοντας τα αποτελέσματα τους.

Είναι γεγονός, ότι τα τελευταία χρόνια, λέξεις όπως μηχανική μάθηση, νευρωνικά δίκτυα, τεχνητή νοημοσύνη και άλλες γίνονται καθημερινά εξώφυλλα των μεγαλύτερων επιστημονικών και όχι μόνο περιοδικών ενώ αποτελούν και έναν από τους μεγαλύτερους τομείς, στον οποίο οι επιστήμονες έχουν επικεντρώσει τις ερευνητικές τους προσπάθειες. Η τεχνητή νοημοσύνη γίνεται όλο και περισσότερο κομμάτι της καθημερινής ζωής των ανθρώπων, αφού μέσα από τις εφαρμογές της

βελτιώνει τη ζωή τους και τους δίνει δυνατότητες που μέχρι πριν λίγα χρόνια συναντούσε κάποιος μόνο σε ταινίες επιστημονικής φαντασίας. Ένα «έξυπνο ρολόι» που μετράει τους παλμούς και την πίεση και προτείνει στον ιδιοκτήτη του τρόπους για να βελτιώσει την υγεία του, μια προσωπική εικονική βοηθός που απαντάει και πραγματοποιεί κάθε επιθυμία του κατόχου της από το να ανοίξει τα φώτα στο σπίτι μέχρι και διαβάσει την εφημερίδα αλλά και ένα αυτοκινούμενο όχημα που σε πηγαίνει στον προορισμό σου χωρίς εσύ να αγγίξεις το τιμόνι, είναι μόνο μερικές από τις εκφάνσεις της μηχανικής μάθησης. Αξιοσημείωτες ακόμα εφαρμογές της τεχνητής νοημοσύνης, είναι και οι παρακάτω:

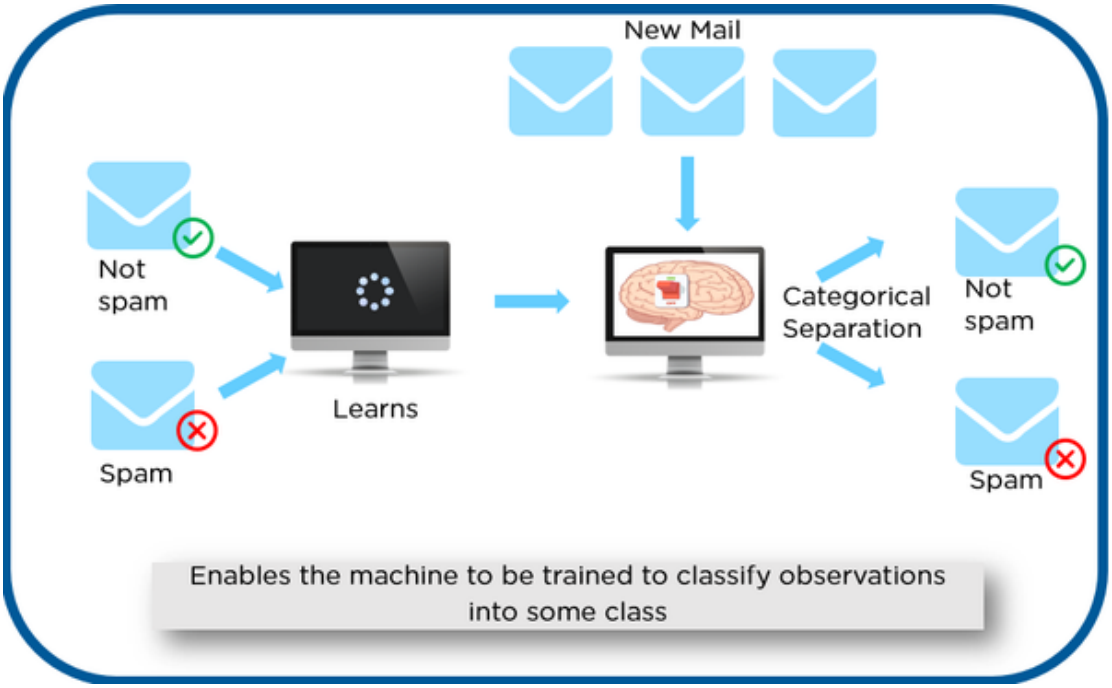
- Πρόβλεψη: Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί και να δώσει πολύ καλά αποτελέσματα σε προβλήματα πρόβλεψης, όπως το να προβλέψει αν ένας πελάτης μιας τράπεζας είναι πιθανό να φύγει ή όχι, αν ένα εξάρτημα είναι ελαττωματικό ή όχι, ποια θα είναι η τιμή μιας μετοχής μετά από το πέρας ενός χρονικού διαστήματος κ.ο.κ. .
- Αναγνώριση εικόνας: Η μηχανική μάθηση μπορεί επίσης να χρησιμοποιηθεί στην αναγνώριση προσώπων και εικόνας, και εφαρμόζεται με μεγάλη αποτελεσματικότητα στα αυτόνομα οχήματα όπου η μηχανική μάθηση είναι υπεύθυνη ώστε να διακρίνει τη διαφορά μεταξύ ενός φαναριού και ενός ανθρώπου.
- Αναγνώριση φωνής: Μια πολύ χρήσιμη εφαρμογή της μηχανικής μάθησης, η αναγνώριση φωνής, που συναντάμε σε εφαρμογές που μετατρέπουν τις προφορικές λέξεις σε κείμενο αλλά και σε εικονικές βοηθούς.
- Ιατρική διάγνωση: Μοντέλα μηχανικής μάθησης μπορούν να εκπαιδευτούν ώστε να αναγνωρίζουν καρκινικά κύτταρα και να προβλέπουν την εμφάνιση ασθενειών.
- Οικονομία και συναλλαγές: Η μηχανική μάθηση μπορεί να χρησιμοποιηθεί, για αναγνώριση κινδύνου σε πιθανές επενδύσεις, αναγνώριση απάτης αλλά και πιστωτικούς ελέγχους.

Η μηχανική μάθηση περιλαμβάνει τρία είδη αλγορίθμων, την επιτηρούμενη μάθηση (supervised learning), τη μη- επιτηρούμενη μάθηση (unsupervised learning) και την ενισχυτική μάθηση (reinforcement learning), τα οποία θα αναλυθούν παρακάτω.



Τα είδη μηχανικής μάθησης και οι εφαρμογές τους. Πηγή: towards science

Επιτηρούμενη μάθηση: Στην επιτηρούμενη μάθηση, ένα σύστημα τεχνητής νοημοσύνης, διαχειρίζεται δεδομένα τα οποία έχουν ετικέτες (labeled data). Συγκεκριμένα κάθε παράδειγμα-εγγραφή (instance), των δεδομένων χαρακτηρίζεται από μια ταμπέλα, η οποία ανταποκρίνεται σε μια πληροφορία για τη συγκεκριμένη εγγραφή. Ο σκοπός λοιπόν ενός μοντέλου τεχνητής νοημοσύνης, επιτηρούμενης μάθησης είναι να κατασκευάσει μια συνάρτηση-μοτίβο, με βάση το οποίο όταν θα δέχεται ένα σύνολο δεδομένων X, θα προβλέπει την ετικέτα του Y. Το μοντέλο πετυχαίνει το σκοπό αυτό αφού αρχικά εκπαιδευτεί σε ένα υποσύνολο των δεδομένων (training set). Στη συνέχεια η απόδοση-ακρίβεια του μοντέλου αξιολογείται με βάση το πόσο καλά πρόβλεψε τις τιμές του υπόλοιπου υποσυνόλου των δεδομένων (validation-test set), με βάση την εκπαίδευση του. Η επιτηρούμενη μάθηση έχει πολλές εφαρμογές στην καθημερινή ζωή. Στην παρακάτω εικόνα βλέπουμε τη διαδικασία κατά την οποία ένα μοντέλο επιτηρούμενης μηχανικής μάθησης εκπαιδεύεται ώστε να αναγνωρίζει αν ένα μήνυμα ηλεκτρονικής αλληλογραφίας είναι ανεπιθύμητο (spam) ή όχι.

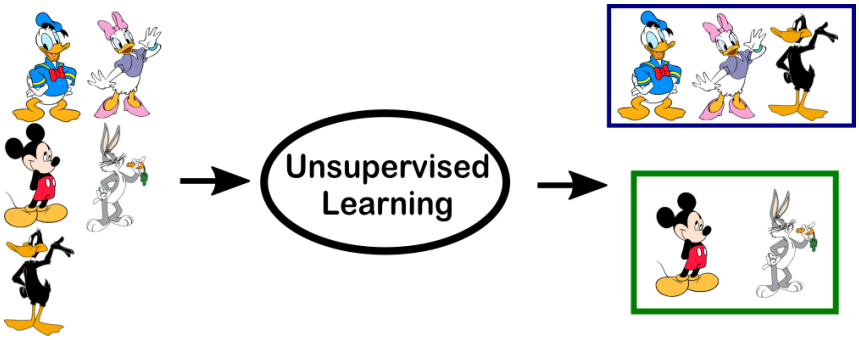


Παράδειγμα μοντέλου επιβλεπόμενης μηχανικής μάθησης. Πηγή: towards science

Να σημειωθεί ότι υπάρχουν δυο τύποι επιτηρούμενης μηχανικής μάθησης και είναι οι παρακάτω:

- Ταξινόμηση (Classification): Ένα πρόβλημα επιτηρούμενης μηχανικής μάθησης καλείται πρόβλημα ταξινόμησης, όταν η τιμή-ετικέτα που προβλέπει αντιπροσωπεύει κατηγορίες, όπως για παράδειγμα στο παραπάνω παράδειγμα: «ανεπιθύμητη αλληλογραφία», «μη ανεπιθύμητη αλληλογραφία» ή «άρρωστος», «υγειής».
- Παλινδρόμηση (Regression): Ένα πρόβλημα επιτηρούμενης μηχανικής μάθησης καλείται πρόβλημα παλινδρόμησης, όταν η τιμή-ετικέτα που προβλέπει παίρνει συνεχείς τιμές, όπως για παράδειγμα είναι το εισόδημα, η τιμή μιας μετοχής κ.λ.π. .

Μη επιτηρούμενη μάθηση: Στη μη επιτηρούμενη μάθηση ένα σύστημα τεχνητής νοημοσύνης, διαχειρίζεται δεδομένα τα οποία δεν έχουν ετικέτες (unlabeled data) και δεν είναι ταξινομημένα. Μέσα από διάφορους αλγορίθμους και χωρίς να προηγηθεί εκπαίδευση το σύστημα, βρίσκει τα κοινά χαρακτηριστικά μεταξύ των δεδομένων και τα διακρίνει σε ομάδες. Στην παρακάτω εικόνα βλέπουμε τη διαδικασία κατά την οποία ένα μοντέλο μη επιτηρούμενης μηχανικής μάθησης διαχωρίζει τις πάπιες από τα υπόλοιπα ζώα χωρίς αυτό να γνωρίζει ότι είναι πάπιες, αλλά πρακτικά εντοπίζοντας κοινά χαρακτηριστικά σε αυτές που διαφέρουν από τα χαρακτηριστικά των άλλων ζώων.

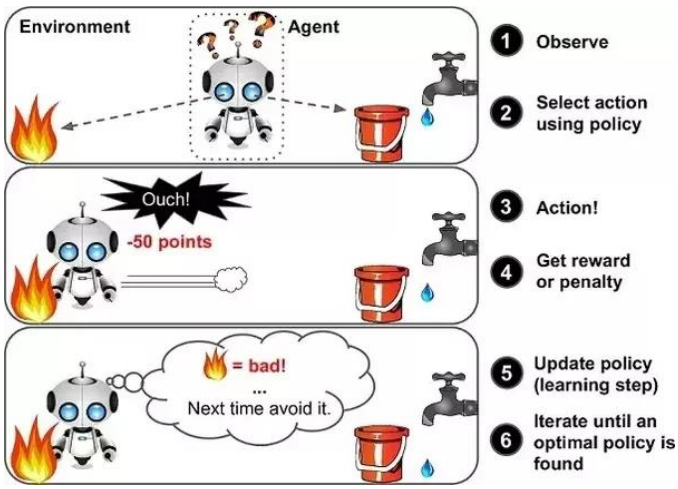


Παράδειγμα μοντέλου μη επιβλεπόμενης μηχανικής μάθησης. Πηγή: towards science

Να σημειωθεί ότι υπάρχουν δυο τύποι μη επιτηρούμενης μηχανικής μάθησης και είναι οι παρακάτω:

- Συσταδοποίηση (Clustering): Ένα πρόβλημα συσταδοποίησης, συνίσταται στην εύρεση (εγγενών) συστάδων με κοινά χαρακτηριστικά, όπως για παράδειγμα, η συσταδοποίηση των πελατών με βάση την καταναλωτική τους συμπεριφορά.
- Association: Ένα πρόβλημα association, συνίσταται, στην εύρεση σχέσεων-κανόνων που περιγράφουν μεγάλο μέρος των δεδομένων, όπως το ότι οι καταναλωτές που αγοράζουν το Χ προϊόν τείνουν να αγοράζουν και το Υ.

Ενισχυτική μάθηση (Reinforcement Learning): Ένας αλγόριθμος ενισχυτικής μάθησης, μαθαίνει αλληλεπιδρώντας με το περιβάλλον του και επιβραβεύεται όταν κάνει σωστές επιλογές ενώ υποβάλλεται σε ποινές όταν παίρνει λάθος αποφάσεις. Ο αλγόριθμος, μαθαίνει στην ουσία χωρίς την παρέμβαση του ανθρώπου, προσπαθώντας να ελαχιστοποιήσει τις ποινές και να μεγιστοποιήσει την επιβράβευση, όπως ένα πρόγραμμα δυναμικού προγραμματισμού. Στην παρακάτω εικόνα βλέπουμε μια αναπαράσταση ενός τέτοιου αλγορίθμου σε απλή μορφή.



Παράδειγμα μοντέλου ενισχυτικής μηχανικής μάθησης. Πηγή: towards science

Είναι φανερό, ότι ο αλγόριθμος επιλέγει μεταξύ δυο επιλογών, ενός κουβά νερού και μιας φωτιάς. Όταν επιλέγει τη φωτιά τότε επιβαρύνεται με ποινή ενώ όταν επιλέγει το νερό επιβραβεύεται, έτσι σταδιακά μαθαίνει να αποφεύγει τη φωτιά. Η ενισχυτική μάθηση έχει πολλές εφαρμογές στη βιομηχανία (ρομποτική) αλλά και στον τομέα της οικονομίας (αξιολόγηση στρατηγικών συναλλαγών).

ΜΕΡΟΣ ΙΙ – Τεχνικές Κατασκευής και Αξιολόγησης Μοντέλων Παλινδρόμησης με Χρήση Μηχανικής Μάθησης

1. Γραμμική παλινδρόμηση

1.1. Εισαγωγή στη Γραμμική Παλινδρόμηση με τη Μέθοδο Ελαχίστων Τετραγώνων

Ας υποθέσουμε ότι γνωρίζουμε τις τιμές κάποιων επεξηγηματικών (predictors) μεταβλητών $X_1, X_2, \dots, X_k$, τις οποίες θα συμβολίζουμε για συντομία με $\mathbf{X}$ και επιθυμούμε να προβλέψουμε την τιμή μιας αποκριτικής (response) μεταβλητής Y.

Θεωρούμε, ότι η Y είναι μια τυχαία μεταβλητή, της οποίας η δεσμευμένη μέση τιμή, δίνεται από τη σχέση:

$$\mathrm{E}[\mathrm{Y}|\mathbf{X}] = \mathrm{g}(\mathbf{X} : \boldsymbol{\theta}),$$

Όπου g συνάρτηση γνωστής μορφής και $\boldsymbol{\theta}=(\theta_1, \theta_2, \dots, \theta_k) \in \Theta$, άγνωστη παράμετρος. Η παραπάνω σχέση, ονομάζεται επιφάνεια παλινδρόμησης (καμπύλη για μονοδιάστατο x), της Y επί της $\mathbf{X}$. Όταν γνωρίζουμε την από κοινού κατανομή των $\mathbf{X}$ και Y, τότε η παραπάνω σχέση προσδιορίζεται ως εξής:

$$\mathrm{g}(\mathbf{X} : \boldsymbol{\theta}) = \mathrm{E}[\mathrm{Y}|\mathbf{X}] = \int_{-\infty}^{+\infty} \mathrm{y} f_{\mathrm{Y}|\mathbf{X}}(y|\mathbf{X})\mathrm{dy},$$

$$\text{με} \qquad f_{\mathrm{Y}|\mathbf{X}}(\mathrm{y}|\mathbf{X}) = \frac{f_{\mathbf{X},\mathrm{Y}}(\mathbf{X},\mathrm{y})}{f_{\mathbf{X}}(\mathbf{X})} \ .$$

Στις περισσότερες περιπτώσεις ωστόσο, η από κοινού κατανομή των $\mathbf{X}$ και Y είναι άγνωστη και έτσι η $\mathrm{E}[\mathrm{Y}|\mathbf{X}]$, προσδιορίζεται μέσω ενός συνόλου παρατηρήσεων $(\boldsymbol{X}_1, Y_1)$, $(\boldsymbol{X}_2, Y_2)$,..., $(\boldsymbol{X}_n, Y_n)$, με τον τρόπο που περιγράφεται παρακάτω. Αρχικά επιλέγεται η συναρτησιακή μορφή της g, λαμβάνοντας υπόψη τον τρόπο εξάρτησης της Y από το $\mathbf{X}$, και στη συνέχεια, από τις παρατηρήσεις $(\boldsymbol{X}_i, Y_i)$ (i=1,2,...n), εκτιμάται η παράμετρος $\boldsymbol{\theta}$ έτσι ώστε να ικανοποιείται κάποιο κριτήριο.

Στην περίπτωση μας το κριτήριο αυτό είναι τα ελάχιστα τετράγωνα (least squares), με βάση τα οποία η τιμή του $\boldsymbol{\theta}$, προσδιορίζεται από την ελαχιστοποίηση του αθροίσματος:

$$\mathrm{Q}(\boldsymbol{\theta}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \{\mathrm{y_i} - \mathrm{g}(\mathbf{x_i}: \boldsymbol{\theta})\}^2.$$

Η τιμή του $\boldsymbol{\theta}$ για την οποία ελαχιστοποιείται το παραπάνω άθροισμα καλείται εκτιμήτρια ελαχίστων τετραγώνων (least squares estimate), και συμβολίζεται με $\widehat{\boldsymbol{\theta}}$. Οι τιμές $\widehat{y_i} = g(\mathbf{x}_i: \widehat{\boldsymbol{\theta}})$, αποτελούν τις προβλέψεις του Y για $\mathbf{X} = \mathbf{x}_i$.

1.2. Πολλαπλή Γραμμική Παλινδρόμηση

Στην περίπτωση της γραμμικής παλινδρόμησης, η συναρτησιακή μορφή της g είναι γραμμική με αποτέλεσμα η δεσμευμένη μέση τιμή της τ.μ. Y, να δίνεται από τη σχέση:

$$E[Y|\mathbf{X}] = \alpha + b_1x_1 + b_2x_2 + \cdots + b_kx_k,$$

Για παρατηρήσεις $(\boldsymbol{X}_i, Y_i)$ (i=1,2,...n), η υπό ελαχιστοποίηση ποσότητα Q(a, $\mathbf{b}$), με $\mathbf{b}=(b_1, \dots, b_k)^T$, θα είναι:

$$Q(a, \mathbf{b}) = \sum_{i=1}^{n} \{y_i - \alpha - b_1x_{i1} - b_2x_{i2} - \cdots - b_kx_{ik}\}^2.$$

Εξισώνοντας, τις μερικές παραγώγους του παραπάνω αθροίσματος ως προς $\alpha, b_1, \dots, b_k$ με το μηδέν, προκύπτει ένα σύστημα η λύση του οποίου, μας δίνει τις εκτιμήτριες ελαχίστων τετραγώνων των $\alpha, b_1, \dots, b_k$.

Αν θεωρήσουμε τον πίνακα σχεδιασμού $X=\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$,

$\boldsymbol{\beta}=(\alpha, b_1, \dots, b_k)^T$ και $\mathbf{y}=(y_1, \dots, y_n)^T$, τότε η εκτιμήτρια ελαχίστων τετραγώνων του διανύσματος $\boldsymbol{\beta}$, δίνεται από τη σχέση:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}.$$

Οπότε και το μοντέλο που εκτιμάμε δίνεται από τη σχέση:

$$\hat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}.$$

1.3. Ερμηνεία των Παραμέτρων στην Πολλαπλή Γραμμική Παλινδρόμηση

Η παράμετρος α, εκφράζει την αναμενόμενη τιμή της τ.μ. Y όταν κάθε συνιστώσα του τυχαίου δείγματος $\mathbf{X}$ είναι μηδέν. Η παράμετρος b_m εκφράζει το πόσο αναμένεται να μεταβληθεί η τ.μ. Y, αν η τ.μ. X_m, μεταβληθεί κατά μια μονάδα και όλες οι άλλες τυχαίες μεταβλητές X_j $με\ m \neq j$, παραμείνουν σταθερές. Το πρόσημο του b_m, εκφράζει τη σχέση εξάρτησης μεταξύ των Y και X_m όταν όλες οι X_j παραμείνουν σταθερές.

Οι ποσότητες $\hat{\varepsilon_i} = y_i - \hat{y}_i$, με $\hat{y}_i = \hat{\alpha} + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k$, (i=1,…,n), ονομάζονται υπόλοιπα και εκφράζουν τις εκτιμησεις των σφαλμάτων των μετρήσεων, ενώ οι ποσότητες $\hat{y}_i$ είναι οι προβλεπόμενες τιμές.

1.4. Αξιολόγηση Μοντέλου Παλινδρόμησης

Υπάρχουν πολλοί τρόποι ώστε να ελέγξουμε αν ένα μοντέλο παλινδρόμησης είναι αποτελεσματικό και αν μπορεί να μας δώσει ακριβείς προβλέψεις. Στην παράγραφο αυτή θα δούμε μερικούς από τους πιο χαρακτηριστικούς τρόπους, που χρησιμοποιήσαμε και στην συγκεκριμένη έρευνα.

1.4.1. Συντελεστής Προσδιορισμού (R^2)

Η ποσότητα $R^2 = 1 - \frac{\sum_{i=1}^{n}\{y_i - \hat{y}_i\}^2}{\sum_{i=1}^{n}\{y_i - \bar{y}_i\}^2}$, καλείται συντελεστής προσδιορισμού (coefficient of determination), παίρνει τιμές στο [0,1] και εκφράζει το ποσοστό της διασποράς της τ.μ. Υ που περιγράφεται μέσω του μοντέλου παλινδρόμησης, δηλαδή φανερώνει το βαθμό στον οποίο οι $\mathbf{X}$ ερμηνεύουν με τη βοήθεια της γραμμικής παλινδρόμησης τη μεταβλητότητα της εξαρτημένης μεταβλητής Υ. Όσο μεγαλύτερη είναι η τιμή του συντελεστή τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των Υ και $\mathbf{X}$, υπό την προυπόθεση βέβαια ότι το πολλαπλό γραμμικό μοντέλο είναι κατάλληλο. Ωστόσο ο συντελεστής αυτός , είναι καλό να μην χρησιμοποιείται ως μέτρο καλής προσαρμογής του μοντέλου στα δεδομένα ή ως μέτρο σύγκρισης δύο μοντέλων και αυτό γιατί αν σε ένα γραμμικό μοντέλο προσθέσουμε μία επεξηγηματική μεταβλητή με ελάχιστη συνεισφορά στη μείωση της αβεβαιότητας μας για την τιμή της μεταβλητής απόκρισης, ο πολλαπλός συντελεστής προσδιορισμού θα αυξηθεί δίνοντάς μας έτσι την εσφαλμένη εντύπωση ότι το νέο πιο πολύπλοκο μοντέλο είναι περισσότερο κατάλληλο. Σε τέτοιες περιπτώσεις είναι προτιμότερο να υπολογίζουμε τον προσαρμοσμένο συντελεστή προσδιορισμού (adjusted coefficient of determination) ο οποίος εκφράζει το ποσοστό της διασποράς της τυχαίας μεταβλητής Υ που εξηγείται με βάση το μοντέλο παλινδρόμησης λαμβάνοντας όμως υπόψη και την πολυπλοκότητα (αριθμός επεξηγηματικό μεταβλητών) του μοντέλου. Έτσι αν σε ένα μοντέλο προσθέσουμε μία επεξηγηματική μεταβλητή, ενδέχεται το νέο πιο πολύπλοκο μοντέλο, να έχει μικρότερη τιμή στον προσαρμοσμένο συντελεστή προσδιορισμού σε σχέση με αυτή του αρχικού του μοντέλου. Τέλος, η σχέση που συνδέει τους δύο συντελεστές προσδιορισμού είναι η εξής:

$$\tilde{R}^2 = R^2 - \frac{(1-R^2)\,p}{n-p-1},$$

όπου p ο αριθμός των επεξηγηματικών μεταβλητών και n ο αριθμός των παρατηρήσεων που έχουμε στη διάθεση μας.

1.4.2. Μέσο Τετραγωνικό Σφάλμα(Mean Squared Error)

Το μέσο τετραγωνικό σφάλμα(MSE) αποτελεί ίσως το πιο αντικειμενικό κριτήριο καλής προσαρμογής ενός μοντέλου παλινδρόμησης. Εκτιμάει την άγνωστη διασπορά του τυχαίου σφάλματος σ^2 και αποτελεί αμερόληπτο εκτιμητή του. Το μέσο τετραγωνικό σφάλμα δίνεται από τη σχέση:

$$\text{MSE}=\frac{1}{n}\sum_{\text{i}=1}^{\text{n}}\{\text{y}_\text{i}-\hat{\text{y}}_\text{i}\}^2.$$

Η θετική τετραγωνική ρίζα του MSE, καλείται τυπικό σφάλμα της παλινδρόμησης(standard error of the regression), και όσο μικρότερη η τιμή της τόσο καλύτερη προσαρμογή έχουμε στα δεδομένα μας. Στο σημείο αυτό αξίζει να σημειωθεί ότι πολλές φορές αντί για το MSE υπολογίζουμε τη ρίζα του, δηλαδή το RMSE(Root Mean Squared Error), καθώς η μετρική αυτή σε αντίθεση με το MSE, έχει την ίδια κλίμακα με το y και έτσι μας δίνει μια καλύτερη εικόνα για το πόσο είναι το σφάλμα του μοντέλου που έχουμε κατασκευάσει.

1.4.3. Μέσο Απόλυτο Σφάλμα (Mean Absolute Error)

Το μέσο απόλυτο σφάλμα, εκφράζει το μέσο μέγεθος των σφαλμάτων σε ένα σύνολο προβλέψεων, ενός μοντέλου παλινδρόμησης, αγνοώντας την κατεύθυνση τους (δηλαδή το αν τα $\hat{y}$ είναι μεγαλύτερα ή μικρότερα από τα y). Είναι στην πράξη ο μέσος όρος των απόλυτων διαφορών μεταξύ των προβλέψεων για το y και των πραγματικών τιμών του y, και δίνεται από τη σχέση:

$$\text{MAE}=\frac{1}{n}\sum_{\text{i}=1}^{\text{n}}|\widehat{y_i}-y_i|.$$

Η βασική διαφορά του MAE από το MSE & RMSE, είναι ότι το πρώτο δε δίνει τόσο βάρος στα μεγάλα σφάλματα, έτσι είναι πιο ανθεκτικό απέναντι σε ακραίες προβλέψεις(outliers), σε αντίθεση με το MSE & RMSE, που λόγω του τετραγώνου στον τύπο τους, δίνουν μεγάλο βάρος στις ακραίες τιμές στις προβλέψεις $\hat{y}$.

1.4.4. Ποσοστιαίο – Σχετικό Σφάλμα(Relative Error)

Το ποσοστιαίο – σχετικό σφάλμα της παλινδρόμησης, εκφράζει τη διαφορά των προβλέψεων του μοντέλου παλινδρόμησης και των πραγματικών τιμών που αντιστοιχούν σε αυτές, προς τις πραγματικές τιμές αυτές και ορίζεται ως:

$$Relative\ Error = \frac{y_{ipredicted} - y_{itrue}}{y_{itrue}}$$

Το σχετικό σφάλμα δεν αποτελεί αυτούσιο μετρική για την αξιολόγηση ενός μοντέλου παλινδρόμησης, αλλά η μέση τιμή του και η διασπορά του αποτελούν δυο καλές μετρικές για το πόση ακρίβεια έχουν οι προβλέψεις που κάνει το

μοντέλο που έχουμε κατασκευάσει. Σε ιδανικές συνθήκες οι δυο τιμές αυτές θέλουμε να πλησιάζουν το απόλυτο μηδέν όσο το δυνατόν περισσότερο, ώστε να έχουμε μεγαλύτερη ακρίβεια σε όλο το φάσμα τιμών της μεταβλητής απόκρισης χωρίς μεγάλες αποκλίσεις.

1.5. Σημαντικότητα Επεξηγηματικών Μεταβλητών του Γραμμικού Μοντέλου Παλινδρόμησης

Άλλο ένα κριτήριο για να ελέγξουμε πόσο καλό είναι το μοντέλο παλινδρόμησης που έχουμε προσαρμόσει είναι η στατιστική σημαντικότητα των επεξηγηματικών μεταβλητών του μοντέλου.

Από την προσαρμογή του μοντέλου παλινδρόμησης, λαμβάνουμε για κάθε συντελεστή του μοντέλου μια p-τιμή, η οποία, αντιστοιχεί στις ακόλουθες υποθέσεις που θέλουμε να ελέγξουμε σε επίπεδο σημαντικότητας έστω α:

$$H_0: a=0 \text{ έναντι της εναλλακτικής } H_1: a\neq 0$$

$$H_0: b_1=0 \text{ έναντι της εναλλακτικής } H_1: b_1\neq 0$$

$$\vdots$$

$$H_0: b_k=0 \text{ έναντι της εναλλακτικής } H_1: b_k\neq 0$$

Με τον πρώτο έλεγχο θέλουμε να δούμε κατά πόσο η αναμενόμενη τιμή της Υ είναι 0 όταν $\mathbf{X}=0$. Πολλές φορές ωστόσο, η τιμή αυτή δεν έχει ερμηνεία, διότι η τιμή $\mathbf{X}=0$ δεν παρατηρείται ποτέ στην πράξη. Με κάθε έναν από τους επόμενους ελέγχους θέλουμε να διαπιστώσουμε κατά πόσο τελικά η αύξηση κατά μια μονάδα της $X_j (j = 1, \dots, k)$ σημαίνει και μεταβολή της αναμενόμενης τιμής της Υ, δηλαδή με άλλα λόγια κατά πόσο η Υ περιγράφεται από τη X_j.

Η p-τιμή, εκφράζει την πιθανότητα το στατιστικό ελέγχου που χρησιμοποιούμε να πάρει σε κάποιο άλλο δείγμα μία τόσο “ακραία” ή και ακόμα περισσότερο ακραία τιμή σε σχέση με αυτή που έχουμε παρατηρήσει, δεχόμενοι την μηδενική υπόθεση. Συνήθως όταν η p-τιμή είναι μικρότερη του επιπέδου σημαντικότητας α, απορρίπτουμε τη μηδενική υπόθεση και αποδεχόμαστε την εναλλακτική. Η πληροφορία λοιπόν που μας παρέχει η p-τιμή, είναι η πιθανότητα να βρούμε ενδείξεις αντίθετες με την μηδενική υπόθεση πιο ισχυρές από αυτές που έχουμε διαθέσιμες αν η μηδενική υπόθεση ίσχυε.

Συνήθως ελέγχουμε σε επίπεδο σημαντικότητας 95% (άρα $\alpha=0.05$), οπότε αν οι p-τιμές των συντελεστών του μοντέλου είναι <0.05 θεωρούμε ότι οι επεξηγηματικές μεταβλητές που αντιστοιχούν σε αυτούς είναι στατιστικά σημαντικές.

1.6. Προϋποθέσεις Πολλαπλού Γραμμικού Μοντέλου

Η εκτίμηση των παραμέτρων με τη μέθοδο ελαχίστων τετραγώνων, προϋποθέτει αρχικά να ικανοποιούνται κάποιες προϋποθέσεις. Αρχικά θα πρέπει να βεβαιωθούμε πριν την προσαρμογή του μοντέλου παλινδρόμησης ότι η σχέση μεταξύ της δεσμευμένης μέσης τιμής της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών μπορεί να περιγραφεί από ένα γραμμικό μοντέλο , «Υπόθεση Γραμμικότητας». Στη συνέχεια, αν θέλουμε να προβούμε σε στατιστική συμπερασματολογία, θα πρέπει να βεβαιωθούμε ότι ικανοποιούνται και οι υπόλοιπες προϋποθέσεις που είναι η «Κανονικότητα των Σφαλμάτων», η «Ομοσκεδατικότητα», και η «Ανεξαρτησία των Σφαλμάτων». Ας δούμε λοιπόν τώρα πως ελέγχουμε κάθε υπόθεση:

- Γραμμικότητα: Ο πιο απλός τρόπος να ελέγξουμε την υπόθεση της γραμμικότητας, είναι για κάθε μια από τις p επεξηγηματικές μεταβλητές, να σχεδιάσουμε το διάγραμμα διασποράς των σημείων (x_i, y_i), $\acute{o}\pi o\upsilon\ i = 1, \dots, n$ (n ο αριθμός των παρατηρήσεων), έτσι ώστε για κάθε μια από αυτές να ελέγξουμε γραφικά αν έχει γραμμική σχέση με τη μεταβλητή απόκρισης, ελέγχοντας αν τα σημεία στο διάγραμμα είναι κοντά στην ευθεία $y = x$. Ωστόσο ο τρόπος αυτός δουλεύει μόνο στην περίπτωση που οι επεξηγηματικές μεταβλητές είναι μεταξύ τους ασυσχέτιστες, γιατί με τον παραπάνω τρόπο ελέγχουμε με κάθε διάγραμμα το απλό γραμμικό μοντέλο $\mathrm{E}\left[\mathrm{Y}\middle|\mathrm{X}_j = x_j\right] = \alpha + \mathrm{b_j x_j}$, όπου j=1,...,p και όχι το $\mathrm{E}[\mathrm{Y}|\mathbf{X}] = \alpha + \mathrm{b_1 x_1} + \mathrm{b_2 x_2} + \cdots + \mathrm{b_k x_k}$. Στην περίπτωση που οι επεξηγηματικές μεταβλητές σχετίζονται μεταξύ τους, πρέπει να ελέγξουμε αν η τιμή της επεξηγηματικής μεταβλητής X_j, j=1,...,p, συνδέεται γραμμικά με τη δεσμευμένη μέση τιμή της Y, αν όλες οι τιμές των υπολοίπων επεξηγηματικών μεταβλητών, συνδέονται γραμμικά με τη δεσμευμένη μέση τιμή της Y. Για τον έλεγχο αυτό, θεωρούμε τη σχέση: $y \approx \hat{a} + \hat{b}_1 x_{i1} + \cdots + p_j\left(x_{ij}\right) + \cdots + \hat{b}_p x_{ip}$, i=1,...,n. Όπου αποδεικνύεται εύκολα ότι, $p_j\left(x_{ij}\right) \approx \hat{b}_j x_{ij} + \widehat{e_i} \equiv P_{ij}$. Οι όροι P_{ij} καλούνται j-μερικά υπόλοιπα (partial residuals) και για να ελέγξουμε την προϋπόθεση της γραμμικότητας όταν έχουμε συσχετισμένες μεταξύ τους μεταβλητές, σχεδιάζουμε τα διαγράμματα διασποράς των σημείων (x_{ij}, P_{ij}), για κάθε μεταβλητή.
- Κανονικότητα των σφαλμάτων: Για τον έλεγχο αυτής της προϋπόθεσης, θεωρούμε τα υπόλοιπα $\widehat{e_i} = y_i - \hat{a} - \hat{b}_1 x_{i1} - \cdots - \hat{b}_p x_{ip}$, που αποτελούν τις εκτιμήσεις των τυπικών σφαλμάτων και είτε μέσα από το ιστόγραμμα τους είτε από κανονικό διάγραμμα (normal Q-Q plot, που συγκρίνει στην πράξη τα ποσοστιαία σημεία της εμπειρικής συνάρτησης κατανομής* των δεδομένων με τα ποσοστιαία σημεία της Τυποποιημένης κατανομής), ελέγχουμε αν υπάρχει κανονικότητα, βλέποντας αν τα παραγόμενα

σημεία βρίσκονται περίπου σε μια ευθεία γραμμή. Αν οι παραπάνω έλεγχοι μας δείξουν απόκλιση από την κανονική κατανομή, τότε συνήθως μετασχηματίζουμε τη μεταβλητή απόκρισης είτε χρησιμοποιώντας το λογάριθμο είτε υψώνοντας την σε μια δύναμη.

*(Αν $x_1, \dots, x_n$ οι τιμές ενός τυχαίου δείγματος, τότε η συνάρτηση $F_n(x) = \frac{\pi\lambda\acute{\eta}\theta o\varsigma\ \tau\omega\nu\ x_i \leq x}{n}$ καλείται εμπειρική συνάρτηση κατανομή)

- Ομοσκεδαστικότητα: Με τον όρο αυτό εννοούμε, ότι η διασπορά της δεσμευμένης κατανομής της μεταβλητής απόκρισης, δοσμένης της τιμής **x** του τυχαίου διανύσματος παραμένει σταθερή ανεξάρτητα της τιμής **x.** Με άλλα λόγια, η διασπορά των τυχαίων σφαλμάτων e_i παραμένει σταθερή για τις διάφορες τιμές **x** του τυχαίου διανύσματος. Για τον έλεγχο αυτής της προϋπόθεσης, χρησιμοποιούμε το διάγραμμα διασποράς, μεταξύ των υπολοίπων $\hat{e}_i$, και των προβλεπόμενων τιμών $\hat{y}_i$. Στο διάγραμμα δε θα πρέπει τα σημεία να εμφανίζουν κάποιο συστηματικό τρόπο συμπεριφοράς, θα πρέπει να είναι διάσπαρτα στο επίπεδο. Αν παρουσιάζουν κάποιο συστηματικό τρόπο συμπεριφοράς, τότε έχουμε ενδείξεις ετεροσκεδαστικότητας, και πρέπει και σε αυτή την περίπτωση να μετασχηματίσουμε τη μεταβλητή απόκρισης.
- Ανεξαρτησία των σφαλμάτων: Τα τυχαία σφάλματα θα πρέπει να είναι ανεξάρτητες τυχαίες μεταβλητές. Όταν η υπόθεση αυτή δεν ικανοποιείται παρουσιάζεται το πρόβλημα της αυτοσυσχέτισης. Την αυτοσυσχέτιση την ελέγχουμε με διάγραμμα διασποράς των υπολοίπων σε σχέση με τη σειρά των δεδομένων, στο οποίο τα σημεία θα πρέπει να συμπεριφέρονται τυχαία.

1.7. Σημεία Επιρροής (Leverage Points)

Σημεία επιρροής ονομάζονται οι παρατηρήσεις που επηρεάζουν σε μεγάλο βαθμό, τη διαμόρφωση της εκτιμημένης συνάρτησης παλινδρόμησης με την έννοια ότι αν αυτές παραληφθούν, τότε τα αποτελέσματα μας θα είναι αρκετά διαφορετικά, δηλαδή, θα υπάρχει σημαντική διαφοροποίηση μεταξύ των τιμών των εκτιμώμενων συντελεστών παλινδρόμησης (εκτιμητές ελαχίστων τετραγώνων). Για αυτό το λόγο, η ύπαρξη των σημείων αυτών μπορεί να έχει αρνητικό αντίκτυπο στην απόδοση-ακρίβεια του μοντέλου παλινδρόμησης. Έτσι είναι σημαντικό πάντα να ελέγχουμε για την ύπαρξη σημείων επιρροής και αφού τα εντοπίσουμε να ερευνήσουμε το λόγο ύπαρξης τους (μπορεί να κρύβουν κάποια πληροφορία) και να αποφασίσουμε αν θα τα αφαιρέσουμε ή όχι ανάλογα με το πρόβλημα. Μια βασική μέθοδος για τον εντοπισμό σημείων επιρροής είναι η απόσταση Cook, η οποία εξετάζει άμεσα την επίδραση της παράληψης ενός σημείου από το δείγμα στην εκτίμηση του διανύσματος παραμέτρων του μοντέλου. Έστω $\hat{\beta}$ η εκτίμηση του διανύσματος των

παραμέτρων, όταν λαμβάνονται υπόψη όλες οι παρατηρήσεις ενώ $\hat{\beta}_i$, το διάνυσμα των εκτιμήσεων που προκύπτει από την αφαίρεση της i-οστής παρατήρησης. Αν μια παρατήρηση είναι σημείο επιρροής, τότε τα $\hat{\beta}$ και $\hat{\beta}_i$, θα διαφοροποιούνται αρκετά μεταξύ τους, έτσι η διαφορά $\hat{\beta} - \hat{\beta}_i$, μπορεί να αποτελέσει ένα μέτρο επιρροής της i-οστής παρατήρησης. Στο γραμμικό μοντέλο, αποδεικνύεται ότι:

$$\hat{\beta} - \hat{\beta}_i = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i}{1 - h_{ii}}$$

Όπου $\mathbf{X}$ ο πίνακας με όλες τις παρατηρήσεις, $\boldsymbol{x}_i$ η i-οστη γραμμή του πίνακα $\mathbf{X}$, e_iτο υπόλοιπο της i-οστης παρατήρησης και h_{ii} το i-οστο διαγώνιο στοιχείο του πίνακα προβολής $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Τελικά η απόσταση Cook, δίνεταιαπό τη σχέση:

$$D_i = \frac{e_i^2 h_{ii}}{pS^2(1 - h_{ii})^2}$$

όπου p=k+1 (k ο αριθμός των επεξηγηματικών μεταβλητών) και $S^2 = \frac{e'e}{n-k-1}$ η αμερόληπτη εκτιμήτρια της διασποράς των τυχαίων σφαλμάτων (n το μέγεθος του δείγματος). Στο σημείο αυτό αξίζει να σημειωθεί ότι οι πληροφορίες που αξιοποιεί ο υπολογισμός της απόστασης Cook, είναι όλες διαθέσιμες από την αρχική ανάλυση με όλες τις παρατηρήσεις. Επίσης, να τονιστεί ότι η απόσταση D_i, ορίζεται μόνο κατά αναλογία και στην πραγματικότητα δεν ακολουθεί την κατανομή $F_{(k+1),(n-k-1)}$, έτσι δε μπορούμε να χρησιμοποιήσουμε τα ποσοστιαία σημεία της κατανομής για τον εντοπισμό παρατηρήσεων με μεγάλη επιρροή. Συνήθως η προσοχή μας εστιάζεται στις παρατηρήσεις με $D_i > 1$ ή $D_i > \frac{4}{n}$ ή $\frac{4}{n-k-1}$.

1.8. Συσχέτιση Μεταβλητών Μοντέλου Παλινδρόμησης

1.8.1. Συντελεστής Συσχέτισης Pearson

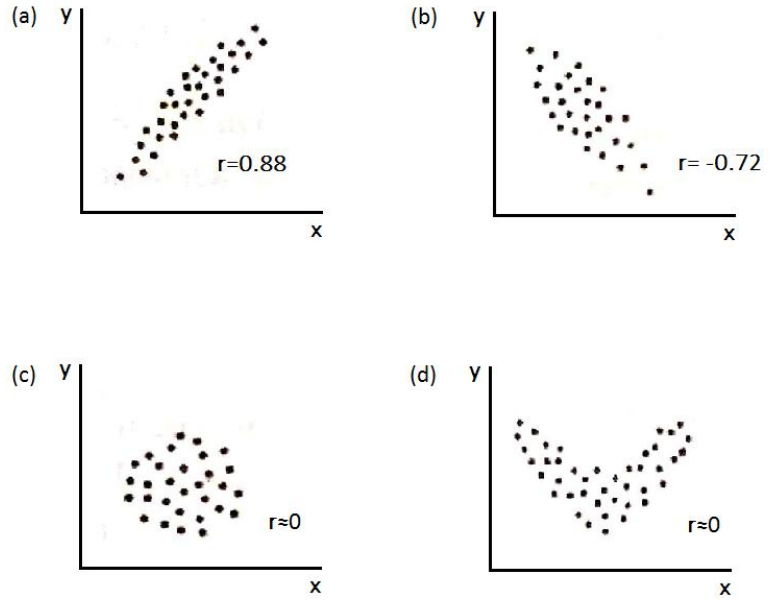
Ο συντελεστής συσχέτισης (correlation coefficient), μεταξύ των τ.μ. X, Y εκφράζει το βαθμό στον οποίο μπορούμε να εκτιμήσουμε γραμμικά τη μια τ.μ. όταν γνωρίζουμε την τιμή της άλλης, και δίνεται από τη σχέση:

$$\rho = Cov(X, Y)/\{V[X]V[Y]\}^{\frac{1}{2}}$$

Ωστόσο, στην πράξη εκτιμάμε τον συντελεστή συσχέτισης από ζεύγη παρατηρήσεων (x_i, y_i) (i=1,2,…,n). Όταν τα ζεύγη προέρχονται από διμεταβλητό κανονικό πληθυσμό τότε η εκτιμήτρια μεγίστης πιθανοφάνειας του ρ δίνεται από τη σχέση:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{n}\{x_i - x\}^2 \sum_{i=1}^{n}\{y_i - y\}^2)^{\frac{1}{2}}}$$

και ονομάζεται δειγματικός συντελεστής συσχέτισης ή συντελεστής συσχέτισης κατά Pearson (Pearson's correlation coefficient). Ο συντελεστής, παίρνει τιμές στο [-1,1] και όσο πιο κοντά είναι η απόλυτη τιμή του στη μονάδα τόσο πιο μεγάλη είναι η γραμμική συσχέτιση μεταξύ των (x_i, y_i), συγκεκριμένα όσο πιο κοντά στο $+1$ τόσο μεγαλύτερη θετική συσχέτιση, ενώ όσο πιο κοντά στο -1 τόσο μεγαλύτερη αρνητική συσχέτιση. Στην περίπτωση όπου η τιμή του είναι ± 1, έχουμε τέλεια γραμμική συσχέτιση και τα (x_i, y_i), βρίσκονται πάνω σε ευθεία γραμμή. Το διάγραμμα διασποράς (scatter plot), δηλαδή η απεικόνιση των (x_i, y_i), στο ορθοκανονικό σύστημα συντεταγμένων, βοηθάει στο να ελέγξουμε γραφικά την ύπαρξη ή όχι γραμμικής συσχέτισης μεταξύ των X και Y. Στο γράφημα που ακολουθεί βλέπουμε μερικά διαγράμματα διασποράς και τους συντελεστές συσχέτισης που αντιστοιχούν σε αυτά.



Παρατηρούμε από το παραπάνω σχήμα, ότι στα διαγράμματα (a),(b) έχουμε γραμμική εξάρτηση μεταξύ των X, Y, ενώ στο (c) δεν υπάρχει εξάρτηση. Ωστόσο, ενδιαφέρον παρουσιάζει το διάγραμμα (d), στο οποίο βλέπουμε ότι υπάρχει εξάρτηση μεταξύ των X, Y όμως όχι γραμμική. Σε τέτοιες περιπτώσεις είναι πολλές φορές αναγκαίο να πραγματοποιήσουμε μετασχηματισμούς στα δεδομένα ώστε να καταλάβουμε τη σχέση εξάρτησης μεταξύ των μεταβλητών.

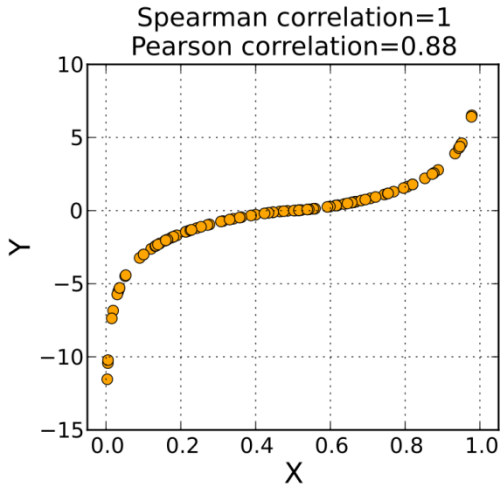
Συμπεραίνουμε λοιπόν πως για να εξάγουμε ολοκληρωμένα συμπεράσματα για τη συσχέτιση δυο μεταβλητών δεν αρκεί μόνο να υπολογίσουμε το συντελεστή συσχέτισης αλλά να κάνουμε και τα διαγράμματα διασποράς καθώς μπορεί δυο μεταβλητές να έχουν χαμηλό συντελεστή συσχέτισης και να έχουν παράλληλα υψηλή μη γραμμική συσχέτιση.

1.8.2. Συντελεστής Συσχέτισης Spearman

Σε περιπτώσεις που τα ζεύγη (x_i, y_i) (i=1,2,…,n), δεν προέρχονται από κανονικό πληθυσμό, χρησιμοποιούμε ως εκτιμήτρια του συντελεστή συσχέτισης r, τον συντελεστή συσχέτισης του Spearman(Spearman's correlation coefficient), ο οποίος προκύπτει από τον συντελεστή συσχέτισης του Pearson αντικαθιστώντας τις τιμές των παρατηρήσεων (x_i, y_i) (i=1,2,…,n) με τους αντίστοιχους βαθμούς αυτών. Για παράδειγμα αν r_{x_i} ο βαθμός της x_i παρατήρησης στο δείγμα των x και r_{y_i} ο βαθμός της y_i παρατήρησης στο δείγμα των y, ο συντελεστής συσχέτισης του Spearman, δίνεται από τη σχέση:

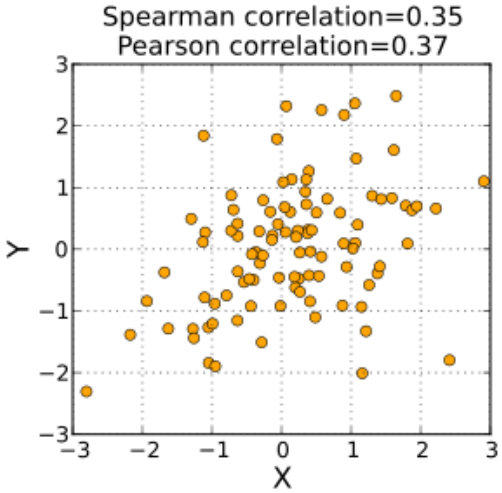
$$r_s = \frac{n\sum_{i=1}^{n} r_{x_i} r_{y_i} - \left(\sum_{i=1}^{n} r_{x_i}\right)\left(\sum_{i=1}^{n} r_{y_i}\right)}{\sqrt{n\sum_{i=1}^{n} r_{x_i}^2 - \left(\sum_{i=1}^{n} r_{x_i}\right)^2}\sqrt{n\sum_{i=1}^{n} r_{y_i}^2 - \left(\sum_{i=1}^{n} r_{y_i}\right)^2}}$$

Ο συντελεστής συσχέτισης Spearman, θεωρείται ως "μη παραμετρικός." Το γεγονός αυτό δηλώνει ότι μια τέλεια συσχέτιση Spearman προκύπτει όταν X και Y σχετίζονται με οποιαδήποτε μονότονη συνάρτηση, κάτι που δεν ισχύει με τη συσχέτιση Pearson, η οποία δίνει τέλεια τιμή μόνο όταν οι X και Y σχετίζονται με μια γραμμική συνάρτηση. Έτσι σε περιπτώσεις που έχουμε μεγάλη συσχέτιση που δεν είναι γραμμική ο Spearman μπορεί να δώσει απάντηση με μεγαλύτερη ακρίβεια για τη συσχέτιση μεταξύ των δυο μεταβλητών. Στις παρακάτω εικόνες βλέπουμε στα διαγράμματα διασποράς διαφορετικών ζευγών X, Y τις διαφορές μεταξύ των τιμών των δυο συντελεστών Pearson και Spearman.

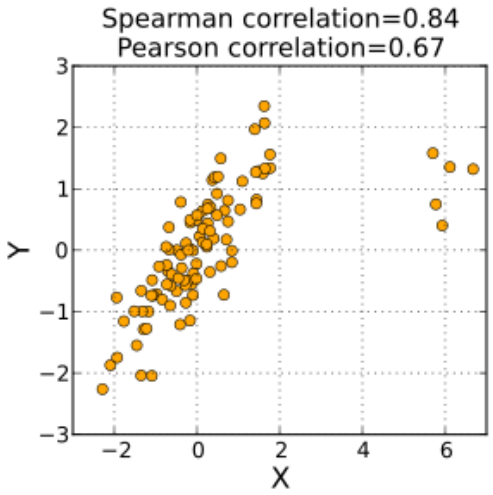


Στο διπλανό διάγραμμα βλέπουμε ότι η τιμή της συσχέτισης Spearman που είναι πολύ κοντά στη μονάδα οφείλεται στο γεγονός ότι οι δύο μεταβλητές που συγκρίνονται σχετίζονται μονοτονικά. Ωστόσο, επειδή η σχέση τους δεν είναι γραμμική δεν έχουμε τέλεια συσχέτιση Pearson.

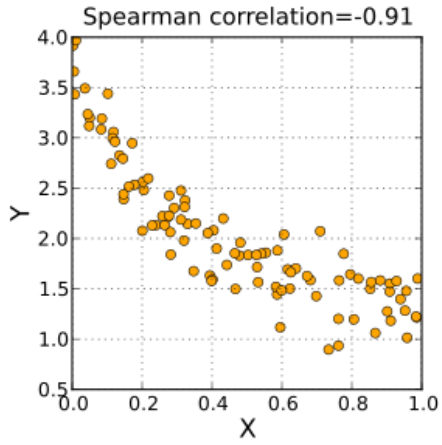
Στο διπλανό διάγραμμα βλέπουμε ότι η τιμή της συσχέτισης Spearman είναι παρόμοια με την αντίστοιχη της συσχέτισης Pearson, καθώς τα

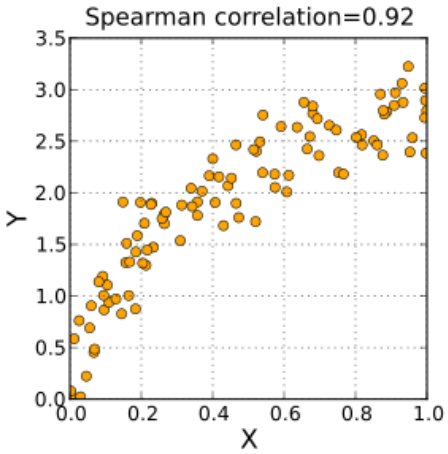


δεδομένα είναι σχεδόν ελλειπτικά κατανεμημένα.



Στο διπλανό διάγραμμα βλέπουμε ότι η συσχέτιση Spearman είναι λιγότερο ευαίσθητη από τη συσχέτιση Pearson σε ισχυρά ακραίες τιμές που βρίσκονται στην ουρά των δύο δειγμάτων.





Από τα δυο τελευταία διαγράμματα διασποράς της προηγούμενης σελίδας, διαπιστώνουμε ότι μια αύξουσα μονότονη τάση μεταξύ των X και Y, οδηγεί σε θετικό συντελεστή συσχετισης Spearman ενώ μια φθίνουσα μονότονη τάση μεταξύ των X και Y, οδηγεί σε αρνητικό συντελεστή συσχέτισης Spearman.

Εν κατακλείδι, για να μπορούμε να εξάγουμε σωστά και ακριβή αποτελέσματα για τη συσχέτιση μεταξύ δυο μεταβλητών, είναι καλό να υπολογίσουμε την τιμή και των δυο συντελεστών συσχέτισης και να απεικονίσουμε τα δεδομένα σε ένα διάγραμμα διασποράς ώστε να έχουμε μια σαφή εικόνα για τον τρόπο που οι δυο μεταβλητές σχετίζονται μεταξύ τους (περιγράφεται η μια από την άλλη).

1.8.3. Συσχέτιση και Παλινδρόμηση

Η συσχέτιση είναι πολύ σημαντικός παράγοντας που επηρεάζει την προσαρμογή και την απόδοση ενός μοντέλου παλινδρόμησης. Γενικά, όταν θέλουμε να προβλέψουμε μια μεταβλητή απόκρισης Y, διαθέτοντας μερικές επεξηγηματικές μεταβλητές $X_1, X_2, \dots, X_k$, είναι επιθυμητό οι επεξηγηματικές μεταβλητές να έχουν μεταξύ τους όσο μικρότερη συσχέτιση γίνεται (έτσι ώστε κάθε μια να παρέχει διαφορετική πληροφορία σε σχέση με τη μεταβλητή απόκρισης και να μην επεξηγούνται η μια από την άλλη) και παράλληλα όσο μεγαλύτερη συσχέτιση γίνεται με την μεταβλητή απόκρισης (έτσι ώστε η πληροφορίες που μας παρέχουν οι επεξηγηματικές μεταβλητές να μας χρησιμεύουν για να περιγράψουμε και να προβλέψουμε τη μεταβλητή απόκρισης).

Ωστόσο στην πράξη και στα προβλήματα του πραγματικού κόσμου η παραπάνω προϋπόθεση είναι δύσκολα επιτεύξιμη. Για το λόγο αυτό όταν διαθέτουμε ένα σύνολο επεξηγηματικών μεταβλητών καλούμαστε να κάνουμε μια αρχική επεξεργασία και ανάλυση ώστε να αποφασίσουμε ποιες θα χρησιμοποιήσουμε για την κατασκευή του μοντέλου παλινδρόμησης. Τέλος, όταν γίνεται λόγος για μεγάλη συσχέτιση συνήθως εννοούμε από 0.50 και πάνω.

1.9. Πολυσυγγραμικότητα Μεταβλητών

Δεν είναι λίγες οι φορές που συμβαίνει να έχουμε μεγάλη συσχέτιση μεταξύ δυο ή περισσότερων επεξηγηματικών μεταβλητών ενός μοντέλου παλινδρόμησης, γεγονός που δημιουργεί το φαινόμενο πολυσυγγραμικότητας (multiconllinearity). Η παρουσία πολυσυγγραμικότητας έχει ως συνέπεια να προκύπτουν μεγάλα σφάλματα των εκτιμητών του μοντέλου παλινδρόμησης $\widehat{\boldsymbol{\beta}}$, με αποτέλεσμα να γίνεται πιο δύσκολη η εκτίμηση της επίδρασης της κάθε επεξηγηματικής μεταβλητής. Επίσης σε τέτοιες περιπτώσεις είναι δύσκολος, και ο εντοπισμός των στατιστικά σημαντικών μεταβλητών διότι η τιμή της ελεγχοσυνάρτησης $\mathrm{t} = \frac{\widehat{\beta}}{se(\widehat{\beta})}$, υπό την H_0: β=0, θα είναι μικρή.

Έτσι στην πράξη όταν έχουμε δυο μεταβλητές που έχουν υψηλή συσχέτιση μεταξύ τους κρατάμε τη μια από τις δυο (συνήθως αυτή που έχει και μεγαλύτερη συσχέτιση με τη μεταβλητή απόκρισης), αφού η ύπαρξη και των δυο μεταβλητών στο μοντέλο παλινδρόμησης όχι μόνο δεν οδηγεί σε κάποια βελτίωση του μοντέλου καθώς η δεύτερη δεν προσφέρει κάποια περισσότερη πληροφορία σε σχέση με την πρώτη (περιγράφονται η μια από την άλλη), αλλά αντίθετα μπορεί να μας οδηγήσει σε λανθασμένα συμπεράσματα για τη σημαντικότητα της κάθε μεταβλητής στο μοντέλο και σε μη ακριβείς προβλέψεις.

Το μεγάλο πρόβλημα όμως εντοπίζεται στην περίπτωση που έχουμε λίγες επεξηγηματικές μεταβλητές με υψηλή συσχέτιση μεταξύ τους και με τη

μεταβλητή απόκρισης, όπου δεν μπορούμε να αφαιρέσουμε μεταβλητές καθώς θα χάσουμε σημαντική πληροφορία.

Στην περίπτωση αυτή υπάρχουν διάφορες μέθοδοι που μπορούν να συνεισφέρουν στο να ξεπεράσουμε το φαινόμενο της πολυσυγγραμικότητας.

1.9.1. Παλινδρόμηση κορυφογραμμής

Η παλινδρόμηση κορυφογραμμής (ridge regression), αποτελεί μια αναλυτική μέθοδο, που μπορεί να συμβάλλει στην επίλυση του προβλήματος πολυσυγγραμικότητας. Κατά τη μέθοδο αυτή η εκτίμηση των παραμέτρων δίνεται από την εξίσωση:

$$\widehat{\boldsymbol{\beta}}_{\mathbf{ridge}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X} + \boldsymbol{\lambda I})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y},$$

για κάποια επιλεγμένη λ>0. Είναι γνωστό, ότι η εκτιμήτρια $\widehat{\boldsymbol{\beta}}$, είναι αμερόληπτη οπότε συμπεραίνουμε ότι η $\widehat{\boldsymbol{\beta}}_{\mathbf{ridge}}$ θα είναι μεροληπτική. Παρόλα αυτά η διαπορά της $\widehat{\boldsymbol{\beta}}_{\mathbf{ridge}}$ θα είναι μικρότερη από αυτή της $\widehat{\boldsymbol{\beta}}$ με αποτέλεσμα το μέσο τετραγωνικό σφάλμα της $\widehat{\boldsymbol{\beta}}_{\mathbf{ridge}}$ είναι πολύ μικρότερο από το αντίστοιχο της $\widehat{\boldsymbol{\beta}}$. Αυτό έχει ως συνέπεια οι εκτιμήσεις του μοντέλου να έχουν μεγαλύτερη ακρίβεια και να μπορούμε να εξάγουμε πιο βέβαια συμπεράσματα για τη στατιστική σημαντικότητα των μεταβλητών.

1.9.2. Μερικά Ελάχιστα Τετράγωνα (Partial Least Squares):

Συχνά, συμβαίνει να έχουμε ένα μεγάλο αριθμό επεξηγηματικών μεταβλητών, για ένα μοντέλο παλινδρόμησης, και πολλές από αυτές τις μεταβλητές μάλιστα να είναι αρκετά συσχετισμένες μεταξύ τους. Σε τέτοιες περιπτώσεις, χρησιμοποιούμε μεθόδους ώστε να μειώσουμε τις διαστάσεις του προβλήματος(τον αριθμό των επεξηγηματικών μεταβλητών). Μια από τις πιο διαδεδομένες μεθόδους που χρησιμοποιούμε στην πολλαπλή γραμμική παλινδρόμηση ώστε να κατασκευάσουμε ένα μοντέλο παλινδρόμησης, μειώνοντας παράλληλα τις διαστάσεις του , είναι η μέθοδος μερικών ελαχίστων τετραγώνων(PLS).

Η μέθοδος PLS, κατασκευάζει έναν αριθμό από γραμμικούς συνδυασμούς $Z_m, \text{όπου } m = 1, \dots, M, \mu\varepsilon\ M \leq p$(p ο αριθμός επεξηγηματικών μεταβλητών του μοντέλου), από τις αρχικές μεταβλητές $x_j, j = 1, \dots, p$, και στη συνέχεια τα Z_m, χρησιμοποιούνται στη θέση των x_j, ως επεξηγηματικές μεταβλητές του μοντέλου. Η μέθοδος, λόγω του ότι δεν έχει σταθερή κλίμακα (not scale invariant) θεωρεί ότι κάθε μεταβλητή x_j είναι τυποποιημένη (έχει μέση τιμή 0 και διασπορά 1) και στη συνέχεια υπολογίζει τους όρους $\widehat{\varphi_{1j}} = \langle x_j, y\rangle, \text{όπου } \langle x, y\rangle = \sum x_i y_i = x^T y$, για κάθε j.Έτσι, κατασκευάζει την παραγόμενη μεταβλητή $z_1 = \sum_j \widehat{\varphi_{1j}} x_j$, όπου αποτελεί την πρώτη κατεύθυνση (πρώτο γραμμικό συνδυασμό) της παλινδρόμησης μερικών ελαχίστων τετραγώνων. Ακόμα, κατά την κατασκευή των υπόλοιπων z_m, οι παραγόμενες

μεταβλητές επιβαρύνονται με βάση την επίδραση τους στο y. Η εκτίμηση για το y, προκύπτει από την παλινδρόμηση του Z_1, παράγοντας ένα συντελεστή $\hat{\theta}_1$ και στη συνέχεια ορθογωνιοποιώντας τα $x_1, \dots, x_p$, με βάση το Z_1. Συνεχίζουμε τη διαδικασία, μέχρι οι $M \leq p$, διευθύνσεις να έχουν υπολογιστεί. Με αυτό τον τρόπο, η PLS, παράγει μια σειρά από ανεξάρτητες διευθύνσεις $Z_1, \dots, Z_M$. Αν ωστόσο κατασκευάσουμε και τις M=p διευθύνσεις, το μοντέλο που παίρνουμε ως λύση είναι ισοδύναμο με το κλασσικό πολλαπλό μοντέλο παλινδρόμησης ελαχίστων τετραγώνων. Επιλέγοντας M<p διευθύνσεις, πραγματοποιούμε παλινδρόμηση με λιγότερους βαθμους ελευθερίας- λιγότερες διαστάσεις. Ο αναλυτικός αλγόριθμος της διαδικασίας, είναι ο παρακάτω:

1. **Δεδομένα:** Τυποποίησε κάθε x_j ώστε να έχει μέση τιμή 0 και διασπορά 1. Θέσε $\hat{y}^{(0)} = \bar{y}1$, και $x_j^{(0)} = x_j, j = 1, \dots, p$.
2. Για $m = 1,2, \dots, p$
   a. $z_m = \sum_j \hat{\varphi}_{mj} x_j^{(m-1)}$, $όπου\ \hat{\varphi}_{mj} = \langle x_j^{(m-1)}, y \rangle$. $j = 1, \dots, p$
   b. $\hat{\theta}_m = \frac{\langle z_m,\ y \rangle}{\langle z_m,\ z_m \rangle}$.
   c. $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.
   d. Ορθογωνοποίησε κάθε $x_j^{(m-1)}$, με βάση το z_m: $x_j^{(m)} = x_j^{(m-1)} - \left[\frac{\langle z_m,\ x_j^{(m-1)} \rangle}{\langle z_m,\ z_m \rangle}\right] z_m,\ \ j = 1,2, \dots, p.$
3. **Αποτελέσματα:** τα προσαρμοσμένα διανύσματα $\{\hat{y}^{(m)}\}_1^p$. Αφού τα $\{z_l\}_1^m$ είναι γραμμικά στα αρχικά x_j, το ίδιο θα είναι και τα $\hat{y}^{(m)} = X\hat{\beta}^{pls}(m)$. Αυτοί οι γραμμικοί συντελεστές, μπορούν να ανακτηθούν από τους μετασχηματισμούς που πραγματοποιεί η μέθοδος PLS.

Το γεγονός ότι η μέθοδος PLS, χρησιμοποιεί τη μεταβλητή απόκρισης y, για να κατασκευάσει τις διευθύνσεις της, μας οδηγεί στο συμπέρασμα, ότι η λύση είναι μια μη γραμμική συνάρτηση του y. Πρακτικά η PLS, αναζητάει διευθύνσεις, που έχουν υψηλή συσχέτιση με τη μεταβλητή απόκρισης. Συγκεκριμένα, αν η μέθοδος έχει κατασκευάσει M διευθύνσεις (οι οποίες καλούνται και κύρια συστατικά – principal components), το m-οστό κύριο συστατικό, v_m, λύνει το πρόβλημα:

$$\max_a Var(X_a), \text{ με } ||\alpha||=1, \alpha^T S v_l = 0, l =, \dots, m-1$$

Όπου S ο πίνακας συνδιασποράς των x_j. Η συνθήκη, $\alpha^T S v_l = 0$, διασφαλίζει, ότι η $Z_m = X_a$ είναι ασυσχέτιστη με όλους τους προηγούμενους συνδυασμούς $Z_l = X_{v_l}$. Η m-οστή κατεύθυνση $\widehat{\varphi_m}$ λύνει το πρόβλημα:

$$\max_a Corr^2(y, X_a) Var(X_a), \text{ με } ||\alpha||=1, \widehat{\varphi_l}^T Sa = 0, l =, \dots, m-1$$

Τέλος, αξίζει να σημειωθεί ότι αν ο πίνακας των επεξηγηματικών μεταβλητών X, είναι ορθογώνιος, τότε η μέθοδος μερικών ελαχίστων τετραγώνων, κατασκευάζει τις εκτιμήσεις ελαχίστων τετραγώνων, μετά από m=1 βήματα. Περεταίρω βήματα, δεν προσφέρουν καμία αλλαγή στο αποτέλεσμα καθώς τα $\widehat{\varphi_{mJ}}$, είναι μηδενικά για m>1.

1.9.3. Κανονικοποίηση(Normalize)

Όταν το φαινόμενο της πολυσυγγραμικότητας οφείλεται στη δομή των δεδομένων, καλείται διαρθρωτική πολυσυγγραμικότητα (structural mulitconllinearity). Το φαινόμενο αυτό παρουσιάζεται συνήθως όταν δημιουργούμε επεξηγηματικές μεταβλητές από άλλες επεξηγηματικές μεταβλητές, όπως παράδειγμα θα μπορούσαμε να κατασκευάσουμε το x^2 από το x. Σε τέτοιες περιπτώσεις η κανονικοποίηση θα μπορούσε να βοηθήσει στην αντιμετώπιση του φαινομένου. Με τη μέθοδο της κανονικοποίησης έχουμε τη δυνατότητα να μετασχηματίσουμε τα δεδομένα μας ώστε αυτά να παίρνουν τιμές στο διάστημα [0,1], έτσι όλες οι επεξηγηματικές μεταβλητές του μοντέλου παλινδρόμησης έχουν το ίδιο σύνολο τιμών. Υπάρχουν πολλοί μέθοδοι για να κανονικοποιήσουμε τα δεδομένα μας με την πιο απλή να είναι η μέθοδος min-max scaler, κατά την οποία τα δεδομένα μετασχηματίζονται με βάση τη σχέση:

$$z = \frac{x - min(x)}{\max(x) - \min(x)}$$

Η κανονικοποίηση είναι αρκετά χρήσιμη καθώς μπορεί να βελτιώσει την απόδοση του μοντέλου παλινδρόμησης, αφού κάνει την εκπαίδευση του μοντέλου λιγότερο ευαίσθητη στην κλίμακα των χαρακτηριστικών μεταβλητών, με αποτέλεσμα να μπορούμε να εκτιμήσουμε καλύτερα τις παραμέτρους του. Ωστόσο, υπάρχουν περιπτώσεις που η κλίμακα των χαρακτηριστικών είναι απαραίτητη για την κατανόηση και καλύτερη εξήγηση του μοντέλου παλινδρόμησης. Σε τέτοιες περιπτώσεις η κανονικοποίηση θα μπορούσε να οδηγήσει σε απώλεια πληροφορίας.

1.9.4. Τυποποίηση(Standardize)

Ένας εναλλακτικός τρόπος να μετασχηματίσω τα δεδομένα μου ώστε να έχουν ίδια κλίμακα(scale) και μονάδες(units), είναι να εφαρμόσω τυποποίηση. Κατά την τυποποίηση τα δεδομένα μετασχηματίζονται με τέτοιο τρόπο ώστε να έχουν μέση τιμή μ=0 και διασπορά σ=1. Ο μετασχηματισμός που εφαρμόζεται σε αυτή την περίπτωση περιγράφεται από την παρακάτω σχέση:

$$z = \frac{x - \mu}{\sigma}$$

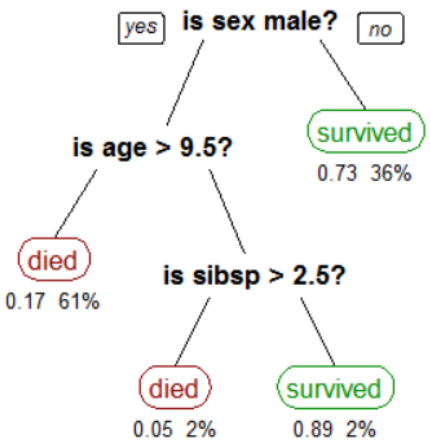
Όπου z η μετασχηματισμένη τιμή, x η αρχική τιμή και μ, σ η μέση τιμή και η διασπορά του αρχικού συνόλου των δεδομένων.

2. Μη Γραμμική Παλινδρόμηση

Μέχρι στιγμής είδαμε πώς μπορούμε μέσω της γραμμικής παλινδρόμησης, να κατασκευάσουμε ένα μοντέλο το οποίο θα μπορεί με τη βοήθεια μερικών επεξηγηματικών μεταβλητών να προβλέψει την τιμή μιας μεταβλητής απόκρισης. Ωστόσο, στα περισσότερα προβλήματα του πραγματικού κόσμου, η σχέση μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής που θέλουμε να προβλέψουμε δεν μπορεί να χαρακτηριστεί από γραμμικότητα, αλλά από μια μονότονη συνάρτηση. Σε τέτοιες περιπτώσεις, ένα γραμμικό μοντέλο δεν θα μπορούσε να έχει καλή απόδοση με αποτέλεσμα να μην έχουμε μεγάλη ακρίβεια στις προβλέψεις μας. Έτσι, σε προβλήματα που η γραμμική παλινδρόμηση δεν καταφέρνει να δώσει καλά αποτελέσματα χρησιμοποιούμε άλλες μεθόδους παλινδρόμησης.

2.1. Δέντρα Απόφασης (Decision Trees)

Τα πιο συνηθισμένα, απλά και αποτελεσματικά εργαλεία που μπορούν να αντιμετωπίσουν το πρόβλημα της μη γραμμικότητας σε ένα μοντέλο παλινδρόμησης είναι τα δέντρα απόφασης (decision trees). Ένα δένδρο απόφασης, αποτελείται από κόμβους και ακμές και η μορφή του μοιάζει με ένα ανάποδο δέντρο. Η λειτουργία ενός δέντρου μπορεί να παρομοιαστεί με τη λειτουργία ενός διαγράμματος ροής. Σε κάθε κόμβο υπάρχει μια συνθήκη (ερώτηση) και οι ακμές που ξεκινούν από το συγκεκριμένο κόμβο αντιπροσωπεύουν τις διαφορετικές απαντήσεις στην ερώτηση και οδηγούν παράλληλα σε μια άλλη ερώτηση ή στην τελική απόφαση. Ο αρχικός κόμβος(αρχική ερώτηση) από τον οποίο ξεκινάνε ακμές αλλά καμιά ακμή δεν οδηγεί σε αυτόν, ονομάζεται ρίζα, ενώ οι κόμβοι που δεν έχουν ακμές να ξεκινούν από αυτούς ονομάζονται φύλλα και αποτελούν τις τελικές αποφάσεις. Όλοι οι υπόλοιποι κόμβοι ονομάζονται κλαδιά. Τα δέντρα στην ουσία υλοποιούν και οπτικοποιούν έναν αλγόριθμο με τον οποίο καλούμαστε να απαντήσουμε σε ένα πρόβλημα, στη δική μας περίπτωση να προβλέψουμε μια τιμή. Ας δούμε για παράδειγμα το παρακάτω δέντρο το οποίο δείχνει την πιθανότητα ένας επιβάτης να επιβιώσει αν βρίσκεται στον Τιτανικό («sibsp»=ο αριθμός των μελών της οικογένειας, αδέρφια και σύζυγοι).



Σχήμα : Παράδειγμα δέντρου απόφασης για τον υπολογισμό της

πιθανότητας επιβίωσης ενός επιβάτη του «Τιτανικού»

2.2. Δέντρα Ταξινόμησης και Παλινδρόμησης (CART)

Η μέθοδος Classification And Regression Tree (CART), που αναπτύχθηκε από το Leo Breiman το 1984, περιλαμβάνει απαραμετρικά μη-γραμμικά μοντέλα πρόβλεψης, τα δέντρα ταξινόμησης(classification trees) και τα δέντρα παλινδρόμησης (regression trees). Όταν η μεταβλητή απόκρισης είναι κατηγορική τότε η μέθοδος παράγει δέντρα ταξινόμησης ενώ όταν είναι συνεχής παράγει δέντρα παλινδρόμησης. Είναι μια step-by-step μέθοδος, που βασίζεται στην ιδέα του αναδρομικού διαμερισμού(recursive partitioning), και σκοπό της έχει τη δημιουργία ενός δέντρου απόφασης μέσω της σταδιακής διαμέρισης κάθε κόμβου σε δυο κόμβους (θυγατρικούς). Ένα βασικό χαρακτηριστικό της μεθόδου CART, είναι ότι ο αλγόριθμος που την υποστηρίζει βασίζεται σε ένα σύνολο ιεραρχικών ερωτήσεων για την κατασκευή των δέντρων απόφασης, με αποτέλεσμα η ερμηνεία και η κατανόηση των αποτελεσμάτων και των τελικών αποφάσεων να είναι αρκετά απλή. Η μέθοδος ξεκινάει από τον κόμβο-ρίζα ο οποίος περιλαμβάνει όλο το σύνολο εκπαίδευσης ενώ κάθε επόμενος κόμβος αντιπροσωπεύει ένα υποσύνολο των μεταβλητών. Κάθε κόμβος που δεν είναι φύλλο (parent) χωρίζεται σε δυο θυγατρικούς κόμβους, και η διαδικασία αυτή ονομάζεται δυαδική διαίρεση (binary split). Η διαίρεση αυτή καθορίζεται από μια συνθήκη για την τιμή μιας μεταβλητής, η οποία ικανοποιείται ή όχι από την παρατηρούμενη τιμή αυτής της μεταβλητής. Η διαδικασία περιγράφεται ως εξής: Όταν οι παρατηρήσεις φτάσουν σε έναν κόμβο τότε αυτές που ικανοποιούν τη συνθήκη του κόμβου για μια συγκεκριμένη μεταβλητή μεταβαίνουν στον κατάλληλο θυγατρικό κόμβο ενώ αυτές που δεν ικανοποιούν τη συνθήκη μεταβαίνουν στον άλλο θυγατρικό κόμβο. Μετά το τέλος της διαδικασίας όλες οι παρατηρήσεις έχουν καταλήξει στο κατάλληλο φύλλο (κόμβοι που δεν έχουν θυγατρικούς), και κάθε φύλλο είναι συνδεδεμένο με μια τιμή (στατιστικό όταν μιλάμε για παλινδρόμηση, κατηγορία κλάσης όταν μιλάμε για ταξινόμηση). Προφανώς μπορεί να υπάρχουν πολλά φύλλα που οδηγούν στην ίδια κλάση/στατιστικό.

Η κατασκευή των δέντρων παλινδρόμησης από τη μέθοδο CART, στηρίζεται σε 3 βασικά στάδια:

1) Τον καθορισμό των ερωτήσεων της μορφής, είναι το $X \leq d$; όπου X είναι μια μεταβλητή και d είναι μια σταθερά . Η απάντηση στην ερώτηση είναι ναι ή όχι.
2) Επιλογή των κατάλληλων κριτηρίων για το διαχωρισμό των μεταβλητών.
3) Παραγωγή συνοπτικών στατιστικών που αφορούν στη μεταβλητή απόκρισης για τους κόμβους φύλλα.

Ο βασικός σκοπός της CART, είναι να δημιουργήσει ένα αποτελεσματικό μοντέλο πρόβλεψης-παλινδρόμησης, που βασίζεται σε δεντρική δομή, με στόχους την όσο το δυνατόν πιο ακριβή πρόβλεψη της μεταβλητής απόκρισης από τιμές των επεξηγηματικών μεταβλητών αλλά και την κατανόηση της σχέσης μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Η μέθοδος πετυχαίνει το σκοπό αυτό αρχικά εντοπίζοντας την ετερογένεια (από άποψη διασποράς της μεταβλητής απόκρισης), που υπάρχει στα δεδομένα και στη συνέχεια προχωράει σε μια εξομάλυνση της ετερογένειας αυτής, διαμερίζοντας αναδρομικά τα δεδομένα σε ομάδες από τελικούς κόμβους (φύλλα του δέντρου), που παρουσιάζουν εσωτερικά μεγαλύτερη ομοιογένεια από ότι οι πρόγονοι κόμβοι. Σε κάθε τελικό κόμβο, η μέση τιμή της μεταβλητής απόκρισης θεωρείτε ως η προβλεπόμενη τιμή. Αν σκοπός του δέντρου παλινδρόμησης είναι η εξήγηση της σχέσης μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών, τότε μέσω των μονοπατιών που ξεκινούν από τη ρίζα του δέντρου και οδηγούν σε ένα τελικό κόμβο μπορούμε εύκολα να κατανοήσουμε τη σχέση αυτή.

2.2.1. Κατασκευή Δέντρου Παλινδρόμησης

Ο τρόπος κατασκευής του δέντρου παλινδρόμησης μοιάζει πολύ με τον τρόπο κατασκευής του δέντρου ταξινόμησης. Τα βήματα της κατασκευής περιγράφονται ως εξής:

1) Ξεκινώντας από τη ρίζα, η CART, πραγματοποιεί κάθε πιθανό διαχωρισμό σε κάθε μια από τις επεξηγηματικές μεταβλητές και παράλληλα εφαρμόζει σε κάθε έναν από αυτούς ένα προκαθορισμένο μέτρο purity, ώστε να αποφασίσει ποιος διαχωρισμός μας δίνει καλύτερο αποτέλεσμα. Στην περίπτωση της παλινδρόμησης αποφασίζει με βάση το ποιος διαχωρισμός δίνει το μικρότερο τετραγωνικό σφάλμα.
2) Στη συνέχεια, αφού εφαρμόσει «goodness-of-split» κριτήρια διαχωρισμού επιλέγει τον καλύτερο διαχωρισμό και χωρίζει το σύνολο των δεδομένων σε δυο θυγατρικούς κόμβους τον αριστερό και τον δεξιό, ανάλογα με το που αυτά ανήκουν.
3) Καθώς η CART είναι μια αναδρομική μέθοδος, τα βήματα (1), (2), επαναλαμβάνονται για κάθε μη τελικό κόμβο, και κατασκευάζεται το μεγαλύτερο δυνατό δέντρο.
4) Τέλος η μέθοδος, εφαρμόζει έναν αλγόριθμο κλαδέματος, στο δέντρο που έχει κατασκευαστεί, με αποτέλεσμα τη δημιουργία μιας σειράς υποδέντρων διαφορετικών διαστάσεων από τα οποία επιλέγεται τελικά το βέλτιστο.

2.2.2. Καθορισμός Ερωτήσεων Διαχωρισμού

Για να πραγματοποιήσει όλους τους πιθανούς διαχωρισμούς η CART, πραγματοποιεί ερωτήσεις της μορφής: Ισχύει ότι $X \leq d$? όπου X, όπου X μια συνεχής επεξηγηματική μεταβλητή και d μια σταθερά στο σύνολο τιμών της X.

Ο αριθμός των διαφορετικών διαχωρισμών της κάθε μεταβλητής, περιορίζεται στον αριθμό των διακριτών τιμών που παίρνει η μεταβλητή στο δείγμα(σύνολο δεδομένων). Για παράδειγμα, αν το δείγμα έχει N παρατηρήσεις και η X είναι συνεχής μεταβλητή, τότε ο μέγιστος αριθμός διαχωρισμών για τη X, είναι N. Η μέθοδος CART, από προεπιλογή, πραγματοποιεί κάθε διαχωρισμό βασιζόμενη σε μια μόνο μεταβλητή τη φορά. Έτσι πρακτικά, η μέθοδος ξεκινάει για κάθε μεταβλητή και υπολογίζει τις πιθανές ερωτήσεις διαχωρισμού ως εξής: Θέτει ως ερώτηση την τιμή της πρώτης παρατήρησης (για τη μεταβλητή αυτή). Για παράδειγμα αν η τιμή της πρώτης παρατήρησης της πρώτης μεταβλητής έστω X_1 είναι 8, η ερώτηση θα είναι $X_1 \leq 8$. Αφού γίνει η ερώτηση οι παρατηρήσεις του συνόλου δεδομένων μοιράζονται στους δυο θυγατρικούς κόμβους που δημιουργούνται, ανάλογα με το αν ικανοποιούν τη συνθήκη της ερώτησης ή όχι. Η διαδικασία επαναλαμβάνεται και για τις υπόλοιπες τιμές των παρατηρήσεων για τη μεταβλητή αυτή και στη συνέχεια εφαρμόζονται κριτήρια καλού διαχωρισμού.

2.2.3. Κανόνες Διαχωρισμού και «goodness –of –split» Κριτήρια (Splitting Rules & Goodness-of-Split Criteria)

Υπάρχουν δυο κανόνες διαχωρισμού (συναρτήσεις καθαρότητας, impurity functions), για ένα δέντρο παλινδρόμησης, οι οποίοι αποτελούν τα κριτήρια για το πόσο καλός είναι ο διαχωρισμός. Ο πρώτος είναι η συνάρτηση Ελαχίστων Τετραγώνων(Least Squares - LS) και ο δεύτερος είναι η συνάρτηση Ελάχιστης Απόλυτης Απόκλισης (Least Absolute Deviation - LAD). Ωστόσο οι μηχανισμοί των δυο κριτηρίων δεν παρουσιάζουν σημαντικές διαφορές. Στη συνέχεια θα παρουσιάσουμε το κριτήριο LS. Σύμφωνα με το κριτήριο LS, η καθαρότητα του κόμβου(node impurity), υπολογίζεται από ένα άθροισμα τετραγώνων (sum of squares), εντός του κόμβου, που περιγράφεται ως εξής:

$$SS(t) = \sum \left(y_{i(t)} - \bar{y}_{(t)}\right)^2 , για\ i = 1,2, \dots, N_t$$

όπου $y_{i(t)} =$ οι τιμές της μεταβλητής απόκρισης στον κόμβο t, και $\bar{y}_{(t)} =$ ο μέσος της μεταβλητής απόκρισης στον κόμβο t. Σε κάθε διαχωρισμό των μεταβλητών s υπολογίζεται το SS(t) και από το διαχωρισμό ένα μέρος των περιπτώσεων μεταβαίνει στον αριστερό θυγατρικό κόμβο (t_L) ενώ το υπόλοιπο μέρος (t_R) στον δεξιό θυγατρικό κόμβο. Στη συνέχεια η αποτελεσματικότητα του διαχωρισμού (goodness-of-split), υπολογίζεται μέσω της συνάρτησης:

$$\varphi(s,t) = SS(t) - SS(t_L) - SS(t_R),$$

όπου το $SS(t_L)$ και $SS(t_R)$, το άθροισμα τετραγώνων του αριστερού και του δεξιού θυγατρικού κόμβου αντίστοιχα.

Έτσι ο καλύτερος διαχωρισμός, προκύπτει για το μεγαλύτερο φ(s,t). Από τους διαφορετικούς διαχωρισμούς που προκύπτουν από μια μεταβλητή σε κάθε

κόμβο, ο κανόνας είναι να επιλέγεται, ο διαχωρισμός εκείνος, που οδηγεί στη μεγαλύτερη μείωση της ακαθαρσίας του πρόγονου κόμβου(maximum reduction in the impurity of the parent node).

Σε αυτό το σημείο αξίζει να σημειωθεί ότι αντί για το SS(t), θα μπορούσε να χρησιμοποιηθεί η βεβαρυμμένη διασπορά των αριστερών και δεξιών κόμβων, όπου τα βάρη προκύπτουν από την αναλογία των περιπτώσεων στον αριστερό και το δεξιό κόμβο, έστω $p(t) = N_t/N$, όπου N_t ο αριθμός των περιπτώσεων που βρίσκονται στον κόμβο t ενώ N ο συνολικός αριθμός των περιπτώσεων. Έτσι, η διασπορά της μεταβλητής απόκρισης στον κόμβο t, δίνεται από τη σχέση:

$$s^2(t) = \frac{1}{N_t}\sum_{i=1}^{N_t}\left[y_i - \bar{y}_{(t)}\right]^2.$$

Έτσι η αποτελεσματικότητα του διαχωρισμού σε αυτή την περίπτωση υπολογίζεται με βάση τη σχέση:

$$\varphi(s,t) = s^2(t) - [p_L s^2(t_L) + p_R s^2(t_R)].$$

Ο καλύτερος διαχωρισμός είναι και σε αυτή την περίπτωση εκείνος για τον οποίο η τιμή της συνάρτησης $\varphi(s,t)$, είναι μέγιστη, δηλαδή η τιμή του $[p_L s^2(t_L) + p_R s^2(t_R)]$ είναι όσο το δυνατόν μικρότερη. Η διαδικασία, διαχωρίζει αποτελεσματικά τις υψηλές τιμές της μεταβλητής απόκρισης από τις χαμηλές, και τις μεταβιβάζει στους θυγατρικούς κόμβους (αριστερό και δεξή), με αποτέλεσμα σε αυτούς να υπάρχει μεγαλύτερη ομοιογένεια από ότι στον πρόγονο κόμβο από τον οποίο προήλθαν. Σε αυτό το σημείο αξίζει να σημειωθεί ότι καθώς κάθε διαχωρισμός, στέλνει παρατηρήσεις στον αριστερό και δεξιό κόμβο η μέση τιμή της μεταβλητής απόκρισης, είναι μικρότερη σε έναν από τους δυο θυγατρικούς κόμβους σε σχέση με τη μέση τιμή στον πρόγονο κόμβο από τον οποίο προήλθαν. Αφού αξιολογηθεί ο κάθε διαχωρισμός για κάθε επεξηγηματική μεταβλητή με την παραπάνω μέθοδο, επιλέγονται για την κατασκευή του δέντρου οι διαχωρισμοί που οδηγούν στα μικρότερα τετραγωνικά σφάλματα. Έτσι για ρίζα του δέντρου επιλέγεται ο διαχωρισμός εκείνος που οδηγεί στο μικρότερο τετραγωνικό σφάλμα (ανεξάρτητα από τη μεταβλητή την οποία αφορά). Το υπόλοιπο δέντρο χτίζεται με τον ίδιο τρόπο, επιλέγοντας κάθε φορά σαν ερώτηση-διαχωρισμό για κάθε θυγατρικό κόμβο που προκύπτει εκείνο που οδηγεί στο μικρότερο τετραγωνικό σφάλμα.

2.2.4. Κριτήρια Διακοπής (Stopping Criteria) & Διαδικασία Κλαδέματος (Pruning)

Στο σημείο αυτό, προκύπτει το ερώτημα, πότε η μέθοδος, σταματάει τη διαδικασία κατασκευής του δέντρου. Είναι φανερό ότι ένα πολύ μεγάλο δέντρο (με πολύ μεγάλο βάθος) θα μπορούσε να οδηγήσει σε υπερπροσαρμογή στα δεδομένα αφού οι τελικοί κόμβοι, τα φύλλα θα είχαν ελάχιστες (μια με δυο

παρατηρήσεις). Έτσι αν το μοντέλο προσπαθούσε να προβλέψει την τιμή νέων παρατηρήσεων, πιθανώς θα είχε μεγάλο σφάλμα καθώς θα βασιζόταν σε πολύ λίγες και συγκεκριμένες παρατηρήσεις για να κάνει την πρόβλεψη του. Για το λόγο αυτό εφαρμόζονται κάποια κριτήρια διακοπής για το πότε να σταματήσει η διαδικασία κατασκευής του δέντρου. Το πιο συνηθισμένο και αποτελεσματικό κριτήριο, είναι το κριτήριο των ελάχιστων παρατηρήσεων ανά κόμβο (minimum samples leaf), με το οποίο υποχρεώνουμε τη διαδικασία να σταματήσει το διαχωρισμό όταν οι παρατηρήσεις που έχουν μεταβεί σε ένα κόμβο είναι λιγότερες από ένα όριο (threshold). Το όριο αυτό συνήθως είναι 10 με 20 παρατηρήσεις. Ένα άλλο κριτήριο τέλος, είναι το μέγιστο βάθος του δέντρου. Οι βέλτιστες τιμές των κριτηρίων αυτών, μπορούν να προσδιοριστούν πειραματικά ελέγχοντας μέχρι πιο βάθος η ακρίβεια προβλέψεων του δέντρου για τα δεδομένα εκπαίδευσης και επαλήθευσης είναι παρόμοια.

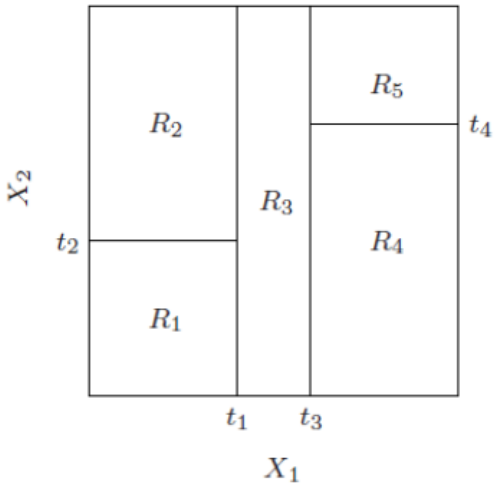
Η μέθοδος εφαρμόζει άλλη μια τεχνική που συμβάλλει στην αποφυγή του overfitting, το κλάδεμα. Αφού η μέθοδος κατασκευάσει το μεγαλύτερο δυνατό δέντρο, εφαρμόζει έναν αλγόριθμο κλαδέματος, πραγματοποιώντας cross-validation ή independent test sample ώστε να υπολογίσει την καλή προσαρμογή-αποτελεσματικότητα του δέντρου. Πρακτικά, το κλάδεμα αφαιρεί είτε τους τελικούς κόμβους που έχουν πολύ λίγες παρατηρήσεις είτε αυτούς που αν αφαιρεθούν ή μειώνουν το συνολικό τετραγωνικό σφάλμα ή δε δημιουργούν καμία διαφορά σε αυτό είτε τέλος τους τελικούς κόμβους για τους οποίους η p-value του στατιστικού ελέγχου με μηδενική υπόθεση ότι οι παρατηρήσεις των κόμβων αυτών ανήκουν σε ίδιους πληθυσμούς είναι <0.05 (κάτι που σημαίνει ότι οι παρατηρήσεις αυτές δε θα έπρεπε να έχουν διαχωριστεί) . Η LS χρησιμοποιεί το μέσο τετραγωνικό σφάλμα (Mean Squared Error - MSE) για να μετρήσει την ακρίβεια των προβλέψεων, με σκοπό να ταξινομήσει τα δέντρα που προέκυψαν από το κλάδεμα με βάση την αποτελεσματικότητα τους και την ακρίβεια τους. Στη συνέχεια το δέντρο με το μικρότερο MSE (μεγαλύτερη ακρίβεια), επιλέγεται ως βέλτιστο. (Στο σημείο αυτό σε μερικές περιπτώσεις εφαρμόζεται στο βέλτιστο δέντρο που προέκυψε ο κανόνας one-standard-error ώστε να προκύψει το βέλτιστο δέντρο). Αν ωστόσο είχαμε χρησιμοποιήσει αντί για τον κανόνα LS, τον LAD, τότε η μέθοδος θα χρησιμοποιούσε τη Μέση Απόλυτη Απόκλιση(Mean Absolute Deviation - MAD) για να βρει το βέλτιστο δέντρο.

Αφού επιλεχθεί το βέλτιστο δέντρο που δημιουργήθηκε από τη διαδικασία του κλαδέματος, η CART, υπολογίζει συνολικά στατιστικά για κάθε τερματικό κόμβο (φύλλο). Συγκεκριμένα στην LS, η CART, υπολογίζει τη μέση τιμή(mean) και την τυπική απόκλιση(standard deviation) της μεταβλητής απόκρισης. Η μέση τιμή του τελικού κόμβου, αποτελεί την πρόβλεψη για τη μεταβλητή απόκρισης, για τις παρατηρήσεις που βρίσκονται στο συγκεκριμένο τελικό κόμβο. Στην LAD, η CART, υπολογίζει τη διάμεσο(median) και τη μέση απόλυτη απόκλιση(average mean absolute deviation). Στη συνέχεια όπως και στην LS, η διάμεσος αποτελεί

την πρόβλεψη για τη μεταβλητή απόκρισης για τις παρατηρήσεις του συγκεκριμένου κόμβου. Οι προβλέψεις αυτές είναι οι τελικές προβλέψεις του μοντέλου για το σύνολο των δεδομένων που του δώσαμε.

2.2.5. Παράδειγμα Κατασκευής Δέντρου Παλινδρόμησης με τη χρήση της CART

Ας δούμε λοιπόν ένα παράδειγμα για να κατανοήσουμε καλύτερα τη μέθοδο που περιγράφηκε παραπάνω. Έστω ότι έχουμε ένα πρόβλημα παλινδρόμησης με μια μεταβλητή απόκρισης Y (συνεχή) και δυο επεξηγηματικές μεταβλητές X_1, X_2. Στο παρακάτω σχήμα βλέπουμε ένα διαχωρισμό του χώρου των μεταβλητών, που προέκυψε από αναδρομικά διαδοχικά χωρίσματα, ως εξής: αρχικά χωρίζεται ο χώρος σε δυο περιοχές και στη συνέχεια μοντελοποιείται η μεταβλητή απόκρισης από τη μέση τιμή του Y στην κάθε περιοχή. Στη συνέχεια , η μια ή και οι δυο αυτές περιοχές χωρίζονται σε δυο ή περισσότερες περιοχές η κάθε μια, αφού πρώτα επιλεχθεί το κατάλληλο διαχωριστικό κριτήριο και το διαχωριστικό σημείο για την καλύτερη προσαρμογή. Η διαδικασία επαναλαμβάνεται μέχρι την εφαρμογή κάποιου κριτηρίου διακοπής.



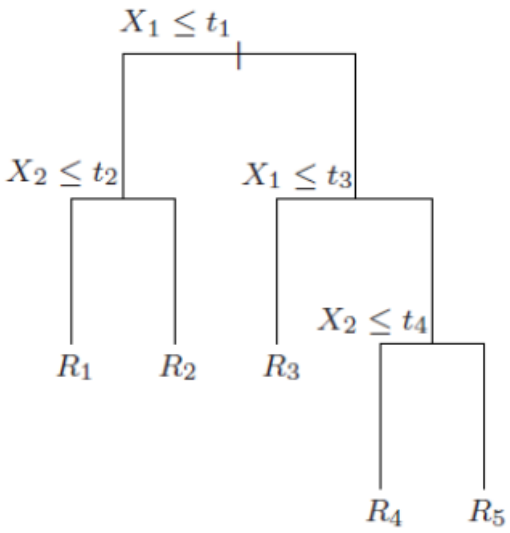
Σχήμα: Γράφημα απεικόνισης διαμέρισης του δισδιάστατου χώρου επεξηγηματικών μεταβλητών, με αναδρομική δυαδική διάσπαση.

Στο παραπάνω σχήμα λοιπόν βλέπουμε, ότι πρώτα χωρίζεται η περιοχή $X_1 = t_1$. Στη συνέχεια, η περιοχή $X_1 \leq t_1$, χωρίζεται σε $X_2 = t_2$ και η περιοχή $X_1 > t_1$, χωρίζεται σε $X_1 = t_3$. Τέλος, η περιοχή $X_1 > t_3$, χωρίζεται σε $X_2 \leq t_4$.

Το αποτέλεσμα της διαδικασίας, είναι η διαμέριση σε πέντε περιοχές R_1, R_2, R_3, R_4, R_5. Το αντίστοιχο μοντέλο παλινδρόμησης προβλέπει την Y με μια σταθερά c_m στην περιοχή R_m, που εκφράζεται ως:

$$\hat{f}(X) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \ \epsilon R_m\}$$

Το παραπάνω μοντέλο μπορεί να περιγραφεί από το παρακάτω δυαδικό δέντρο απόφασης.



Σχήμα : Απεικόνιση Δέντρου απόφασης για το παραπάνω παράδειγμα.

Το σύνολο εκπαίδευσης βρίσκεται στην κορυφή-ρίζα του δέντρου. Οι παρατηρήσεις που ικανοποιούν τη συνθήκη σε κάθε κόμβο μεταβαίνουν στον αριστερό κλάδο, και οι άλλες στη δεξιά διακλάδωση. Οι τερματικοί κόμβοι ή τα φύλλα του δένδρου αντιστοιχούν στις περιοχές.

Ας γενικεύσουμε όμως τώρα το παράδειγμα και ας υποθέσουμε ότι έχουμε p επεξηγηματικές μεταβλητές, με μεταβλητή απόκρισης την Y και N παρατηρήσεις, δηλαδή έχουμε N ζεύγη παρατηρήσεων (x_i, y_i) με $x_i = (x_{i1}, \dots, x_{ip})$ $για$ $i = 1, \dots, N$. Η μέθοδος CART αποφασίζει αυτόματα σχετικά με το διαχωρισμό των μεταβλητών και τα διαχωριστικά σημεία. Έτσι αν πραγματοποιηθεί διαχωρισμός σε M περιοχές $R_1, \dots, R_M$, η απόκριση μοντελοποιείται ως μια σταθερά c_m σε κάθε περιοχή, με τη μορφή :

$$\hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \ \epsilon \ R_m\}$$

Με κριτήριο λοιπόν την ελαχιστοποίηση του αθροίσματος τετραγώνων:

$$\sum (y_i - f(x_i)^2$$

Βλέπουμε ότι η καλύτερη $\widehat{c_m}$ για κάθε περιοχή, είναι ο μέσος όρος των y_i στην περιοχή R_m:

$$\widehat{c_m} = average(y_i | x_i \epsilon \ R_m)$$

Ωστόσο λόγω του ότι η καλύτερη δυαδική διαμέριση, όσον αφορά το άθροισμα ελαχίστων τετραγώνων είναι υπολογιστικά ανέφικτη, ειδικά όταν οι διαστάσεις του προβλήματος είναι αρκετά μεγάλες, η μέθοδος προχωράει χρησιμοποιώντας έναν άπλειστο αλγόριθμο. Ξεκινώντας με όλο το σύνολο εκπαίδευσης, εξετάζουμε μια διασπασμένη μεταβλητή j και το σημείο διαχωρισμού s, και καθορίζουμε το ζεύγος των ημι-επιπέδων:

$$R_1(j,s) = \{X | X_j \leq s\} \quad \& \quad R_2(j,s) = \{X | X_j > s\}$$

Στη συνέχεια αναζητάμε τη διασπασμένη μεταβλητή(splitting variable) j και το σημείο διαχωρισμού (split point) s, που ικανοποιούν τη σχέση:

$$\min_{j,s}\{\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \}$$

Για κάθε επιλογή j & s, η εσωτερική ελαχιστοποίηση, ικανοποιείται, για :

$$\widehat{c_1} = average(y_i | x_i \epsilon R_1(j,s)) \ \& \ \widehat{c_2} = average(y_i | x_i \epsilon R_2(j,s))$$

Με τον τρόπο αυτό επιλέγεται το βέλτιστο ζεύγος διαχωρισμού (j, s). Αφού αυτό βρεθεί, τα δεδομένα διαχωρίζονται στις δυο περιοχές και η διαδικασία αυτή του διαχωρισμού επαναλαμβάνεται στις δυο περιοχές που θα προκύψουν και σε κάθε άλλη περιοχή που θα προκύψει στη συνέχεια.

Αφού κατασκευαστεί το δέντρο παλινδρόμησης και ο αλγόριθμος τερματιστεί, εφαρμόζεται η μέθοδος κλαδέματος ώστε να καθορίσει για το δέντρο το βέλτιστο βάθος-μέγεθος. Συγκεκριμένα ας θεωρήσουμε, ένα υποδέντρο $T \subset T_0$, ως ένα δέντρο που προκύπτει από το κλάδεμα του T_0. Συμβολίζουμε τους τερματικούς κόμβους με το δείκτη m, με τον κόμβο m να αντιπροσωπεύει την περιοχή R_m. Έστω |T| ο αριθμός των τερματικών κόμβων στο T. Θέτουμε :

$$\widehat{c_m} = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

$$Q_m = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \widehat{c_m})^2$$

Ακόμα, ορίζουμε το κριτήριο κόστους περιπλοκότητας του κλαδέματος(cost complexity criterion)

$$C_a(T) = \sum_{m=1}^{|T|} N_m \ Q_m(T) \ + a|T|$$

Στη συνέχεια, για κάθε α, βρίσκουμε το υποδέντρο $T_\alpha \subset T_0$, για την ελαχιστοποίηση του $C_a(T)$. Η ρυθμιστική παράμετρος $\alpha \geq 0$ έρχεται να λύσει το δίλημμα μεταξύ του μεγέθους του δέντρου και της καλής προσαρμογής του στα δεδομένα. Μεγάλες τιμές του α, οδηγούν σε μικρότερα δέντρα T_α και το αντίστροφο. Ενώ για α=0, λαμβάνουμε το αρχικό δέντρο πλήρους μεγέθους.

Για κάθε α είναι εύκολο να δείξουμε ότι υπάρχει ένα μοναδικό μικρότερο υπόδεντρο T_α που ελαχιστοποιεί το $C_a(T)$. Για να το βρούμε, χρησιμοποιούμε το κλάδεμα του πιο αδύναμου κρίκου (weakest link pruning): έχουμε διαδοχικά συμπτύξει τον εσωτερικό κόμβο που παράγει τη μικρότερη ανά κόμβο αύξηση στο άθροισμα $\sum_{m=1}^{|T|} N_m \ Q_m(T)$, και συνεχίζουμε μέχρι να παράξουμε το δέντρο με έναν κόμβο (ρίζα). Αυτό δίνει μία (πεπερασμένη) ακολουθία από υπόδενδρα η οποία περιέχει το T_α [Βλ. Breiman(1984) για περισσότερες λεπτομέρειες στην απόδειξη αυτών]. Η εκτίμηση του α επιτυγχάνεται με πέντε- ή δέκα- φορές διασταυρωμένη επικύρωση (5 or 10-Fold Cross Validation): επιλέγουμε την τιμή $\hat{\alpha}$ που ελαχιστοποιεί το άθροισμα των τετραγώνων της διασταυρωμένης επικύρωσης. Το τελικό δέντρο που προκύπτει είναι το $T_{\hat{\alpha}}$.

2.3. Μέθοδοι Ενδυνάμωσης (Boosting)

Η ιδέα του boosting, προέκυψε ως μια τεχνική για βελτίωση μοντέλων στο χώρο της επιβλεπόμενης εκμάθησης (supervised learning). Με τον όρο supervised learning, εννοούμε, την αυτόματη εκπαίδευση ενός αλγορίθμου, που βασίζεται σε δεδομένα που έχουν παρατηρηθεί και ξέρουμε την τιμή απόκρισης τους, με σκοπό την έγκυρη και ακριβή πρόβλεψη για δεδομένα που δεν γνωρίζουμε την τιμή απόκρισης τους. Η βασική ερώτηση που άνοιξε δρόμο για την ανάπτυξη της ιδέας του boosting, ήταν κατά πόσο ένα αδύναμο εργαλείο εκπαίδευσης θα μπορούσε να μετατραπεί σε ένα ισχυρό εργαλείο εκπαίδευσης. Αδύναμο εργαλείο, θα θεωρούσαμε ένα μοντέλο που προβλέπει με 50% ακρίβεια την τιμή μιας μεταβλητής απόκρισης ενώ ισχυρό ένα μοντέλο που προβλέπει με 99% ακρίβεια. Το γεγονός όμως, ότι κάθε αδύναμος εκτιμητής μπορεί να βελτιωθεί (boosted) επαναληπτικά, και να μετατραπεί σε ισχυρό εκτιμητή, έκανε τους ερευνητές Schaphire και Freund να θεωρούν ότι θα μπορούσε να υπάρξει μια μέθοδος που να ικανοποιεί την παραπάνω ιδέα.

Ο σκοπός λοιπόν του αλγορίθμου boosting, είναι να συνδυάσει την πληροφορία και το αποτέλεσμα που προκύπτει από μια επαναληπτική εκπαίδευση ενός αδύναμου εκτιμητή, ώστε να κατασκευάσει έναν ισχυρό εκτιμητή. Ωστόσο, το να εκπαιδεύεις έναν αδύναμο εκτιμητή με τα ίδια δεδομένα πολλές φορές (επαναληπτικά), δεν μπορεί να προκαλέσει αλλαγές στην απόδοση του. Για το λόγο αυτό η μέθοδος boosting κάνει κάτι λίγο πιο εξεζητημένο, σε κάθε επανάληψη αντί να προσπαθεί να βελτιώσει τον ίδιο τον εκτιμητή, δίνει διαφορετικά βάρη στις παρατηρήσεις που περιλαμβάνουν τα δεδομένα εκπαίδευσης, με αποτέλεσμα σε κάθε επανάληψη m, να βρίσκει μια καινούρια

λύση $\hat{h}^{[m]}$ από τα δεδομένα. Ακόμη η μέθοδος, σε κάθε επανάληψη επιβαρύνει περισσότερο(με μεγαλύτερο βάρος) και επικεντρώνεται, στις παρατηρήσεις στις οποίες το μοντέλο-εκτιμητής παρουσιάζει μικρότερη ακρίβεια-σφάλμα, δηλαδή είναι πιο δύσκολο να τις προβλέψει. Έτσι, σε κάθε επανάληψη $m = 1, \dots, m_{stop}$, το διάνυσμα $\boldsymbol{w} = w_1^{[m]}, \dots, w_n^{[m]}$,περιέχει τα ξεχωριστά βάρη όλων των παρατηρήσεων που βασίζονται στο πόσο εύκολα ή όχι έγινε η πρόβλεψη της συγκεκριμένης παρατήρησης. Στο τέλος των επαναλήψεων, όλα τα αποτελέσματα των επαναλήψεων του αδύναμου εκτιμητή, συνδυάζονται, για την παραγωγή μιας πιο ακριβούς πρόβλεψης, δηλαδή ενός ισχυρού εκτιμητή.

2.4. Adaboost

Η μέθοδος Adaboost (Adaptive Boosting), είναι μια boosting μέθοδος που αναπτύχθηκε από τους Saphire και Freund. Η ιδέα πίσω από τη μέθοδο της Adaboost είναι η εξής: κάθε παρατήρηση (instance) των δεδομένων εκπαίδευσης, λαμβάνει ένα αρχικό βάρος w_i, το οποίο συμβολίζει τη σχετική σημαντικότητα της. Στη συνέχεια κατασκευάζεται ο πρώτος αδύναμος εκτιμητής (π.χ. δέντρο απόφασης), και υπολογίζεται το προσαρμοσμένο σφάλμα της κάθε παρατήρησης ξεχωριστά καθώς και το συνολικό σφάλμα ως το βεβαρυμμένο άθροισμα των επιμέρους σφαλμάτων. Τέλος, υπολογίζεται ο όρος β^t που καθορίζει το πόσο θα συνεισφέρει ο εκτιμητής αυτός στο τελικό αποτέλεσμα και έτσι τελειώνει μια επανάληψη. Μετά από κάθε επανάληψη της μεθόδου, τα βάρη των παρατηρήσεων αναπροσαρμόζονται και οι παρατηρήσεις για τις οποίες το μοντέλο είχε μεγαλύτερο σφάλμα-απόκλιση επιβαρύνονται περισσότερο. Ο επόμενος εκτιμητής που θα κατασκευαστεί θα εκπαιδευτεί σε ένα διαφορετικό σύνολο δεδομένων από ότι ο πρώτος. Παρόλο που το μέγεθος του συνόλου δεδομένων θα είναι ίδιο, οι παρατηρήσεις που θα περιλαμβάνει αυτό επιλέγονται ως εξής: κάθε μία παρατήρηση επιλέγεται ως τυχαίο δείγμα από την κατανομή των βαρών των παρατηρήσεων. Με τον τρόπο αυτό παρατηρήσεις με μεγάλο βάρος που δεν εκτιμήθηκαν με μεγάλη ακρίβεια από τον πρώτο εκτιμητή είναι πιο πιθανό να επιλεγούν για τον καινούργιο εκτιμητή σε σχέση με παρατηρήσεις που εκτιμήθηκαν με μεγαλύτερη ακρίβεια. Έτσι ο καινούργιος εκτιμητής θα περιέχει σε μεγαλύτερο ποσοστό παρατηρήσεις που δεν εκτιμήθηκαν σωστά από τον πρώτο εκτιμητή (κάποιες παρατηρήσεις είναι πολύ πιθανό να υπάρχουν περισσότερες από μία φορές). Τα βάρη των παρατηρήσεων στη συνέχεια αρχικοποιούνται ξανά. Έτσι το σφάλμα του κάθε εκτιμητή επηρεάζει την κατασκευή του επόμενου και η διαδικασία συνεχίζεται με κάθε δέντρο να κατασκευάζεται με την πληροφορία που παίρνει από το προηγούμενο με αποτέλεσμα η διαδικασία της μάθησης να επικεντρώνεται στις παρατηρήσεις που είναι πιο δύσκολο να προβλεφτούν με ακρίβεια. Οι τελικές προβλέψεις προκύπτουν ως βεβαρημένη διάμεσος των προβλέψεων του κάθε δέντρου, με βάρη για κάθε δέντρο το $\ln\left(\frac{1}{\beta_t}\right)$.

Η εκτίμηση της μεθόδου για μια παρατήρηση x_i δίνεται από μια υπόθεση h_t, και το σφάλμα της πρόβλεψης ορίζεται ως $e_i = |y_i - h_t(x_i)|$. Ωστόσο η μέθοδος για να εφαρμόσει τις αναπροσαρμογές στα βάρη των παρατηρήσεων υπολογίζει ένα προσαρμοσμένο σφάλμα e_i' το οποίο, καλείται συνάρτηση κόστους(loss function) και εκφράζει τη σχέση κάθε σφάλματος e_i, με το μέγιστο σφάλμα $D = max_i\,|e_i|\,, για\ i = 0, \dots, n\ όπου\ n\ ο\ αριθμός\ των\ παρατηρήσεων$, παίρνοντας τιμές, στο [0,1]. Συγκεκριμένα, υπάρχουν τρεις συναρτήσεις κόστους που χρησιμοποιούνται και είναι οι παρακάτω: $e_i' = \frac{e_i}{D}\ γραμμική(linear)$, $e_i' = \frac{e_i^2}{D^2}\ τετραγωνική(square)$, $e_i' = 1 - exp\left(-\frac{e_i}{D}\right)\ εκθετική(exponential)$. Ο βαθμός τώρα στον οποίο οι παρατηρήσεις x_i επιβαρύνονται σε κάθε επανάληψη t, εξαρτάται από το πόσο μεγάλο είναι το σφάλμα της h_t για τη x_i σε σχέση με το σφάλμα της χειρότερης παρατήρησης.

2.4.1. Αλγόριθμος

Ο αλγόριθμος της Adaboost είναι ο ακόλουθος:

Δεδομένα (Input): Το σύνολο των δεδομένων T, μεγέθους n, ο μέγιστος αριθμός επαναλήψεων N, και ένας βασικός αλγόριθμος εκμάθησης που καλούμε ως «Learner» και αποτελεί τον αδύναμο εκτιμητή (π.χ. δέντρο απόφασης). Αν δεν προσδιοριστούν τα αρχικά βάρη, θεωρούμε το αρχικό διάνυσμα βαρών w^1, τέτοιο ώστε $w_i^1 = \frac{1}{n}\ για\ 1 \le i \le n$.

Για $t = 1, \dots, N$:

1. Εκπαίδευσε τον Learner με το σύνολο εκπαίδευσης T, χρησιμοποιώντας τα βάρη w^t και κατασκεύασε την υπόθεση $h_t: X \to R$
2. Υπολόγισε το προσαρμοσμένο σφάλμα $e_i'^t$, για κάθε παρατήρηση, υπολογίζοντας πρώτα το $D_t = max_j\left|y_j - h_t\left(x_j\right)\right|, για\ j = 1, \dots, n$ και στη συνέχεια το $e_i'^t = |y_ι - h_t(x_i)|/D_t$
3. Υπολόγισε το προσαρμοσμένο σφάλμα της h_t: $e_t = \sum_{i=1}^{n} e_i^t w_i^t\,, και\ αν\ e_i^t \ge 0.5\ σταμάτα\ και\ θέσε\ N = t - 1$
4. Θέσε $\beta^t = \frac{e_t}{1-e_t}$.
5. Ανανέωσε το διάνυσμα βαρών:

$$w_i^{t+1} = \frac{w_i^t \beta_t^{1-e_i^t}}{Z_t}, όπου\ Z_t\ \ μια\ σταθερά\ κανονικοποίησης$$

Αποτελέσματα(Output): Η υπόθεση :

$$h_f(x) = η\ επιβαρυμένη\ διάμεσος\ της\ h_t(x), για\ 1 \le t \le N,$$
$$χρησιμοποιώντας\ το \ln\left(\frac{1}{\beta_t}\right)\ ως\ βάρος\ για\ την\ υπόθεση\ h_t.$$

Στο σημείο αυτό αξίζει να σημειωθεί πως η Adaboost σε γενικές γραμμές είναι μια μέθοδος που δίνει πολύ καλύτερα αποτελέσματα σε σχέση με μια απλή μέθοδο παλινδρόμησης. Ακόμη, ως βασικό εκτιμητή μπορούμε να χρησιμοποιήσουμε οποιοδήποτε μοντέλο εκπαίδευσης επιθυμούμε. Συνήθως η Adaboost υλοποιείται με δέντρα ταξινόμησης ως βασικούς εκτιμητές, και συνήθως για μικρό βάθος αφού μην ξεχνάμε ότι ο βασικός εκτιμητής πρέπει να είναι και αδύναμος. Τέλος, πρέπει να σημειωθεί, ότι το βασικό μειονέκτημα της Adaboost, είναι το γεγονός, ότι επειδή επικεντρώνεται στις παρατηρήσεις που είναι πιο δύσκολο να προβλεφτούν με ακρίβεια, που συνήθως είναι ακραίες τιμές(outliers), παρουσιάζει χαμηλή απόδοση σε δεδομένα με πολύ θόρυβο(noise), δηλαδή δεδομένα με πολλές ακραίες τιμές ή τιμές που δεν παρέχουν καμία πληροφορία για το πρόβλημα που προσπαθούμε να επιλύσουμε.

2.5. Gradient Boosting

Η μέθοδος Gradient Boosting είναι μια επαναληπτική μέθοδος που όπως και άλλες boosting μέθοδοι, βασίζεται στο συνδυασμό αδύναμων εκτιμητών (weak learners) για την κατασκευή ενός ισχυρού εκτιμητή. Ας υποθέσουμε ότι διαθέτουμε για την κατασκευή ενός μοντέλου παλινδρόμησης, ένα δείγμα (σύνολο εκπαίδευσης – training set) της μορφής $(y_i, \boldsymbol{x}_i), i = 1 \dots N$, όπου y μια μεταβλητή απόκρισης (response) και $\mathbf{x}=\{x_1, x_2, \dots, x_k\}$ ένα σύνολο επεξηγηματικών μεταβλητών. Σκοπός μας, είναι να δημιουργήσουμε μια συνάρτηση $F(x)$, που αντιστοιχίζει το $\mathbf{x}$ με το y, έτσι ώστε μετά από αυτή την αντιστοίχιση να ελαχιστοποιείται μια συγκεκριμένη συνάρτηση κόστους (loss function), όπως το μέσο τετραγωνικό σφάλμα: $\frac{1}{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\{\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}\}^2$, όπου n ο αριθμός των συνολικών παρατηρήσεων που διαθέτουμε στο σύνολο εκπαίδευσης, $\mathrm{y_i}$**:** οι εκτιμήσεις της μεταβλητής απόκρισης που έδωσε η $F(x)$, για δεδομένα $\mathbf{x}$, και $\hat{\mathrm{y}}_\mathrm{i}$, οι πραγματικές τιμές των εκτιμήσεων αυτών. Σε κάθε επανάληψη m, όπου $1 \leq m \leq M$, η μέθοδος, υποθέτει ότι υπάρχει ένα αδύναμο – με μικρή ακρίβεια(μεγάλη απόκλιση), μοντέλο F_m, που προβλέπει τη μεταβλητή y, και προχωράει στη βελτίωση του μοντέλου αυτού, κατασκευάζοντας ένα νέο καλύτερο με μικρότερο σφάλμα μοντέλο F_{m+1} χάρη σε έναν επιπλέον εκτιμητή h και ορίζεται όπως παρακάτω: $F_{m+1}(x) = F_m(x) + h(x)$. Ως καλύτερη επιλογή για το h θεωρείται εκείνη για την οποία ισχύει:

$$F_{m+1}(x) = F_m(x) + h(x) = y \text{ ή } \textit{ισοδύναμα } h(x) = y - F_m(x).$$

Συνεπώς, η gradient boosting προσαρμόζει τον όρο h με βάση το υπόλοιπο $y - F_m(x)$, έτσι ώστε κάθε όρος F_{m+1} να προσπαθεί να διορθώσει το σφάλμα του προηγούμενου όρου F_m, βασιζόμενη στην ιδέα ότι τα υπόλοιπα $y - F_m(x)$,

είναι αρνητικές κλίσεις του τετραγωνικού σφάλματος (συνάρτηση κόστους – loss function), $\frac{1}{2}\left(y - F_m(x)\right)^2$. Έτσι η gradient boosting βασίζεται στην έννοια της απότομης καθόδου (gradient descent).

Πιο αναλυτικά, η μέθοδος gradient boosting αναζητάει μια προσέγγιση $\hat{F}(x)$, στη μορφή βεβαρυμμένων αθροισμάτων από συναρτήσεις h, από μια κλάση H που περιλαμβάνει κάποιους αδύναμους εκτιμητές (weak learners), και έχει τη μορφή:

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + const$$

Έτσι η μέθοδος προσπαθεί να βρει μια προσέγγιση $\hat{F}$ που ελαχιστοποιεί τη μέση τιμή μιας συνάρτησης κόστους για το σύνολο εκπαίδευσης και το πετυχαίνει αυτό, ξεκινώντας από ένα μοντέλο που περιλαμβάνει μια σταθερή συνάρτηση $F_0(x)$, και σταδιακά το επεκτείνει ακολουθώντας μια άπλειστη τεχνική (greedy algorithm):

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

$$F_m(x) = F_{m+1}(x) + \arg\min_{\mathrm{h_m} \epsilon \mathrm{H}} [\sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h_m(x_i))]$$

Πρακτικά όμως η εύρεση της βέλτιστης συνάρτησης h σε κάθε βήμα, για μια συνάρτηση κόστους L, είναι ένα πρόβλημα υπολογιστικά αδύνατο. Έτσι καθίσταται επιτακτική η ανάγκη για την εύρεση ενός πιο αποτελεσματικού και πρακτικού τρόπου για την επίλυση του παραπάνω προβλήματος βελτιστοποίησης, η τεχνική της απότομης καθόδου (gradient descent) φαίνεται να αποτελεί την ιδανική λύση. Αν θεωρήσουμε την περίπτωση που οι H, είναι συνεχείς διαφορικές συναρτήσεις στο R, θα ανανεώναμε το μοντέλο σύμφωνα με τις παρακάτω σχέσεις :

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L\left(y_i, F_{m-1}(x_i)\right)$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - \gamma \nabla_{F_{m-1}} L\left(y_i, F_{m-1}(x_i)\right))$$

Ωστόσο στη διακριτή περίπτωση, όπου το σύνολο H είναι πεπερασμένο, επιλέγουμε τη συνάρτηση h, όσο πιο κοντά στην κλίση(gradient) του L, για την οποία ο συντελεστής γ, μπορεί να υπολογιστεί με την εφαρμογή της μεθόδου line search [36], στις παραπάνω εξισώσεις.

2.5.1. Αλγόριθμος:

Ο αλγόριθμος της gradient boosting περιγράφεται όπως παρακάτω:
Δεδομένα (Input): Το σύνολο των δεδομένων εκπαίδευσης $\{(x_i, y_i)\}$, για $i = 1, \dots, n$, μια παραγωγίσιμη συνάρτηση κόστους $L(y, F(x))$, ο αριθμός επαναλήψεων M.

Αλγόριθμος:

1. Αρχικοποίησε το μοντέλο με μια σταθερή τιμή:
   $F_0(x) = \arg\min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$.
2. Για m=1 μέχρι M:
    1. Υπολόγισε τα ψευδοϋπόλοιπα (pseudo-residuals):
       $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_m-\mathbf{1}(x)}$ $για\ i = 1, \dots, n.$
    2. Προσάρμοσε ένα βασικό αδύναμο εκτιμητή (π.χ. δέντρο), $h_m(x)$, στα ψευδοϋπόλοιπα και εκπαίδευσε τον με το σύνολο εκπαίδευσης $\{(x_i, r_{im})\}$, για $i = 1, \dots, n$.
    3. Υπολόγισε τον όρο γ_m, λύνοντας το παρακάτω πρόβλημα βελτιστοποίησης:
       $$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L\big(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\big).$$
    4. Ανανέωσε το μοντέλο: $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.

Αποτελέσματα(Output): $F_M(x)$.

2.5.2. Gradient Boosting για Δέντρα

Η gradient boosting εφαρμόζεται κατά κύριο λόγο σε δέντρα απόφασης (κυρίως σε CART δέντρα), συγκεκριμένου βάθους, τα οποία χρησιμοποιούνται ως βασικοί εκτιμητές(base learners). Για τη συγκεκριμένη εφαρμογή της Gradient Boosting, ο Friedman [16], πρότεινε μια μετατροπή στον αλγόριθμο της μεθόδου, που βελτιώνει την ποιότητα της προσαρμογής στα δεδομένα, του κάθε βασικού εκτιμητή. Γενικά, η gradient boosting, στο m-βήμα, προσαρμόζει ένα δέντρο απόφασης (decision tree) $h_m(x)$, στα ψευδο-υπόλοιπα. Έστω J_m, ο αριθμός των φύλλων του δέντρου. Το δέντρο χωρίζει το σύνολο των δεδομένων σε J_mδιακριτές περιοχές $R_{1m}, \dots, R_{J_m m}$, και εκτιμάει-προβλέπει μια σταθερή τιμή σε κάθε περιοχή. Χρησιμοποιώντας μια δείκτρια συνάρτηση, το αποτέλεσμα του $h_m(x)$, για δεδομένο x, μπορεί να γραφτεί ως το παρακάτω άθροισμα:

$$h_m(x) = \sum_{j=1}^{J_m} b_{jm} \mathbf{1}_{R_{jm}}(x)$$

όπου $\mathbf{1}_{R_{jm}}(x) = \begin{cases} 1\ \alpha\nu\ x \in R_{jm} \\ 0\ \alpha\nu\ x \notin R_{jm} \end{cases}$, b_{jm} είναι η τιμή που εκτιμάται-προβλέπεται στην περιοχή R_{jm}. Στη συνέχεια, οι συντελεστές b_{jm}, πολλαπλασιάζονται με μια τιμή γ_m, που επιλέγεται με βάση τη μέθοδο line search [36], έτσι ώστε να ελαχιστοποιείται η συνάρτηση κόστους, και το ανανεωμένο μοντέλο προκύπτει ως εξής:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} b_{jm}\mathbf{1}_{R_{jm}}(x)$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Στην παραπάνω λοιπόν διαδικασία, ο Friedman, πρότεινε μια αλλαγή, έτσι ώστε ο αλγόριθμος να επιλέγει διαφορετική βέλτιστη τιμή γ_{jm} για κάθε μια από τις περιοχές του δέντρου, αντί για την ίδια γ_m σε κάθε περιοχή για όλο το δέντρο. Οι συντελεστές b_{jm} που προκύπτουν από τη διαδικασία προσαρμογής του δέντρου μπορούν να απορριφθούν, και το τελικό μοντέλο προκύπτει από τα παρακάτω βήματα:

$$F_m(x) = F_{m-1}(x) - \sum_{j=1}^{J_m} \gamma_{jm}\mathbf{1}_{R_{jm}}(x)$$

$$\gamma_m = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

2.5.3. Συντελεστής Συρρίκνωσης (Shrinkage)

Ένα σημαντικό κομμάτι της μεθόδου gradient boosting, αποτελεί η κανονικοποίηση (regularization) μέσω της συρρίκνωσης(shrinkage), που περιλαμβάνει τη μετατροπή του μοντέλου με τον παρακάτω τρόπο:

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm}\mathbf{1}_{R_{jm}}(x), \quad 0 < \nu \leq 1$$

όπου η παράμετρος ν, ονομάζεται βαθμός εκμάθησης(learning rate). Εμπειρικά, έχει παρατηρηθεί ότι μοντέλα της gradient boosting στα οποία έχει εφαρμοστεί συρρίκνωση για μικρούς βαθμούς εκμάθησης (όπως ν<0.1) , τείνουν να είναι αρκετά πιο γενικευμένα σε σχέση με μοντέλα στα οποία δεν έχει εφαρμοστεί συρρίκνωση(ν=1). Με άλλα λόγια η συρρίκνωση για σχετικά μικρά ν, συμβάλλει στην αποφυγή του φαινομένου της υπερπροσαρμογής (overfitting) στα δεδομένα εκπαίδευσης. Έτσι ένα μοντέλο στο οποίο έχει εφαρμοστεί

συρρίκνωση, μπορεί να είναι πολύ πιο αποδοτικό σε καινούρια δεδομένα από ότι ένα μοντέλο στο οποίο δεν έχει εφαρμοστεί συρρίκνωση. Ωστόσο, αξίζει να σημειωθεί ότι η εφαρμογή της συρρίκνωσης σε ένα μοντέλο gradient boosting, οδηγεί σε αύξηση του υπολογιστικού χρόνου(computational time), αφού χαμηλοί βαθμοί εκμάθησης απαιτούν περισσότερες επαναλήψεις. Τέλος, αξίζει να σημειωθεί ότι δυο ακόμα σημαντικές παράμετροι που επηρεάζουν ένα μοντέλο παλινδρόμησης gradient boosting, είναι το μέγεθος των δέντρων (βάθος) και ο αριθμός των δέντρων που θα κατασκευαστούν (αριθμός αδύναμων εκτιμητών). Δεν πρέπει να ξεχνάμε ότι η gradient boosting, είναι μια boosting μέθοδος, δηλαδή συνδυάζει την πληροφορία αδύναμων εκτιμητών άρα το βάθος των δέντρων δεν πρέπει να είναι πολύ μεγάλο. Στις περισσότερες εφαρμογές ένα βάθος από 1 μέχρι 5 θεωρείται αρκετά ικανοποιητικό. Ωστόσο, αυτό δεν αποτελεί κανόνα που να απαγορεύει και μεγαλύτερα βάθη. Το ίδιο ισχύει και με τον αριθμό των εκτιμητών-δέντρων, που γενικά δε θα πρέπει να είναι πολύ μεγάλος. Όσο μεγαλύτερες είναι αυτές οι παράμετροι τόσο περισσότερο, κινδυνεύουμε να οδηγηθούμε σε φαινόμενο υπερπροσαρμογής. Στην πράξη η καλύτερη προσέγγιση για την επιλογή των παραμέτρων για την gradient boosting, για ένα συγκεκριμένο πρόβλημα είναι να ελέγξουμε για το συγκεκριμένο πρόβλημα ποιες είναι οι καλύτερες παράμετροι, δηλαδή ποιες οδηγούν σε καλύτερες προβλέψεις εξασφαλίζοντας παράλληλα κάθε φορά τη γενικότητα του μοντέλου δηλαδή αποφεύγοντας το φαινόμενο υπερπροσαρμογής.

2.5.4. Περιληπτική Περιγραφή του Αλγορίθμου

Αρχικά η μέθοδος ξεκινάει υπολογίζοντας το μέσο όρο της μεταβλητής απόκρισης y (για όλες τις παρατηρήσεις) και θέτει αυτόν ως αρχική εκτίμηση $\hat{y}$ γι' αυτές. Στη συνέχεια υπολογίζει τα ψευδοϋπόλοιπα $(y - \hat{y})$ για κάθε παρατήρηση και εκπαιδεύει πάνω σε αυτά το πρώτο δέντρο απόφασης (το οποίο θα προβλέπει από τις επεξηγηματικές μεταβλητές το ψευδοϋπόλοιπο για κάθε παρατήρηση). Αφού κατασκευαστεί το δέντρο, η πρόβλεψη για κάθε παρατήρηση θα είναι ο μέσος όρος των εκτιμήσεων του δέντρου (ψευδοϋπολοίπων) που βρίσκονται στον κόμβο που αντιστοιχεί η παρατήρηση αυτή. Στη συνέχεια οι τιμές των αρχικών ψευδοϋπολοίπων αντικαθίστανται από τις προβλέψεις τους. Η πρόβλεψη του μοντέλου για κάθε y προκύπτει ως άθροισμα του μέσου όρου των y και του ψευδοϋπόλοιπου που αντιστοιχεί στην παρατήρηση αυτή. Ωστόσο για να αποφύγουμε το overfitting προσθέτουμε έναν όρο, το ρυθμό εκμάθησης ο οποίος καθορίζει τη συνεισφορά του νέου δέντρου στην τελική πρόβλεψη. Έτσι η νέα πρόβλεψη δίνεται από το άθροισμα της μέσης τιμής της y με την τιμή του ψευδοϋπόλοιπου επί το ρυθμό εκμάθησης. Με τον τρόπο αυτό παίρνουμε κάθε φορά μία λίγο καλύτερη πρόβλεψη χωρίς να κινδυνεύουμε από overfitting (σύμφωνα με το Friedman [36], τα πειράματα δείχνουν ότι μικρά βήματα προς τη σωστή κατεύθυνση οδηγούν σε πιο ακριβείς προβλέψεις). Η τεχνική αυτή της μεθόδου είναι αντίστοιχη της απότομης καθόδου (grand descent) και από εκει παίρνει και το όνομά της. Η διαδικασία

συνεχίζεται, χτίζοντας το επόμενο δέντρο το οποίο θα εκπαιδευτεί στα ψευδοϋπόλοιπα που προκύπτουν ως η διαφορά των πραγματικών τιμών της μεταβλητής απόκρισης και των προβλέψεων που έκανε το προηγούμενο δέντρο. Η τελική πρόβλεψη κάθε φορά προκύπτει ως το άθροισμα του μέσου όρου των πραγματικών τιμών με τα ψευδοϋπόλοιπα του κάθε δέντρου για την παρατήρηση αυτή πολλαπλασιασμένα κάθε φορά με το ρυθμό εκμάθησης. Η διαδικασία σταματάει όταν φτάσουμε έναν καθορισμένο αριθμό δέντρων ή όταν τα ψευδοϋπόλοιπα δεν αλλάζουν σημαντικά.

2.6. Adaboost vs Gradient Boosting

Οι δυο boosting μέθοδοι που αναλύθηκαν παραπάνω βασίζονται στην ίδια ιδέα, τον συνδυασμό αδύναμων εκτιμητών(weak learners), συνήθως δέντρων ταξινόμησης, για τη δημιουργία δυνατών εκτιμητών(strong learners). Η διαφορά τους ωστόσο εντοπίζεται στον τρόπο με τον οποίο κατασκευάζουν και βελτιώνουν τους εκτιμητές μέσα από την επαναληπτική διαδικασία.

Η Adaboost, ξεκινάει κατασκευάζοντας τους εκτιμητές και στη συνέχεια αυξάνει την απόδοση-ακρίβεια τους επιβαρύνοντας(μέσω αυξημένων βαρών) περισσότερο τις παρατηρήσεις που προβλέπονται δυσκολότερα (παρατηρήσεις που το μοντέλο έχει μεγαλύτερο σφάλμα) σε κάθε ένα από αυτούς. Έτσι οι εκτιμητές αυτοί συνδυάζονται και με βάση το βάρος τους συνεισφέρουν ανάλογα στην κατασκευή του ισχυρού εκτιμητή.

Από την άλλη η Gradient Boosting, κατασκευάζει κάθε εκτιμητή εκπαιδεύοντας τον με τα υπόλοιπα του προηγούμενου εκτιμητή και κάθε φορά η τελική πρόβλεψη του μοντέλου προκύπτει από το αποτέλεσμα του κάθε εκτιμητή πολλαπλασιασμένο επί το ρυθμό εκμάθησης. Έτσι πρακτικά κάθε εκτιμητής βελτιώνει σταδιακά την ακρίβεια του μοντέλου αποφεύγοντας την υπερπροσαρμογή.

2.7. Υπερπροσαρμογή στα δεδομένα (Overfitting the Data)

Έγινε λόγος στην προηγούμενη παράγραφο για υπερπροσαρμογή στα δεδομένα, σε περίπτωση ενός πολύπλοκου σχεδιασμού του δέντρου απόφασης, όπως για παράδειγμα ενός μεγάλου βάθους. Με τον όρο υπερπροσαρμογή στα δεδομένα (overfitting), καλούμε το φαινόμενο κατά το οποίο ένα μοντέλο (παλινδρόμησης ή ταξινόμησης), ενώ προσαρμόζεται πολύ καλά στο τρέχον σύνολο δεδομένων που διαθέσαμε για την εκπαίδευση του, παρουσιάζει μεγαλύτερο σφάλμα(χειρότερη απόδοση) όταν έχει να διαχειριστεί ένα διαφορετικό σύνολο δεδομένων. Με απλά λόγια το μοντέλο που έχει κατασκευαστεί παρουσιάζει πρόβλημα γενίκευσης, δηλαδή εξαρτάται σε μεγάλο βαθμό από τα δεδομένα που του δόθηκαν.

Υπάρχουν πολλοί τρόποι να διαπιστώσουμε αν ένα μοντέλο εμφανίζει πρόβλημα υπερπροσαρμογής. Ο πιο εύκολος είναι να συγκρίνουμε την απόδοση του μοντέλου για τα δεδομένα που χρησιμοποιήσαμε για την εκπαίδευση του (training set) και για τα δεδομένα που χρησιμοποιήσαμε για την επαλήθευση (testing set – validation set). Αν η απόδοση-ακρίβεια του μοντέλου στα δεδομένα εκπαίδευσης είναι πολύ καλύτερη από την αντίστοιχη των δεδομένων επαλήθευσης, τότε είναι πολύ πιθανό να εμφανιστεί το φαινόμενο υπερπροσαρμογής. Όπως είδαμε ακόμη οι πολύπλοκοι σχεδιασμοί οδηγούν σε υπερπροσαρμογή, έτσι καλό είναι να ξεκινάμε από απλά μοντέλα (π.χ. δέντρα με μικρό βάθος) και να προχωράμε σε αλλαγές μόνο όσο βλέπουμε σημαντικές βελτιώσεις στην ακρίβεια του μοντέλου. Ωστόσο αφού εντοπίσουμε το φαινόμενο overfitting στο μοντέλο που έχουμε προσαρμόσει υπάρχουν αρκετοί τρόποι με τους οποίους, μπορούμε να το αντιμετωπίσουμε, με βασικότερους τους παρακάτω:

2.7.1. Κ-Φορές Διασταυρωμένη Επικύρωση (K-Fold Cross - Validation)

Η διασταυρωμένη επικύρωση είναι μια αρκετά αποτελεσματική μέθοδος αντιμετώπισης του overfitting, και λειτουργεί ως εξής: αρχικά χωρίζουμε τα δεδομένα σε Κ διαφορετικά μέρη, από τα οποία το ένα θα λειτουργήσει ως σύνολο επαλήθευσης ενώ τα υπόλοιπα ως σύνολο εκπαίδευσης. Επαναλαμβάνουμε τη διαδικασία Κ φορές έτσι ώστε κάθε φορά ένα διαφορετικό μέρος να αποτελέσει το σύνολο επαλήθευσης. Με τον τρόπο αυτό όλα τα δεδομένα θα έχουν χρησιμοποιηθεί και ως σύνολο επαλήθευσης και ως σύνολο εκπαίδευσης. Σε κάθε μια από τις Κ επαναλήψεις υπολογίζουμε την απόδοση του μοντέλου και στο τέλος παίρνουμε το μέσο όρο αυτής, έτσι πετυχαίνουμε πολύ μεγαλύτερη ακρίβεια και αποτελεσματικότητα στο μοντέλο καθώς η απόδοση του δεν εξαρτάται πλέον από την επιλογή των δεδομένων. Στο σημείο αυτό αξίζει να σημειωθεί, ότι η μέθοδος μπορεί να χρησιμοποιηθεί και για τη σύγκριση μοντέλων μεταξύ τους. Όταν έχουμε δυο ή περισσότερα μοντέλα και θέλουμε να τα συγκρίνουμε για ένα σύνολο δεδομένων μπορούμε να πραγματοποιήσουμε K-Fold Cross Validation σε κάθε ένα από αυτά και υπολογίζοντας την απόδοση του κάθε μοντέλου να κρατήσουμε τελικά το καλύτερο.

2.7.2. Εκπαίδευση με όσο το δυνατόν Περισσότερα Δεδομένα

Είναι φανερό, ότι όσο μεγαλύτερος, είναι ο όγκος των δεδομένων τόσο περισσότερη πληροφορία μπορεί να αντλήσει ένα μοντέλο από αυτά. Ωστόσο, αυτό δε συμβαίνει πάντα καθώς μπορεί ο όγκος των δεδομένων να είναι μεγάλος αλλά παράλληλα να είναι μεγάλος και ο θόρυβος σε αυτά τα δεδομένα (noisy data). Με την έννοια του θορύβου εννοούμε δεδομένα που δεν παρέχουν καμία πληροφορία για το μοντέλο όπως πολύ ακραίες ή άκυρες τιμές(π.χ. αν μετράμε το ύψος μιας ομάδας ανθρώπων το να υπάρχουν τιμές όπως 5μ

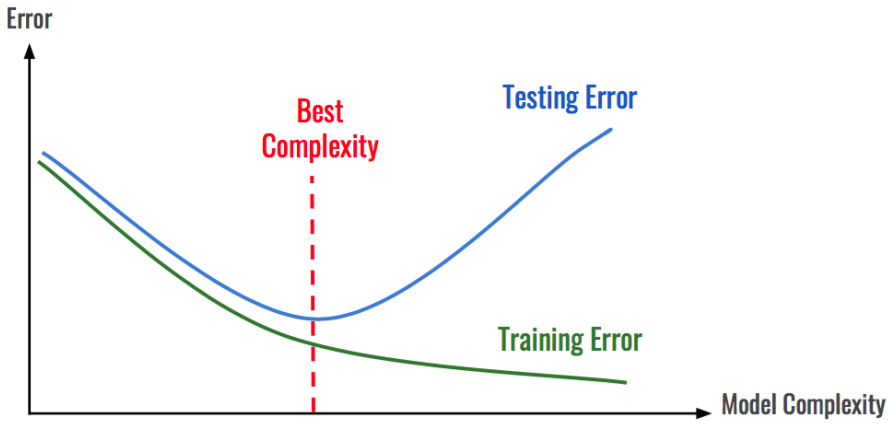
προφανώς δεν προσφέρουν καμία πληροφορία καθώς είναι εσφαλμένες και μπορεί αν επηρεάσουν την απόδοση του μοντέλου), αλλά και αλλοιωμένα δεδομένα (corrupted data).

2.7.3. Μείωση Διαστάσεων – Αφαίρεση Μεταβλητών (Dimension Reduction – Feature Removal)

Είναι γεγονός ότι όσο περισσότερες είναι οι επεξηγηματικές μεταβλητές (περισσότερες διαστάσεις) τόσο πιο περίπλοκο είναι το μοντέλο που κατασκευάζουμε, άρα τόσο μεγαλύτερη και η πιθανότητα να εμφανίσει overfitting. Είναι λοιπόν σημαντικό να περιορίζουμε τον τον αριθμό των μεταβλητών ενός μοντέλου και να αφαιρούμε αυτές που παρέχουν μικρή ή καθόλου πληροφορία. Άλλωστε ο μεγάλος αριθμός μεταβλητών σε ένα μοντέλο παλινδρόμησης για παράδειγμα σημαίνει ότι για να κάνουμε μια πρόβλεψη για μια τιμή θα χρειαστεί να συλλέξουμε πολύ περισσότερη πληροφορία. Έτσι είναι σημαντικό να μειώσουμε τις διαστάσεις του μοντέλου όσο περισσότερο γίνεται. Στη συνέχεια, σε επόμενη παράγραφο θα δούμε μερικούς τρόπους μείωσης διαστάσεων του μοντέλου.

2.7.4. Εύρεση Ορίου Υπερπροσαρμογής

Ο πιο διαδεδομένος και αποτελεσματικός τρόπος για να αποφύγουμε και να αντιμετωπίσουμε φαινόμενα υπερπροσαρμογής, είναι να δούμε πως επηρεάζεται η απόδοση - ακρίβεια του μοντέλου (πόσο μειώνεται το σφάλμα ή η οποιαδήποτε μετρική χρησιμοποιούμε για την αξιολόγηση του μοντέλου μας), καθώς μεταβάλλουμε κάποια παράμετρο, όπως ο αριθμός των επαναλήψεων ή το βάθος του δέντρου ή η πολυπλοκότητα ενός νευρωνικού δικτύου κ.λ.π. . Αν λοιπόν αναπαραστήσουμε γραφικά το πώς μεταβάλλεται το σφάλμα συναρτήσει της μεταβολής μιας παραμέτρου, θα πάρουμε ένα γράφημα σαν το παρακάτω:



Σχήμα : Συνάρτηση που δείχνει πως μεταβάλλεται το σφάλμα ενός μοντέλου σε σχέση με την πολυπλοκότητα του. Πηγή: https://hackernoon.com/memorizing-

is-not-learning-6-tricks-to-prevent-overfitting-in-machine-learning-820b091dc42

Όπως φαίνεται λοιπόν στο παραπάνω γράφημα ξεκινώντας από ένα αρκετά απλό μοντέλο, όσο αυξάνουμε την πολυπλοκότητα, το μοντέλο φαίνεται να έχει όλο καλύτερη απόδοση μέχρι ενός ορίου. Όταν ξεπεραστεί το όριο αυτό παρατηρούμε ότι η απόδοση του μοντέλου για τα δεδομένα εκπαίδευσης αρχίζει και διαφοροποιείται από την αντίστοιχη απόδοση του σε δεδομένα καινούρια διαφορετικά από αυτά που χρησιμοποιήθηκαν για την εκπαίδευση του. Το όριο αυτό αντιπροσωπεύει τη βέλτιστη πολυπλοκότητα του μοντέλου, και οριοθετεί δυο περιοχές την περιοχή υποπροσαρμογής (underfitting area) στα αριστερά του, όπου το μοντέλο είναι αρκετά απλό και έχει χαμηλή ακρίβεια – μεγάλο σφάλμα και την περιοχή υπερπροσαρμογής (overfitting area) στα δεξιά του όπου το μοντέλο είναι αρκετά περίπλοκο και παρουσιάζει αυξημένο σφάλμα σε διαφορετικά δεδομένα από αυτά που εκπαιδεύτηκε. Έτσι λοιπόν είναι σημαντικό όταν κατασκευάζουμε ένα μοντέλο τεχνητής νοημοσύνης να εξασφαλίζουμε με τον παραπάνω τρόπο την παρόμοια απόδοση του τόσο στα δεδομένα που έχει εκπαιδευτεί (training set) όσο και σε νέα δεδομένα, διαφορετικά από αυτά που χρησιμοποιήθηκαν για την εκπαίδευση του (test set). Σε πολλές περιπτώσεις στην πραγματικότητα καλούμαστε να θυσιάσουμε λίγο από την απόδοση και την ακρίβεια του μοντέλου ώστε να έχουμε ένα μοντέλο πιο γενικευμένο που μπορεί να ανταπεξέλθει με μεγαλύτερη ακρίβεια σε καινούρια δεδομένα.

2.8. Κ Nearest Neighbors Παλινδρόμηση

Ο αλγόριθμος Κ Nearest Neighbors, είναι ένας από τους πιο απλούς αλγορίθμους παλινδρόμησης και ταξινόμησης. Στην περίπτωση της παλινδρόμησης μπορεί να φανεί εξαιρετικά αποτελεσματικός όταν η μεταβλητή απόκρισης είναι συνεχής και όχι διακριτή. Ο αλγόριθμος βασίζεται σε μια μη παραμετρική τεχνική και λειτουργεί ως εξής:

- Αποθηκεύει το σύνολο δεδομένων εκπαίδευσης (test set), χωρίς να πραγματοποιεί κάποια λειτουργία ή υπολογισμό πάνω σε αυτό
- Στη συνέχεια για κάθε μια παρατήρηση του συνόλου επαλήθευσης, ο αλγόριθμός, αναγνωρίζει τις Κ «κοντινότερες» παρατηρήσεις σε αυτήν (τις Κ παρατηρήσεις που έχουν τα περισσότερα κοινά χαρακτηριστικά με αυτήν), με βάση κάποια μετρική, συνήθως μια συνάρτηση απόστασης
- Στη συνέχεια για αυτές τις Κ παρατηρήσεις, είτε βρίσκει απλά το μέσο όρο της μεταβλητής απόκρισης τους και τον θέτει ως πρόβλεψη, είτε υπολογίζει έναν αντίστροφα βεβαρυμμένο μέσο όρο με βάση την απόσταση τους και θέτει αυτόν ως πρόβλεψη

Οι βασικότερες συναρτήσεις απόστασης που χρησιμοποιεί η μέθοδος, για να βρει τις Κ κοντινότερες παρατηρήσεις είναι οι παρακάτω:

- Ευκλείδια: $\sqrt{\sum_{i=1}^{k}(x_i - x_i')^2}$
- Manhattan: $\sum_{i=1}^{k}|x_i - x_i'|$
- Minskowski: $\left(\sum_{i=1}^{k}(|x_i - x_i'|)^q\right)^{\frac{1}{q}}$

όπου x, x' δυο διαφορετικές παρατηρήσεις και k, ο αριθμός των επεξηγηματικών τους μεταβλητών. Οι παραπάνω αποστάσεις χρησιμοποιούνται στην περίπτωση που οι επεξηγηματικές μεταβλητές είναι συνεχείς. Όταν είναι κατηγορικές, χρησιμοποιείται η απόσταση Hamming, που εκφράζει τον αριθμό των κατηγορικών μεταβλητών που έχουν διαφορετική τιμή, μεταξύ δυο παρατηρήσεων.

- Hamming: $D_H = \sum_{i=1}^{k}|x_i - x_i'|$ $όπου\ αν\ x = y \rightarrow D = 0,\ ενώ\ αν\ x \neq y \rightarrow D = 1$

Στο σημείο αυτό, αξίζει να σημειωθεί, ότι μιας και η μέθοδος μετράει αποστάσεις μεταξύ των παρατηρήσεων με βάση τη διαφορά των τιμών των μεταβλητών τους, είναι λογικό να δημιουργούνται προβλήματα όταν έχουμε διαφορετική μετρική κλίμακα (measurement scale) μεταξύ των μεταβλητών ή όταν οι επεξηγηματικές μεταβλητές περιέχουν και κατηγορικές και συνεχείς μεταβλητές. Για παράδειγμα αν μια επεξηγηματική μεταβλητή εκφράζει το ετήσιο εισόδημα σε ευρώ ενώ μια άλλη την ηλικία σε χρόνια, τότε το εισόδημα θα έχει πολύ μεγαλύτερη επιρροή στην διαμόρφωση της απόστασης σε σχέση με την ηλικία. Σε τέτοιες περιπτώσεις ενδείκνυται να πραγματοποιούμε προτού τρέξουμε τον αλγόριθμο κανονικοποίηση ή τυποποίηση στις επεξηγηματικές μεταβλητές.

Πολύ σημαντικό ρόλο στην απόδοση-ακρίβεια της μεθόδου διαδραματίζει η παράμετρος Κ. Και εδώ, δεν υπάρχουν συγκεκριμένοι κανόνες για την επιλογή του Κ. Για το λόγο αυτό θα πρέπει σε κάθε περίπτωση να αναζητάμε το Κ που ταιριάζει περισσότερο στα δεδομένα και στο πρόβλημα που καλούμαστε να επιλύσουμε. Η πιο σίγουρη τεχνική για την επιλογή του κατάλληλου Κ, είναι να τρέξουμε τον αλγόριθμο πολλές φορές για διαφορετικά Κ, και να κρατήσουμε το Κ εκείνο για το οποίο το σφάλμα της μεθόδου στα δεδομένα επαλήθευσης είναι μικρότερο, δηλαδή το Κ για το οποίο η μέθοδος έχει μεγαλύτερη ακρίβεια σε δεδομένα που δεν έχει ξαναδεί. Ωστόσο, πρέπει να έχει κάποιος υπόψη του, ότι όσο το Κ μικραίνει και πλησιάζει τη μονάδα, οι προβλέψεις γίνονται λιγότερο ευσταθής και επηρεάζονται περισσότερο από τυχαίες διακυμάνσεις του δείγματος.

Η ελαχιστοποίηση της υπολογιστικής πολυπλοκότητας της μεθόδου, αποτελεί αντικείμενο έρευνας στο χώρο της μηχανικής μάθησης τα τελευταία χρόνια. Υπάρχουν τρεις βασικές υλοποιήσεις της μεθόδου, όσο αφορά στον τρόπο εύρεσης των κοντινότερων γειτόνων και κάθε μια έχει διαφορετική υπολογιστική πολυπλοκότητα. Η απλούστερη υλοποίηση με τη μεγαλύτερη όμως πολυπλοκότητα, καλείται «brute-force», και περιλαμβάνει τον υπολογισμό όλων των αποστάσεων μεταξύ των παρατηρήσεων του συνόλου δεδομένων. Για N παρατηρήσεις με D επεξηγηματικές μεταβλητές (διάσταση), η τεχνική αυτή απαιτεί $O[DN^2]$ χρόνο για να υλοποιηθεί και ενώ σε μικρά σύνολα δεδομένων είναι αρκετά αποτελεσματική, όσο τα δεδομένα μεγαλώνουν (δηλαδή όσο μεγαλώνει το N), η τεχνική γίνεται υπολογιστικά αδύνατη.

Για την επίλυση του προβλήματος της μεγάλης υπολογιστικής πολυπλοκότητας της brute-force για μεγάλα δεδομένα, αναπτύχθηκε μια μέθοδος που βασίζεται σε αρχιτεκτονική που χρησιμοποιεί δεντρικές δομές δεδομένων για τους υπολογισμούς της. Οι μέθοδος αυτή (K-D Tree methods), προσπαθεί να μειώσει τον απαιτούμενο αριθμό αποστάσεων που πρέπει να υπολογιστούν από τον αλγόριθμο, βασιζόμενη στην παρακάτω λογική: Αν ένα σημείο A έχει μεγάλη απόσταση από ένα σημείο B, και το σημείο B είναι πολύ κοντά στο σημείο C, τότε γνωρίζουμε ότι τα σημεία A, C, έχουν και αυτά μεγάλη απόσταση μεταξύ τους, χωρίς να χρειάζεται να την υπολογίσουμε. Με αυτό τον τρόπο, το υπολογιστικό κόστος της αναζήτησης των κοντινότερων γειτόνων μπορεί να μειωθεί σε $O[DNlog(N)]$. Έχουμε λοιπόν πολύ μεγάλη βελτίωση της πολυπλοκότητας ειδικά για μεγάλα N σε σχέση με τη brute-force. Η μέθοδος πετυχαίνει τη βελτίωση αυτή με τη βοήθεια μιας δυαδικής δεντρικής δομής δεδομένων που χωρίζει αναδρομικά τον παραμετρικό χώρο σε περιοχές με συγκεκριμένες παρατηρήσεις. Έτσι η κάθε αναζήτηση απαιτεί χρόνο $O[\log(N)]$. Ωστόσο, όσο οι διαστάσεις (D), αυξάνονται τόσο αυξάνεται η πολυπλοκότητα, με αποτέλεσμα και αυτή η τεχνική για μεγάλα D να γίνεται υπολογιστικά αδύνατη.

Για το λόγο αυτό σχεδιάστηκε η μέθοδος, Ball Tree, η οποία στηρίζεται σε μια δεντρική αρχιτεκτονική από την οποία έχει πάρει το όνομα της. Ενώ τα KD Trees χωρίζουν τα δεδομένα κατά μήκος των καρτεσιανών αξόνων, τα Ball Trees, χωρίζουν τα δεδομένα με τη βοήθεια εμφωλευμένων σφαιρών. Αυτό ενώ από τη μια οδηγεί σε μεγαλύτερο κόστος (υπολογιστικό) κατά την κατασκευή του δέντρου σε σχέση με την KD Tree, εξασφαλίζει από την άλλη, την αποδοτικότητα και αποτελεσματικότητα της μεθόδου σε δεδομένα με μεγάλες διαστάσεις D. Η μέθοδος χωρίζει αναδρομικά τα δεδομένα σε κόμβους, που κάθε ένας από αυτούς χαρακτηρίζεται από ένα κέντρο C και μια ακτίνα r, τέτοια ώστε κάθε παρατήρηση που ανήκει στον κόμβο να βρίσκεται μέσα στη σφαίρα που ορίζεται από τα C και r. Ο αριθμός των πιθανών σημείων για την αναζήτηση των κοντινότερων γειτόνων περιορίζεται από τη χρήση της τριγωνικής ανισότητας: $|x + y| \leq |x| + |y|$. Με τον τρόπο αυτό, ο υπολογισμός της

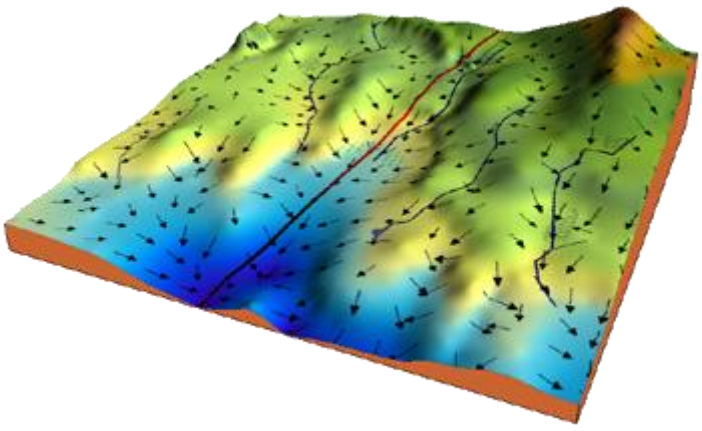
απόστασης, μεταξύ ενός σημείου και του κέντρου είναι επαρκής για τον καθορισμό ενός άνω και κάτω ορίου των αποστάσεων όλων των σημείων που βρίσκονται στον ίδιο κόμβο με το σημείο αυτό.

2.9. Μέθοδοι Βελτιστοποίησης – Ελαχιστοποίησης

2.9.1. Μέθοδος Απότομης Καθόδου (Gradient Descent):

Η Gradient Descent είναι μια μέθοδος βελτιστοποίησης, που βασίζεται σε έναν επαναληπτικό αλγόριθμο με σκοπό να εντοπίσει το ελάχιστο μιας συνάρτησης, επιλέγοντας κάθε φορά την κατεύθυνση της απότομης καθόδου όπως αυτή καθορίζεται από την κλίση σε κάποιο σημείο της συνάρτησης.

Η λειτουργία της μεθόδου με απλά λόγια μπορεί να περιγραφεί με τη βοήθεια του παρακάτω προβλήματος. Ας θεωρήσουμε ένα τρισδιάστατο γράφημα το οποίο αναπαριστά μια συνάρτηση κόστους για την οποία θέλουμε να βρούμε το ελάχιστο της. Σκοπός μας είναι να μετακινηθούμε από το βουνό στη δεξιά γωνία του γραφήματος(υψηλό κόστος – high cost) προς το βαθύ μπλε της θάλασσας στο αριστερό μέρος του γραφήματος(χαμηλό κόστος – low cost). Τα βέλη δείχνουν την κατεύθυνση της απότομης καθόδου (αρνητική κλίση), σε κάθε σημείο, δηλαδή την κατεύθυνση κατά την οποία η συνάρτηση κόστους μειώνεται με γρηγορότερο βαθμό.



Έτσι ξεκινάμε από την κορυφή του βουνού και κατεβαίνουμε προς την κατεύθυνση που υποδεικνύει η αρνητική κλίση(negative gradient), πηγαίνοντας στο επόμενο σημείο, στο οποίο επαναϋπολογίζουμε με βάση τις συντεταγμένες του την αρνητική κλίση και συνεχίζουμε στην κατεύθυνση που αυτή υποδεικνύει. Συνεχίζουμε τη διαδικασία επαναληπτικά μέχρι να φτάσουμε στο βυθό του γραφήματος ή σε ένα σημείο στο οποίο δε μπορούμε να συνεχίσουμε να κατεβαίνουμε άλλο(τοπικό ελάχιστο).

Η gradient descent, βασίζεται στην παρατήρηση, ότι, αν έχουμε μια πολυδιάσταστη συνάρτηση $F(x)$, που είναι ορισμένη και παραγωγίσιμη σε μια γειτονιά (κοντά) ενός σημείου α, τότε η $F(x)$, μειώνεται γρηγορότερα αν μετακινηθούμε κατά την κατεύθυνση της αρνητικής κλίσης της F στο α. Συνεπώς, αν:

$$a_{n+1} = a_n - \gamma \nabla F(a_n), \gamma\iota\alpha\ \gamma \in R\ \alpha\rho\kappa\epsilon\tau\acute{\alpha}\ \mu\iota\kappa\rho\acute{o},$$

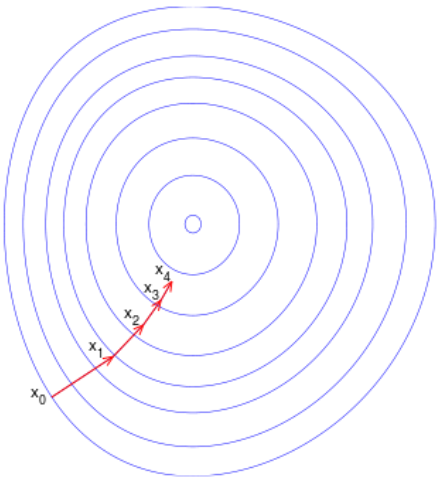
τότε $F(a_n) \geq F(a_{n+1})$. Με άλλα λόγια, ο όρος, $\gamma \nabla F(a)$ έχει αφαιρεθεί από το α, γιατί, θέλουμε να προχωρήσουμε αντίθετα με την κλίση προς το ελάχιστο της συνάρτησης. Έτσι ξεκινάμε από ένα αρχικό x_0, το οποίο θεωρούμε ως τοπικό ελάχιστο της F, και θεωρούμε τα $x_1, x_2, \ldots$, τέτοια ώστε

$$x_{n+1} = x_n - \gamma_n \nabla F(x_n), n \geq 0.$$

Έτσι προκύπτει μια μονότονη ακολουθία, $F(x_0) \geq F(x_1) \geq F(x_2) \geq \cdots$, η οποία συγκλίνει στο τοπικό ελάχιστο που θέλουμε να βρούμε. Εδώ αξίζει να σημειωθεί ότι το βήμα γ μπορεί να αλλάζει σε κάθε επανάληψη. Με διάφορες υποθέσεις για τη συνάρτηση F, (π.χ. F κυρτή και ∇F Lipschitz) και με συγκεκριμένες επιλογές του γ, προκύπτει ένα γ_n που ορίζεται ως εξής:

$$\gamma_n = \frac{|(x_n - x_{n-1})^T [\nabla F(x_n) - \nabla F(x_{n-1})]|}{\|\nabla F(x_n) - \nabla F(x_{n-1})\|^2}$$

και συγκλίνει με βεβαιότητα σε ένα τοπικό ελάχιστο(local minimum). Ωστόσο, όταν η συνάρτηση, είναι κυρτή το τοπικό ελάχιστο στο οποίο συγκλίνει ο παραπάνω όρος, είναι και το ολικό ελάχιστο της συνάρτησης(global minimum). Η παραπάνω διαδικασία για μια πολυδιάστατη συνάρτηση περιγράφεται από την παρακάτω εικόνα:



Πηγή: https://en.wikipedia.org/wiki/Gradient_descent

Στο παραπάνω παράδειγμα θεωρούμε μια πολυδιάστατη F, που το γράφημα της έχει σχήμα μπολ. Οι μπλε γραμμές, είναι οι ισοϋψείς καμπύλες(contour lines), στις οποίες η F έχει σταθερή τιμή. Το κόκκινο βέλος, δείχνει σε κάθε σημείο την κατεύθυνση της αρνητικής κλίσης στο σημείο αυτό. Βλέπουμε λοιπόν, ότι η μέθοδος της απότομης καθόδου, μας οδηγεί στον πάτο του μπολ (ελάχιστο), το σημείο στο οποίο η συνάρτηση F, παίρνει την ελάχιστη τιμή της.

Ρυθμός εκμάθησης: Σημαντικό μέρος στην λειτουργία της μεθόδου διαδραματίζει ο ρυθμός εκμάθησης (learning rate), γ. Το γ καθορίζει το μέγεθος κάθε βήματος που θα κάνει η μέθοδος. Ένας μεγάλος ρυθμός εκμάθησης, σημαίνει ότι σε κάθε βήμα μπορούμε να καλύψουμε μεγαλύτερο έδαφος, όμως ρισκάρουμε να προσπεράσουμε το ελάχιστο, καθώς η πλαγιά-κλίση (slope) σε κάθε σημείο αλλάζει. Από την άλλη, ένας μικρός ρυθμός εκμάθησης, σημαίνει ότι θα έχουμε μεγαλύτερη ακρίβεια στον υπολογισμό του ελαχίστου και ότι δεν κινδυνεύουμε να το προσπεράσουμε. Ωστόσο λαμβάνοντας υπόψη ότι ο υπολογισμός της κλίσης καταναλώνει κάποιο χρόνο, χαμηλοί ρυθμοί εκμάθησης απαιτούν πολύ χρόνο για να φτάσουμε στο ελάχιστο της συνάρτησης.

2.9.2. Στοχαστική Απότομη Κάθοδος (Stochastic Gradient Descent SGD)

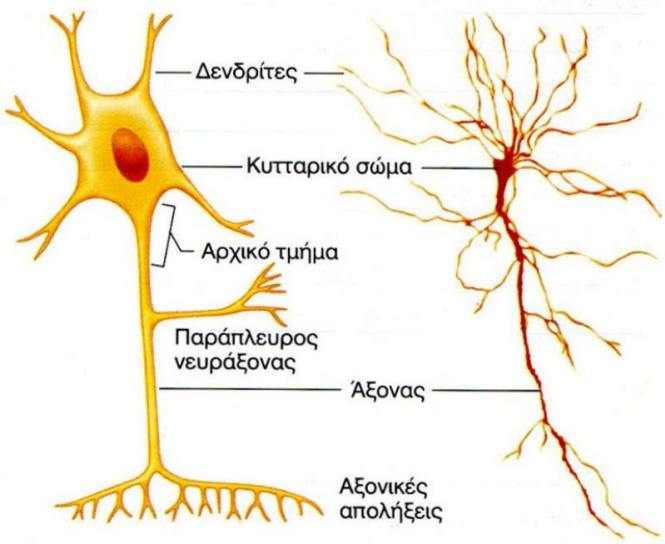
Έστω τώρα ότι έχουμε ένα μοντέλο μηχανικής μάθησης (π.χ. ένα μοντέλο παλινδρόμησης) και θέλουμε να βρούμε για ποιες τιμές ελαχιστοποιείται η συνάρτηση κόστους με την οποία αξιολογούμε το μοντέλο. Η μέθοδος gradient descent που αναλύσαμε παραπάνω για να υπολογίσει την κλίση της συνάρτησης σε κάθε σημείο και την τιμή της βασίζεται σε όλο το σύνολο των δεδομένων που διαθέτουμε για την εκπαίδευση του. Ωστόσο όταν το σύνολο αυτό είναι πολύ μεγάλο, τότε η μέθοδος θα χρειαστεί να καταναλώσει πολύ χρόνο για τον υπολογισμό του ελαχίστου. Σε τέτοιες περιπτώσεις λοιπόν χρησιμοποιούμε μια παραλλαγή της μεθόδου, ώστε να λύσουμε το πρόβλημα ελαχιστοποίησης που επιθυμούμε, που ονομάζεται στοχαστική απότομη κάθοδος. Η μέθοδος αυτή για να υπολογίσει την κλίση και την κατεύθυνση της σε κάθε σημείο χρησιμοποιεί μόνο μια παρατήρηση (ή ένα μέρος τον παρατηρήσεων). Συνεπώς σε μεγάλο αριθμό δεδομένων είναι πολύ πιο γρήγορη και αποδοτική καθώς χρειάζεται αρκετά λιγότερο χρόνο για να δώσει αποτέλεσμα και γιατί αποφεύγει τα τοπικά ελάχιστα σε αντίθεση με την απλή gradient descent. Αυτό συμβαίνει γιατί η stochastic gradient descent, επιλέγει με τυχαίο τρόπο την παρατήρηση(ή το μέρος των παρατηρήσεων), στις οποίες θα βασιστεί με αποτέλεσμα να αποφεύγει τοπικά ελάχιστα στα οποία θα κατέληγε η απλή μορφή της μεθόδου. Κάτι που είναι φανερό και από το γεγονός ότι αν επαναλάβουμε τη μέθοδο για το ίδιο σύνολο δεδομένων αυτή θα δώσει διαφορετικό αποτέλεσμα (λόγω της τυχαιότητας). Έτσι στην πράξη για τους παραπάνω λόγους προτιμάτε η χρήση της stochastic gradient descent (sgd) από την απλή gradient descent.

2.10. Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks)

2.10.1. Εισαγωγή στην Έννοια των Νευρωνικών Δικτύων

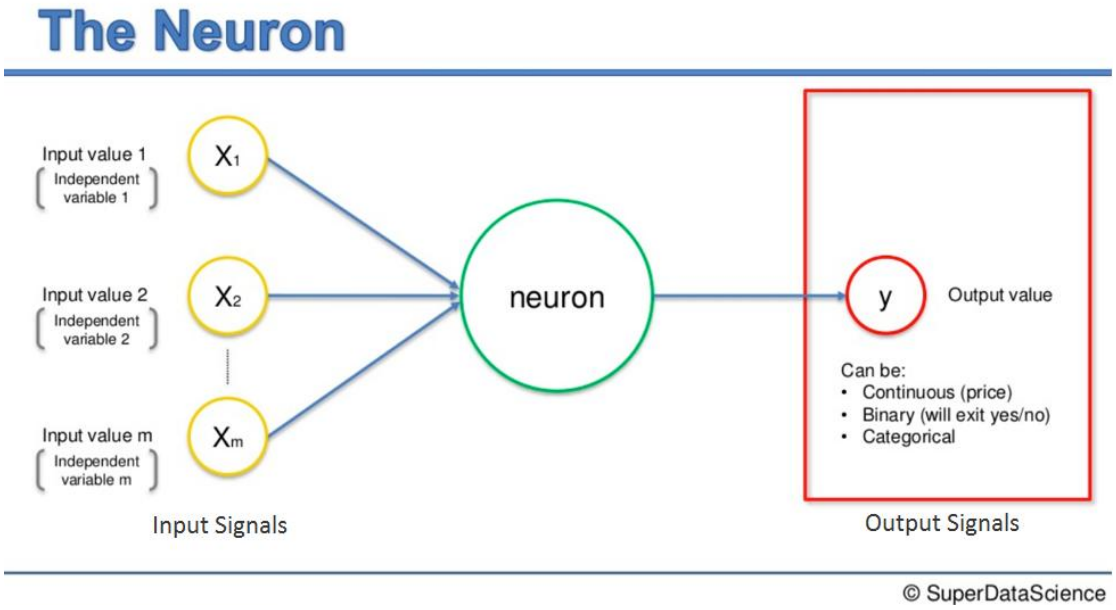
Η προσπάθεια των επιστημόνων να κατασκευάσουν έναν αλγόριθμο-μοντέλο πρόβλεψης, που μιμείται τον τρόπο λειτουργίας του πολύπλοκου και εξαιρετικά ευφυούς νευρικού συστήματος του ανθρώπινου οργανισμού, οδήγησε στην ανάπτυξη των νευρωνικών δικτύων. Η αρχιτεκτονική των νευρωνικών δικτύων θυμίζουν σε πολύ μεγάλο βαθμό την αρχιτεκτονική του νευρικού συστήματος του ανθρώπινου οργανισμού.

Οι νευρώνες αποτελούν τη δομική και λειτουργική μονάδα του νευρικού συστήματος του ανθρώπινου οργανισμού, και έχουν την ιδιότητα να αντιδρούν σε συγκεκριμένες μεταβολές του περιβάλλοντος στέλνοντας πληροφορίες από και προς το υπόλοιπο νευρικό σύστημα με σκοπό την προσαρμογή του οργανισμού στις νέες συνθήκες περιβάλλοντος. Οι νευρώνες αποτελούνται από τον πυρήνα, τους δενδρίτες, τον νευράξονα και τις αξονικές απολίξεις (Βλέπε παρακάτω εικόνα). Οι νευρώνες ενός νευρικού συστήματος είναι συνδεδεμένοι ως εξής: ο άξονας του κάθε νευρώνα είναι συνδεδεμένος μέσω των απολίξεων με τους δενδρίτες ενός άλλου νευρώνα με τη βοήθεια των συνάψεων. Όταν συμβεί μια αλλαγή στο περιβάλλον, το ερέθισμα πυροδοτεί ένα ηλεκτροχημικό σήμα κατά μήκος του άξονα. Στο σημείο αυτό αξίζει να σημειωθεί ότι για να ενεργοποιήσει ένα ερέθισμα το νευρώνα η ένταση του θα πρέπει να υπερβεί μια συγκεκριμένη τιμή. Στην περίπτωση που ένα ερέθισμα υπερβεί την τιμή αυτή τότε μεταδίδεται από τον άξονα του νευρώνα στους δενδρίτες του γειτονικού νευρώνα μέσω των συνάψεων με τη βοήθεια νευροδιαβιβαστών (χημικές ουσίες που εκκρίνονται από τις αξονικές απολίξεις για τη μετάδοση της πληροφορίας). Έτσι η πληροφορία μεταδίδεται από νευρώνα σε νευρώνα με σκοπό να αναλυθεί και να επεξεργασθεί και να δοθούν οι κατάλληλες εντολές στα εκτελεστικά όργανα.



2.10.2. Ο Τεχνητός Νευρώνας

Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks), ήταν η απάντηση των επιστημόνων στο παρακάτω ερώτημα: είναι δυνατόν να κατασκευαστούν κάποια πρότυπα μοντέλα του ανθρώπινου νευρωνικού συστήματος τα οποία θα έχουν τα ίδια χαρακτηριστικά και θα μπορούν από μόνα τους να επιτελούν τις ίδιες λειτουργίες με τον ίδιο τρόπο που γίνονται στα βιολογικά νευρωνικά δίκτυα. Η βασική διαφορά των τεχνιτών νευρωνικών δικτύων από τα βιολογικά, είναι ότι παρόλο που και τα δυο είδη μαθαίνουν και γίνονται πιο αποδοτικά με την εξάσκηση και την εμπειρία, τα τεχνητά δεν ακολουθούν κάποιους προκαθορισμένους κανόνες κάτι που είναι χαρακτηριστικό των υπολογιστών. Ο βασικός σκοπός των νευρωνικών δικτύων είναι να επιτελούν από μόνα τους συγκεκριμένες διεργασίες όπως π.χ. να αναγνωρίζουν εικόνες, να προβλέπουν την τιμή μιας μετοχής, να ταξινομούν μια πληροφορία κ.ο.κ. . Η αρχιτεκτονική των τεχνητών νευρωνικών δικτύων θυμίζει σε μεγάλο βαθμό την αρχιτεκτονική των βιολογικών, αφού και αυτά αποτελούνται από ένα σύνολο νευρώνων που ενώνονται μεταξύ τους. Η δομή του κάθε νευρώνα φαίνεται στο παρακάτω σχήμα:

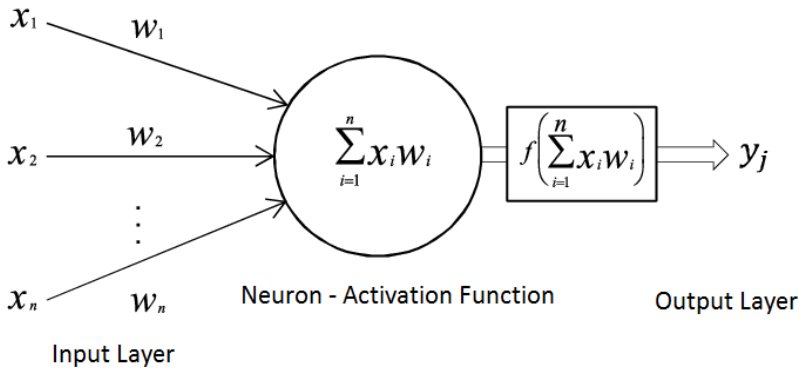


Όπως και ο βιολογικός νευρώνας, ο τεχνητός νευρώνας αποτελείται από τα σήματα εισόδου (αντίστοιχοι δενδρίτες) από τα οποία έρχεται η πληροφορία και κάθε ένα από αυτά αντιστοιχεί σε μια επεξηγηματική μεταβλητή (αντίστοιχα στις αισθήσεις). Στη συνέχεια έχουμε το νευρώνα στον οποίο γίνεται η επεξεργασία της πληροφορίας και τέλος έχουμε το σήμα εξόδου (αντίστοιχες απολίξεις). Το σήμα εξόδου αποτελεί την απόκριση του νευρώνα και μπορεί να είναι συνεχής τιμή (όταν για παράδειγμα προσπαθούμε να προβλέψουμε την τιμή μιας μετοχής), δυαδική τιμή (όταν θέλουμε για παράδειγμα να διαπιστώσουμε αν ένας ασθενής έχει μια ασθένεια ή όχι), ή κατηγορική τιμή

(όταν θέλουμε για παράδειγμα να διαπιστώσουμε αν μια εικόνα απεικονίζει ζώο, άνθρωπο ή πράγμα).

2.10.3. Η Λειτουργία του Νευρώνα

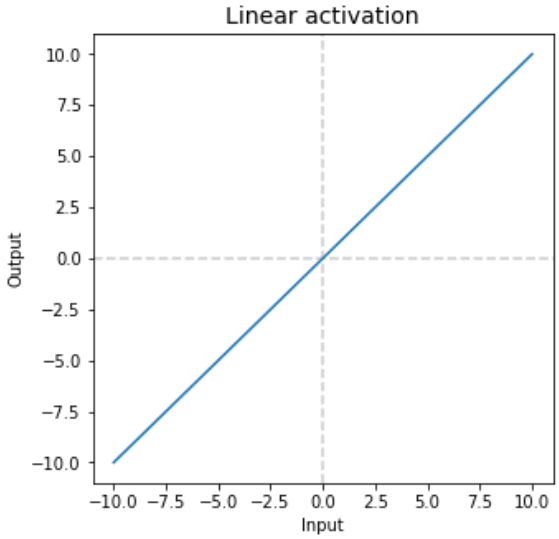
Η λειτουργία του νευρώνα περιγράφεται ως εξής: Η πληροφορία από τις μεταβλητές απόκρισης(σήμα εισόδου) φτάνει στο στρώμα εισόδου (Input Layer) και από εκεί επιβαρύνονται μέσω κάποιων αρχικοποιημένων βαρών. Στη συνέχεια η πληροφορία οδηγείται στο νευρώνα στον οποίο υπολογίζεται το άθροισμα $\sum w_i x_i$, όπου x_i η τιμή της i-οστής επεξηγηματικής μεταβλητής και w_i η το βάρος που αντιστοιχεί σε αυτήν. Έπειτα η τιμή μιας συνάρτησης ενεργοποίησης f για την τιμή του αθροίσματος, θα ενεργοποιήσει ή όχι το νευρώνα, δηλαδή θα στείλει ή όχι την τιμή y_i στο στρώμα εξόδου (Output Layer).



2.10.4. Συναρτήσεις Ενεργοποίησης (Activation Functions)

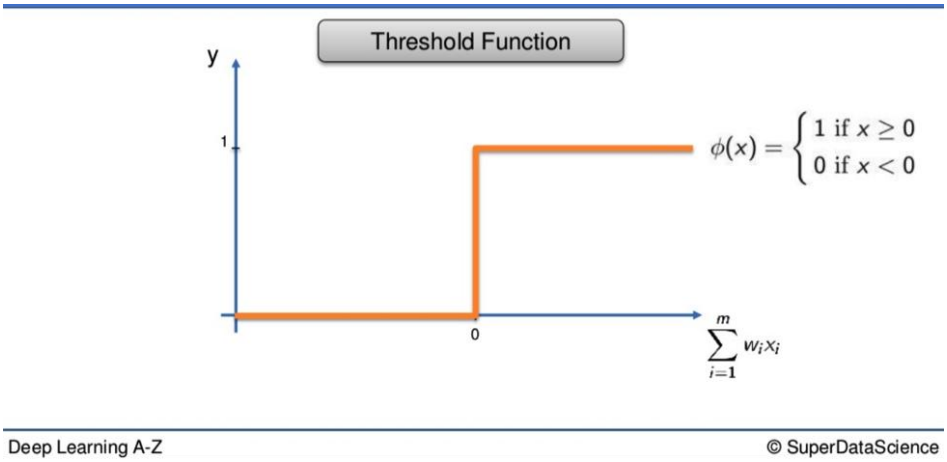
Η συνάρτηση ενεργοποίησης αποτελεί μια βασική παράμετρο σε ένα τεχνητό νευρωνικό δίκτυο, καθώς καθορίζει το αν θα ενεργοποιηθεί ή όχι ο νευρώνας, δηλαδή αν θα παράγει ή όχι output. Γενικά, υπάρχουν πολλές διαφορετικές συναρτήσεις ενεργοποίησης και δεν υπάρχει κάποιος συγκεκριμένος αυστηρός κανόνας σχετικά με το ποια χρησιμοποιούμε σε κάθε περίπτωση. Ωστόσο, υπάρχουν κάποιοι εμπειρικοί κανόνες που έχουν προκύψει μετά από πολλά πειράματα και δοκιμές οι οποίοι αναλύονται παρακάτω. Στο σημείο αυτό βέβαια, αξίζει να σημειωθεί, ότι πάντα πρέπει να επιλέγουμε μια συνάρτηση ενεργοποίησης που ταιριάζει στο πρόβλημα που έχουμε να λύσουμε. Οι βασικές συναρτήσεις ενεργοποίησης, είναι οι παρακάτω: η γραμμική συνάρτηση (linear function), η βηματική συνάρτηση (threshold), η ανορθωτική γραμμική (rectifier), η λογιστική σιγμοειδής (sigmoid) και η υπερβολική εφαπτομένη (hyperbolic tangent).

2.10.4.1. Γραμμική Συνάρτηση (linear or Identity fuction)



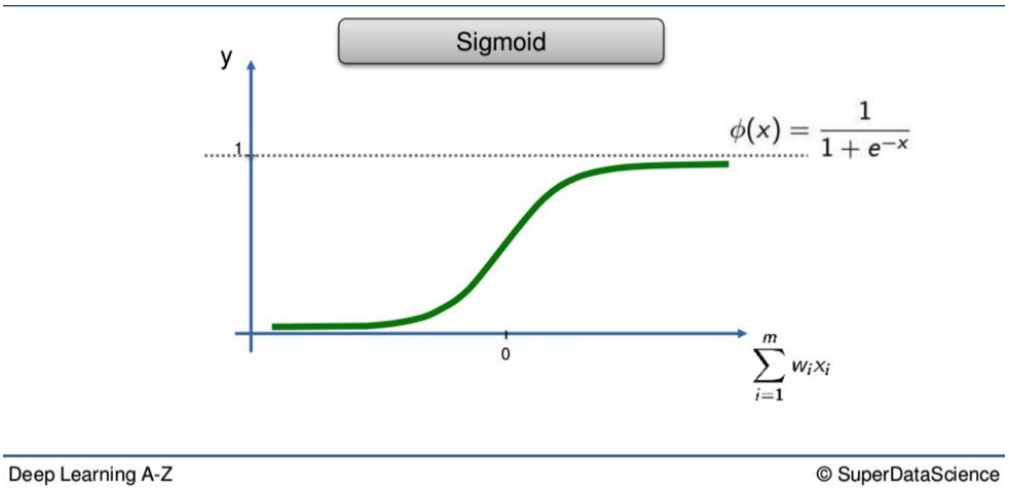
Η γραμμική συνάρτηση y=x, είναι η πιο απλή συνάρτηση ενεργοποίησης. Για κάθε τιμή του αθροίσματος $\sum w_i x_i$, αυτή επιστρέφει το ίδιο το άθροισμα. Ωστόσο η χρήση της συνδέεται με ένα βασικό πρόβλημα. Η παράγωγος της γραμμικής συνάρτησης είναι πάντα σταθερός αριθμός και ανεξάρτητη από την τιμή του Χ. Έτσι όταν το νευρωνικό δίκτυο θα χρησιμοποιήσει τη μέθοδο της απότομης καθόδου (gradient descent), για να ελαχιστοποιήσει το κόστος η κλίση θα είναι σταθερή και ανεξάρτητη από το Χ. Έτσι η εκπαίδευση του δικτύου πραγματοποιείται για μια σταθερή τιμή ανεξάρτητη του βεβαρυμμένου αθροίσματος. Άλλο ένα βασικό πρόβλημα προκύπτει όταν έχω ένα νευρωνικό δίκτυο με πολλά κρυφά στρώματα με πολλούς νευρώνες που συνδέονται μεταξύ τους, και έχω χρησιμοποιήσει τη γραμμική συνάρτηση ως συνάρτηση ενεργοποίησης. Κάθε στρώμα ενεργοποιείται από τη γραμμική συνάρτηση, η τιμή της οποίας (δηλαδή το βεβαρυμμένο άθροισμα), περνάει ως τιμή εισόδου στο επόμενο στρώμα κ.ο.κ. . Αυτό όμως έχει σαν αποτέλεσμα, η γραμμική συνάρτηση του τελευταίου στρώματος να μην είναι τίποτα άλλο παρά η γραμμική συνάρτηση της εισόδου του πρώτου στρώματος (ο συνδυασμός γραμμικών συναρτήσεων είναι γραμμική συνάρτηση), με απλά λόγια τα στρώματα αυτά μπορούν να αντικατασταθούν από ένα μοναδικό, δηλαδή έχουμε χάσει τη δυνατότητα να προσθέσουμε στρώματα στο νευρωνικό δίκτυο. Αυτοί είναι και οι λόγοι που η γραμμική συνάρτηση δε χρησιμοποιείται ως συνάρτηση ενεργοποίησης σε κανένα κρυφό στρώμα. Ωστόσο χρησιμοποιείται ως συνάρτηση ενεργοποίησης στο στρώμα εξόδου των νευρωνικών δικτύων που χρησιμοποιούνται για προβλήματα παλινδρόμησης, αφού σε τέτοιες περιπτώσεις, θέλουμε να έχουμε ως έξοδο του δικτύου μια τιμή που να μην έχει όρια (πρόβλεψη μοντέλου), κάτι που είναι χαρακτηριστικό της συνάρτησης αυτής αφού έχει πεδίο τιμών το $(-inf, inf)$.

2.10.4.2. Βηματική Συνάρτηση (threshold function):



Η βηματική συνάρτηση είναι μια από τις πιο απλές συναρτήσεις ενεργοποίησης. Όπως φαίνεται και στην παραπάνω εικόνα, όταν το βεβαρυμένο άθροισμα είναι μικρότερο του μηδενός τότε η συνάρτηση επιστρέφει την τιμή μηδέν αλλιώς την τιμή 1. Είναι λοιπόν φανερό ότι μια τέτοια συνάρτηση θα μπορούσε να φανεί χρήσιμη σε περιπτώσεις όπου θέλουμε να προβλέψουμε μια δίτιμη τιμή(ναι ή όχι, μαύρο ή άσπρο κ.λ.π.) με ένα δυαδικό ταξινομητή (binary classifier).

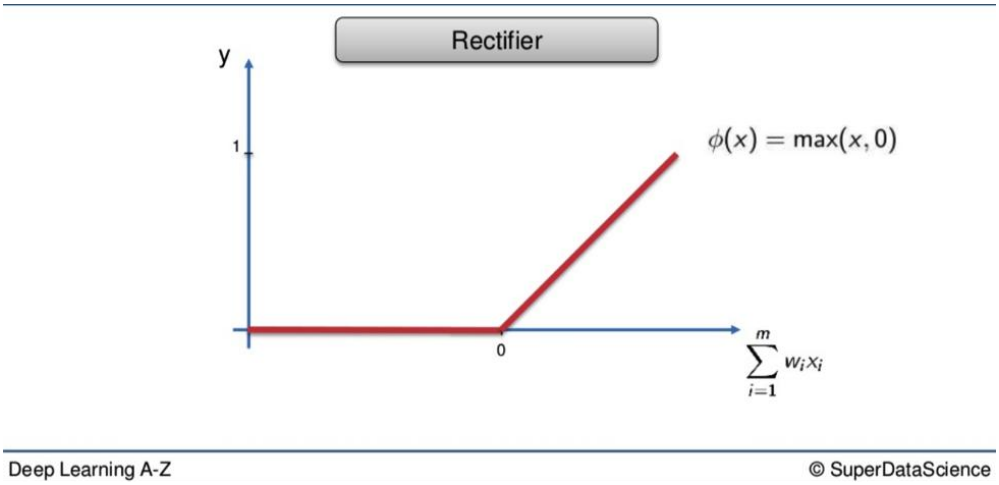
2.10.4.3. Σιγμοειδής Συνάρτηση (sigmoid or logistic):



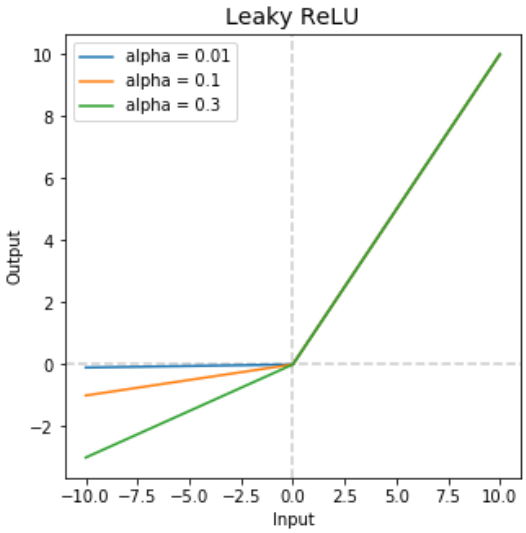
Η σιγμοειδής συνάρτηση, για τιμές μεγαλύτερες του μηδενός επιστρέφει άσσο, ενώ για τιμές από το μηδέν και κάτω επιστρέφει μηδέν. Ωστόσο η βασική της διαφορά από τη βηματική συνάρτηση, είναι ότι λόγω της καμπυλότητας της είναι πολύ πιο κατάλληλη για την περιγραφή πιθανοτήτων, αφού εκτός από τις τιμές 0, 1 παίρνει και όλες τις ενδιάμεσες τιμές. Για το λόγο αυτό, όπως θα δούμε και παρακάτω χρησιμοποιείται κατά κύριο λόγο, ως συνάρτηση ενεργοποίησης στο στρώμα εξόδου ενός νευρωνικού δικτύου, όταν αυτό επιλύει ένα πρόβλημα ταξινόμησης (classification). Σε ένα τέτοιο πρόβλημα θέλουμε η έξοδος του

νευρωνικού δικτύου να αντιστοιχεί στην πιθανότητα μια συγκεκριμένη εγγραφή να ανήκει σε κάθε μια από τις υπάρχουσες κλάσεις (αν έχουμε τρεις κλάσεις θέλουμε η έξοδος του νευρωνικού δικτύου να είναι τρεις πιθανότητες) και η σιγμοειδής συνάρτηση εξυπηρετεί τέλεια αυτό το σκοπό. Ωστόσο, η σιγμοειδής συνάρτηση, συνδέεται με ένα πρόβλημα γνωστό ως εξαφανιζόμενη κλίση (vanishing gradient). Όπως κάποιος μπορεί να παρατηρήσει και από τη γραφική της παράσταση, η σιγμοειδής συνάρτηση από κάποιες τιμές του αθροίσματος $\sum w_i x_i$ και μετά τείνει να επιστρέφει την ίδια (με ελάχιστη διαφορά) τιμή. Αυτό σημαίνει πρακτικά ότι από ένα σημείο και μετά η κλίση (gradient) της συνάρτησης κόστους με βάση την οποία αξιολογείται η απόδοση του δικτύου εξαφανίζεται, δηλαδή το δίκτυο δε μπορεί να εκπαιδευτεί περεταίρω ή εκπαιδεύεται με πολύ αργό ρυθμό.

2.10.4.4. Ανορθωτική Γραμμική Συνάρτηση (rectifier – ReLU function):



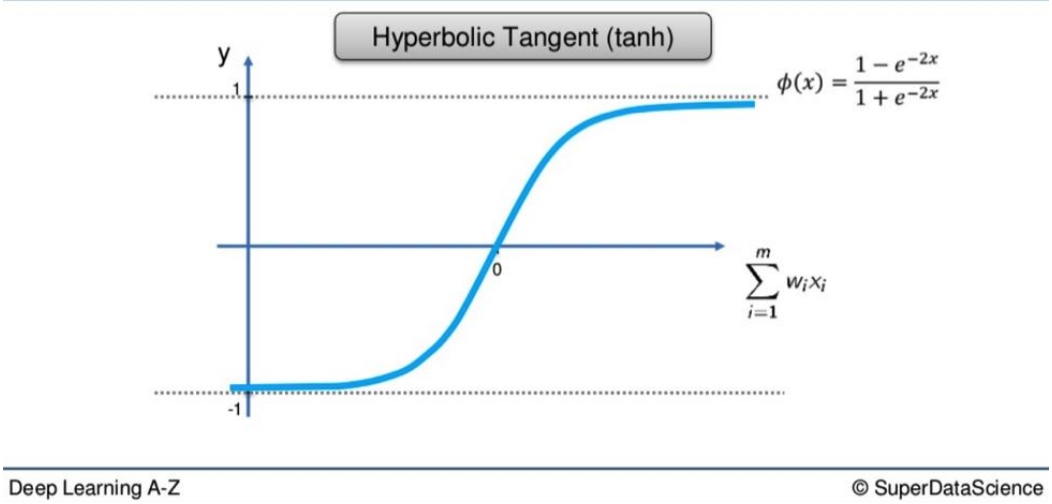
Η ανορθωτική γραμμική συνάρτηση είναι ίσως η πιο διαδεδομένη συνάρτηση που χρησιμοποιείται στα νευρωνικά δίκτυα λόγω της απλότητας, της αποτελεσματικότητας και της βασικής της ιδιότητας να μηδενίζει το βεβαρυμμένο άθροισμα όταν αυτό έχει τιμή μικρότερη του μηδενός και να επιστρέφει την ίδια την τιμή του όταν αυτό είναι μεγαλύτερο του μηδενός. Ωστόσο, η ανορθωτική συνάρτηση, συνδέεται με ένα βασικό πρόβλημα. Για αρνητικές τιμές του $\sum w_i x_i$, η συνάρτηση επιστρέφει πάντα μηδέν, κάτι που μπορεί να οδηγήσει ένα δίκτυο στο να μη μπορεί να διαχειριστεί αποτελεσματικά τις αρνητικές τιμές εισόδου. Αυτό συμβαίνει γιατί για αρνητικές τιμές του αθροίσματος η κλίση-gradient παραμένει μηδενική ,οι νευρώνες τείνουν να απενεργοποιούνται και τα βάρη να μην ενημερώνονται αφού παγιδεύονται σε ένα τοπικό ελάχιστο. Το πρόβλημα αυτό, είναι γνωστό ως «dying relu». Ωστόσο, μικρές παραλλαγές της μεθόδου δίνουν λύση στο πρόβλημα αυτό, δίνοντας μια μικρή κλίση (που διαμορφώνεται από μια παράμετρο alpha) στην οριζόντια γραμμή της γραφικής παράστασης τηε ανορθωτικής συνάρτησης. Μια από αυτές τις παραλλαγές ονομάζεται Leaky ReLU, και η γραφική της παράσταση φαίνεται στην παρακάτω εικόνα:



Πηγή: towards data science

Τέλος, αξίζει να σημειωθεί ότι η ReLU έχει μικρότερο υπολογιστικό κόστος (όταν υπόκεινται σε βελτιστοποίηση) σε σχέση με τη σιγμοειδή και την υπερβολική εφαπτομένη και έτσι είναι πιο αποτελεσματική και γρήγορη σε βαθιά νευρωνικά δίκτυα με πολλά κρυφά στρώματα και νευρώνες (deep learning networks).

2.10.4.5. Υπερβολική Εφαπτομένη (TanH)

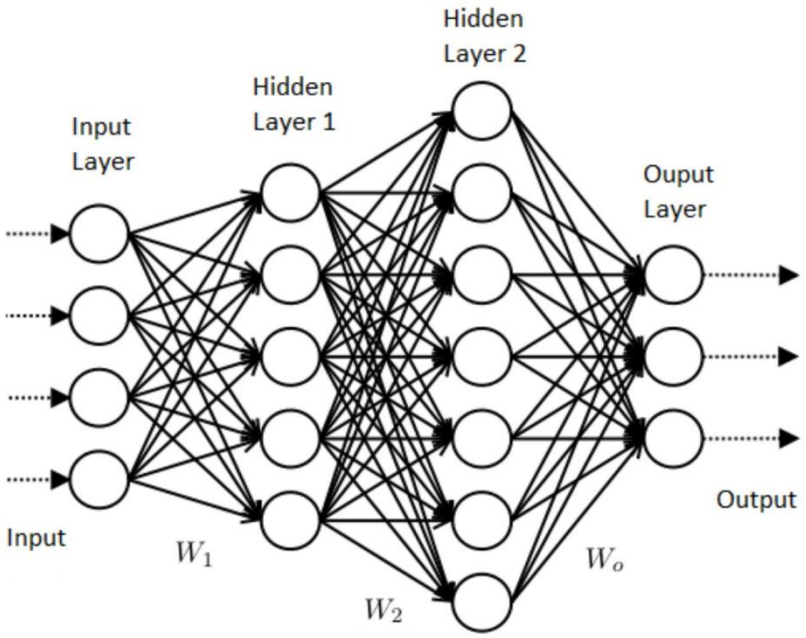


Η μορφή της υπερβολικής εφαπτομένης, θυμίζει σε μεγάλο βαθμό τη σιγμοειδή συνάρτηση, με βασική διαφορά ότι έχει πεδίο τιμών το $(-1,1)$. Όπως η σιγμοειδής έτσι και η υπερβολική εφαπτομένη είναι κατάλληλη για να εκφράσει πιθανότητες λόγω του πεδίου τιμών της. Ωστόσο υπερτερεί σε σχέση με τη σιγμοειδή σε δυο βασικά χαρακτηριστικά. Το πρώτο είναι ότι σε περιπτώσεις που τα δεδομένα είναι κοντά στο μηδέν και γενικά η τιμή του βεβαρυμμένου αθροίσματος είναι κοντά στο μηδέν τότε έχουμε μεγαλύτερες κλίσεις όπως φαίνεται και από τη γραφική παράσταση, που οδηγούν σε καλύτερη εκπαίδευση του δικτύου και το δεύτερο ότι αποφεύγει το σφάλμα στον υπολογισμό των

κλίσεων λόγω του πεδίου τιμών της. Τέλος αξίζει να σημειωθεί ότι και αυτή η συνάρτηση είναι επιρρεπής στο πρόβλημα της εξαφανιζόμενης κλίσης(vanishing gradient).

2.10.5. Η Εκμάθηση του Νευρωνικού Δικτύου

Όπως ακριβώς και στον ανθρώπινο οργανισμό, ένας νευρώνας δε μπορεί να ανταποκριθεί από μόνος του αποτελεσματικά στα ερεθίσματα που λαμβάνει. Έτσι και ένα τεχνητό νευρωνικό δίκτυο όσο πιο πολλούς νευρώνες έχει τόσο πιο αποτελεσματικό είναι, ειδικά όταν μιλάμε για δύσκολα προβλήματα. Η δομή ενός νευρωνικού δικτύου φαίνεται στο παρακάτω σχήμα:



Ένα νευρωνικό δίκτυο αποτελείται από το στρώμα εισόδου (Input Layer), το οποίο έχει τόσους νευρώνες όσες οι επεξηγηματικές μεταβλητές του μοντέλου, τα κρυφά στρώματα (Hidden Layers) που τόσο ο αριθμός τους όσο και ο αριθμός των νευρώνων τους ποικίλει ανάλογα με την αρχιτεκτονική του νευρωνικού δικτύου. Τέλος, έχουμε το στρώμα εξόδου (Output Layer), από το οποίο προκύπτει το αποτέλεσμα του δικτύου και ο αριθμός των νευρώνων του ποικίλει ανάλογα με το πρόβλημα που επιλύει το νευρωνικό δίκτυο. Αν χρησιμοποιείται για ένα πρόβλημα παλινδρόμησης τότε το στρώμα εξόδου θα έχει ένα νευρώνα που θα παράγει την πρόβλεψη του δικτύου για μια τιμή απόκρισης ενώ αν χρησιμοποιείται για πρόβλημα ταξινόμησης θα έχει τόσους νευρώνες όσες οι κλάσεις στις οποίες θέλουμε να ταξινομήσουμε έναν αριθμό παρατηρήσεων.

Οι νευρώνες κάθε στρώματος συνδέονται με εκείνους του επόμενου στρώματος, με τη βοήθεια συνάψεων-ακμών. Κάθε ακμή έχει και ένα βάρος. Σε κάθε

νευρώνα εκτελείται η λειτουργία που περιγράφηκε παραπάνω. Πώς όμως το νευρωνικό δίκτυο «μαθαίνει»;

Ας υποθέσουμε ότι έχουμε ένα νευρωνικό δίκτυο που επιλύει ένα πρόβλημα παλινδρόμησης. Αρχικά, οι παρατηρήσεις εισέρχονται στο στρώμα εισόδου. Από εκεί, μεταβαίνουν στους νευρώνες του πρώτου κρυφού στρώματος αφού επιβαρυνθούν με τα αρχικά βάρη των ακμών. Στους νευρώνες υπολογίζεται το βεβαρυμμένο άθροισμα και η τιμή της συνάρτησης ενεργοποίησης του νευρώνα για το άθροισμα αυτό καθορίζει το αν ο νευρώνας θα ενεργοποιηθεί ή όχι. Στην συνέχεια οι τιμές που προκύπτουν από τη διαδικασία σε κάθε νευρώνα θα αποτελέσουν τις τιμές εισόδου για το επόμενο κρυφό στρώμα και η διαδικασία επαναλαμβάνεται μέχρι να φτάσουμε στο τελευταίο κρυφό στρώμα. Η διαδικασία κατά την οποία οι νευρώνες από αριστερά προς τα δεξιά ενεργοποιούνται με βάση την τιμή εισόδου τους και παράγεται το αποτέλεσμα του νευρωνικού δικτύου, ονομάζεται εμπρόσθια τροφοδότηση (forward propagation). Οι τιμές του κάθε νευρώνα του τελευταίου στρώματος, καταλήγουν στο στρώμα εξόδου, και έτσι διαμορφώνεται η πρόβλεψη του νευρωνικού δικτύου $\hat{y}$ για τις συγκεκριμένες παρατηρήσεις. Στη συνέχεια υπολογίζεται το σφάλμα της πρόβλεψης μέσω μιας συνάρτησης κόστους. Η συνάρτηση κόστους μπορεί να είναι οποιαδήποτε συνάρτηση επιθυμούμε να χρησιμοποιήσουμε για να αξιολογήσουμε και να εκπαιδεύσουμε το μοντέλο μας. Στην περίπτωση ενός μοντέλου παλινδρόμησης, συνήθως χρησιμοποιείται το μέσο τετραγωνικό σφάλμα MSE ($C = \frac{1}{n}\sum(\hat{y}_i - y_i)^2$). Σκοπός είναι, η τιμή αυτής της συνάρτησης να ελαχιστοποιηθεί δηλαδή η πρόβλεψη για το y να είναι όσο το δυνατόν πιο κοντά στην πραγματική του τιμή. Μια χαμηλή τιμή της συνάρτησης κόστους συνεπάγεται με μεγάλη ακρίβεια του νευρωνικού δικτύου. Αφού λοιπόν υπολογιστεί η τιμή της συνάρτησης κόστους, τα βάρη ανανεώνονται κατάλληλα και η διαδικασία επαναλαμβάνεται με στόχο η νέα τιμή της συνάρτησης κόστους να είναι μικρότερη. Η διαδικασία επαναλαμβάνεται μέχρι η τιμή της συνάρτησης κόστους να πλησιάσει το 0 ή να μη βελτιώνεται άλλο. Τότε ξέρουμε ότι τα βάρη των ακμών-συνάψεων είναι τα βέλτιστα. Όταν όλα τα δεδομένα περάσουν από το νευρωνικό δίκτυο τότε έχει ολοκληρωθεί μια εποχή (epoch). Το πόσα δεδομένα διαχειρίζεται κάθε φορά το νευρωνικό δίκτυο καθορίζεται από το χρήστη με μια μεταβλητή που καλείται batch. Αν το batch είναι 5 αυτό σημαίνει ότι 5 παρατηρήσεις τη φορά θα αποτελούν την είσοδο του νευρωνικού δικτύου. Αφού οι 5 αυτές παρατηρήσεις επεξεργασθούν από το νευρωνικό δίκτυο θα παραχθούν 5 προβλέψεις, μια για κάθε μια από αυτές και η συνάρτηση κόστους θα υπολογιστεί για το σύνολο των πέντε αυτών προβλέψεων. Στη συνέχεια θα ανανεωθούν τα βάρη του δικτύου, και η διαδικασία συνεχίζεται με την είσοδο των επόμενων 5 παρατηρήσεων.

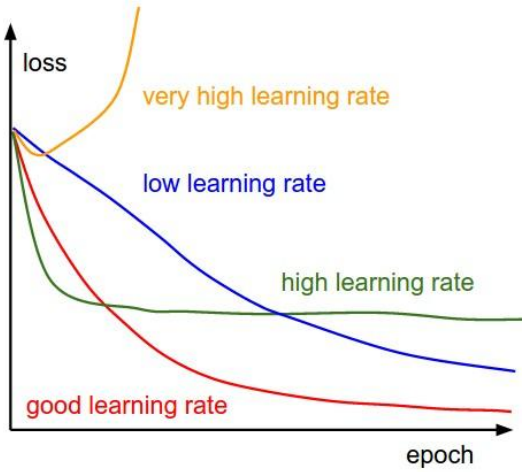
2.10.6. Διαδικασία Ενημέρωσης Βαρών

Είδαμε ότι αφού κάνει μια πρόβλεψη το νευρωνικό δίκτυο ενημερώνονται στη συνέχεια τα βάρη των συνάψεων μεταξύ του στρώματος εισόδου και του πρώτου κρυφού στρώματος, και στη συνέχεια και τα υπόλοιπα βάρη με τη λειτουργία των νευρώνων. Η διαδικασία αυτή κατά την οποία τα αποτελέσματα του νευρωνικού δικτύου αξιολογούνται και ενημερώνονται τα βάρη του σταδιακά, ονομάζεται Backpropagation.

Σκοπός μας όπως είδαμε είναι να ελαχιστοποιήσουμε τη συνάρτηση κόστους βρίσκοντας τα κατάλληλα βάρη για τις συνάψεις. Ωστόσο το να υπολογίσουμε τη συνάρτηση κόστους για κάθε διαφορετικό συνδυασμό βαρών είναι υπολογιστικά αδύνατο. Έτσι χρησιμοποιούμε ποιο έξυπνες μεθόδους ελαχιστοποίησης με την πιο γνωστή να είναι αυτή της απότομης καθόδου (gradient descent).

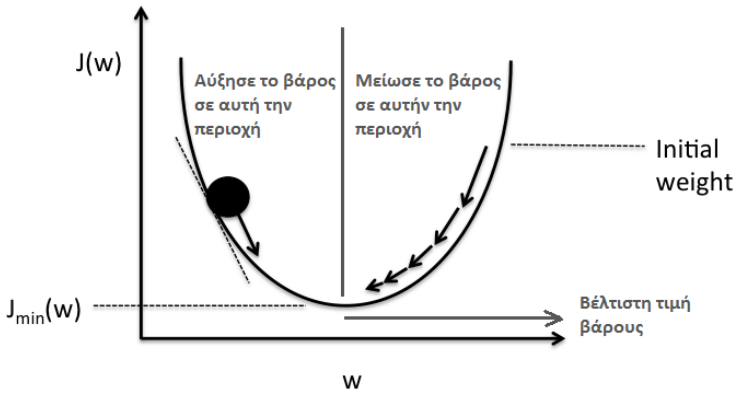
Η gradient descent ελαχιστοποιεί τη συνάρτηση κόστους ως εξής:

- Αρχικά υπολογίζει την παράγωγο της συνάρτησης κόστους ως προς κάθε βάρος και την πολλαπλασιάζει με ένα σταθερό αριθμό, ο οποίος καλείται ρυθμός εκμάθησης (learning rate): $\frac{d(loss)}{d(weight)} * learning\ rate$
- Ο ρυθμός εκμάθησης καθορίζει το ρυθμό με τον οποίο ενημερώνονται τα βάρη ενός νευρωνικού δικτύου, δηλαδή με απλά λόγια το πόσο πολύ ή λίγο θα μεταβληθεί το κάθε βάρος. Η τιμή της μεταβλητής αυτής καθορίζεται από εκείνον που σχεδιάζει το νευρωνικό δίκτυο και δεν ακολουθεί κάποιο κανόνα. Ωστόσο, μικρές τιμές της οδηγούν το νευρωνικό δίκτυο σε πολύ αργή εκπαίδευση και συνεπώς αργή σύγκλιση, ενώ μεγάλες τιμές μπορεί να οδηγήσουν το νευρωνικό στο να μη βρει τις βέλτιστες τιμές για τα βάρη με αποτέλεσμα να αποκλίνει. Στο παρακάτω γράφημα βλέπουμε, πως μεταβάλλεται η τιμή της συνάρτησης κόστους για διαφορετικό ρυθμό εκμάθησης κατά τη διάρκεια εκμάθησης του νευρωνικού δικτύου.



Effect of various learning rates on convergence (http://cs231n.github.io/neural-networks-3/)

- Η παράγωγος $\frac{d(loss)}{d(weight)} * learning\ rate$, είναι η κλίση (gradient) και καθορίζει την ανανέωση των βαρών ως εξής:
    - Έστω ότι για ένα βάρος w η βέλτιστη τιμή είναι w= 2. Στην περίπτωση αυτή η συνάρτηση κόστους θα είναι μηδέν, αφού η πρόβλεψη του δικτύου $\hat{y}$ θα ταυτίζεται με την πραγματική τιμή y.
    - Αν w<2, τότε η συνάρτηση κόστους θα έχει μια θετική τιμή, αλλά η παράγωγος θα είναι αρνητική, κάτι που σημαίνει, ότι μια αύξηση στο βάρος θα μειώσει την τιμή της συνάρτησης κόστους.
    - Αν w=2, τότε η τιμή της συνάρτησης κόστους θα είναι 0, και έχουμε βρει τη βέλτιστη τιμή του βάρους για το μοντέλο μας.
    - Αν w<2, τότε η συνάρτηση κόστους θα έχει πάλι μια θετική τιμή αλλά τώρα η παράγωγος θα είναι θετική, κάτι που σημαίνει, ότι μια αύξηση στο βάρος θα αυξήσει την τιμή της συνάρτησης κόστους ακόμα περισσότερο.
- Με αυτόν τον τρόπο λοιπόν ενημερώνονται τα βάρη του νευρωνικού δικτύου με τη βοήθεια της μεθόδου της απότομης καθόδου. Η διαδικασία φαίνεται στην παρακάτω εικόνα:



Στο σημείο αυτό αξίζει να σημειωθεί πως ο ρυθμός εκμάθησης καθορίζει το πόσο γρήγορα ή αργά κατεβαίνουμε κατά μήκος της «πλαγιάς». Το τελικό βάρος δίνεται από τη σχέση: $w_{ij}(t+1) = w_{ij}(t) - \frac{\partial C}{\partial w_{ij}} l$, όπου C η συνάρτηση κόστους, l ο ρυθμός εκμάθησης και w_{ij} ένα συγκεκριμένο βάρος που καθορίζεται από τα i,j.

Ωστόσο, όπως είδαμε στην προηγούμενη παράγραφο η χρήση της gradient descent συνδέεται με δυο βασικά μειονεκτήματα. Το γεγονός ότι για μεγάλο αριθμό δεδομένων, λόγω της επαναληπτικότητας της, η μέθοδος θα καθυστερήσει πολύ να συγκλίνει και το γεγονός ότι κινδυνεύει να παγιδευτεί σε

κάποιο τοπικό ελάχιστο. Για το λόγο αυτό κατά βάση χρησιμοποιούμε τη stochastic gradient descent ως μέθοδο βελτιστοποίησης για την ανανέωση των βαρών.

2.10.7. Αλγόριθμος Λειτουργίας Νευρωνικού Δικτύου

Ας δούμε λοιπόν συνοπτικά, τα βήματα σε ψευδοκώδικα που εκτελεί ένα νευρωνικό δίκτυο για να κάνει προβλέψεις του για έναν αριθμό παρατηρήσεων:

[1] Αρχικοποίησε με τυχαίο τρόπο τα βάρη των συνάψεων με μικρές τιμές κοντά στο 0.
[2] Τοποθέτησε τις πρώτες παρατηρήσεις (τόσες όσες καθορίζονται από το batch που έχουμε ορίσει) στους νευρώνες του στρώματος εισόδου. Κάθε επεξηγηματική μεταβλητή (feature – independent variable), τοποθετείται στον κατάλληλο νευρώνα που ανήκει.
[3] Forward Propagation: Από τα αριστερά προς τα δεξιά, οι νευρώνες ενεργοποιούνται και παράγεται η πρόβλεψη του νευρωνικού δικτύου.
[4] Σύγκρινε την πρόβλεψη του νευρωνικού δικτύου $\hat{y}$ με την πραγματική της τιμή y, και υπολόγισε την τιμή της συνάρτησης κόστους (σφάλμα πρόβλεψης).
[5] Back Propagation: Από αριστερά προς τα δεξιά ενημέρωσε τα βάρη σύμφωνα με την τιμή της συνάρτησης κόστους. Ο ρυθμός εκμάθησης καθορίζει το πόσο πολύ ή λίγο θα διορθωθούν τα βάρη. Τα βάρη διορθώνονται με βάση το πόσο είναι υπεύθυνα για το παραγόμενο σφάλμα.
[6] Επανάλαβε τα βήματα 1-5 για κάθε batch παρατηρήσεων μέχρι να περάσουν όλα τα δεδομένα από το νευρωνικό δίκτυο.
[7] Όταν όλα τα δεδομένα περάσουν από το νευρωνικό δίκτυο τότε έχει πραγματοποιηθεί μια εποχή. Πραγματοποίησε και άλλες εποχές.

2.10.8. Υπερεκπαίδευση Νευρωνικού Δικτύου (Overfitting)

Όπως κάθε τεχνική μηχανικής μάθησης έτσι και τα νευρωνικά δίκτυα είναι σε κάποιο βαθμό ευαίσθητα στο φαινόμενο της υπερεκπαίδευσης. Όπως είδαμε σε προηγούμενη παράγραφο δυο βασικές τεχνικές ώστε να αποφύγουμε τέτοια φαινόμενα είναι το να εκπαιδεύουμε το μοντέλο μας σε μεγάλο αριθμό δεδομένων και να προσπαθούμε να έχει όσο το δυνατόν πιο απλή αρχιτεκτονική. Ωστόσο, ένα νευρωνικό δίκτυο για να κατασκευαστεί πρέπει να προσδιοριστεί ένας αριθμός υπερπαραμέτρων (hyperparameters), μερικές τιμές των οποίων πολλές φορές μπορούν να οδηγήσουν σε φαινόμενα overfitting.

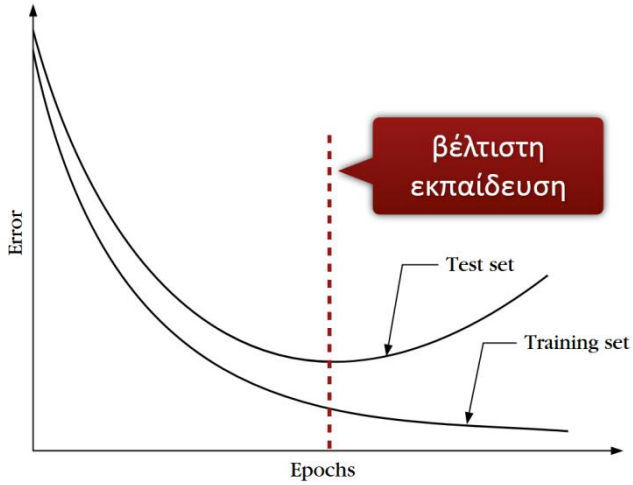
Υπερεκπαίδευση έχουμε όταν η απόδοση-ακρίβεια ενός νευρωνικού δικτύου παρουσιάζει σημαντική διαφορά όταν αυτό εφαρμόζεται στα δεδομένα εκπαίδευσης (training set), σε σχέση με όταν αυτό εφαρμόζεται στα δεδομένα

επαλήθευσης (test set). Στο σημείο αυτό αξίζει να σημειωθεί πως δεν υπάρχουν συγκεκριμένες τιμές για τις οποίες ένα μοντέλο παρουσιάζει overfitting. Η τιμή μιας παραμέτρου για ένα συγκεκριμένο νευρωνικό δίκτυο που είναι αρκετά γενικευμένο μπορεί να οδηγεί ένα διαφορετικό νευρωνικό δίκτυο σε υπερεκπαίδευση. Για το λόγο αυτό είναι πολύ σημαντικό πάντα να σαρώνουμε όσο το δυνατόν μεγαλύτερο μέρος του φασικού χώρου των υπερπαραμέτρων ώστε να δούμε ποιες είναι κατάλληλες για το μοντέλο μας, από τη σκοπιά της γενίκευσης του. Πρέπει να θυμόμαστε άλλωστε ότι ένα μοντέλο μηχανικής μάθησης πρέπει να μπορεί να «σκέφτεται» και να παίρνει τις σωστές αποφάσεις σε οποιαδήποτε δεδομένα του δοθούν. Έτσι παρατηρείται το γεγονός πολλές φορές να θυσιάζουμε λίγο από την απόδοση του μοντέλου ώστε να εξασφαλίσουμε τη γενίκευση του.

Παρακάτω θα δούμε μερικές από τις υπερπαραμέτρους του νευρωνικού δικτύου και πότε αυτές οδηγούν σε υπερπροσαρμογή:

- Ρυθμός εκμάθησης: Όπως είδαμε ένας χαμηλός ρυθμός εκμάθησης μπορεί να καθυστερήσει πολύ το νευρωνικό δίκτυο από τη σύγκλιση αλλά επίσης μπορεί να οδηγήσει και σε overfitting, αφού η μέθοδος βελτιστοποίησης μπορεί να «παγιδευτεί» σε τοπικά ελάχιστα του συγκεκριμένου σετ δεδομένων. Ένας πιο μεγάλος ρυθμός εκμάθησης μπορεί να συμβάλλει στην αποφυγή του φαινομένου αφού, τα βήματα που θα κάνει η μέθοδος βελτιστοποίησης κατά την ενημέρωση των βαρών θα είναι μεγαλύτερα αποφεύγοντας έτσι τοπικά ελάχιστα.
- Μέγεθος batch: Το μέγεθος των batch, δηλαδή του αριθμού των παρατηρήσεων που παίρνει ως είσοδο κάθε φορά το νευρωνικό, επίσης μπορεί να οδηγήσει σε overfitting. Αξίζει να σημειωθεί ότι το μέγεθος των batch συνδέεται κατά μια έννοια με το ρυθμό εκμάθησης, αφού όσο μεγαλύτερο είναι το batch, τόσο περισσότερα δεδομένα εισέρχονται στο νευρωνικό ανά φορά, και τόσο μεγαλύτερα περιθώρια έχουμε να αυξήσουμε το ρυθμό εκμάθησης χωρίς να κινδυνεύουμε από overfitting. Άρα όσο μικρότερο το batch, τόσο μικρότερος και ο ρυθμός εκμάθησης και άρα μεγαλύτερη πιθανότητα για overfitting. Ωστόσο, αν θέσουμε ένα μεγάλο batch, είναι πιθανό ο υπολογιστής να μη μπορεί να το διαχειριστεί υπολογιστικά, αφού δε θα μπορεί να εκτελέσει για τόσες πολλές παρατηρήσεις παράλληλους υπολογισμούς.
- Εποχές: Ο αριθμός των εποχών ενός νευρωνικού δικτύου επίσης μπορεί να οδηγήσει σε υπερεκπαίδευση όταν είναι πολύ μεγάλος. Ένας μεγάλος αριθμός εποχών οδηγεί το δίκτυο να συνεχίσει να εκπαιδεύεται στα δεδομένα εκπαίδευσης και να συνεχίζουν να ανανεώνονται τα βάρη ακόμα και όταν αυτό έχει συγκλίνει με

αποτέλεσμα αυτό να προσαρμόζεται πολύ περισσότερο στα δεδομένα αυτά. Για το λόγο αυτό είναι σημαντικό πάντα να ελέγχουμε τη γραφική παράσταση της συνάρτησης κόστους συναρτήσει των εποχών για τα δεδομένα εκπαίδευσης και επαλήθευσης, ώστε να επιλέγουμε το βέλτιστο αριθμό εποχών (εκεί που η συνάρτηση κόστους παρουσιάζει χαμηλή τιμή και η απόδοση του δικτύου στα δεδομένα εκπαίδευσης και επαλήθευσης είναι παρόμοια).



2.10.9. Αρχιτεκτονική Νευρωνικού Δικτύου

Όπως έχει ήδη αναφερθεί δεν υπάρχει κάποιος επιστημονικά αποδεδειγμένος κανόνας σχετικά με την αρχιτεκτονική (αριθμός των κρυφών στρωμάτων και των νευρώνων τους) που πρέπει να έχει ένα νευρωνικό δίκτυο. Ωστόσο, υπάρχουν κάποιοι εμπειρικοί κανόνες και πρακτικές που είναι αποτελεσματικές σε πολλές περιπτώσεις. Μερικοί από αυτούς είναι οι παρακάτω:

- Η αρχιτεκτονική του νευρωνικού δικτύου, επηρεάζεται από το μέγεθος του συνόλου εκπαίδευσης. Όσο πιο μικρό είναι αυτό, τόσο πιο απλή αρχιτεκτονική πρέπει να έχει το νευρωνικό δίκτυο (λίγα κρυφά στρώματα με λίγους νευρώνες). Αυτό γιατί μια πιο περίπλοκη αρχιτεκτονική λόγω του μικρού αριθμού δεδομένων θα μπορούσε να οδηγήσει σε υπερεκπαίδευση.
- Ξεκινάμε πάντα με μια απλή αρχιτεκτονική (π.χ. ένα κρυφό στρώμα με λίγους νευρώνες) και σταδιακά προσθέτουμε νευρώνες ή/και στρώματα ελέγχοντας παράλληλα την απόδοση-ακρίβεια του δικτύου με το σύνολο επαλήθευσης.
- Δοκιμάζουμε ως αριθμό των νευρώνων των κρυφών στρωμάτων το μέσο όρο των νευρώνων του προηγούμενου και του επόμενου στρώματος.

Οι παραπάνω κανόνες ίσως να αποτελούν ένα μπούσουλα για κάποιον που θέλει να κατασκευάσει ένα νευρωνικό δίκτυο, σε καμία περίπτωση όμως δεν είναι απόλυτοι. Μπορεί η βέλτιστη αρχιτεκτονική ενός νευρωνικού δικτύου για ένα συγκεκριμένο πρόβλημα να μην υπακούει σε κανένα από τους παρακάτω κανόνες. Για αυτό είναι πολύ σημαντικό πάντα να δοκιμάζουμε διάφορες αρχιτεκτονικές από απλές μέχρι περίπλοκες ώστε να είμαστε σίγουροι για το ποια είναι αυτή που ταιριάζει περισσότερο στο πρόβλημα που έχουμε να λύσουμε.

2.10.10. Βαθιά Μάθηση (Deep Learning)

Η βαθιά μάθηση, είναι μια τεχνική μηχανικής μάθησης, που κάνει έναν υπολογιστή να σκέφτεται με παρόμοιο τρόπο με τον άνθρωπο. Είναι η τεχνολογία που κρύβεται πίσω από τα αυτόνομα αυτοκίνητα, και τους επιτρέπει να αναγνωρίζουν τα σήματα οδικής κυκλοφορίας και να διαχωρίζουν τους πεζούς από τους φανοστάτες κ.λ.π.. Είναι ακόμη το κλειδί των εικονικών βοηθών (virtual assistants) σε πολλές έξυπνες συσκευές, όπως τηλέφωνα, ρολόγια και τηλεοράσεις. Η βαθιά μάθηση αποκτάει όλο και περισσότερη φήμη λόγω των αποτελεσμάτων που προσφέρει. Τα μοντέλα βαθιάς μάθησης, μπορούν να επιτύχουν απίστευτη απόδοση-ακρίβεια, ακόμα και να φτάσουν σε απόδοση ανθρωπίνων επιπέδων. Στη συγκεκριμένη διπλωματική δε θα επεκταθούμε στον τομέα αυτό της μηχανικής μάθησης ωστόσο θεωρήσαμε πως είναι αρκετά χρήσιμο μιας και αναλύσαμε τη λειτουργία ενός νευρωνικού δικτύου να αναφερθούμε συνοπτικά και στα βαθιά νευρωνικά δίκτυα.

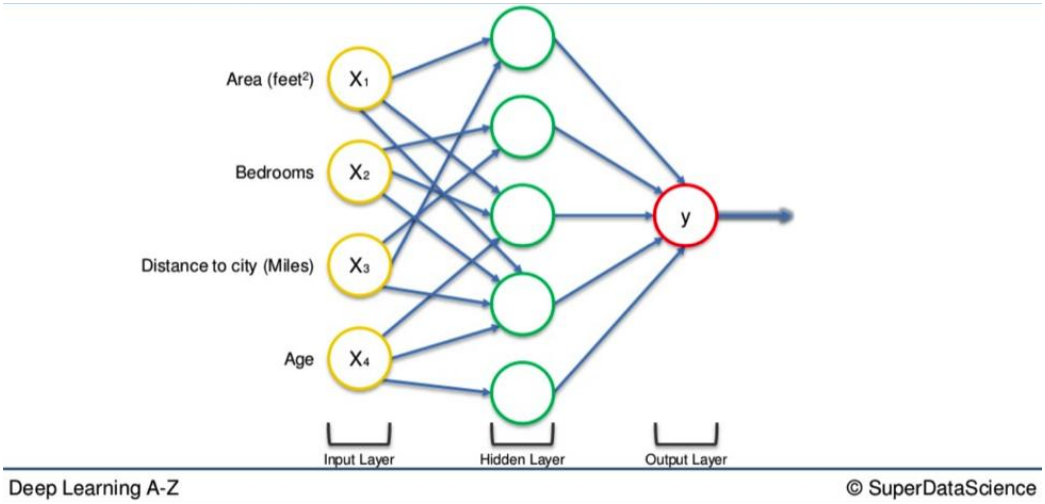
Με τον όρο deep learning, εννοούμε νευρωνικά δίκτυα με πολλαπλά κρυφά στρώματα και νευρώνες. Ένα απλό τεχνητό νευρωνικό δίκτυο, μπορεί να έχει 2-3 κρυφά στρώματα, όταν ένα βαθύ νευρωνικό δίκτυο μπορεί να έχει 150 κρυφά στρώματα. Για την εκπαίδευση ενός deep neural network χρησιμοποιούνται μεγάλα σύνολα δεδομένων, ενώ το δίκτυο έχει παράλληλα τη δυνατότητα να μαθαίνει από μόνο του τις επεξηγηματικές μεταβλητές και να κατανοεί τη σημαντικότητα τους, χωρίς έτσι να χρειάζεται ο χρήστης να κάνει προηγουμένως εξαγωγή των σημαντικών μεταβλητών (feature extraction). Ωστόσο αξίζει να σημειωθεί ότι ένα βαθύ νευρωνικό δίκτυο χρειάζεται πολύ περισσότερο χρόνο για την εκπαίδευση του σε σχέση με ένα απλό νευρωνικό δίκτυο.

2.10.11. Ερμηνεία Νευρωνικού Δικτύου

Όπως και σε κάθε μοντέλο μηχανικής μάθησης έτσι και στα νευρωνικά δίκτυα, πολύ σημαντικό ρόλο διαδραματίζει και η ερμηνεία τους. Η ερμηνεία των νευρωνικών δικτύων, έγκειται στην κατανόηση των βαρών και των συνάψεων μεταξύ των διαφόρων στρωμάτων του δικτύου.

Είναι σημαντικό αρχικά να σημειωθεί ότι μεταξύ δυο στρωμάτων του δικτύου, δεν είναι απαραίτητο όλοι οι νευρώνες του ενός στρώματος να συνδέονται με συνάψεις με όλους τους νευρώνες του άλλου στρώματος. Στην πράξη μέσα από την εκπαίδευση του νευρωνικού δικτύου και την ενημέρωση των βαρών, τα βάρη πολλών συνάψεων μηδενίζονται με αποτέλεσμα οι συνάψεις αυτές να απορρίπτονται. Οι συνάψεις που παραμένουν μετά το πέρας της διαδικασίας της εκπαίδευσης μαζί με τα βάρη τους αποτελούν το κλειδί στην ερμηνεία του νευρωνικού δικτύου. Ας δούμε λοιπόν μέσα από ένα παράδειγμα πως λειτουργεί αυτό.

Έστω ότι έχουμε κατασκευάσει ένα νευρωνικό δίκτυο, για να λύσουμε ένα πρόβλημα παλινδρόμησης. Συγκεκριμένα θέλουμε να προβλέψουμε την τιμή ενός σπιτιού (dependent variable), βασιζόμενοι σε κάποιες επεξηγηματικές μεταβλητές (features), που είναι το εμβαδόν του (Area), ο αριθμός των δωματίων (Bedrooms), η απόσταση από την πόλη (Distance to city) και η ηλικία του σπιτιού (Age). Μετά λοιπόν από πειραματισμούς καταλήγουμε σε ένα νευρωνικό δίκτυο με ένα κρυφό στρώμα και 5 νευρώνες, το εκπαιδεύουμε και στο παρακάτω σχήμα βλέπουμε την εικόνα του μετά την εκπαίδευση. Όπως είναι φανερό δε συνδέονται όλοι οι νευρώνες του στρώματος εισόδου με όλους τους νευρώνες του κρυφού στρώματος. Αυτό σημαίνει ότι τα βάρη μερικών συνάψεων μηδενίστηκαν αφού το νευρωνικό δίκτυο κατά την εκπαίδευση του τις θεώρησε μη σημαντικές. Ας παρατηρήσουμε με λίγο περισσότερη προσοχή λοιπόν τη μορφή του δικτύου και ας προσπαθήσουμε να ερμηνεύσουμε τα αποτελέσματά του.



Ας επικεντρωθούμε αρχικά στον πρώτο νευρώνα του κρυφού στρώματος. Βλέπουμε ότι έρχονται συνάψεις από το στρώμα εισόδου μόνο από τη μεταβλητή Εμβαδόν και Απόσταση από την πόλη ενώ οι συνάψεις από τα Δωμάτια και την Ηλικία για το συγκεκριμένο νευρώνα έχουν μηδενιστεί. Αυτό μπορεί να σημαίνει, ότι τα μεγάλα σπίτια τείνουν να είναι πιο φτηνά όσο απομακρυνόμαστε από την πόλη. Οπότε έτσι συμπεραίνουμε ότι αυτός ο

νευρώνας είναι πιθανό να ψάχνει συγκεκριμένες κατοικίες που είναι μεγάλες αλλά όχι πολύ μακριά από την πόλη. Στη συνέχεια αυτός ο νευρώνας λαμβάνοντας υπόψη αυτές τις δυο μεταβλητές χρησιμοποιώντας τη συνάρτηση ενεργοποίησης η οποία θα λάβει με τη σειρά της υπόψη και τα βάρη των συνάψεων των δυο νευρώνων εισόδου θα παράγει την πρόβλεψη του. Εκεί έγκειται και η «δύναμη ενός νευρωνικού δικτύου», αφού κάθε νευρώνας λαμβάνει υπόψη διαφορετικό συνδυασμό επεξηγηματικών μεταβλητών με διαφορετικά βάρη και κάνει την πρόβλεψη του. Ενώ η τελική πρόβλεψη προκύπτει από το συνδυασμό όλων των προβλέψεων μαζί, όπως ακριβώς συμβαίνει και με τον άνθρωπο (όταν έχει ένα αντικείμενο μπροστά του οι διαφορετικές αισθήσεις του θα στείλουν την πρόβλεψη τους για το τι είναι το αντικείμενο αυτό και ο συνδυασμός τους θα καθορίσει την απόφαση του ανθρώπου). Ας δούμε όμως και τον τρίτο νευρώνα στον οποίο καταλήγουν συνάψεις από το εμβαδόν της κατοικίας, τον αριθμό των δωματίων και την ηλικία της κατοικίας. Αυτό δείχνει πως το νευρωνικό δίκτυο κατάφερε να αναγνωρίσει το πρότυπο που ακολουθεί η τιμή σύγχρονων κατοικιών που έχουν μεγάλο εμβαδόν και αριθμό δωματίων ανεξάρτητα από την απόσταση τους από την πόλη. Τα βάρη των συνάψεων μπορούν επίσης να μας δώσουν πληροφορία για το πώς επηρεάζει κάθε μεταβλητή την πρόβλεψη που κάνει ο συγκεκριμένος νευρώνας. Με αυτό λοιπόν τον τρόπο μπορεί κάποιος να ερμηνεύσει τις αποφάσεις που παίρνει ένα νευρωνικό δίκτυο και να κατανοήσει τη σχέση μεταξύ των επεξηγηματικών μεταβλητών του μοντέλου.

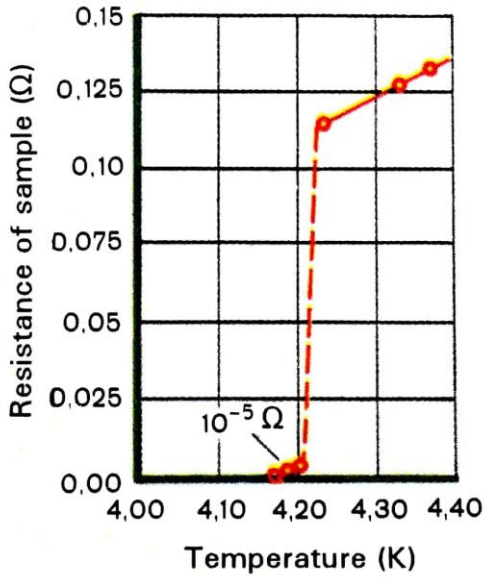
ΜΕΡΟΣ ΙΙΙ – Επιστημονικό υπόβαθρο του προβλήματος

1. Εισαγωγή στην υπεραγωγιμότητα

1.1. Η Ανακάλυψη

Με τον όρο υπεραγωγιμότητα, χαρακτηρίζουμε, ένα συνδυασμό αξιοσημείωτων ηλεκτρικών και μαγνητικών ιδιοτήτων που παρουσιάζονται σε συγκεκριμένους αγωγούς, όταν αυτοί ψύχονται σε εξαιρετικά χαμηλές θερμοκρασίες, όπως αυτές που παρατηρήθηκαν για πρώτη φορά το 1908, όταν φυσικός Heike Kamerlingh Onnes, κατάφερε να υγροποιήσει το ήλιο, στο πανεπιστήμιο του Λέιντεν. Με τον τρόπο αυτό, μπορούσε να δημιουργήσει θερμοκρασίες έως περίπου 1 βαθμό Κ.

Μια από τις πρώτες έρευνες που ο Onnes, κατάφερε να διεξάγει στις συνθήκες αυτών των τόσο χαμηλών θερμοκρασιών που μόλις είχαν προσεγγιστεί, ήταν μια μελέτη που αφορούσε στην διακύμανση της ηλεκτρικής αντίστασης των μετάλλων σε σχέση με τη θερμοκρασία. Μέχρι τότε ήταν γνωστό, για πολλά χρόνια, ότι η αντίσταση των μετάλλων μειώνεται όταν αυτά ψύχονται κάτω από θερμοκρασίες δωματίου, ωστόσο δεν ήταν γνωστό ποια είναι η οριακή τιμή της αντίστασης, αν η θερμοκρασία πλησίαζε το απόλυτο 0. Ύστερα από πειραματισμούς με πλατίνα, ο Onnes διαπίστωσε, ότι αφού αυτή ψυχθεί, η αντίσταση της πέφτει σε μια χαμηλή τιμή η οποία εξαρτάται από την καθαρότητα του δείγματος. Μέχρι εκείνη τη χρονική στιγμή το πιο καθαρό διαθέσιμο μέταλλο ήταν ο υδράργυρος, και έτσι πάνω σε μια προσπάθεια κατανόησης της συμπεριφοράς ενός τέτοιου μετάλλου, ο Onnes, υπολόγισε την αντίσταση του καθαρού υδραργύρου. Ανακάλυψε έτσι, ότι σε πολύ χαμηλές θερμοκρασίες, η αντίσταση, έγινε ανυπολόγιστα μικρή, κάτι που ήταν λογικό, ωστόσο σύντομα διαπίστωσε ακόμα (1911) ότι ο τρόπος με τον οποίο η αντίσταση εκμηδενίστηκε ήταν εντελώς απροσδόκητος. Συγκεκριμένα, αντί η αντίσταση να μειώνεται σταδιακά (με ομαλό τρόπο), καθώς η θερμοκρασία μειωνόταν και πλησίαζε το απόλυτο 0, η αντίσταση, έπεσε ακαριαία στους 4.15 Κ, και κάτω από αυτή τη θερμοκρασία η ηλεκτρική αντίσταση του υδραργύρου εκμηδενίστηκε. Η συμπεριφορά της αντίστασης, συναρτήσει της θερμοκρασίας, φαίνεται στο Σχήμα 1.

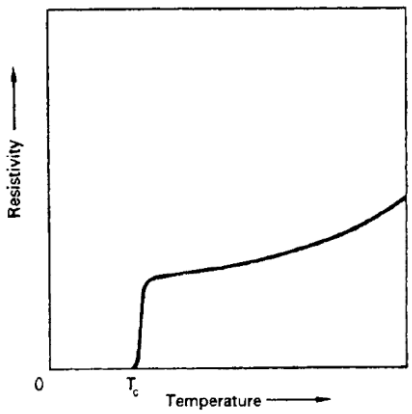


Επιπρόσθετα, αυτή η απότομη μετάβαση του υλικού σε μια κατάσταση με μηδενική αντίσταση, δεν συνέβη μόνο όταν ο υδράργυρος ήταν απόλυτα καθαρός αλλά και όταν ήταν σε μεγάλο βαθμό ακάθαρτος. Το γεγονός αυτό, οδήγησε τον Ολλανδό φυσικό, στο συμπέρασμα ότι κάτω από τους 4.15 K, ο υδράργυρος, περνάει σε μια νέα κατάσταση, με νέες ιδιότητες, αρκετά διαφορετικές από εκείνες που είχε μέχρι την ψύξη του. Η νέα αυτή κατάσταση ονομάστηκε κατάσταση «υπεραγωγιμότητας».

Τα υλικά που παρουσιάζουν το φαινόμενο υπεραγωγιμότητας, όταν ψυχθούν επαρκώς, ονομάζονται υπεραγωγοί. Μέχρι σήμερα, περίπου το ήμισυ των μεταλλικών στοιχείων και ένα σύνολο κραμάτων, έχει αποδειχτεί ότι γίνονται υπεραγωγοί σε χαμηλές θερμοκρασίες, δηλαδή κάτω από περίπου 25 K. Ωστόσο, υπάρχουν και υλικά, όπως ορισμένα μεταλλικά κεραμικά οξείδια, που μετατρέπονται σε υπεραγωγούς σε αρκετά μεγαλύτερες θερμοκρασίες, περίπου 100 K.

1.2. Κρίσιμη Θερμοκρασία (T_c)

Η θερμοκρασία κάτω από την οποία ένα υλικό, μετατρέπεται σε υπεραγωγό (εκμηδενισμός ηλεκτρικής αντίστασης), ονομάζεται κρίσιμη θερμοκρασία, και συμβολίζεται με T_c. Η θερμοκρασία αυτή είναι διαφορετική για κάθε υλικό. Το διπλανό γράφημα δείχνει πως μειώνεται η αντίσταση καθώς η θερμοκρασία ενός υλικού πέφτει και πλησιάζει την κρίσιμη τιμή.

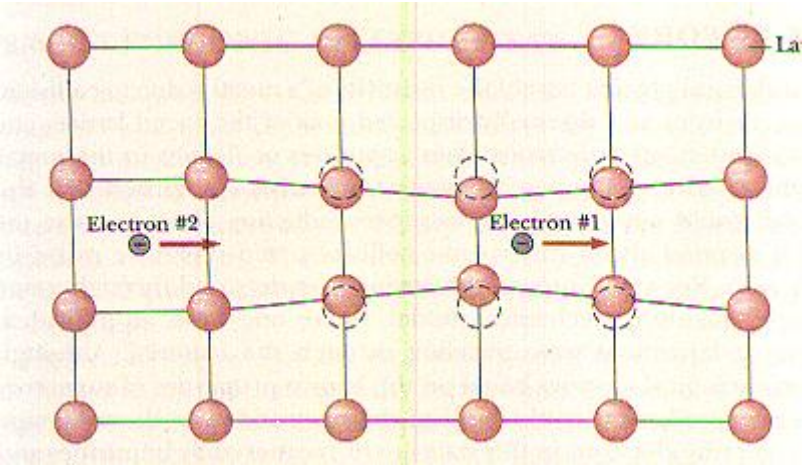


1.3. Το Φαινόμενο της Υπεραγωγιμότητας

Η ηλεκτρική αντίσταση όλων των μετάλλων και κραμάτων, μειώνεται όταν αυτά ψύχονται. Για να καταλάβουμε γιατί συμβαίνει αυτό αρκεί να αναλογιστούμε τι είναι αυτό που προκαλεί αντίσταση σε έναν αγωγό. Η ηλεκτρική αντίσταση σε έναν αγωγό πηγάζει από τη σκέδαση των ηλεκτρονίων καθώς διαδίδονται μέσα σε αυτόν και συναντούν αποκλίσεις από την τέλεια περιοδικότητα. Η σκέδαση αυτή οφείλεται είτε στις προσμίξεις που πιθανόν υπάρχουν στο μέταλλο (αντίσταση που δεν εξαρτάται από την θερμοκρασία), είτε στις ταλαντώσεις των πλεγματικών θέσεων.

Η ταχύτητα με την οποία πραγματοποιούνται οι ταλαντώσεις των πλεγματικών θέσεων ενός υλικού αποτελεί μέτρο της θερμοκρασίας του. Όσο μεγαλύτερη είναι η θερμοκρασία του, τόσο μεγαλύτερη είναι η ταχύτητα των ταλαντώσεων αυτών και συνεπώς τόσο μεγαλύτερη είναι και η αντίσταση. Όσο μειώνεται η θερμοκρασία, τόσο μειώνεται και η αντίσταση, καθώς οι ταλαντώσεις των

θετικών ιόντων επιβραδύνουν. Το 1957, οι φυσικοί John Bardeen, Leon N. Cooper και J. Robert Schrieffer πρότειναν μια εξήγηση του μηχανισμού που υπόκειται της υπεραγωγιμότητας στα μέταλλα, διατυπώνοντας μια θεωρία που φέρει τα αρχικά των επωνύμων τους, BCS. Η θεωρία τους βασίζεται, στο ζευγάρωμα των ηλεκτρονίων σε «ζεύγη Cooper». Σε ένα υλικό που βρίσκεται σε κατάσταση υπεραγωγιμότητας, ένα ηλεκτρόνιο καθώς κινείται στο εσωτερικό του, αλληλεπιδρά με θετικά φορτισμένα άτομα (κατιόντα), δημιουργώντας τοπικές παραμορφώσεις του φορτίου, δηλαδή περιοχές με μεγαλύτερη πυκνότητα θετικού φορτίου γύρω τους. Παράλληλα, το ηλεκτρόνιο που ακολουθεί, έλκεται από αυτήν την τοπική θετική πυκνότητα φορτίου που διαδίδεται μαζί με το πρώτο ηλεκτρόνιο. Με τον τρόπο αυτό τα ηλεκτρόνια έλκονται έμμεσα το ένα με το άλλο και σχηματίζουν ένα ζεύγος Cooper, το οποίο ρέει ανεμπόδιστα μέσα από τον υπεραγωγό (Βλέπε Σχήμα 2). Η κατάσταση αυτή των δύο ηλεκτρονίων είναι μια δέσμια κατάσταση, και τα ζεύγη αυτά είναι οι φορείς του ρεύματος κατά την υπεραγωγιμότητα.



Σε αυτό το σημείο, αξίζει να σημειωθεί, ότι ένα ζεύγος Cooper είναι πιο σταθερό ενεργειακά από ένα μεμονωμένο ηλεκτρόνιο. Αυτό συμβαίνει, επειδή το ζεύγος Cooper είναι πιο ανθεκτικό στις σκεδάσεις που προκαλούνται από τις ταλαντώσεις του πλέγματος, καθώς η έλξη του κάθε ηλεκτρονίου με το ζεύγος του βοηθάει και τα δύο να μην αποκλίνουν από την πορεία τους. Τα ζεύγη Cooper κινούνται μέσα στο πλέγμα, σχετικά ανεπηρέαστα από τις θερμικές ταλαντώσεις, κάτω από την κρίσιμη θερμοκρασία. Ωστόσο, η θεωρία BCS προβλέπει μια θεωρητικά μέγιστη τιμή για την κρίσιμη θερμοκρασία, της τάξης των 30-40K, καθώς πάνω από αυτήν, η θερμική ενέργεια θα απαιτούσε αλληλεπιδράσεις ηλεκτρονίων-φωνονίων πολύ υψηλής ενέργειας, για να δημιουργηθούν και να παραμείνουν σταθερά τα ζεύγη Cooper.

1.4. Μαγνητικές Ιδιότητες Υπεραγωγών

Εκτός από την ανακάλυψη ότι όταν ένα υλικό μετατρέπεται σε υπεραγωγό, η αντίσταση του εκμηδενίζεται, ο Onnes προχώρησε και σε μια ακόμα

αξιοσημείωτη ανακάλυψη. Συγκεκριμένα, παρατήρησε, ότι αν εφαρμοστεί σε έναν υπεραγωγό ένα αρκετά ισχυρό μαγνητικό πεδίο, η υπεραγωγιμότητα εξαφανίζεται (επαναφορά ηλεκτρικής αντίστασης) και το υλικό επιστρέφει στην αρχική του κατάσταση. Για κάθε υλικό, για κάθε θερμοκρασία κάτω από την κρίσιμη, υπάρχει μια αντίστοιχη κρίσιμη τιμή του εφαρμοζόμενου εξωτερικού μαγνητικού πεδίου, B_c που αν την ξεπεράσουμε, το υλικό ξαναγυρνάει στην συνηθισμένη του αγώγιμη κατάσταση. Σε όσο πιο χαμηλή θερμοκρασία βρισκόμαστε, τόσο πιο ισχυρό μαγνητικό πεδίο χρειάζεται για να καταστρέψει την υπεραγωγιμότητα.

Το 1933, οι φυσικοί Meissner και OschenfeldIt παρατήρησαν για πρώτη φορά την δεύτερη σπουδαία ιδιότητα των υπεραγωγών – τον τέλειο διαμαγνητισμό. Συγκεκριμένα, ανακάλυψαν ότι ένα υπεραγώγιμο υλικό σε θερμοκρασία κάτω της κρίσιμης, όταν βρίσκεται στο εσωτερικό ενός μαγνητικού πεδίου, απωθεί όλες τις δυναμικές γραμμές του πεδίου εκτός της μάζας του, με αποτέλεσμα να εμφανίζονται απωστικές δυνάμεις ανάμεσα σε αυτό και στο μαγνήτη. Βασική βέβαια προϋπόθεση για να συμβεί αυτό είναι το μαγνητικό πεδίο να μην έχει ξεπεράσει σε ισχύ, την κρίσιμη τιμή B_c για την αντίστοιχη θερμοκρασία, γιατί αλλιώς η υπεραγωγιμότητα θα χαθεί. Το φαινόμενο αυτό είναι γνωστό σήμερα ως φαινόμενο Meissner.

1.5. Τύποι Υπεραγωγών

Για πολλά χρόνια, οι επιστήμονες, θεωρούσαν ότι όλοι οι υπεραγωγοί, συμπεριφέρονται με βάση το ίδιο μοτίβο. Ωστόσο, πλέον είναι πια αποδεδειγμένο ότι υπάρχουν δυο είδη υπεραγωγών, ο τύπος-I και ο τύπος-II. Τα περισσότερα στοιχεία παρουσιάζουν, υπεραγωγιμότητα τύπου-I, ενώ τα κράματα(alloys), γενικά παρουσιάζουν υπεραγωγιμότητα τύπου-II. Οι δυο τύποι έχουν αρκετές κοινές ιδιότητες μεταξύ τους, όμως παρουσιάζουν σημαντικές διαφοροποιήσεις όσο αφορά στη μαγνητική τους συμπεριφορά.

Οι υπεραγωγοί τύπου I είναι εκείνοι που απωθούν τελείως από το εσωτερικό τους τα εφαρμοζόμενα μαγνητικά πεδία. Τα πιο συνηθισμένα και απλά υπεραγώγιμα υλικά, είναι τύπου I. Ενώ, υπεραγωγοί τύπου II, είναι εκείνοι οι οποίοι αποβάλλουν τελείως από το εσωτερικό τους τα μικρής έντασης μαγνητικά πεδία, αλλά δεν αποβάλλουν εξ ολοκλήρου τα εφαρμοζόμενα μαγνητικά πεδία μεγάλης έντασης. Ο διαμαγνητισμός τους δεν είναι τέλειος αλλά μερικός στα ισχυρά μαγνητικά πεδία. Το Νιόβιο είναι ένα παράδειγμα ενός στοιχειώδους υπεραγωγού τύπου II.

1.6. Υπεραγωγοί Υψηλών Θερμοκρασιών

Το 1986, μια ανακάλυψη προκάλεσε μεγάλες εξελίξεις στον κλάδο της υπεραγωγιμότητας. Ο Alex Müller και ο Georg Bednorz, ερευνητές στο IBM Research Laboratory στο Rüschlikon της Ελβετίας, δημιούργησαν μια

εύθραυστη κεραμική ένωση που γινόταν υπεραγωγός στην υψηλότερη μέχρι τότε κρίσιμη θερμοκρασία, 30 K. Αυτό όμως που έκανε αυτή την ανακάλυψη τόσο ξεχωριστή, ήταν το γεγονός, ότι τα κεραμικά υλικά είναι κατά βάση μονωτές, και συνεπώς δεν άγουν καθόλου την ηλεκτρική ενέργεια. Έτσι, οι ερευνητές δεν τα υπολόγιζαν ως πιθανούς υπεραγωγούς. Σε έρευνες που ακολούθησαν, πάνω στην αγωγιμότητα μεταλλικών οξειδίων ανακαλύφθηκαν υπεραγώγιμα μεταλλικά οξείδια με T_c μέχρι και 100 K και B_c 50-100T. Οι εξελίξεις αυτές, οδήγησαν στη δημιουργία μιας νέας έννοιας στον τομέα των υπεραγωγών, τους υπεραγωγούς υψηλών θερμοκρασιών (HTSC). Οι υπεραγωγοί αυτοί, είναι κεραμικά υλικά, και έχουν μεγάλο ερευνητικό ενδιαφέρον, καθώς για την ψύξη τους μπορεί να χρησιμοποιηθεί πλέον υγρό άζωτο (αντί για υγρό Ήλιο), υλικό που είναι πολύ φθηνότερο και πιο εύκολα επιτεύξιμο. Συνεπώς μπορούμε να έχουμε όλα τα πλεονεκτήματα ενός υπεραγωγού με λιγότερο κόπο, μεγαλύτερη ασφάλεια και χαμηλότερο κόστος. Η μελέτη των υπεραγωγών υψηλών θερμοκρασιών είναι υπό εξέλιξη, καθώς η θεωρία BCS στη σημερινή της μορφή δεν είναι απόλυτα επαρκής για την περιγραφή τους. Ακόμα, η μελέτη αυτή δίνει ελπίδες για περαιτέρω αύξηση της μέγιστης T_c, η οποία σήμερα έχει φθάσει στους 203K, με τους επιστήμονες να πιστεύουν ότι είναι μεγάλη η πιθανότητα για επίτευξη υπεραγωγιμότητας σε θερμοκρασίες δωματίου(273K).

1.7. Εφαρμογές των Υπεραγωγών

Οι υπεραγωγοί έχουν πολλές εφαρμογές χάρη σε δυο βασικά τους χαρακτηριστικά, την έλλειψη ηλεκτρικής αντίστασης, που τους κάνει μοναδικούς στη μεταφορά ενέργειας σε μεγάλες αποστάσεις (λόγω έλλειψης θερμικών απωλειών), αλλά και της εκδήλωσης ενός έντονου διαμαγνητισμού. Έτσι μερικές από τις σημαντικότερες εφαρμογές των υπεραγωγών είναι οι ακόλουθες:

- Ηλεκτρομαγνήτες: Για επίτευξη μαγνητικών πεδίων υψηλής έντασης. (πυρηνικοί αντιδραστήρες, μαγνητοϋδροδυναμικές γεννήτριες, μαγνητικοί τομογράφοι, κ.λπ.)
- Ηλεκτρικές μηχανές: Υπεραγώγιμος επαγωγέας, για εξασφάλιση υψηλών μαγνητικών εντάσεων και την εξοικονόμηση όγκου.
- Μεταφορά ηλεκτρικής ενέργειας: Χρήση υπεραγώγιμων συρμάτων, για μεταφορά ηλεκτρικού ρεύματος υψηλής πυκνότητας χωρίς απώλειες.
- Μαζικές μεταφορές: Εφαρμογή υπεραγώγιμων υλικών για την κατασκευή του "απωθούμενου τραίνου", όπου η δράση ισχυρών μαγνητικών πεδίων είναι απαραίτητη για την ανάπτυξη μεγάλων απωστικών δυνάμεων.

- Ηλεκτρονικοί υπολογιστές: Η χρήση των υπεραγώγιμων υλικών στην τεχνολογία των micro-switches, θα φέρει μεγάλη βελτίωση στην ταχύτητα επεξεργασίας σήματος.
- Ιατρική: Χρήση μαγνητικών τομογράφων και άλλων ευαίσθητων μαγνητικών οργάνων, που απαιτούν μαγνητικά πεδία της τάξης του Tesla, με εξαιρετική σταθερότητα και ομοιομορφία, ιδιότητες που προσφέρουν οι υπεραγώγιμοι μαγνήτες.
- Φυσική Στοιχειωδών Σωματιδίων: Μεγάλης ισχύος ηλεκτρομαγνήτες στους επιταχυντές σωματιδίων (SSC: Superconductor Super Collider).

2. Επιστημονική περιγραφή Επεξηγηματικών μεταβλητών

2.1. Ατομική Μάζα (Atomic Mass)

Η ατομική μάζα εκφράζει τη μάζα ενός ατόμου. Η μονάδα μέτρησης της είναι η ατομική μονάδα μάζας (amu) όπου 1 μονάδα ατομικής μάζας ορίζεται ως το 1/12 της μάζας ενός ατόμου άνθρακα ^{12}C . Ο ^{12}C είναι το ισότοπο του άνθρακα που έχει 6 πρωτόνια και 6 νετρόνια στον πυρήνα του και ζυγίζει $1{,}66 \cdot 10^{-24} g$. Ως εκ τούτου, 1 amu $= 1{,}66 \cdot 10^{-24} g = 1{,}66 \cdot 10^{-27} g$ Kg. Η σχετική ατομική μάζα ή το ατομικό βάρος, ορίζεται ως ο αριθμός που δείχνει πόσες φορές είναι μεγαλύτερη η μάζα ενός ατόμου του από το 1/12 της μάζας του ατόμου του άνθρακα ^{12}C. Για ένα άτομο, τα πρωτόνια και τα νετρόνια του πυρήνα του, αντιπροσωπεύουν σχεδόν τη συνολική του μάζα.

2.2. Ενέργεια Ιονισμού (Ionization Energy)

Η ενέργεια ιονισμού, ορίζεται ως η ελάχιστη ενέργεια που απαιτείται για την απόσπαση του ασθενέστερα συγκρατούμενου ηλεκτρονίου ενός ελεύθερου ατόμου ή μορίου, που βρίσκεται στη θεμελιώδη του κατάσταση και σε αέρια φάση, με αποτέλεσμα το σχηματισμό ενός μονοσθενούς κατιόντος. Η μονάδα μέτρησης της ενέργειας ιονισμού είναι τα kilojoules ανά mole (kJ / mol), και είναι η ποσότητα ενέργειας που απαιτείται, ώστε όλα τα άτομα σε ένα mole ουσίας να χάσουν ένα ηλεκτρόνιο το καθένα. Η πρώτη ενέργεια ιονισμού (First Ionization Energy) εκφράζει το έργο που απαιτείται για την απομάκρυνση του πρώτου ηλεκτρονίου από ένα ουδέτερο άτομο ή μόριο και είναι πάντα μικρότερη από τη δεύτερη η οποία με τη σειρά της είναι μικρότερη από την Τρίτη κ.ο.κ. . Είναι ακόμα γνωστό, πως για τα άτομα των στοιχείων μιας περιόδου του περιοδικού πίνακα η ηλεκτροαρνητικότητα (δηλαδή η τάση πρόσληψης ηλεκτρονίων) αυξάνεται από τα αριστερά προς τα δεξιά. Αυτό

συμβαίνει γιατί τα στοιχεία που βρίσκονται στα δεξιά του Περιοδικού Πίνακα έχουν αυξημένο δραστικό πυρηνικό φορτίο, συγκρατώντας ισχυρά τα εξωτερικά τους ηλεκτρόνια. Συνεπώς, και η ενέργεια 1^{ου} ιονισμού θα αυξάνεται από τα αριστερά προς τα δεξιά. Ομοίως, εξηγείται η αύξηση της ενέργειας 1^{ου} ιονισμού από κάτω προς τα πάνω κατά μήκος των ομάδων του Περιοδικού Πίνακα.

2.3. Ατομική Ακτίνα (Atomic Radius)

Η ατομική ακτίνα ενός χημικού στοιχείου είναι ένα μέτρο του μεγέθους των ατόμων του. Ορίζεται συνήθως, ως η μέση ή τυπική απόσταση από το κέντρο του πυρήνα του ατόμου μέχρι την πιο απομακρυσμένη στιβάδα (κέλυφος) ηλεκτρονίων του. Είναι γνωστό, ότι δεν υπάρχει καμία βεβαιότητα για τη θέση των ηλεκτρονίων σε κάθε χρονική στιγμή, καθώς δεν έχουν συγκεκριμένες τροχιές, ούτε σαφώς καθορισμένες περιοχές. Αντίθετα, οι θέσεις τους μπορούν να περιγραφούν ως κατανομές πιθανοτήτων που μειώνονται βαθμιαία καθώς απομακρύνεται κανείς από τον πυρήνα, χωρίς απότομη αποκοπή. Ακόμη, το είδος του δεσμού ή το σύστημα κρυστάλλωσης μίας ένωσης στην οποία συμμετέχει ένα άτομο και γενικότερα οι συνθήκες μέτρησης (πυκνότητα, θερμοκρασία) μπορεί να οδηγούν σε μεταβολές της ακτίνας του. Για αυτό οι συνθήκες κάτω από τις οποίες γίνεται η μέτρηση της ατομικής ακτίνας, πρέπει να είναι καθορισμένες. Είναι λοιπόν φανερό το γιατί είναι αρκετά δύσκολος ο υπολογισμός της ατομικής ακτίνας με ακρίβεια. Σε αυτό το σημείο αξίζει να σημειωθεί, ότι υπάρχουν διάφοροι μη ισοδύναμοι ορισμοί της ατομικής ακτίνας. Τρεις ευρέως χρησιμοποιούμενοι ορισμοί της ατομικής ακτίνας είναι: η ακτίνα Van der Waals, η μεταλλική ακτίνα και η ομοιοπολική ακτίνα. Ανάλογα με τον ορισμό, ο όρος μπορεί να εφαρμόζεται είτε σε απομονωμένα άτομα είτε σε άτομα σε συμπυκνωμένη ύλη, συνδεδεμένα ομοιοπολικά σε μόρια είτε σε ιονισμένες και διεγερμένες καταστάσεις. Συγκεκριμένα, η ακτίνα Van der Waals, είναι το μισό της απόστασης που μπορούν να πλησιάσουν οι πυρήνες δυο ομοίων ατόμων που συνδέονται με δυνάμεις Van Der Waals, η μεταλλική ακτίνα είναι το μισό της απόστασης που μπορούν να πλησιάσουν οι πυρήνες δυο ομοίων ατόμων που συνδέονται ομοιοπολικά με απλό δεσμό ενώ τέλος, η ομοιοπολική ακτίνα είναι το μισό της απόστασης που μπορούν να πλησιάσουν οι πυρήνες δυο ομοίων ατόμων που συνδέονται ομοιοπολικά με απλό δεσμό. Με βάση τους περισσότερους ορισμούς, οι ακτίνες των απομονωμένων ουδέτερων ατόμων κυμαίνονται μεταξύ 30 και 300 μm (τρισεκατομμυρίων ενός μέτρου) ή μεταξύ 0,3 και 3 angstroms. Επομένως, η ακτίνα ενός ατόμου είναι περισσότερο από 10.000 φορές μεγαλύτερη από την ακτίνα του πυρήνα. Το μέγεθος των ατόμων είναι περιοδική συνάρτηση του ατομικού τους αριθμού(ο αριθμός των πρωτονίων του πυρήνα ενός ατόμου) και επηρεάζεται από τις αλληλεπιδράσεις τόσο μεταξύ πυρήνα και ηλεκτρονίων όσο και μεταξύ των ηλεκτρονίων, δηλαδή από τις τιμές Z*, n* και l. Όσο αυξάνεται ο ατομικός αριθμός, το μέγεθος των

ατόμων κατά μήκος μιας περιόδου του περιοδικού πίνακα ελαττώνεται, ενώ αντίθετα κατά μήκος μιας ομάδας αυξάνεται.

Κατά μήκος μιας ομάδας του περιοδικού πίνακα το δραστικό πυρηνικό φορτίο Z*, αυξάνεται (ή μένει σταθερό) με την αύξηση του ατομικού αριθμού. Η ταυτόχρονη όμως αύξηση του δραστικού κύριου κβαντικού αριθμού n* των ηλεκτρονίων σθένους κυριαρχεί και έχει ως τελικό αποτέλεσμα την αύξηση της απόστασης των ηλεκτρονίων σθένους από τον πυρήνα και κατά συνέπεια την αύξηση του μεγέθους των ατόμων προς αυτή την κατεύθυνση. Κατά μήκος μιας περιόδου ο δραστικός κύριος κβαντικός αριθμός n* των ηλεκτρονίων σθένους, παραμένει σταθερός. Η παράλληλη όμως με την αύξηση του ατομικού αριθμού, αύξηση του δραστικού πυρηνικού φορτίου Z*, οδηγεί σε αύξηση των ελκτικών δυνάμεων του πυρήνα στα ηλεκτρόνια σθένους με αποτέλεσμα την ελάττωση του μεγέθους των ατόμων κατά την κατεύθυνση αυτή.

2.4. Πυρηνική Πυκνότητα (Nuclear Density)

Πυρηνική πυκνότητα είναι η πυκνότητα (λόγος μάζας ανά μονάδα όγκου) του πυρήνα ενός ατόμου με μέσο όρο περίπου $2{,}3 \times 10^{17}\ \frac{kg}{m^3}$. Η πυρηνική πυκνότητα για έναν τυπικό πυρήνα ατόμου, μπορεί να υπολογιστεί περίπου από το μέγεθος του πυρήνα, το οποίο μπορεί να προσεγγιστεί με βάση τον αριθμό των πρωτονίων και νετρονίων που περιέχει, δεδομένου ότι ο ατομικός πυρήνας φέρει το μεγαλύτερο μέρος της μάζας του ατόμου. Η ακτίνα ενός τυπικού πυρήνα, από την άποψη του αριθμού των νουκλεονίων, δίνεται από τη σχέση $R = A^{\frac{1}{3}}R_0$,όπου A ο αριθμός ατομικής μάζας (ο συνολικός αριθμός πρωτονίων και νετρονίων (γνωστά ως νουκλεόνια) σε έναν ατομικό πυρήνα), και $R_0 = 1{,}25\ fm\ (10^{-15}\ m)$ με τυπικές αποκλίσεις έως και 0,2 fm από αυτή την τιμή. Έτσι η πυκνότητα του πυρήνα δίνεται από τη σχέση: $n = \frac{A}{\frac{4}{3}\pi R^3}$

2.5. Ηλεκτρονιακή Συγγένεια (Electron Affinity)

Ηλεκτρονιακή συγγένεια ενός ατόμου ή ενός μορίου, ορίζεται ως η ποσότητα ενέργειας που απελευθερώνεται ή καταναλώνεται όταν ένα ηλεκτρόνιο προστίθεται στη στοιβάδα σθένους ενός ουδέτερου ατόμου ή μορίου που βρίσκεται σε θεμελιώδη κατάσταση και σε αέρια φάση, με αποτέλεσμα το σχηματισμό ανιόντος. Θεωρητικά η ηλεκτρονιακή συγγένεια είναι το αντίθετο της ενέργειας ιονισμού. Η έκλυση ή απορρόφηση ενέργειας για τον σχηματισμό του ιόντος καθορίζεται από το εάν οι ελκτικές δυνάμεις μεταξύ του πυρήνα του ατόμου και του προσλαμβανομένου ηλεκτρονίου, υπερισχύουν των απωστικών δυνάμεων μεταξύ των ηλεκτρονίων του ατόμου και του προσλαμβανόμενου

ηλεκτρονίου αντίστοιχα. Η πρόσληψη και δεύτερου ηλεκτρονίου από το αρνητικό ιόν που σχηματίσθηκε (**δεύτερη ηλεκτρονιακή συγγένεια**) απαιτεί πάντα μεγαλύτερη προσφορά (απορρόφηση) ενέργειας για να ξεπερασθούν οι απωστικές δυνάμεις μεταξύ αρνητικού ιόντος και προσλαμβανόμενου ηλεκτρονίου. Η ηλεκτρονική συγγένεια μεταβάλλεται περιοδικά με την αύξηση του ατομικού αριθμού με τον ίδιο τρόπο και για τους ίδιους λόγους που μεταβάλλεται η ενέργεια ιονισμού. Κατά μήκος μιας ομάδας του περιοδικού πίνακα κυριαρχεί η αύξηση της τιμής του n* η οποία μειώνει τις ελκτικές δυνάμεις του πυρήνα στο προσλαμβανόμενο ηλεκτρόνιο και ως εκ τούτου προκαλεί ελάττωση της ηλεκτρονικής συγγένειας ενός ατόμου, ενώ κατά μήκος μιας περιόδου η αύξηση του Z* αυξάνει τις ελκτικές δυνάμεις και προκαλεί αύξηση της ηλεκτρονικής συγγένειας ενός ατόμου.

2.6. Ενθαλπία Σύντηξης (Fusion Heat)

Η ενθαλπία σύντηξης μιας ουσίας, επίσης γνωστή και ως θερμότητα σύντηξης, είναι η μεταβολή στην ενθαλπία της που προκύπτει από την παροχή ενέργειας, συνήθως θερμότητας, με σκοπό την αλλαγή της κατάστασης της από στερεή σε υγρή, υπό σταθερή πίεση. Για παράδειγμα, κατά την τήξη 1 kg πάγου (στους 0 ° C), απορροφάται 333,55 kJ ενέργειας χωρίς μεταβολή θερμοκρασίας. Η θερμότητα στερεοποίησης (όταν μια ουσία αλλάζει από υγρή σε στερεή) είναι ίση και αντίθετη.

2.7. Θερμική Αγωγιμότητα (Thermal Conductivity)

Ως θερμική αγωγιμότητα ορίζεται η χαρακτηριστική ιδιότητα της ύλης που εκφράζει το πόσο εύκολο ή δύσκολο είναι μεταδοθεί θερμότητα διαμέσου ενός υλικού. Η θερμική αγωγιμότητα μετριέται μέσω του "συντελεστή αγωγιμότητας" ο οποίος διαφέρει από υλικό σε υλικό και εκφράζει την ποσότητα θερμότητας (σε Watt) που περνά από τις απέναντι πλευρές ενός υλικού, πάχους ενός μέτρου, όταν η διαφορά θερμοκρασίας μεταξύ των επιφανειών αυτών είναι ίση με ένα βαθμό Κέλβιν 1^{0}K.
Ο συντελεστής αγωγιμότητας (λ) ενός υλικού, μετριέται σε βατ ανά μέτρο και βαθμό κέλβιν (W/mk) και επηρεάζεται από τη φύση του υλικού, τη δομή, τη θερμοκρασία, την υγρασία και την πίεση του.
Η θερμική αγωγιμότητα είναι υψηλή στα υλικά τα οποία αποκαλούνται θερμοαγώγιμα, όπως είναι τα μέταλλα και είναι χαμηλή στα υλικά που αποκαλούνται θερμομονωτικά, γι αυτό όσο μικρότερος είναι ο συγκεκριμένος συντελεστής ενός υλικού τόσο καλύτερη θερμομόνωση έχει.

2.8. Σθένος (Valence)

Το σθένος, εκφράζει τον αριθμό των ηλεκτρονίων που χρειάζονται για να γεμίσει το εξωτερικό κέλυφος (στιβάδα) ενός ατόμου. Ωστόσο, στο γενικότερο ορισμό του το σθένος εκφράζει τον αριθμό των δεσμών που σχηματίζει ένα άτομο. (Για παράδειγμα ο σίδηρος, μπορεί να έχει σθένος 2 ή σθένος 3). Ο συνολικός αριθμός των δεσμών στους οποίους μπορεί να συμμετέχει ένα άτομο είναι ίσος με τον αριθμό των μη ζευγαρωμένων ηλεκτρονίων. Τα μη ζευγαρωμένα ηλεκτρόνια είναι τα ελεύθερα ηλεκτρόνια του εξωτερικού κελύφους ενός ατόμου, τα οποία συνδέονται σε ζεύγη με τα εξωτερικά ηλεκτρόνια ενός άλλου ατόμου.

2.9. Number of Elements

Ο αριθμός των στοιχείων του υπεραγωγού.

ΜΕΡΟΣ IV – Κατασκευή Μοντέλου Παλινδρόμησης για Πρόβλεψη της Κρίσιμης Θερμοκρασίας Υπεραγωγών

1. Περιγραφή των Δεδομένων και των Επεξηγηματικών Μεταβλητών

Τα δεδομένα που χρησιμοποιήθηκαν για την κατασκευή του μοντέλου, αντλήθηκαν από το αποθετήριο (UCI – Machine Learning Repository)[38], και αποτελούνται, από δυο αρχεία excel.

Το πρώτο αρχείο, περιέχει τις τιμές για 82 χαρακτηριστικά που αντιστοιχούν σε 21263 υπεραγωγούς. Το δεύτερο αρχείο, το οποίο βρίσκεται σε αντιστοιχία με το πρώτο, περιλαμβάνει τη χημική σύσταση του κάθε ενός από τους παραπάνω υπεραγωγούς, δηλαδή το σύνολο των χημικών ουσιών που υπάρχουν στον υπεραγωγό συμπεριλαμβανομένης και της περιεκτικότητας του σε αυτές.

Το πακέτο δεδομένων προέρχεται από τη βάση δεδομένων υπεραγώγιμων υλικών (Superconducting Material Database maintained by Japan's National Institute for Materials Science (NIMS) [39]). Ωστόσο, η τελική τους μορφή διαμορφώθηκε ύστερα από πρώτη επεξεργασία τους από τον Kam Hamidieh, στα πλαίσια της έρευνας, με τίτλο «A data-driven statistical model for predicting the critical temperature of a superconductor» [21]. Η έρευνα είχε επίσης σκοπό την κατασκευή ενός μοντέλου που μπορεί να προβλέψει την κρίσιμη θερμοκρασία ενός υπεραγωγού βασισμένο στη χημική του σύσταση. Για την κατασκευή του μοντέλου χρησιμοποιήθηκαν οι παρακάτω 9 μεταβλητές (που αφορούν στην χημική σύσταση του υπεραγωγού): η ατομική μάζα (atomic mass), πρώτη ενέργεια ιονισμού (first ionization energy), ατομική ακτίνα (atomic radius), πυρηνική πυκνότητα (nuclear energy), ηλεκτρονιακή συγγένεια (electron affinity), ενθαλπία σύντηξης (fusion heat), θερμική αγωγιμότητα (thermal conductivity), σθένος (valence), αριθμός στοιχείων του υπεραγωγού(number of elements). Ωστόσο, για την κάθε μια από τις πρώτες 8 μεταβλητές, κατασκευάστηκαν 10 υπομεταβλητές με σκοπό να δοθεί έμφαση-βάρος στη χημική σύσταση του κάθε υπεραγωγού. Συγκεκριμένα, οι υπομεταβλητές που κατασκευάστηκαν για κάθε μια από τις παραπάνω μεταβλητές ήταν οι εξής: Mean, Weighted mean, Geometric mean, Weighted geometric mean, Entropy, Weighted entropy, Range, Weighted range, Standard deviation, Weighted standard deviation και ο τρόπος υπολογισμού τους φαίνεται στον παρακάτω πίνακα.

Μεταβλητές	Τρόπος Υπολογισμού	Ενδεικτική τιμή
Mean	$\mu = \frac{t_1 + t_2}{2}$	35.5
Weighted mean	$v = (p_1 t_1) + (p_2 t_2)$	44.43
Geometric mean	$(t_1 t_2)^{1/2}$	33.23
Weighted geometric mean	$t_1^{p_1} + t_2^{p_2}$	43.21
Entropy	$-w_1 ln(w_1) - w_2 ln(w_2)$	0.63
Weighted entropy	$-Aln(A) - Bln(B)$	0.26
Range	$t_1 - t_2, \quad (t_1 > t_2)$	25
Weighted range	$p_1 t_1 - p_2 t_2$	37.86
Standard deviation	$[(1/2)((t_1 - \mu)^2 + (t_2 - \mu)^2)]$	12.5
Weighted standard deviation	$[p_1(t_1 - v)^2 + p_2(t_2 - v)^2]^{\frac{1}{2}}$	8.75

Οι βοηθητικές μεταβλητές $t_1, t_2, w_1, w_2, p_1, p_2,$ που παρουσιάζονται στον παραπάνω πίνακα, και ο τρόπος χρήσης τους περιγράφονται μέσω του παρακάτω παραδείγματος (Παράδειγμα 1).

Παράδειγμα 1 :

Έστω ότι έχουμε το υλικό με χημική σύσταση: Re_6Zr_1 και θέλουμε να εξάγουμε τις τιμές για τις παραπάνω 10 υπομεταβλητές, για τη μεταβλητή θερμική αγωγιμότητα.

Αρχικά γνωρίζουμε τη θερμική αγωγιμότητα για κάθε ένα στοιχείο ξεχωριστά. Η θερμική αγωγιμότητα για το Ρήνιο είναι $t_1 = 48\ W/(mK)$ ενώ για το Ζιρκόνιο $t_2 = 23W/(mK)$.

Η αναλογία των στοιχείων που προκύπτει από τη χημική σύσταση(περιεκτικότητα σε κάθε στοιχείο) στο υλικό είναι αντίστοιχα: $p_1 = \frac{6}{6+1} = \frac{6}{7}$ και $p_2 = \frac{1}{6+1} = \frac{1}{7}$.

Επίσης, η αναλογία θερμικής αγωγιμότητας κάθε στοιχείου του υλικού, δίνεται από τη σχέση: $w_1 = \frac{t_1}{t_1+t_2} = \frac{48}{71}, w_2 = \frac{t_2}{t_1+t_2} = \frac{23}{71}$.

Τέλος, υπολογίζουμε τους συντελεστές $A = \frac{p_1 w_1}{p_1 w_1 + p_2 w_2} \approx 0.926$ & $B = \frac{p_2 w_2}{p_1 w_1 + p_2 w_2} \approx 0.074$.

Χρησιμοποιώντας τις τιμές των παραπάνω βοηθητικών μεταβλητών, υπολογίζουμε για κάθε μια από τις παραπάνω δέκα υπομεταβλητές την τιμή τους και παίρνουμε το αποτέλεσμα που φαίνεται στον πίνακα για κάθε μια.

Επαναλαμβάνουμε την ίδια διαδικασία για κάθε μια από τις υπόλοιπες οκτώ μεταβλητές (για κάθε μια υπολογίζουμε τις τιμές για τις δέκα υπομεταβλητές) και έτσι τελικά έχουμε τιμή για $8 \times 10 = 80$ μεταβλητές $+1$ (αριθμό στοιχείων στον υπεραγωγό), με αποτέλεσμα να έχουμε τελικά 81 επεξηγηματικές μεταβλητές που θα μας χρησιμεύσουν στο να κατασκευάσουμε το μοντέλο που θα προβλέπει την μεταβλητή απόκρισης, που είναι η κρίσιμη θερμοκρασία (T_c).

2. Προπαρασκευή των Δεδομένων (Preprocessing Data)

Αφού διαμορφώσαμε μια καλή ιδέα για το ποιες είναι οι αρχικές επεξηγηματικές μεταβλητές του μοντέλου που θέλουμε να κατασκευάσουμε και τι αντιπροσωπεύει η κάθε μια από αυτές, ας προχωρήσουμε στην επεξεργασία και σε μια πρώτη ανάλυση τους. Στο σημείο αυτό αξίζει να σημειωθεί ότι πριν κατασκευάσουμε οποιοδήποτε μοντέλο παλινδρόμησης και πριν εφαρμόσουμε οποιαδήποτε τεχνική μηχανικής μάθησης, θα πρέπει πρώτα να σιγουρευτούμε ότι τα δεδομένα που έχουμε στη διάθεση μας ικανοποιούν κάποιες προϋποθέσεις και βρίσκονται σε μια κατάλληλη προς επεξεργασία μορφή. Η διαδικασία κατά την οποία τα δεδομένα επεξεργάζονται έτσι ώστε να ικανοποιούν τις προϋποθέσεις αυτές, ονομάζεται προπαρασκευή δεδομένων (preprocessing data).

Η προπαρασκευή δεδομένων έχει διάφορα στάδια και περιλαμβάνει ένα πλήθος τεχνικών που μπορούν να χρησιμοποιηθούν για την επίτευξη του σκοπού της. Μερικά από τα σημαντικότερα θα περιγραφούν στη συνέχεια.

Η διαδικασία ξεκινάει αρχικά ελέγχοντας αν όλα τα δεδομένα που αφορούν σε κάθε μεταβλητή έχουν την ίδια μορφή και αν στα δεδομένα υπάρχουν κενές ή εσφαλμένες τιμές. Λόγω αστοχίας κάποιου αισθητήρα ή λόγω σφάλματος ενός παρατηρητή ή ενός προγράμματος συμβαίνει σε πολλές περιπτώσεις να εντοπιστούν κενές τιμές (blanks or corrupted values), εσφαλμένες τιμές (distorted values) ή ακραίες τιμές (outliers). Οι τιμές αυτές θεωρούνται ως θόρυβος στα δεδομένα (noise) και θα πρέπει να εντοπιστούν και να αντιμετωπιστούν με τον κατάλληλο κάθε φορά τρόπο πριν προχωρήσουμε στην επεξεργασία των δεδομένων, καθώς σε διαφορετική περίπτωση μπορούν να

επηρεάσουν σε μεγάλο βαθμό την απόδοση και την ακρίβεια ενός μοντέλου μηχανικής μάθησης.

Γενικά, όταν έχουμε κενές, εσφαλμένες ή ακραίες τιμές σε κάποιες μεταβλητές μπορούμε να επέμβουμε με δυο τρόπους. Ο πρώτος τρόπος είναι να διαγράψουμε ολόκληρες τις εγγραφές (σειρές) που περιέχουν τις κενές τιμές αυτές, δηλαδή να διαγράψουμε όλες τις τιμές όλων των μεταβλητών για τις εγγραφές που έχουν κενές τιμές. Ο δεύτερος τρόπος είναι να αντικαταστήσουμε την κενή τιμή με κάποια μετρική (μέσο όρο, διάμεσο) της συγκεκριμένης μεταβλητής. Ωστόσο, ο κάθε τρόπος συνδέεται με κάποια μειονεκτήματα. Ο πρώτος στην περίπτωση που έχουμε λίγα δεδομένα και πολλές κενές τιμές μπορεί να μας οδηγήσει σε σημαντική μείωση του αριθμού των δεδομένων επηρεάζοντας έτσι την απόδοση και την γενίκευση του μοντέλου ενώ ο δεύτερος τρόπος υποθέτει μια τιμή για την κενή τιμή, που μπορεί να απέχει αρκετά από την πραγματική της τιμή. Στην πράξη συνήθως χρησιμοποιούμε την πρώτη τεχνική.

Το ερώτημα όμως που προκύπτει στο σημείο αυτό, είναι πως εντοπίζουμε αυτές τις τιμές. Είναι πολύ σημαντικό πάντα να κοιτάμε τα δεδομένα που διαθέτουμε αναλυτικά και να ελέγχουμε για κάθε μεταβλητή κάποιες μετρικές όπως ποιο είναι το πεδίο τιμών της, ποια η μέγιστη και η ελάχιστη τιμή, ποιος ο μέσος ποια η διάμεσος κ.λ.π. . Έτσι για παράδειγμα αν έχουμε μια μεταβλητή που αντιπροσωπεύει το ύψος και έχουμε μέγιστη τιμή για τη συγκεκριμένη μεταβλητή 5 μέτρα τότε προφανώς πρόκειται για μια ακραία ή εσφαλμένη τιμή η οποία θα πρέπει να διαγραφεί η να αντικατασταθεί. Ακόμα θα πρέπει να ελέγξουμε αν η μορφή των δεδομένων είναι σωστή, δηλαδή μπορεί για κάποιο λόγο οι τιμές της μεταβλητής «ηλικία» για παράδειγμα, να έχουν τη μορφή «23;». Σε αυτή την περίπτωση θα πρέπει να αφαιρέσουμε το ερωτηματικό από τις παρατηρήσεις ώστε να είμαστε σε θέση να τις επεξεργαστούμε σωστά.

Στη δική μας περίπτωση με τη βοήθεια της βιβλιοθήκης «pandas» της Python διαπιστώσαμε ότι δεν έχουμε καθόλου κενές τιμές ενώ οι τιμές των μεταβλητών φάνηκαν αρκετά λογικές χωρίς να υπάρχουν εσφαλμένες ή ακραίες τιμές. Στους παρακάτω πίνακες βλέπουμε τα αποτελέσματα των ελέγχων αυτών μέσω της Python.

```
number_of_elements                False
mean_atomic_mass                  False
wtd_mean_atomic_mass              False
gmean_atomic_mass                 False
wtd_gmean_atomic_mass             False
entropy_atomic_mass               False
wtd_entropy_atomic_mass           False
range_atomic_mass                 False
wtd_range_atomic_mass             False
std_atomic_mass                   False
wtd_std_atomic_mass               False
mean_fie                          False
wtd_mean_fie                      False
gmean_fie                         False
wtd_gmean_fie                     False
entropy_fie                       False
wtd_entropy_fie                   False
range_fie                         False
wtd_range_fie                     False
std_fie                           False
wtd_std_fie                       False
mean_atomic_radius                False
wtd_mean_atomic_radius            False
gmean_atomic_radius               False
wtd_gmean_atomic_radius           False
entropy_atomic_radius             False
wtd_entropy_atomic_radius         False
range_atomic_radius               False
wtd_range_atomic_radius           False
std_atomic_radius                 False
                                  ...
```

Στον παραπάνω πίνακα βλέπουμε για κάθε μεταβλητή ξεχωριστά αν περιέχει κενές τιμές ή όχι. Παρατηρούμε ότι για όλες τις μεταβλητές η τιμή είναι False άρα δεν έχουμε να διαχειριστούμε κάποια κενή τιμή.

	number_of_elements	mean_atomic_mass	wtd_mean_atomic_mass	gmean_atomic_mass	wtd_gmean_atomic_mass	entropy_atomic_mass	wtd_entropy_atomic_m
count	21263.000000	21263.000000	21263.000000	21263.000000	21263.000000	21263.000000	21263.000
mean	4.115224	87.557631	72.988310	71.290627	58.539916	1.165608	1.063
std	1.439295	29.676497	33.490406	31.030272	36.651067	0.364930	0.401
min	1.000000	6.941000	6.423452	5.320573	1.960849	0.000000	0.000
25%	3.000000	72.458076	52.143839	58.041225	35.248990	0.966676	0.775
50%	4.000000	84.922750	60.696571	66.361592	39.918385	1.199541	1.146
75%	5.000000	100.404410	86.103540	78.116681	73.113234	1.444537	1.359
max	9.000000	208.980400	208.980400	208.980400	208.980400	1.983797	1.958

8 rows × 82 columns

Στον παραπάνω πίνακα βλέπουμε για κάθε μεταβλητή ξεχωριστά μερικές μετρικές όπως τον αριθμό των εγγραφών, τη μέση τιμή της, την τυπική απόκλιση, την ελάχιστη και τη μέγιστη τιμή καθώς και τις τιμές των τεταρτημορίων. Έτσι μπορούμε να εντοπίσουμε ακραίες τιμές, ή εσφαλμένες τιμές, καθώς μια ακραία τιμή θα εμφανιστεί ως μέγιστη τιμή και θα απέχει πολύ από το μέσο όρο. Ωστόσο για να σιγουρευτούμε ότι μια τιμή αποτελεί ακραία τιμή μπορούμε να δούμε και πόσες φορές εμφανίζονται μεγάλες τιμές σαν αυτή στα δεδομένα, αν αυτές είναι σχετικά λίγες τότε πολύ πιθανόν να είναι ακραία τιμή. Όμως αν το πλήθος των τιμών που είναι εξίσου μεγάλες με την ακραία τιμή

δεν είναι ανεπαίσθητο, τότε ίσως αυτές οι ακραίες τιμές να σημαίνουν κάτι για τον πληθυσμό που έχουμε να περιγράψουμε, να περιέχουν δηλαδή κάποια πληροφορία. Τέλος, θα πρέπει να σιγουρευτούμε ότι ο τύπος κάθε μεταβλητής που διαθέτουμε είναι ο αναμενόμενος, δηλαδή αν έχουμε μια μεταβλητή που αντιπροσωπεύει εισόδημα για παράδειγμα και η python την αναγνωρίζει σαν χαρακτήρα και όχι σαν αριθμητική τότε είναι πολύ πιθανό να υπάρχουν σε κάποιες τιμές της χαρακτήρες ή σύμβολα που δεν θα επιτρέψουν στη συνέχεια την επεξεργασία της.

```
Data columns (total 82 columns):
number_of_elements                21263 non-null int64
mean_atomic_mass                  21263 non-null float64
wtd_mean_atomic_mass              21263 non-null float64
gmean_atomic_mass                 21263 non-null float64
wtd_gmean_atomic_mass             21263 non-null float64
entropy_atomic_mass               21263 non-null float64
wtd_entropy_atomic_mass           21263 non-null float64
range_atomic_mass                 21263 non-null float64
wtd_range_atomic_mass             21263 non-null float64
std_atomic_mass                   21263 non-null float64
wtd_std_atomic_mass               21263 non-null float64
mean_fie                          21263 non-null float64
wtd_mean_fie                      21263 non-null float64
gmean_fie                         21263 non-null float64
wtd_gmean_fie                     21263 non-null float64
entropy_fie                       21263 non-null float64
wtd_entropy_fie                   21263 non-null float64
range_fie                         21263 non-null float64
wtd_range_fie                     21263 non-null float64
std_fie                           21263 non-null float64
wtd_std_fie                       21263 non-null float64
mean_atomic_radius                21263 non-null float64
wtd_mean_atomic_radius            21263 non-null float64
gmean_atomic_radius               21263 non-null float64
wtd_gmean_atomic_radius           21263 non-null float64
entropy_atomic_radius             21263 non-null float64
wtd_entropy_atomic_radius         21263 non-null float64
range_atomic_radius               21263 non-null int64
wtd_range_atomic_radius           21263 non-null float64
std atomic radius                 21263 non-null float64
```

Στον παραπάνω πίνακα βλέπουμε για κάθε μεταβλητή ξεχωριστά τον τύπο της και όπως φαίνεται είναι όλες αριθμητικές όπως ήταν αναμενόμενο άρα δε χρειάζεται κάποια περεταίρω ενέργεια.

3. Επιλογή Μεταβλητών (Feature Selection)

Η επιλογή των μεταβλητών που θα χρησιμοποιήσουμε στο μοντέλο παλινδρόμησης διαδραματίζει σημαντικό ρόλο στην απόδοση του. Όπως έχει αναφερθεί κατά την κατασκευή ενός μοντέλου μηχανικής μάθησης (είτε παλινδρόμησης είτε ταξινόμησης), είναι σημαντικό οι επεξηγηματικές μεταβλητές (features), που θα χρησιμοποιηθούν από το μοντέλο να είναι όσο το δυνατόν λιγότερες και όσο το δυνατόν πιο περιεκτικές σε πληροφορία γίνεται. Γενικά, ένας μεγάλος αριθμός μεταβλητών σημαίνει ότι όταν θα χρησιμοποιήσουμε το μοντέλο για να κάνουμε μια πρόβλεψη σε καινούρια δεδομένα θα πρέπει να συλλέξουμε γι αυτά μεγάλο όγκο πληροφορίας κάτι που μεταφράζεται σε κόστος και χρόνο. Επίσης σε πολλές περιπτώσεις παρατηρείται δυο ή περισσότερες μεταβλητές να παρέχουν ίδια σχεδόν πληροφορία, δηλαδή η

μια να επεξηγεί την άλλη. Σε τέτοιες περιπτώσεις συνηθίζουμε να κρατάμε τη μία από τις μεταβλητές αυτές (συνήθως αυτή που έχει μεγαλύτερη συσχέτιση με τη μεταβλητή απόκρισης) καθώς οι υπόλοιπες δε μας προσφέρουν κάποια επιπλέον πληροφορία οπότε δε χρειάζεται να τις συμπεριλάβουμε στο μοντέλο. Αντίστοιχα, αφαιρούμε και τις μεταβλητές που έχουν μικρή συσχέτιση με τη μεταβλητή απόκρισης καθώς δεν παρέχουν κάποια πληροφορία που θα μπορούσε να βοηθήσει στην πρόβλεψη της. Έτσι μένουν μόνο οι μεταβλητές που έχουν μεγάλη συσχέτιση με τη μεταβλητή απόκρισης (την επεξηγούν σε μεγάλο βαθμό) και μικρή συσχέτιση μεταξύ τους (η κάθε μια παρέχει διαφορετική πληροφορία για τη μεταβλητή απόκρισης). Ωστόσο, παρατηρούνται πιο σπάνια περιπτώσεις που όλες οι μεταβλητές απόκρισης μπορεί να έχουν μεγάλη συσχέτιση μεταξύ τους ή πολύ μικρή συσχέτιση με τη μεταβλητή απόκρισης. Σε τέτοιες περιπτώσεις εφαρμόζουμε διαφορετικές τεχνικές ανάλογα την περίπτωση και το πρόβλημα. Η επιλογή των κατάλληλων μεταβλητών για το μοντέλο μπορεί να συμβάλλει στη:

- Βελτίωση της απόδοσης: Αφαιρούνται οι μεταβλητές που δεν σχετίζονται με τη μεταβλητή απόκρισης καθώς και οι μεταβλητές που δεν προσφέρουν κάποια παραπάνω πληροφορία και μπορεί να «παραπλανούν» το μοντέλο, με αποτέλεσμα την αύξηση της απόδοσης του.
- Μείωση Υπερπροσαρμογής: Η αφαίρεση των περιττών μεταβλητών συνδέεται με μείωση της υπερπροσαρμογής αφού το μοντέλο αποφασίζει χωρίς να επηρεάζεται από το θόρυβο αυτών.
- Μείωση χρόνου εκπαίδευσης: Λιγότερα δεδομένα σημαίνει μείωση της πολυπλοκότητας του αλγορίθμου και συνεπώς μείωση του χρόνου εκπαίδευσης του μοντέλου.

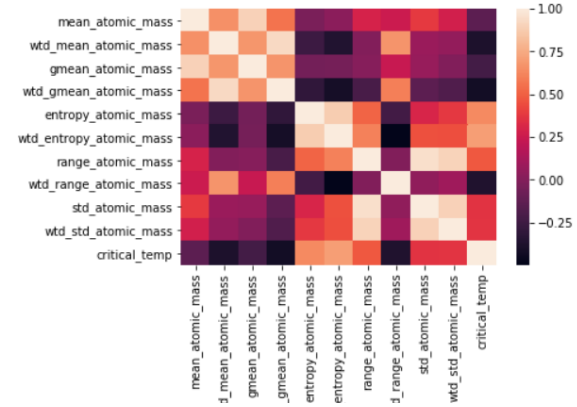
Υπάρχουν πολλές διαφορετικές μέθοδοι για επιλογή επεξηγηματικών μεταβλητών, ωστόσο οι περισσότερες στηρίζονται σε δυο βασικές τεχνικές. Η πρώτη τεχνική βασίζεται στον πίνακα συσχέτισης (correlation matrix), που απεικονίζει τις συσχετίσεις μεταξύ όλων των μεταβλητών του μοντέλου, στον heatmap (πίνακας που δείχνει συσχετίσεις μεταξύ των μεταβλητών και τις χρωματίζει σε μια κλίμακα ανάλογα με το πόσο συσχετισμένες είναι) καθώς και στα διαγράμματα διασποράς (scatterplots) μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Με τη βοήθεια των δυο πινάκων αυτών μπορούμε να εντοπίσουμε τις μεταβλητές που έχουν μεγαλύτερη συσχέτιση με τη μεταβλητή απόκρισης και μικρότερη συσχέτιση με τις υπόλοιπες επεξηγηματικές μεταβλητές και να επιλέξουμε αυτές για το μοντέλο μας. Η δεύτερη τεχνική εντοπίζει το πως ανταποκρίνεται το μοντέλο σε διαφορετικές ομάδες επεξηγηματικών μεταβλητών, δηλαδή αξιολογεί το πόσο επηρεάζεται η απόδοση του μοντέλου από την αφαίρεση μιας ή περισσότερων επεξηγηματικών μεταβλητών. Αν η αφαίρεση μιας μεταβλητής οδηγεί σε μεγάλη μείωση στην απόδοση του μοντέλου, τότε καταλαβαίνουμε ότι η μεταβλητή αυτή έχει μεγάλη σημασία για το μοντέλο και πρέπει να τη συμπεριλάβουμε στην κατασκευή του. Ωστόσο, πρέπει να σημειωθεί στο σημείο αυτό ότι η τεχνική αυτή έχει μεγάλο υπολογιστικό και χρονικό κόστος όταν οι επεξηγηματικές μεταβλητές του μοντέλου είναι πολλές (πρακτικά από 50 και πάνω) και τείνει να μην ανταποκρίνεται με το βέλτιστο τρόπο όταν οι επεξηγηματικές μεταβλητές έχουν μεγάλη συσχέτιση μεταξύ τους.

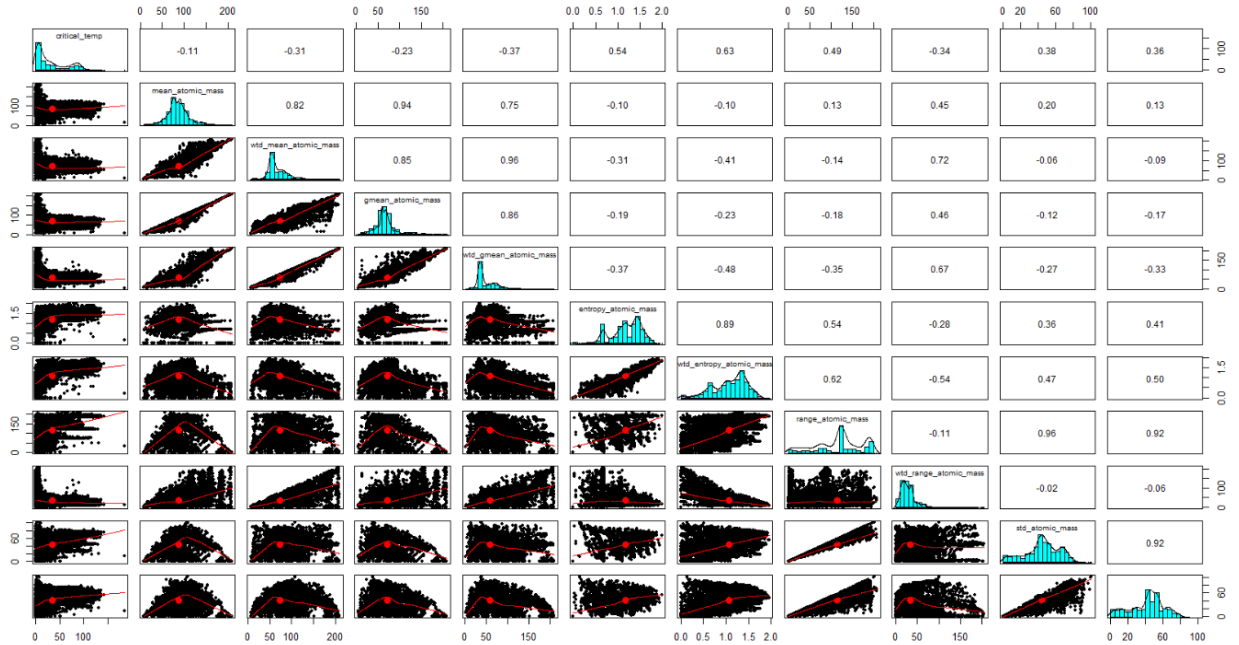
Στο δικό μας πρόβλημα, ξεκινήσαμε με την πρώτη τεχνική, ωστόσο ο μεγάλος αριθμός επεξηγηματικών μεταβλητών του μοντέλου (81), δεν επιτρέπει την εξερεύνηση του πίνακα συσχέτισης μεταξύ τους με το μάτι. Αυτό γιατί θα πρέπει για κάθε μεταβλητή να εντοπίσουμε τις μεταβλητές εκείνες που έχουν μεγάλη συσχέτιση με αυτή και να τις διαγράψουμε. Έτσι κατασκευάσαμε έναν αλγόριθμο ο οποίος θα πραγματοποιούσε τη διαδικασία αυτή και λειτουργεί ως εξής:

- Σαρώνει τον πίνακα συσχέτισης ψάχνοντας κάθε φορά την επεξηγηματική μεταβλητή που έχει μεγαλύτερη συσχέτιση με τη μεταβλητή απόκρισης και τη θεωρεί ως μεταβλητή του μοντέλου.
- Για τη μεταβλητή αυτή εντοπίζει όλες τις μεταβλητές που έχουν μεγάλη συσχέτιση με αυτή και τις αφαιρεί.
- Η διαδικασία τελειώνει μέχρι όλες οι μεταβλητές του μοντέλου να έχουν συμπεριληφθεί στο μοντέλο ή να έχουν αφαιρεθεί από αυτό.

Το αποτέλεσμα όμως του αλγορίθμου ανέδειξε ένα βασικό πρόβλημα που είχαμε να αντιμετωπίσουμε και μέχρι εκείνη τη στιγμή δεν γνωρίζαμε την ύπαρξη του. Αυτό της πολυσυγγραμικότητας. Συνειδητοποιήσαμε λοιπόν ότι οι μεταβλητές που είχαν μεγάλη συσχέτιση με τη μεταβλητή απόκρισης είχαν και μεγάλη συσχέτιση μεταξύ τους, με αποτέλεσμα να μη μπορούμε να βρούμε έστω δυο μεταβλητές που να ικανοποιούν τις προϋποθέσεις του αλγορίθμου. Έτσι θα έπρεπε να λειτουργήσουμε με διαφορετικό τρόπο. Μετά από σκέψη καταλήξαμε στο να εφαρμόσουμε την τεχνική σε κάθε κατηγορία μεταβλητών ξεχωριστά. Κάθε δεκάδα των πρώτων 80 μεταβλητών αφορά σε ένα συγκεκριμένο χαρακτηριστικό. Έτσι για κάθε χαρακτηριστικό ταξινομήσαμε τις μεταβλητές που το αφορούν σε φθίνουσα σειρά ως προς τη συσχέτιση με τη μεταβλητή απόκρισης και επιλέξαμε εκείνη με τη μεγαλύτερη που ταυτόχρονα είχε και μεγάλη συσχέτιση και με τις υπόλοιπες μεταβλητές του ίδιου χαρακτηριστικού , κάτι που σήμαινε ότι αν κρατούσαμε μόνο αυτή δε θα χάναμε πολύ πληροφορία. Με τον τρόπο αυτό επιλέξαμε μια μεταβλητή για κάθε χαρακτηριστικό και κρατήσαμε ακόμα και την μεταβλητή number of elements, η οποία δεν είχε υπομεταβλητές. Τελικά, προέκυψαν 9 επεξηγηματικές μεταβλητές τις οποίες αποφασίσαμε να χρησιμοποιήσουμε στο τελικό μοντέλο μας. Αξίζει να σημειωθεί στο σημείο αυτό, πως το φαινόμενο της πολυσυγγραμικότητας δεν είχε λυθεί όμως είχαμε καταφέρει να μειώσουμε τις μεταβλητές από 81 σε 9 κρατώντας τις πιο περιεκτικές σε πληροφορία.

	mean_atomic_mass	wtd_mean_atomic_mass	gmean_atomic_mass	wtd_gmean_atomic_mass	entropy_atomic_mass	wtd_entropy_atomic_mas
mean_atomic_mass	1.000000	0.657422	0.898693	0.556521	-0.029640	0.02985
wtd_mean_atomic_mass	0.657422	1.000000	0.674872	0.929282	-0.261537	-0.36404
gmean_atomic_mass	0.898693	0.674872	1.000000	0.670704	-0.054027	-0.04983
wtd_gmean_atomic_mass	0.556521	0.929282	0.670704	1.000000	-0.304005	-0.41344
entropy_atomic_mass	-0.029640	-0.261537	-0.054027	-0.304005	1.000000	0.87799
wtd_entropy_atomic_mass	0.029850	-0.364040	-0.049831	-0.413440	0.877999	1.00000
range_atomic_mass	0.296669	-0.003112	0.009285	-0.209862	0.511423	0.60606
wtd_range_atomic_mass	0.255878	0.667977	0.235233	0.599222	-0.233330	-0.49600
std_atomic_mass	0.384081	0.077167	0.068679	-0.136276	0.310221	0.45546
wtd_std_atomic_mass	0.280025	0.056468	-0.005269	-0.181343	0.376932	0.44746
critical_temp	-0.139523	-0.380931	-0.229568	-0.422197	0.632410	0.70737





Στις παραπάνω εικόνες βλέπουμε τον πίνακα συσχέτισης, το heatmap, και τα διαγράμματα διασποράς για τις υπομεταβλητές του χαρακτηριστικού «ατομική μάζα».

Αυτό που παρατηρούμε κοιτάζοντας τα διαγράμματα διασποράς τόσο γι αυτό το χαρακτηριστικό όσο και για τα υπόλοιπα, είναι ότι η σχέση μεταξύ των μεταβλητών δεν είναι πάντα γραμμική (τόσο μεταξύ τους όσο και σε σχέση με

τη μεταβλητή απόκρισης). Ωστόσο υπάρχει μια συστηματικότητα στα διαγράμματα που δηλώνει κάποια σχέση (μεγάλες συσχετίσεις) όπως έδειξε και ο πίνακας διασποράς. Για το λόγο αυτό υπολογίσαμε και τις συσχετίσεις spearman εκτός από pearson, ώστε να ανακαλύψουμε και μη γραμμικές συσχετίσεις μεταξύ των μεταβλητών και του target.

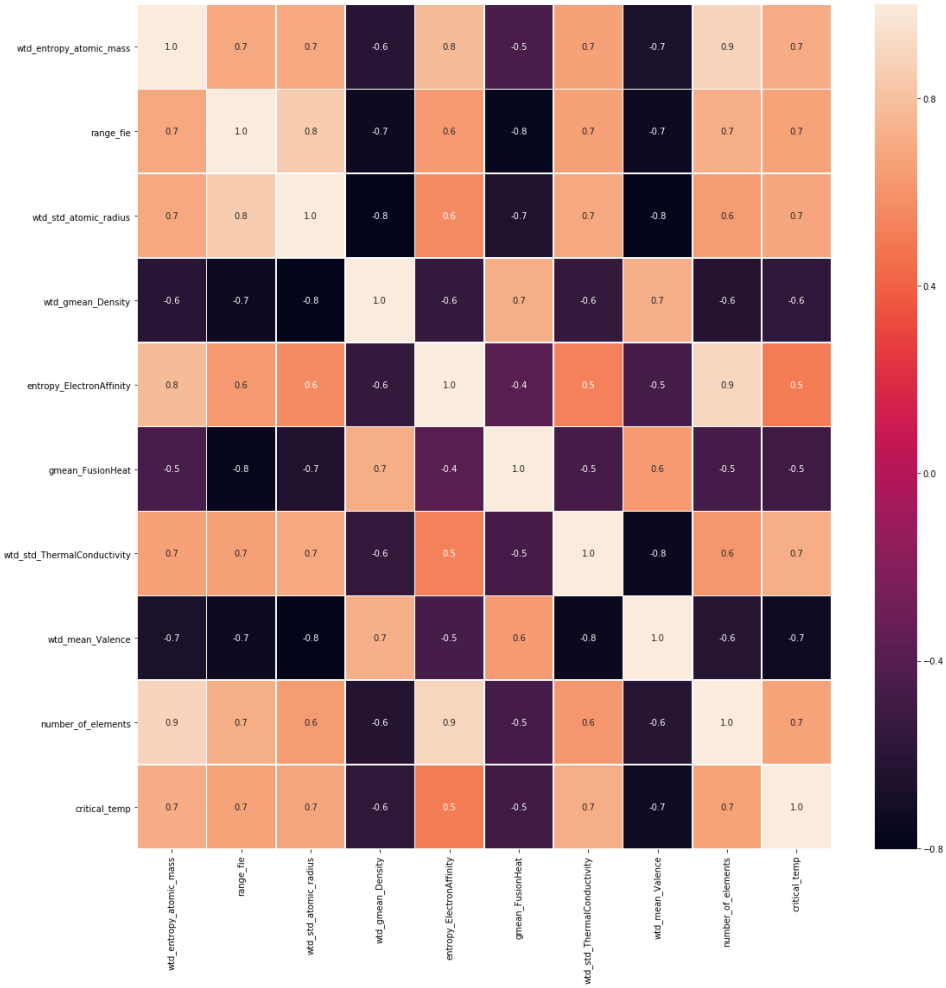
```
wtd_entropy_atomic_mass    0.707374
entropy_atomic_mass        0.632410
range_atomic_mass          0.476012
wtd_std_atomic_mass        0.366411
std_atomic_mass            0.360476
mean_atomic_mass          -0.139523
gmean_atomic_mass         -0.229568
wtd_range_atomic_mass     -0.367985
wtd_mean_atomic_mass      -0.380931
wtd_gmean_atomic_mass     -0.422197
Name: critical_temp, dtype: float64
```

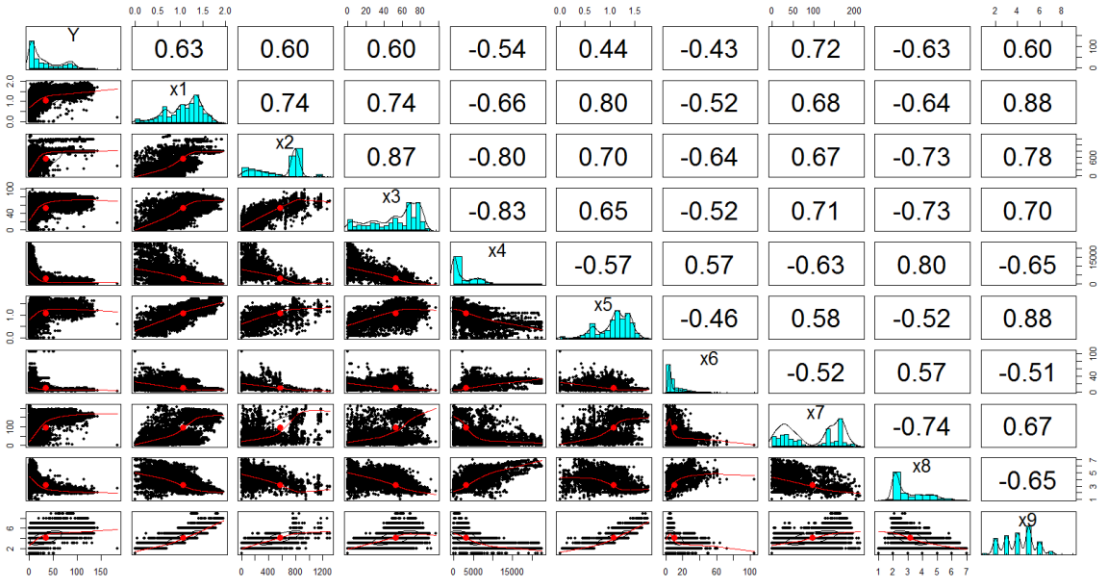
Στην παραπάνω εικόνα βλέπουμε για το χαρακτηριστικό «ατομική μάζα», τις μεταβλητές που το αφορούν σε φθίνουσα διάταξη ανάλογα με τη συσχέτιση τους με τη μεταβλητή απόκρισης που είναι η κρίσιμη θερμοκρασία. Στη συγκεκριμένη περίπτωση για το συγκεκριμένο χαρακτηριστικό επιλέχθηκε η μεταβλητή wtd_entropy_atomic_mass, να χρησιμοποιηθεί για το μοντέλο αφού είχε τη μεγαλύτερη συσχέτιση 0.71 .

Με τον ίδιο τρόπο επιλέχθηκαν και οι υπόλοιπες μεταβλητές του μοντέλου παλινδρόμησης. Έτσι τελικά επιλέχθηκαν οι παρακάτω εννιά επεξηγηματικές μεταβλητές: wtd entropy atomic mass (ατομική μάζα), range fie (πρώτη ενέργεια ιονισμού), wtd std atomic radius (ατομική ακτίνα), wtd gmean density (πυρηνική πυκνότητα), entropy electron affinity (ηλεκτρονιακή συγγένεια), gmean fusion heat (ενθαλπία σύντηξης), wtd std thermal conductivity (θερμική αγωγιμότητα), wtd mean valence (σθένος), numOfElements (αριθμός στοιχείων), για να περιγράψουν τη μεταβλητή απόκρισης critical temperature (κρίσιμη θερμοκρασία).

Στις παρακάτω εικόνες βλέπουμε για τις τελικές επεξηγηματικές μεταβλητές του μοντέλου τον πίνακα συσχέτισης, το heatmap, τα διαγράμματα διασποράς και τα ιστογράμματα τόσο μεταξύ τους όσο και σε σχέση με τη μεταβλητή απόκρισης.

	wtd_entropy_atomic_mass	range_fie	wtd_std_atomic_radius	wtd_gmean_Density	entropy_ElectronAffinity	gmean_FusionHeat	wtd_std
wtd_entropy_atomic_mass	1.000000	0.688637	0.694359	-0.611298	0.769771	-0.454748	
range_fie	0.688637	1.000000	0.846315	-0.745375	0.622983	-0.773094	
wtd_std_atomic_radius	0.694359	0.846315	1.000000	-0.801837	0.565636	-0.652826	
wtd_gmean_Density	-0.611298	-0.745375	-0.801837	1.000000	-0.552342	0.714349	
entropy_ElectronAffinity	0.769771	0.622983	0.565636	-0.552342	1.000000	-0.390548	
gmean_FusionHeat	-0.454748	-0.773094	-0.652826	0.714349	-0.390548	1.000000	
wtd_std_ThermalConductivity	0.657957	0.650259	0.694851	-0.553544	0.523462	-0.476555	
wtd_mean_Valence	-0.669737	-0.744380	-0.791531	0.725080	-0.470515	0.632642	
number_of_elements	0.889554	0.712877	0.637451	-0.629210	0.896565	-0.472912	
critical_temp	0.707222	0.663204	0.678140	-0.581154	0.506272	-0.504388	





Είναι αρκετά εμφανής από τα στιγμιότυπα, η ύπαρξη πολυσυγγραμικότητας μεταξύ των μεταβλητών αλλά και η μη γραμμική σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών.

4. Γραμμική Παλινδρόμηση

4.1. Μέθοδος Ελαχίστων Τετραγώνων

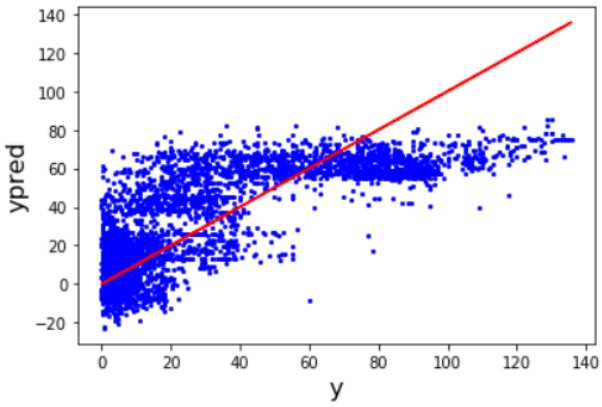
Αφού λοιπόν επιλέξαμε τις μεταβλητές που θα χρησιμοποιηθούν στο μοντέλο παλινδρόμησης ξεκινάμε την κατασκευή του. Δοκιμάζουμε αρχικά το πιο απλό μοντέλο παλινδρόμησης, το γραμμικό, για να δούμε πως ανταποκρίνεται στο πρόβλημα που έχουμε να λύσουμε. Με τη βοήθεια της μεθόδου ελαχίστων τετραγώνων λαμβάνουμε τους συντελεστές του γραμμικού μοντέλου (εκτιμήτριες ελαχίστων τετραγώνων, a , $b_1, \dots, b_k$) καθώς και διάφορα αποτελέσματα στατιστικής συμπερασματολογίας, που φαίνονται στην παρακάτω εικόνα.

```
Intercept:
 -0.0892800547303807
Coefficients:
 [ 1.88542061e+01  1.17840357e-02  6.33345828e-02  7.88306717e-04
 -3.94984653e+01  4.43535513e-02  2.44283319e-01 -3.55913241e+00
  7.64820170e+00]
                            OLS Regression Results
==============================================================================
Dep. Variable:          critical_temp   R-squared:                       0.602
Model:                            OLS   Adj. R-squared:                  0.601
Method:                 Least Squares   F-statistic:                     3565.
Date:                Wed, 17 Jul 2019   Prob (F-statistic):               0.00
Time:                        08:58:51   Log-Likelihood:                -95528.
No. Observations:               21263   AIC:                         1.911e+05
Df Residuals:                   21253   BIC:                         1.912e+05
Df Model:                           9
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                          -0.0893      1.213     -0.074      0.941      -2.467       2.289
wtd_entropy_atomic_mass        18.8542      0.863     21.836      0.000      17.162      20.547
range_fie                       0.0118      0.001      9.308      0.000       0.009       0.014
wtd_std_atomic_radius           0.0633      0.015      4.118      0.000       0.033       0.093
wtd_gmean_Density               0.0008   8.11e-05      9.717      0.000       0.001       0.001
entropy_ElectronAffinity      -39.4985      0.939    -42.073      0.000     -41.339     -37.658
gmean_FusionHeat                0.0444      0.020      2.166      0.030       0.004       0.084
wtd_std_ThermalConductivity     0.2443      0.004     61.442      0.000       0.236       0.252
wtd_mean_Valence               -3.5591      0.248    -14.369      0.000      -4.045      -3.074
number_of_elements              7.6482      0.313     24.442      0.000       7.035       8.262
==============================================================================
Omnibus:                       93.138   Durbin-Watson:                   0.768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              110.925
Skew:                           0.098   Prob(JB):                     8.18e-25
Kurtosis:                       3.294   Cond. No.                     4.18e+04
==============================================================================
```

Βλέπουμε αρχικά ότι τόσο η τιμή του συντελεστή προσδιορισμού όσο και του διορθωμένου-προσαρμοσμένου συντελεστή προσδιορισμού δεν είναι πολύ κοντά στη μονάδα (0.60). Ακόμα, η μεγάλη τιμή του στατιστικού ελέγχου F (3565) κάτω από τη μηδενική υπόθεση $b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = b_8 = b_9 = 0$, και η αντίστοιχη τιμή της p-value για τον έλεγχο αυτό που είναι σχεδόν μηδενική, δηλώνουν ότι έχουμε σοβαρές ενδείξεις για να απορρίψουμε τη μηδενική υπόθεση αυτή. Επίσης, παρατηρούμε ότι οι p-values των επεξηγηματικών μεταβλητών για το στατιστικό έλεγχο με μηδενική υπόθεση $H_0: \beta_\kappa = 0$, είναι όλες < 0.05 (κάτι που φαίνεται και από τα αντίστοιχα διαστήματα εμπιστοσύνης που δεν περιέχουν το μηδέν). Έτσι συμπεραίνουμε ότι μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0 και ότι όλες οι

επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές για το μοντέλο. Η υψηλή τιμή όμως της p-value του στατιστικού ελέγχου με μηδενική υπόθεση $H_0: \alpha = 0$, $0.94 > 0.05$, δηλώνει ότι δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι ο σταθερός όρος είναι ίσος με το μηδέν.

Αφού λοιπόν κατασκευάσαμε ένα μοντέλο γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων, το χρησιμοποιούμε για να προβλέψουμε την κρίσιμη θερμοκρασία για κάθε υπεραγωγό των δεδομένων, δίνοντας ως είσοδο στο μοντέλο μόνο τις τιμές των επεξηγηματικών μεταβλητών. Το μοντέλο παράγει τις προβλέψεις $\hat{y}$ (ypred) τις οποίες συγκρίνουμε με τις πραγματικές τους τιμές y. Στο παρακάτω γράφημα απεικονίζονται για κάθε υπεραγωγό τα σημεία (y, $\hat{y}$) καθώς και η ευθεία y=$\hat{y}$ ($ypred$).



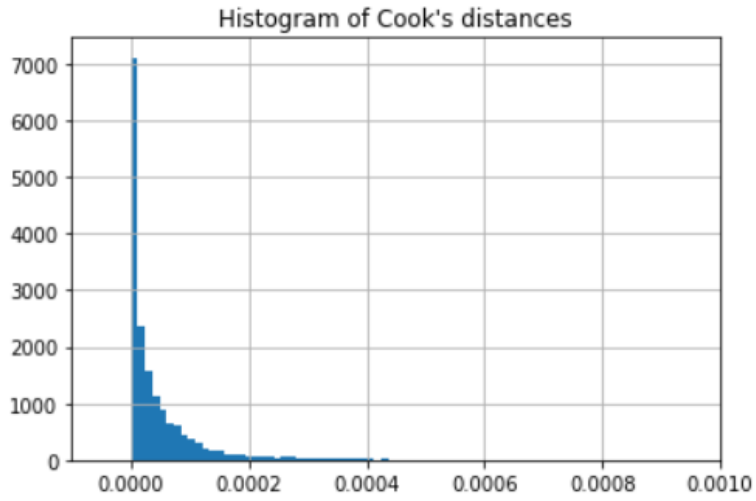
Σε ιδανικές συνθήκες θα θέλαμε όλα τα σημεία να πέφτουν ακριβώς πάνω στην ευθεία ή να απέχουν πολύ λίγο από αυτή, γιατί τότε οι προβλέψεις μας θα είχανε μεγάλη ακρίβεια. Ωστόσο κάτι τέτοιο δε συμβαίνει και ειδικά για ειδικά για ένα διάστημα τιμών (y<30 & y>80) του y, δηλαδή της κρίσιμης θερμοκρασίας οι προβλέψεις του μοντέλου έχουν μεγάλο σφάλμα. Η χαμηλή ακρίβεια του μοντέλου, φαίνεται και από τις υψηλές τιμές που παρουσίασαν οι μετρικές (μέσο τετραγωνικό σφάλμα) MSE, (μέσο απόλυτο σφάλμα) MAE, RMSE, μέση τιμή των σχετικών σφαλμάτων (mean of Relative Errors), διασπορά των σχετικών σφαλμάτων (variance of Relative Errors). Στο σημείο αυτό αξίζει να σημειωθεί πως οι δυο τελευταίες μετρικές υπολογίστηκαν και μετά την αφαίρεση ακραίων τιμών (outliers) των σχετικών σφαλμάτων (συγκεκριμένα αφαιρέθηκαν όλες οι τιμές για τις οποίες το σχετικό σφάλμα ήταν >10 ή <-10). Συγκεκριμένα για το γραμμικό μοντέλο παλινδρόμησης οι τιμές των μετρικών με τις οποίες αξιολογήσαμε το μοντέλο μας υπολογίστηκαν ως:

- MSE=465.27
- RMSE=21.57, ισοδυναμεί με σφάλμα ±21.57K
- MAE=17.08
- R^2=0.60
- Mean of Relative Errors = 19.73

- Variance of Relative Errors=958.75
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.47
- Variance of Relative Errors (without outliers) =2.02

Συμπεραίνουμε λοιπόν πως η απόδοση του γραμμικού μοντέλου δεν είναι καλή αφού ένα σφάλμα ± 21.57K, είναι αρκετά μεγάλο. Ωστόσο, η κακή απόδοση αυτή μπορεί να οφείλεται σε διάφορους λόγους. Στην ύπαρξη σημείων επιρροής (leverage points), στο φαινόμενο πολυσυγγραμικότητας μεταξύ των μεταβλητών αλλά και στην έλλειψη γραμμικής σχέσης μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Ας εξετάσουμε λοιπόν όλες αυτές τις περιπτώσεις ξεχωριστά για να κατανοήσουμε ποιος παράγοντας επηρεάζει πιθανώς την απόδοση του μοντέλου.

Αρχικά εντοπίζουμε τα σημεία επιρροής υπολογίζοντας την απόσταση Cook για κάθε παρατήρηση και παράγουμε το ιστόγραμμα των αποστάσεων που φαίνεται στην παρακάτω εικόνα:

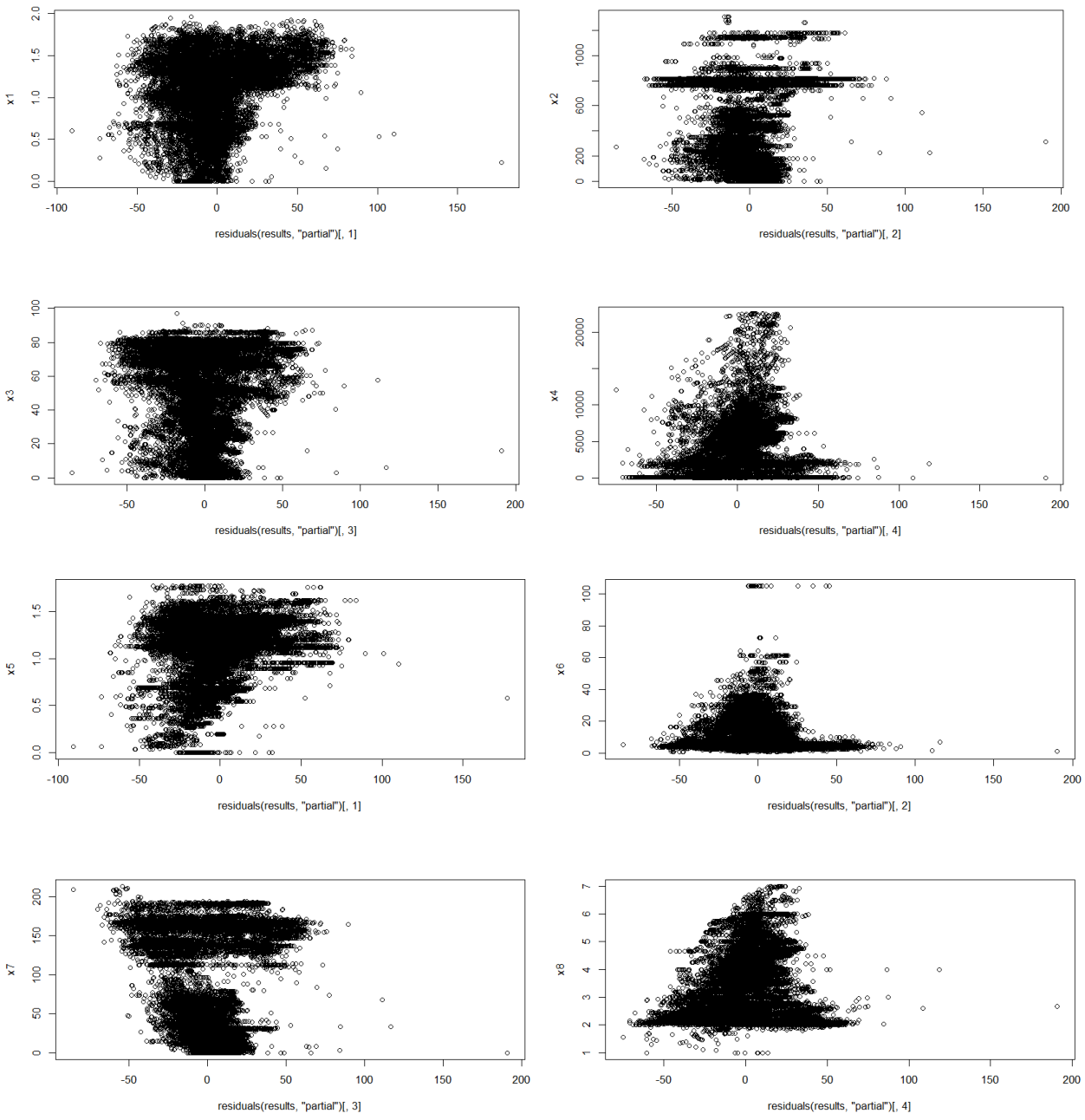


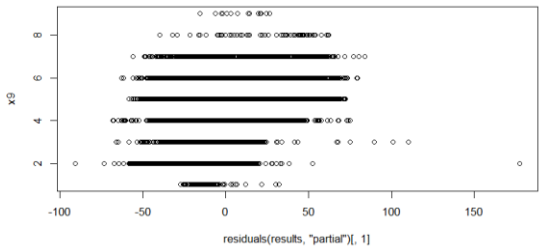
Όπως είναι φανερό οι περισσότερες αποστάσεις είναι πολύ κοντά στο μηδέν. Θα θεωρήσουμε ως το όριο (threshold), από το οποίο και πέρα μια παρατήρηση θεωρείται σημείο επιρροής το $D_{Cook} > 4/n$, που στην δική μας περίπτωση είναι $D_{Cook} > \frac{4}{21263} = 0.0002$. Αφαιρούμε τις τιμές αυτές και ξαναπροσαρμόζουμε το μοντέλο. Όμως, ακόμα και μετά από αυτή την επέμβαση στα δεδομένα, η απόδοση του μοντέλου παραμένει σχεδόν αμετάβλητη. Έτσι συμπεραίνουμε πως δεν ευθύνονται τα σημεία επιρροής για την χαμηλή ακρίβεια του μοντέλου.

Στη συνέχεια για να ελέγξουμε την επιρροή της πολυσυγγραμικότητας στην απόδοση του μοντέλου, πραγματοποιούμε παλινδρόμηση κορυφογραμμής (ridge regression), όμως και πάλι η απόδοση του μοντέλου παραμένει σχεδόν αμετάβλητη.

Έτσι λοιπόν καταλήγουμε στο συμπέρασμα πως ο λόγος που το μοντέλο έχει τόσο χαμηλή ακρίβεια είναι κυρίως ότι η σχέση μεταξύ των επεξηγηματικών

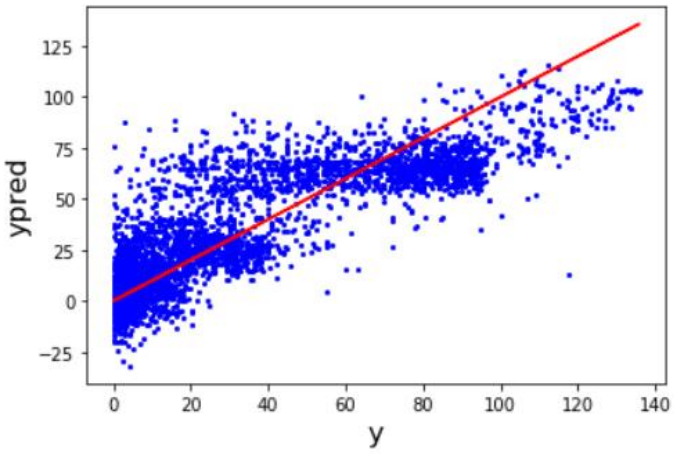
μεταβλητών και της μεταβλητής απόκρισης δεν είναι γραμμική και για το λόγο αυτό δεν είναι αποτελεσματική η εφαρμογή ενός γραμμικού μοντέλου παλινδρόμησης. Αυτό μπορούμε να το διαπιστώσουμε και από τον έλεγχο προϋποθέσεων του γραμμικού μοντέλου και συγκεκριμένα από έλεγχο γραμμικότητας. Λόγω του ότι οι μεταβλητές παρουσιάζουν πολυσυγγραμικότητα, δηλαδή είναι συσχετισμένες μεταξύ τους θα παράγουμε τα διαγράμματα διασποράς των μερικών υπολοίπων για όλες τις επεξηγηματικές μεταβλητές, τα οποία φαίνονται στην παρακάτω εικόνα:





Όπως βλέπουμε από τα παραπάνω διαγράμματα η υπόθεση της γραμμικότητας, δε φαίνεται λογική.

Στο σημείο αυτό αξίζει να σημειωθεί πως προσαρμόσαμε το γραμμικό μοντέλο παλινδρόμησης και σε όλο το σύνολο των μεταβλητών (81), ώστε να δούμε πως θα μεταβληθεί η απόδοση του. Τα αποτελέσματα της προσαρμογής φαίνονται παρακάτω:

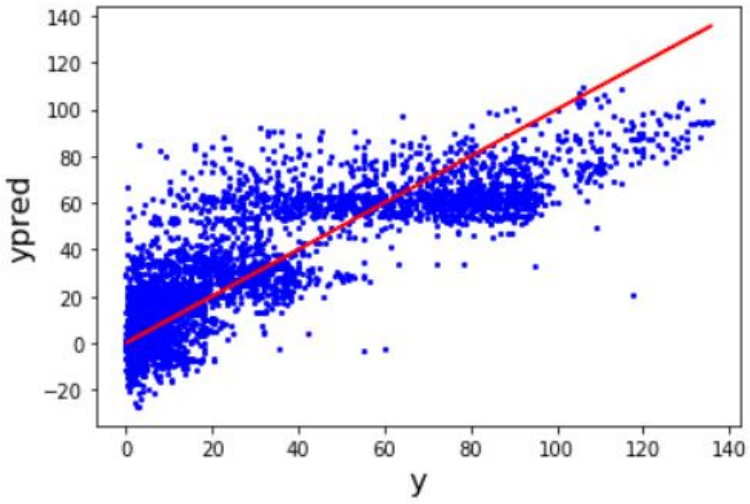


- MSE=318.28
- RMSE=17.84, ισοδυναμεί με σφάλμα $\pm$17.84K
- MAE=13.64
- R^2=0.73
- Mean of Relative Errors = 25.17
- Variance of Relative Errors=774.11
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.38
- Variance of Relative Errors (without outliers) =1.91

Όπως είναι φανερό από τα παραπάνω αποτελέσματα η απόδοση του μοντέλου για όλο το σύνολο των μεταβλητών είναι αισθητά καλύτερη, κάτι που είναι λογικό αφού οι παραπάνω 72 μεταβλητές που προσθέσαμε στις 9 που είχαμε κρατήσει για το μοντέλο, συνεισφέρουν σε αυτό με πρόσθετη πληροφορία η οποία συμβάλλει στην βελτίωση του.

4.2. Μέθοδος Μερικών Ελαχίστων Τετραγώνων

Τέλος, δοκιμάσαμε τη μέθοδο μερικών ελαχίστων τετραγώνων κρατώντας τα πρώτα 9 συστατικά της (πρώτες 9 διευθύνσεις) για να προσαρμόσουμε ένα μοντέλο παλινδρόμησης. Τα αποτελέσματα που πήραμε ήταν τα ακόλουθα:



- MSE=356.50
- RMSE=18.88, ισοδυναμεί με σφάλμα ±18.88 K
- MAE=14.67
- R^2=0.69
- Mean of Relative Errors = 3.29
- Variance of Relative Errors=849.34
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.40
- Variance of Relative Errors (without outliers) =2.02

Από τα αποτελέσματα συμπεραίνουμε, ότι η μέθοδος έχει λίγο χειρότερη απόδοση από ότι τα ελάχιστα τετράγωνα με όλες τις μεταβλητές, ωστόσο όμως έχει καλύτερη απόδοση από τα ελάχιστα τετράγωνα με τις 9 καλύτερες μεταβλητές που επιλέχθηκαν για το μοντέλο. Αυτό είναι λογικό γιατί παρόλο που και στα μερικά ελάχιστα τετράγωνα και στα ελάχιστα τετράγωνα με 9 μεταβλητές έχουμε τον ίδιο αριθμό μεταβλητών, οι επεξηγηματικές μεταβλητές των μερικών ελαχίστων τετραγώνων ανταποκρίνονται σε γραμμικούς συνδυασμούς (διαφορετικές διευθύνσεις) όλων των αρχικών επεξηγηματικών μεταβλητών με αποτέλεσμα έτσι να περιέχουν πληροφορία από όλες τις μεταβλητές του μοντέλου σε αντίθεση με τα κανονικά ελάχιστα τετράγωνα στα οποία διαθέτουμε πληροφορία μόνο για τις μεταβλητές τις οποίες έχουμε κρατήσει στο μοντέλο. Αξίζει ακόμα να σημειωθεί ότι από τις δυο μεθόδους με τις 9 μεταβλητές εμείς θα επιλέγαμε εκείνη των ελαχίστων τετραγώνων καθώς η μέθοδος μερικών ελαχίστων τετραγώνων απαιτεί τη συλλογή πληροφορίας για όλες τις μεταβλητές οπότε μπορεί να έχουμε μειώσει τις διαστάσεις του προβλήματος (dimensionality reduction) σε 9 μεταβλητές εισόδου, δεν έχουμε

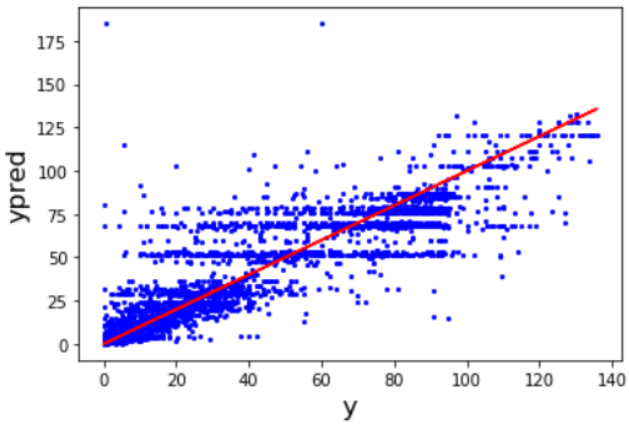
όμως μειώσει τον αριθμό των επεξηγηματικών μεταβλητών (οι 9 μεταβλητές εισόδου περιέχουν πληροφορία από όλες τις μεταβλητές, 81). Αυτή είναι και η βασική διαφορά του όρου μείωση διαστάσεων (dimensionality reduction) από τον όρο επιλογή μεταβλητών (feature selection). Τέλος ένα βασικό μειονέκτημα των μεθόδων που πραγματοποιούν μείωση διαστάσεων, άρα και της partial least squares, είναι ότι επειδή οι μεταβλητές που δημιουργούν (συστατικά-διευθύνσεις) είναι γραμμικοί συνδυασμοί όλων των μεταβλητών, δε μπορούμε να πραγματοποιήσουμε στατιστική συμπερασματολογία σε αυτές.

5. Μη Γραμμική Παλινδρόμηση

Η αδυναμία της γραμμικής παλινδρόμησης, να επιλύσει το πρόβλημα και να δημιουργήσει ένα δυνατό μοντέλο πρόβλεψης της κρίσιμης θερμοκρασίας, μας οδήγησε στην αναζήτηση μοντέλων μη γραμμικής παλινδρόμησης, ώστε αυτά να αναγνωρίσουν και να εντοπίσουν τις σχέσεις και το μοτίβο που συνδέονται οι μεταβλητές απόκρισης με την επεξηγηματική μεταβλητή. Στο πλαίσιο αυτό δοκιμάστηκαν αρκετές μέθοδοι ώστε να εντοπίσουμε ποια βρίσκει την καλύτερη λύση για το πρόβλημα μας.

5.1. Δέντρο Απόφασης

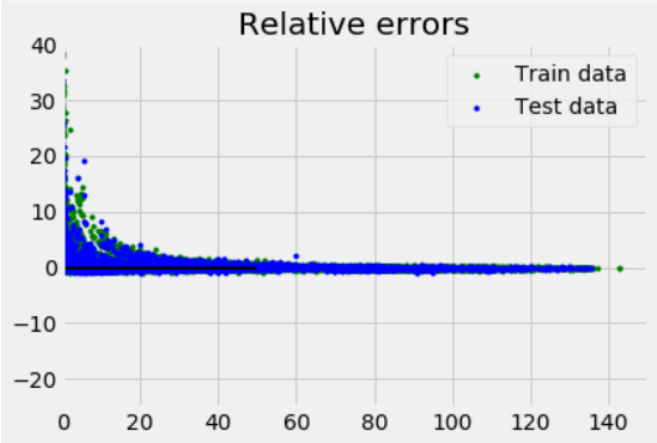
Αρχικά, δοκιμάσαμε να κατασκευάσουμε ένα απλό δέντρο απόφασης για να δούμε πως αυτό ανταποκρίνεται στα δεδομένα και τι ακρίβεια μπορεί να πετύχει. Έτσι, αφού βρήκαμε τις βέλτιστες παραμέτρους ώστε να έχουμε τη μεγαλύτερη δυνατή ακρίβεια κρατώντας παράλληλα τη γενίκευση του μοντέλου, δηλαδή την ίδια απόδοση του στα δεδομένα με τα οποία εκπαιδεύτηκε και στα δεδομένα επαλήθευσης, καταλήξαμε σε ένα δέντρο με βάθος 10 και ελάχιστο αριθμό παρατηρήσεων που απαιτούνται στα φύλλα του δέντρου να είναι 1. Τα αποτελέσματα που πήραμε ήταν τα ακόλουθα:



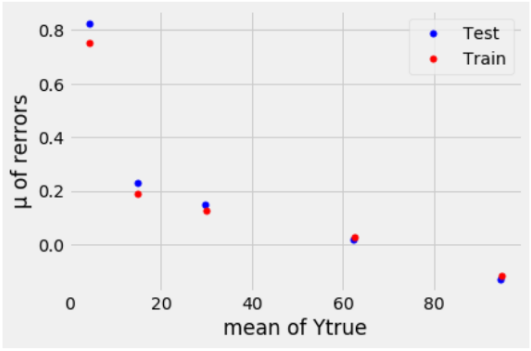
- MSE=201.72
- RMSE=14.20, ισοδυναμεί με σφάλμα ±14.20 K
- MAE=8.41

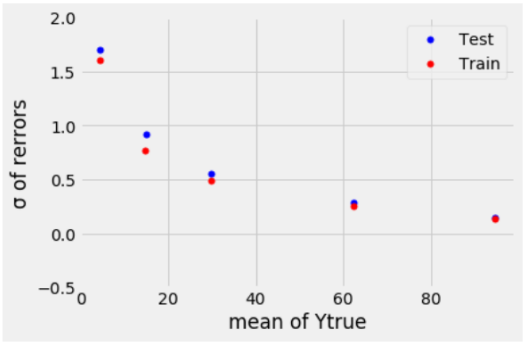
- R^2=0.82
- Mean of Relative Errors = 18.41
- Variance of Relative Errors=530.74
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.33
- Variance of Relative Errors (without outliers) =1.16

Όπως φαίνεται από τα παραπάνω αποτελέσματα βλέπουμε σαν μια πρώτη εικόνα ότι έχουμε αρκετά καλύτερη απόδοση του μοντέλου σε σχέση με τα προηγούμενα γραμμικά. Ενώ το σφάλμα μειώθηκε μόλις σε $\pm$14.20 K από $\pm$21.57K που είχαμε με τη μέθοδο ελαχίστων τετραγώνων και όλα αυτά χρησιμοποιώντας μόνο τις 9 μεταβλητές που κρατήσαμε για το μοντέλο μας. Στη συνέχεια βλέπουμε για τα δεδομένα εκπαίδευσης και επαλήθευσης το διάγραμμα διασποράς των σχετικών σφαλμάτων προς την κρίσιμη θερμοκρασία.



Παρατηρούμε ότι τα μεγαλύτερα σχετικά σφάλματα, παρουσιάζονται για τιμές της κρίσιμης θερμοκρασίας στο διάστημα 0-20K. Στη συνέχεια για μια πιο λεπτομερή αξιολόγηση του μοντέλου, αφαιρούμε τις ακραίες τιμές των σχετικών σφαλμάτων, δηλαδή αυτές που είναι >10 ή <-10 και παράλληλα σπάμε το διάγραμμα σε πέντε διαστήματα της κρίσιμης θερμοκρασίας([0,10], (10,20], (20,40], (40,80], (80,140]), και για κάθε ένα από αυτά βλέπουμε στα παρακάτω διαγράμματα τη μέση τιμή και τη διασπορά των σχετικών σφαλμάτων αντίστοιχα για τα δεδομένα επαλήθευσης και εκπαίδευσης ξεχωριστά.

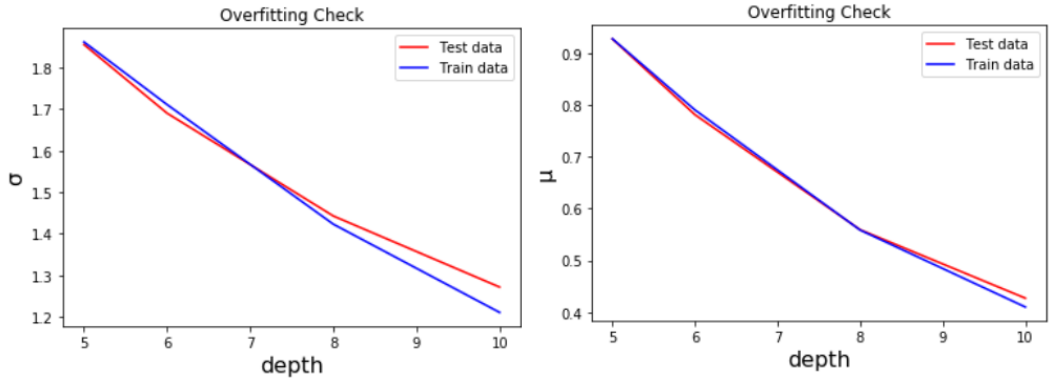




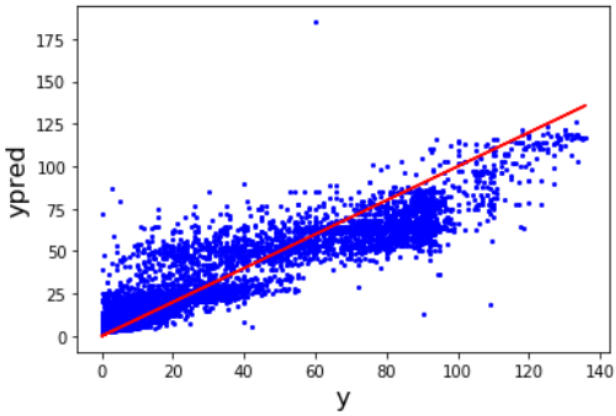
Από τα διαγράμματα, παρατηρούμε αυτό που είδαμε και στο αρχικό διάγραμμα διασποράς ότι τα μεγαλύτερα σφάλματα παρουσιάζονται για το διάστημα 0-20K, αλλά και ότι όσο αυξάνεται η κρίσιμη θερμοκρασία η μέση τιμή και η διασπορά των σχετικών σφαλμάτων τείνει να γίνεται αρνητική. Ακόμα βλέπουμε ότι τα σημεία για τα δεδομένα εκπαίδευσης και επαλήθευσης, είναι πολύ κοντά μεταξύ τους, άρα δεν έχουμε πρόβλημα υπερπροσαρμογής και το μοντέλο μας είναι αρκετά γενικευμένο. Να σημειωθεί τέλος ότι για μεγαλύτερες τιμές βάθους το μοντέλο είχε αρχίσει να πάσχει από υπερπροσασμογή και τα σημεία απείχαν αρκετά μεταξύ τους. Ας δούμε λοιπόν πως θα βελτιωθεί η απόδοση του μοντέλου χρησιμοποιώντας τεχνικές boosting με δέντρα απόφασης.

5.2. Adaboost

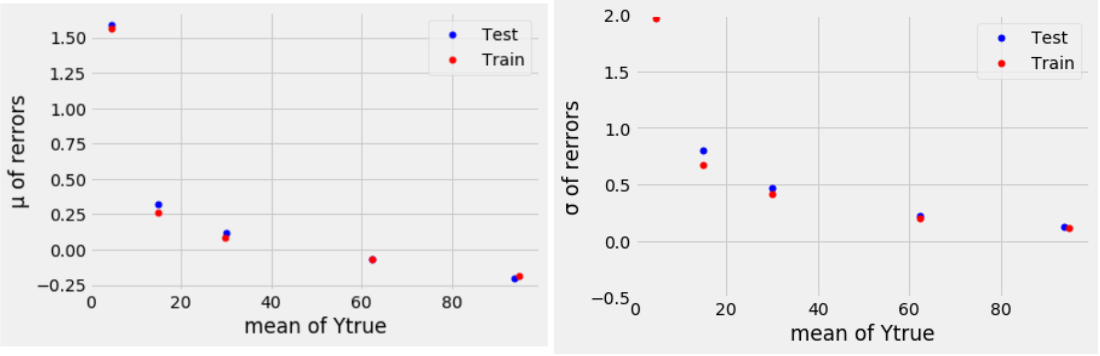
Για τη μέθοδο Adaboost, χρησιμοποιήθηκε ως αδύναμος εκτιμητής ένα δέντρο παλινδρόμησης. Αφού βρήκαμε τις βέλτιστες παραμέτρους ώστε να έχουμε τη μεγαλύτερη δυνατή ακρίβεια κρατώντας παράλληλα τη γενίκευση του μοντέλου, δηλαδή την ίδια απόδοση του στα δεδομένα με τα οποία εκπαιδεύτηκε και στα δεδομένα επαλήθευσης, καταλήξαμε στο να χρησιμοποιήσουμε για τη μέθοδο 70 δέντρα με μέγιστο βάθος 9 και ρυθμό εκμάθησης 0.5. Η επιλογή των βέλτιστων παραμέτρων έγινε με τη βοήθεια αλγορίθμων οι οποίοι επέστρεφαν τις τιμές για τις οποίες το μοντέλο είχε την καλύτερη απόδοση και για τα δεδομένα επαλήθευσης και για τα δεδομένα εκπαίδευσης με βάση μια μετρική. Στο παρακάτω διάγραμμα βλέπουμε τη μέση τιμή και τη διασπορά των σχετικών σφαλμάτων αντίστοιχα για τα δεδομένα επαλήθευσης και εκπαίδευσης ξεχωριστά για διαφορετικά βάθη.



Βλέπουμε ότι μέχρι μέγιστο βάθος 9 η απόδοση του μοντέλου για τα δεδομένα επαλήθευσης και εκπαίδευσης ταυτίζεται. Τα αποτελέσματα που πήραμε από το ήταν τα ακόλουθα:



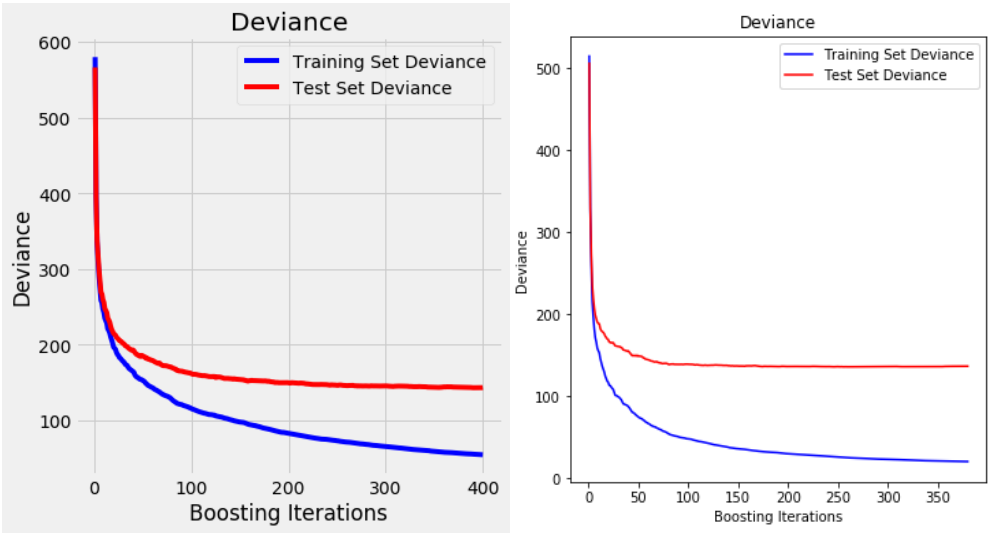
- MSE=186.42
- RMSE=13.65, ισοδυναμεί με σφάλμα ±13.65 K
- MAE=9.59
- R^2=0.84
- Mean of Relative Errors = 16.20
- Variance of Relative Errors=527.70
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.40
- Variance of Relative Errors (without outliers) =1.29



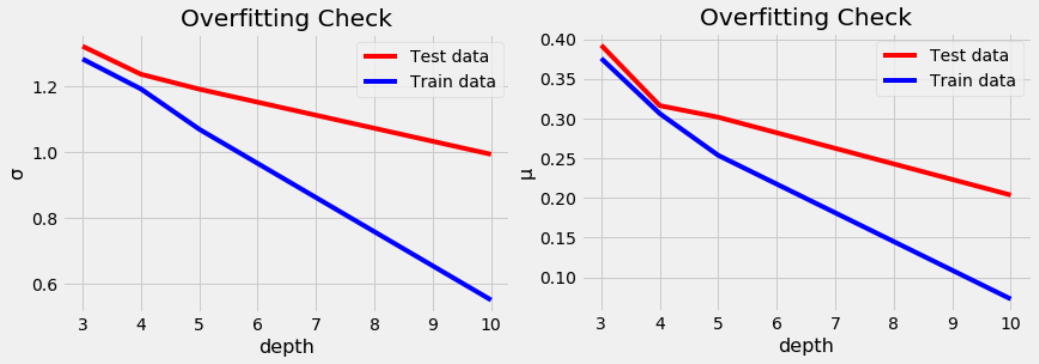
Από τα αποτελέσματα, παρατηρούμε και εδώ ότι τα μεγαλύτερα σφάλματα παρουσιάζονται για το διάστημα 0-20K, αλλά και ότι όσο αυξάνεται η κρίσιμη θερμοκρασία η μέση τιμή και η διασπορά των σχετικών σφαλμάτων τείνει να γίνεται αρνητική. Ακόμα βλέπουμε ότι τα σημεία για τα δεδομένα εκπαίδευσης και επαλήθευσης, είναι πολύ κοντά μεταξύ τους, άρα δεν έχουμε πρόβλημα υπερπροσαρμογής και το μοντέλο μας είναι αρκετά γενικευμένο. Να σημειωθεί επίσης ότι για μεγαλύτερες τιμές βάθους το μοντέλο είχε αρχίσει να πάσχει από υπερπροσασμογή και τα σημεία απείχαν αρκετά μεταξύ τους. Τέλος βλέπουμε ότι το μοντέλο παρουσιάζει λίγο καλύτερη απόδοση σε σχέση με το απλό δέντρο απόφασης. Στη συνέχεια θα δούμε την υλοποίηση της μεθόδου Gradient Boosting.

5.3. Gradient Boosting

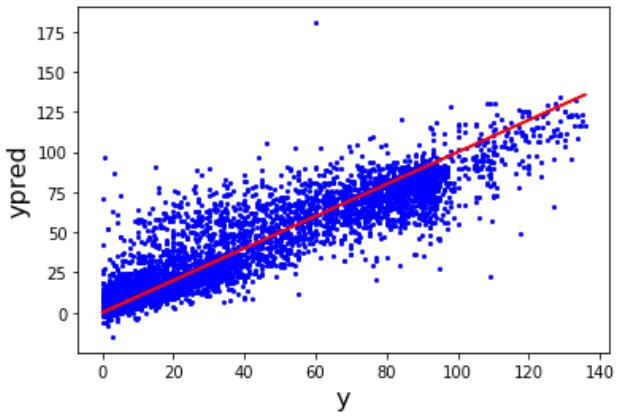
Για τη μέθοδο Gradient Boosting, χρησιμοποιήθηκε ως αδύναμος εκτιμητής ένα δέντρο παλινδρόμησης. Αφού βρήκαμε τις βέλτιστες παραμέτρους ώστε να έχουμε τη μεγαλύτερη δυνατή ακρίβεια κρατώντας παράλληλα τη γενίκευση του μοντέλου, δηλαδή την ίδια απόδοση του στα δεδομένα με τα οποία εκπαιδεύτηκε και στα δεδομένα επαλήθευσης, καταλήξαμε στο να χρησιμοποιήσουμε για τη μέθοδο 110 δέντρα με μέγιστο βάθος 3 και ρυθμό εκμάθησης 0.45. Στα παρακάτω διαγράμματα παρουσιάζονται τα αποτελέσματα των αλγορίθμων που χρησιμοποιήθηκαν ώστε να βρεθούν οι βέλτιστες τιμές των παραμέτρων. Οι αλγόριθμοι στην ουσία κρατούσαν σταθερές τις τιμές κάποιων παραμέτρων και μετέβαλλαν τις υπόλοιπες αξιολογώντας με τις μετρικές που χρησιμοποιούμε το πως επηρεάζεται η απόδοση του μοντέλου κάθε φορά.



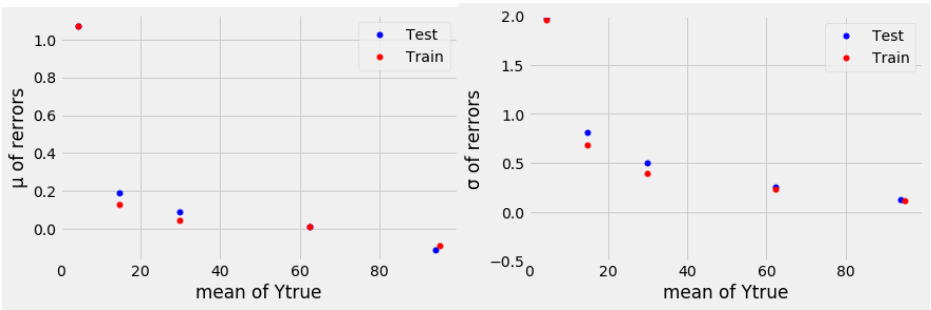
Στο πρώτο γράφημα βλέπουμε για μέγιστο βάθος 3 και βαθμό εκμάθησης 0.45 ότι ο βέλτιστος αριθμός δέντρων, είναι 100 δέντρα περίπου αφού μέχρι εκεί η απόδοση του μοντέλου για το σύνολο δεδομένων εκπαίδευσης και για το σύνολο επαλήθευσης είναι παρόμοια, ενώ στη συνέχεια αυξάνεται η απόκλιση τους. Στο δεύτερο γράφημα όπου παρουσιάζεται η ίδια πληροφορία για μέγιστο βάθος 5, παρατηρούμε ότι ο βέλτιστος αριθμός εκτιμητών είναι 50 δέντρα όμως ακόμα και για αυτά η απόκλιση του μοντέλου για τα δεδομένα εκπαίδευσης είναι αρκετά μεγαλύτερη από ότι για βάθος 3.



Στα δυο παραπάνω γραφήματα βλέπουμε πως μεταβάλλεται η μέση τιμή και η διασπορά των σχετικών σφαλμάτων (όταν έχουμε αφαιρέσει τις ακραίες τιμές relative errors>10 & relative errors ≤ -10), για διαφορετικό μέγιστο βάθος των δέντρων. Είναι φανερό ότι από βάθος 4 το μοντέλο αρχίζει να αποκλίνει και να υπερεκπαιδεύεται γι' αυτό και παρουσιάζει πολύ καλύτερη απόδοση στα δεδομένα εκπαίδευσης σε σχέση με τα δεδομένα επαλήθευσης. Παρακάτω βλέπουμε τα αποτελέσματα του μοντέλου:



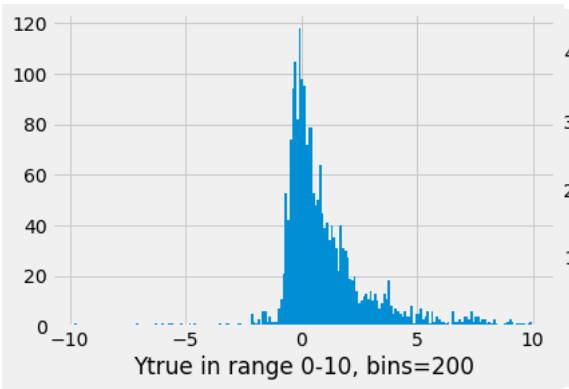
- MSE=158.52
- RMSE=12.59, ισοδυναμεί με σφάλμα ±12.59 K
- MAE=8.26
- R^2=0.86
- Mean of Relative Errors = 14.69
- Variance of Relative Errors=500.44
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.39
- Variance of Relative Errors (without outliers) =1.32

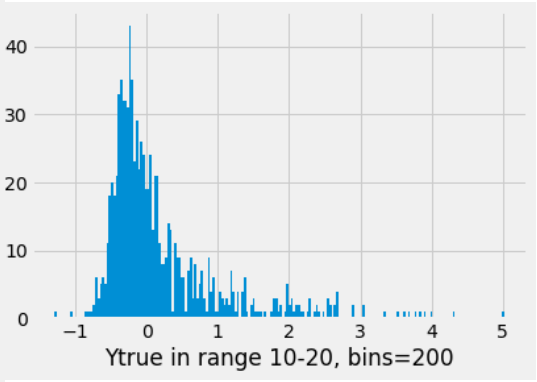


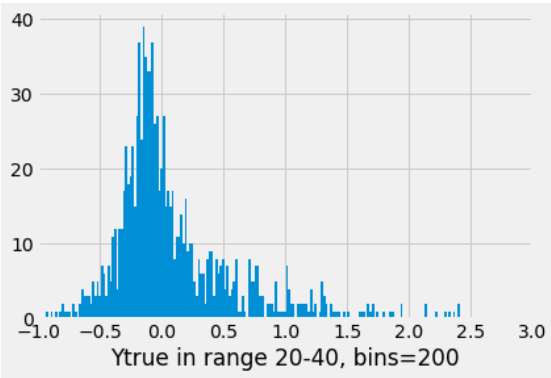
Παρατηρούμε ότι το μοντέλο είναι μέχρι στιγμής το καλύτερο που έχουμε βρει ωστόσο με όχι πολύ μεγάλη διαφορά από τα προηγούμενα. Ακόμα όπως βλέπουμε από τα διαγράμματα η απόδοση του μοντέλου για τα δεδομένα εκπαίδευσης και επαλήθευσης είναι αρκετά παρόμοια και έτσι συμπεραίνουμε ότι το μοντέλο είναι αρκετά γενικευμένο και μπορεί να δουλέψει αποτελεσματικά και σε καινούρια δεδομένα που δεν έχει ξαναδεί. Στο συμπέρασμα αυτό καταλήγουμε και αν πραγματοποιήσουμε 10 K Cross Validation όπου παίρνουμε τα παρακάτω αποτελέσματα:

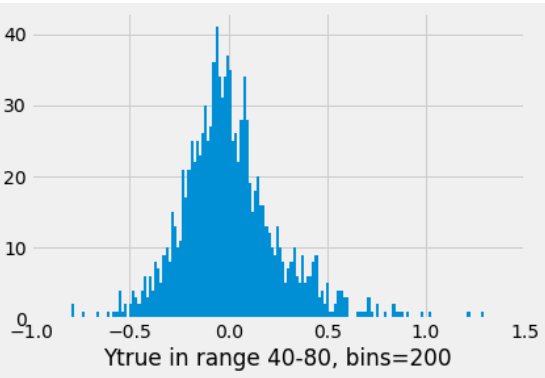
```
array([0.87447864, 0.85236941, 0.8609592 , 0.88131456, 0.86536494,
       0.87951829, 0.85266235, 0.86562184, 0.87679721, 0.87527896])
```

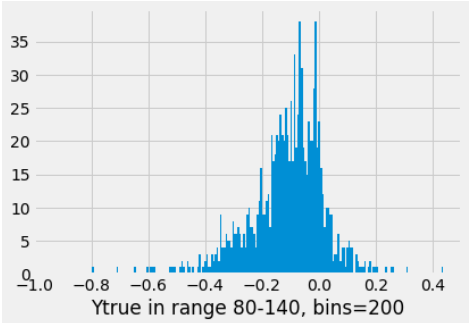
Όπως φαίνεται οι τιμές είναι πολύ κοντά μεταξύ τους, άρα μπορούμε να πειστούμε ότι το μοντέλο είναι γενικευμένο. Τέλος στα παρακάτω διαγράμματα βλέπουμε για τα 5 διαστήματα τιμών στα οποία χωρίσαμε την κρίσιμη θερμοκρασία από τα οποία προέκυψαν τα προηγούμενα δυο διαγράμματα το ιστόγραμμα των σχετικών σφαλμάτων.



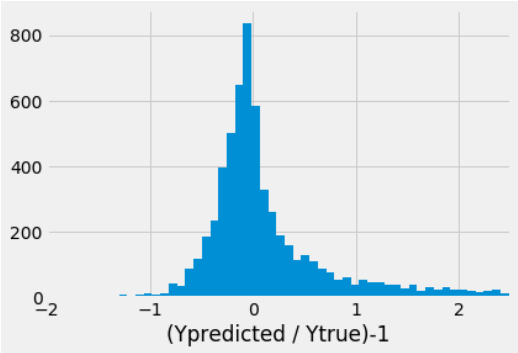








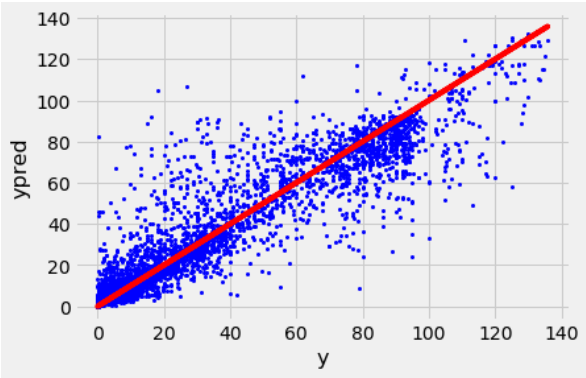
Για όλο το διάστημα τιμών το αντίστοιχο ιστόγραμμα είναι το παρακάτω:



Από τα παραπάνω ιστογράμματα παρατηρούμε ότι οι τιμές που βρήκαμε για τη μέση τιμή και τη διασπορά των σχετικών σφαλμάτων σε κάθε περιοχή φαίνονται λογικές, αφού οι περισσότερες τιμές του ιστογράμματος είναι κοντά στο μηδέν ωστόσο μερικές μεγάλες τιμές των σχετικών σφαλμάτων προκαλούν μια αύξηση στην τελική μέση τιμή και διασπορά. Στο σημείο αυτό αξίζει να σημειωθεί ότι για να ελέγξουμε με εναλλακτικό τρόπο τη μέση τιμή και τη διασπορά σε κάθε περιοχή προσαρμόσαμε μια Johnson SU κατανομή στα σχετικά σφάλματα (όχι κανονική γιατί είχαμε μεγάλες ουρές, οπότε θέλαμε μια κατανομή που να είναι γύρω από το μηδέν και να έχει μεγάλες ουρές) και παρατηρήσαμε ότι οι εκτιμήσεις της μέσης τιμής και της διασποράς της της κατανομής ήταν πολύ κοντά στις τιμές που υπολογίσαμε για κάθε περιοχή. Τέλος, τρέξαμε την Gradient Boosting και για όλο το σύνολο των μεταβλητών (81 μεταβλητές), και πετύχαμε RMSE=11.50, δηλαδή μικρότερο σφάλμα σε σχέση με εκείνο της μεθόδου όταν χρησιμοποιεί μόνο 9 μεταβλητές (12.59 K), κάτι το οποίο ήταν λογικό και οφείλεται στην πληροφορία που περιέγραφαν οι μεταβλητές που δεν χρησιμοποιήθηκαν.

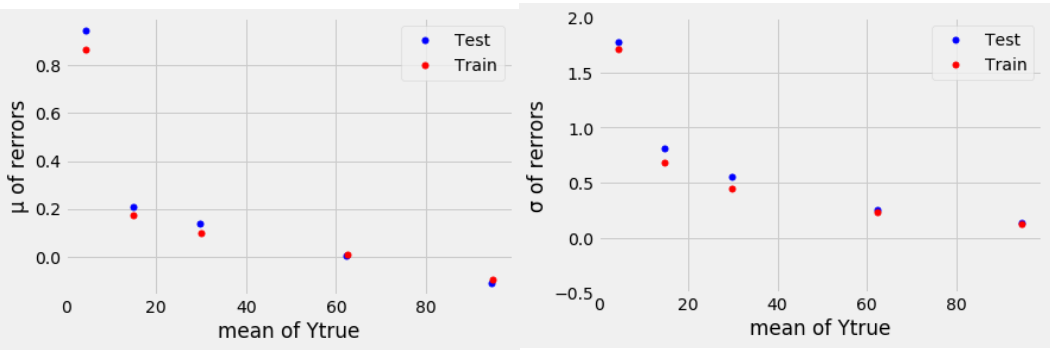
5.4. KN-Neighbors

Μια μέθοδος που επίσης είχε πολύ καλή απόδοση ήταν η KN-Neighbors, την οποία υλοποιήσαμε για 8 γείτονες και βρίσκοντας την απόσταση Minkowski μεταξύ των παρατηρήσεων, πήραμε τα παρακάτω αποτελέσματα:



- MSE=166.49

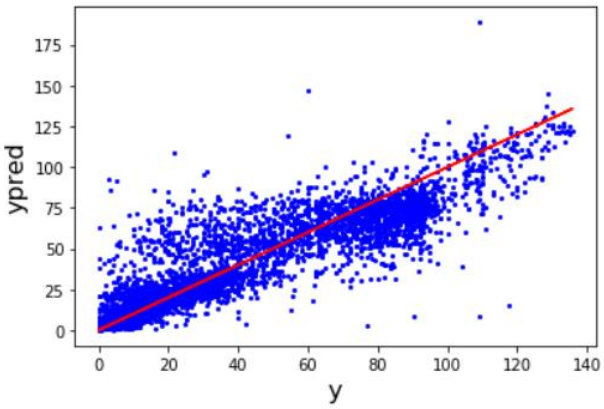
- RMSE=12.90, ισοδυναμεί με σφάλμα $\pm$12.90 K
- MAE=7.59
- R^2=0.86
- Mean of Relative Errors = 9.54
- Variance of Relative Errors=274.62
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.36
- Variance of Relative Errors (without outliers) =1.20



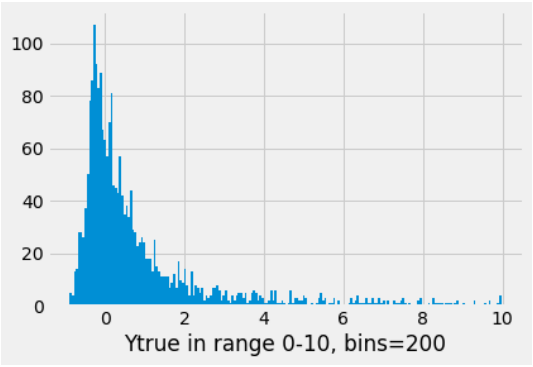
Παρατηρούμε ότι η απόδοση του μοντέλου είναι αρκετά παρόμοια με την Gradient Boosting, κάτι που μας δείχνει τη δύναμη της μεθόδου. Να σημειωθεί εδώ ότι η μέθοδος αυτή είναι από τις πιο απλές στον τρόπο που κάνει τις προβλέψεις της αλλά έχει και πολύ λίγες παραμέτρους που πρέπει να προσδιοριστούν πριν την εφαρμογής της με αποτέλεσμα να είναι πιο δύσκολο να παρουσιάσει φαινόμενα υπερπροσαρμογής.

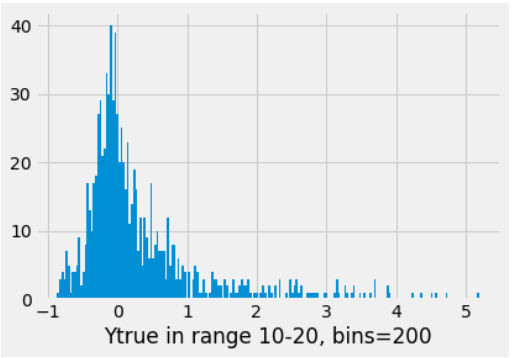
5.5. Νευρωνικό Δίκτυο

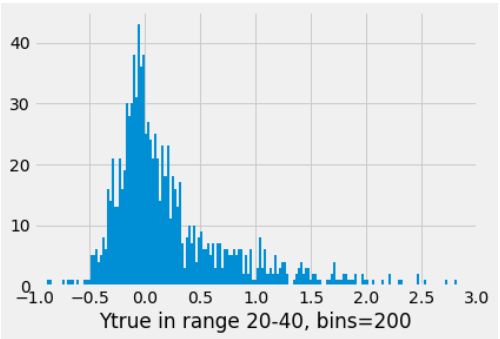
Σχεδιάσαμε, τέλος, ένα νευρωνικό δίκτυο με σκοπό να δούμε κατά πόσο τελικά θα μπορούσαμε να βελτιώσουμε τις προβλέψεις μας πριν επιλέξουμε το τελικό μας μοντέλο. Χρησιμοποιώντας τη βιβλιοθήκη Keras της python, αφού πειραματιστήκαμε με διάφορες τεχνικές, ξεκινώντας από τις πιο απλές (ένα κρυφό στρώμα με τόσους νευρώνες όσες οι μεταβλητές απόκρισης), καταλήξαμε στο ότι η καλύτερη αρχιτεκτονική για το νευρωνικό δίκτυο που επέτρεπε στο μοντέλο να είναι όσο πιο γενικευμένο γίνεται θυσιάζοντας και λίγο από την ακρίβεια του, ήταν 9 κρυφά στρώματα με 140,120,80,60,40,30,20,15,10 νευρώνες το καθένα αντίστοιχα. Ως συνάρτηση ενεργοποίησης για κάθε στρώμα χρησιμοποιήθηκε η ανορθωτική (ReLU) ενώ ως μέθοδος βελτιστοποίησης για την ανανέωση των βαρών η gradient descent. Τέλος, ως συνάρτηση κόστους του μοντέλου ορίστηκε το MSE. Τα αποτελέσματα του νευρωνικού δικτύου ήταν τα παρακάτω:

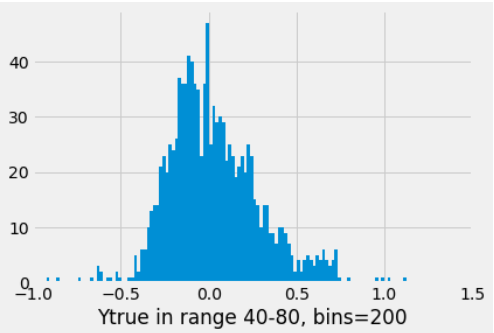


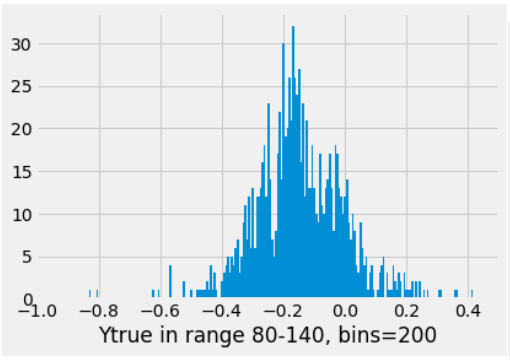
- MSE=179.92
- RMSE=13.41, ισοδυναμεί με σφάλμα ±13.41 K
- MAE=7.59
- R^2=0.84
- Mean of Relative Errors = 9.45
- Variance of Relative Errors=386.71
- Mean of Relative Errors (in range (-10,10)-without outliers) = 0.25
- Variance of Relative Errors (without outliers) =1.09

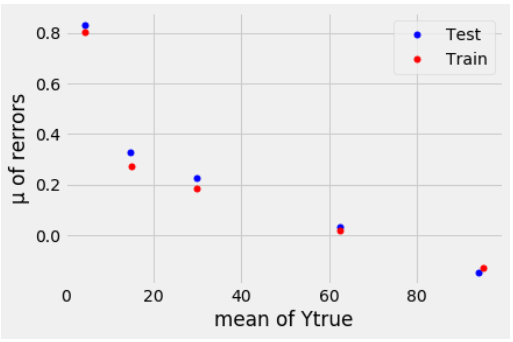


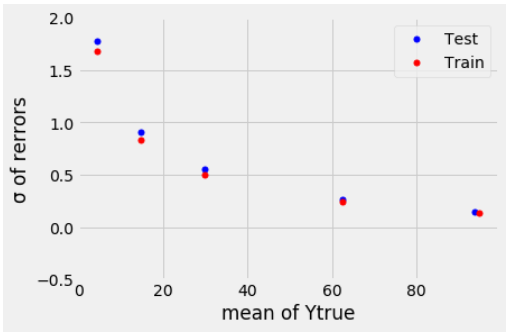




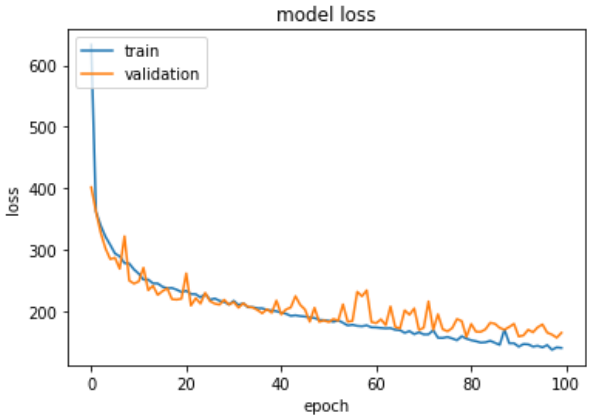






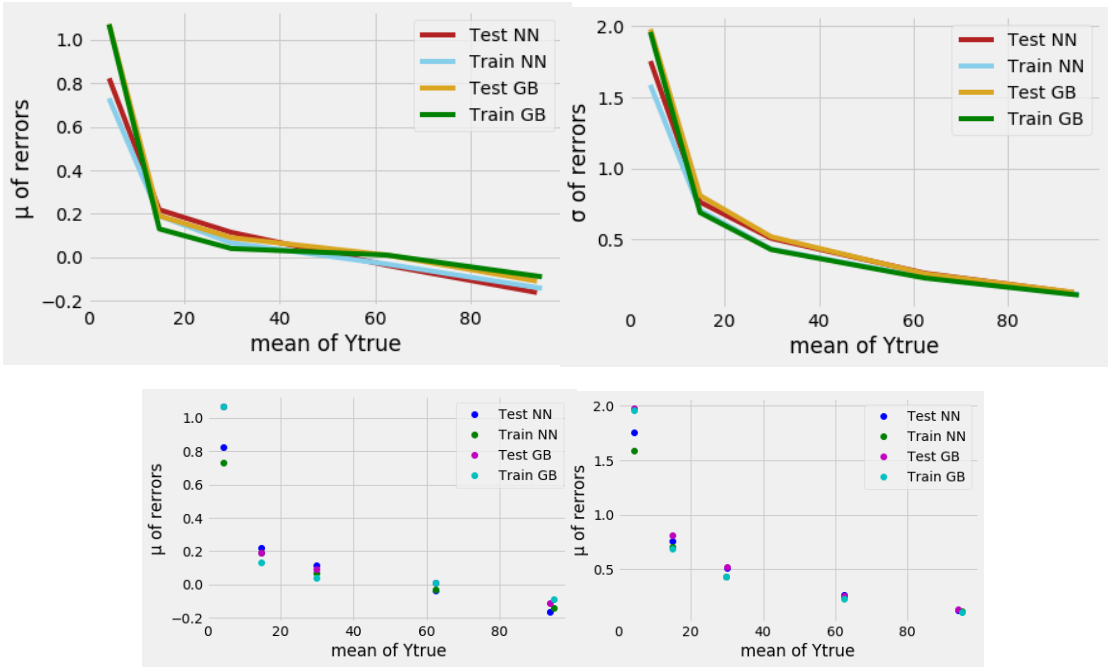


Από τα παραπάνω αποτελέσματα, συμπεραίνουμε ότι ενώ η απόδοση του νευρωνικού δικτύου, ήταν αρκετά καλή δε μπόρεσε να ξεπεράσει την απόδοση της Gradient Boosting, χωρίς ωστόσο να έχει μεγάλη απόκλιση από αυτή. Στο παρακάτω γράφημα βλέπουμε πως μεταβάλλεται η συνάρτηση κόστους (στη συγκεκριμένη περίπτωση το MSE), σε κάθε εποχή. Βλέπουμε ότι μέχρι τις 100 εποχές στις οποίες εκπαιδεύσαμε το νευρωνικό δίκτυο, η απόδοση του μοντέλου στα δεδομένα εκπαίδευσης σε σχέση με τα δεδομένα επαλήθευσης έχει πολύ μικρή απόκλιση. Άρα μπορούμε να πιστεύουμε ότι δεν έχουμε υπερπροσαρμογή στα δεδομένα κάτι το οποίο επαληθεύτηκε και με cross validation.



Τέλος, στα επόμενα γραφήματα βλέπουμε τη σύγκριση του νευρωνικού δικτύου με τη gradient boosting, μέσα από τα διαγράμματα μέσης τιμής και διασποράς των σχετικών σφαλμάτων για διαφορετικά διαστήματα της κρίσιμης θερμοκρασίας, από τα οποία παρατηρούμε ότι στο διάστημα 0-20K το νευρωνικό δίκτυο φαίνεται να αποδίδει καλύτερα ενώ στο υπόλοιπο διάστημα

καλύτερη απόδοση έχει η Gradient Boosting, χωρίς ωστόσο να έχει μεγάλη διαφορά από το νευρωνικό δίκτυο.



6. Σύνοψη Αποτελεσμάτων και Σύγκριση με Αντίστοιχη Έρευνα

Όπως είδαμε στις προηγούμενες σελίδες, το καλύτερο μοντέλο πρόβλεψης της κρίσιμης θερμοκρασίας ενός υπεραγωγού είναι εκείνο που κατασκευάστηκε με τη χρήση της μεθόδου Gradient Boosting. Το μοντέλο καταφέρνει να προβλέψει την κρίσιμη θερμοκρασία με σφάλμα μόλις ± 12.59 K ενώ παράλληλα διατηρεί τη γενικότητα του και ανταποκρίνεται με ίδια αποτελεσματικότητα και σε καινούρια δεδομένα. Επίσης, για να κάνει τις προβλέψεις του χρησιμοποιεί μόλις 9 μεταβλητές από τις 81 που είναι διαθέσιμες.

Στο σημείο αυτό αξίζει να σημειωθεί πως σε αντίστοιχη έρευνα που είχε πραγματοποιηθεί από τον Kam Hamidieh, στο τμήμα στατιστικής του πανεπιστημίου της Πενσυλβάνια, με τίτλο «A data-driven statistical model for predicting the critical temperature of a superconductor» [21], είχε κατασκευαστεί μοντέλο με μέθοδο Boosting (xgboost), το οποίο προέβλεπε με σφάλμα ± 9.5 K, ωστόσο όμως, χρησιμοποιούσε και τις 81 μεταβλητές για να κάνει τις προβλέψεις του με αποτέλεσμα να διαθέτει περισσότερη πληροφορία που εξηγεί τη μεγαλύτερη ακρίβεια.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Christopher Bishop, (2006), Pattern Recognition and Machine Learning

[2] Sergios Theodoridis, Konstantinos Koutroumbas, (1998), Pattern Recognition

[3] Jerome Friedman, Trevor Hastie, Robert Tibshirani, (2008), The Elements of Statistical Learning, Data Mining, Inference and Prediction

[4] Julian Avila, Trent Hauck, (2017), scikit-learn Cookbook, Second Edition, Packt

[5] Giuseppe Bonaccorso, (2017), Machine Learning Algorithms: A reference guide to popular algorithms for data science and machine learning, Packt

[6] Christopher Bourez, (2017), Deep Learning with Theano, Packt

[7] Jerome Friedman, Trevor Hastie, Robert Tibshirani, (2008), The Elements of Statistical Learning, Data Mining, Inference and Prediction

[8] Norman Matloff, (2017), Statistical Regression and Classification, From Linear Models to Machine Learning, CRC Press

[9] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer

[10] Xavier Glorot, Antoine Bordes, Yoshua Bengio, Deep Sparse Rectifier Neural Networks

[11] Robert E. Schapire, Yoav Freund, (2012), Boosting Foundations and Algorithms, The MIT Press

[12] Andreas Mayr, Harald Binder, Olaf Gefeller, Matthias Schmid, (2014), The Evolution of Boosting Algorithms, From Machine Learning to Statistical Modelling

[13] David Pardoe, Peter Stone, (2010), Boosting for Regression Transfer, The University of Texas at Austin

[14] Harris Drucker, Improving Regressors using Boosting Techniques, Monmouth University

[15] D.P. Solomatine, D.L. Shrestha, A Boosting Algorithm for Regression Problems

[16] Jerome H. Friedman, (1999), Stochastic Gradient Boosting

[17] Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, Huiwen Wang, (2010), Handbook of Partial Least Squares, Concepts, Methods and Applications, Springer

[18] Φουσκάκης Δημήτρης, (2013), Ανάλυση Δεδομένων με χρήση της R, Τσότρας

[19] Οικονόμου Π,Καρώνη Χ., (2010), Στατιστικά μοντέλα παλινδρόμησης, Συμεών

[20] Κοκολάκης Γιώργος, Φουσκάκης Δημήτρης, (2009), Στατιστική θεωρία και εφαρμογές, Συμεών

[21] Kam Hamidieh, (2018), A data-driven statistical model for predicting the critical temperature of a superconductor, Statistics Department,

University of Pennsylvania, https://doi.org/10.1016/j.commatsci.2018.07.052
[22] Michael Tinkham, (1996), Introduction to Superconductivity
[23] A.C. Rose Innes, E.H. Rhoderick, (1992), Introduction to Superconductivity, Pergamon
[24] V L Ginzburg, E A Andryushin, (2004), Superconductivity, World Scientific
[25] Αλικαρίδης Φιλάρετος, Ασκληπιακό Πάρκο Ιατρικής Σχολής Πανεπιστημίου Αθηνών, http://panacea.med.uoa.gr/topic.aspx?id=927
[26] https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08
[27] https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient
[28] https://elitedatascience
[29] https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html
[30] https://towardsdatascience.com/deep-study-of-a-not-very-deep-neural-network-part-2-activation-functions-fd9bd8d406fc
[31] https://medium.com/the-theory-of-everything/understanding-activation-functions-in-neural-networks-9491262884e0
[32] https://towardsdatascience.com/activation-functions-and-its-types-which-is-better-a9a5310cc8f
[33] https://www.superdatascience.com/blogs/the-ultimate-guide-to-artificial-neural-networks-ann
[34] http://googlegalaxyscience.com/radius-of-atom
[35] http://panagiotisathanasopoulos.gr/wp-content/uploads/2013/08/xal_math17_2013.pdf
[36] https://en.wikipedia.org/wiki/Line_search
[37] https://en.wikipedia.org/wiki/
[38] https://archive.ics.uci.edu/ml/index.php
[39] https://supercon.nims.go.jp/index_en.html