ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Τεχνικές εξόρυξης δεδομένων και Λογιστική Παλινδρόμηση στην αξιολόγηση πιστοληπτικής ικανότητας (credit scoring)

ΤΟΠΟΥΖΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

Επιβλέπων: Κουκουβίνος Χρήστος, Καθηγητής Ε.Μ.Π.

Αθήνα , Ιούλιος 2012

Περίληψη

Η ανάγκη αναζήτησης πληροφοριών από τεράστιες βάσεις δεδομένων , όπου οι μέχρι πρότινος κλασσικές μέθοδοι της στατιστικής δεν αποδεικνύονταν επαρκείς και ικανοποιητικές , οδήγησε στη διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Πρόκειται για μια σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς κλάδους όπως: η οικονομία , η βιοστατιστική, η δημογραφία και η μετεωρολογία. Στην παρούσα διπλωματική εργασία μελετήθηκαν οι εξής τεχνικές: τα Δέντρα Αποφάσεων, τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) καθώς και η Λογιστική Παλινδρόμηση.

Το πρώτο κεφάλαιο περιλαμβάνει μια σύντομη εισαγωγή στις βασικές έννοιες του data mining , η οποία εξηγεί τις δύο βασικές κατηγορίες του, καθώς επίσης και τους τομείς στους οποίους εφαρμόζεται. Στη συνέχεια γίνεται ανάλυση της KDD διαδικασίας και τέλος μια μικρή εισαγωγή και μαθηματική περιγραφή του προβλήματος ταξινόμησης.

Στο δεύτερο κεφάλαιο αναλύονται οι τεχνικές εξόρυξης γνώσης. Στα Δέντρα Αποφάσεων μελετήθηκε ο Αλγόριθμος C&RT και στα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ) αναλύθηκαν τα κύρια χαρακτηριστικά τους. Τέλος , παρουσιάζεται η μέθοδος της Λογιστικής Παλινδρόμησης και τρόποι με τους οποίους εκτιμούνται οι παράμετροι των μοντέλων.

Το τρίτο κεφάλαιο εξετάζει και παρουσιάζει τη χρήση των τεχνικών εξόρυξης δεδομένων για την κατασκευή μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας (credit scoring). Αρχικά δίνεται ένας επίσημος ορισμός της βαθμολόγησης πιστοληπτικής ικανότητας , στη συνέχεια περιγράφεται η χρησιμότητα και τα πλεονεκτήματα της, καθώς και ορισμένες από τις εφαρμογές της. Το μεγαλύτερο μέρος του κεφαλαίου περιλαμβάνει την εφαρμογή που πραγματοποιήθηκε πάνω σε πραγματικά οικονομικά δεδομένα με τη βοήθεια του προγράμματος Clementine SPSS, ενός λογισμικού εξόρυξης δεδομένων. Τα δεδομένα αυτά εφαρμόστηκαν στον Αλγόριθμο C&RT, στα νευρωνικά δίκτυα και στο μοντέλο της Λογιστικής Παλινδρόμησης με σκοπό την πρόβλεψη και τη σύγκριση των αποτελεσμάτων των μεθόδων στους παράγοντες της ακρίβειας (classification accuracy), της ευαισθησίας (sensitivity), της ειδικότητας (specificity), της θετικής προγνωστικής αξίας (positive predictive value) και της αρνητικής προγνωστικής αξίας (negative predictive value). Τέλος, στον επίλογο συνοψίζονται τα αποτελέσματα και εξάγονται κάποια συμπεράσματα που προκύπτουν από την ανάλυση των αποτελεσμάτων μας.

Abstract

The need for extracting information from large databases, where up until recently the classical statistical methods were proven to be insufficient and not satisfactory enough, led to the development of the process of Data Mining (Data Mining). Data Mining process constitutes of a series of techniques based on the developing of various algorithms. These techniques can be applied in many different fields such as economics, biostatistics, demography and meteorology. This thesis examines the following techniques: Decision Trees, Artificial Neural Networks (ANN) and binary logistic regression.

The first chapter entails a brief introduction to basic concepts of data mining elaborating on its two basic categories and examines the fields in which it can be applied. It then analyze the KDD process and finally it gives a brief introduction and description of the mathematical problem of classification.

The second chapter analyzes the data mining techniques. For the Decision trees the C & RT algorithm was studied and for the Artificial Neural Networks (ANN) there main characteristics were examined. Finally, the logistic regression method is presented and means by which the parameters of the models are estimated.

The third chapter illustrates and discusses the use of data mining techniques to build credit scoring models. Initially it is given a formal definition of credit scoring followed by a description of its efficacy and advantages along with some of its applications. The biggest part of the chapter is dedicated in describing the application made on real financial data with the help of the Clementine SPSS, a data mining software. These data were applied in the C & RT algorithm, the neural networks and at the model of logistic regression to predict and compare the results of the methods in the factors of classification accuracy, sensitivity, specificity, the positive predictive value and the negative predictive value. Closing the thesis, results are summarized and conclusions from the analysis of our results are drawn.

Ευχαριστίες

Η εκπόνηση της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε υπό την επίβλεψη του Καθηγητή του Ε.Μ.Π., κ. Χρήστου Κουκουβίνου, τον οποίο θα ήθελα να ευχαριστήσω θερμά για τη δυνατότητα που μου έδωσε να ασχοληθώ με ένα θέμα το οποίο ανήκει στα ερευνητικά μου ενδιαφέροντα.

Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στην υποψήφια διδάκτορα Χριστίνα Παρπούλα, για την πολύτιμη βοήθεια της και το συνεχές ενδιαφέρον κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας.

Θα ήθελα επίσης να εκφράσω την ευγνωμοσύνη μου στους γονείς μου για την διαρκή τους υποστήριξη , που επέτρεψε την επιτυχή διεκπεραίωση των σπουδών μου. Τέλος, αισθάνομαι την ανάγκη να ευχαριστήσω τους φίλους και συμφοιτητές μου, καθώς και την Άννα μου για την αμέριστη βοήθεια και τη συμπαράστασή τους.

Τοπούζης Κωνσταντίνος

Αθήνα , 2012

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1: DATA MINING 9

1.1 ΕΙΣΑΓΩΓΗ 9

1.2 ΔΙΑΔΙΚΑΣΙΑ KDD 11

1.3 ΕΦΑΡΜΟΓΕΣ ΤΟΥ DATA MINING 13

1.4 ΤΑΞΙΝΟΜΗΣΗ 15

1.4.1 Εισαγωγή ταξινόμησης 15

1.4.2 Μαθηματική περιγραφή του προβλήματος ταξινόμησης 15

ΚΕΦΑΛΑΙΟ 2: ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ 17

2.1 ΔΕΝΤΡΑ ΑΠΟΦΑΣΕΩΝ 17

2.1.1 Αλγόριθμος CART 20

2.1.1.1 Πεδία συχνότητας και πεδία βάρους 21

2.1.1.1.1 Πεδία συχνότητας 21

2.1.1.1.2 Πεδία βάρους 22

2.1.1.2 Κατασκευή ενός CART δέντρου 22

2.1.1.3 Διαχείριση κενών-ελλειπουσών τιμών 24

2.1.1.4 Μέτρα μή καθαρότητας 26

2.1.1.5 Κανόνες διακοπής – ολοκλήρωσης της διαδικασίας 29

2.1.1.6 Κέρδη-κόστη 30

2.1.1.7 Priors πιθανότητες 31

2.1.1.8 Διαδικασία κλαδέματος 32

2.2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ 35

2.2.1 Νευρώνας-Λειτουργία του βιολογικού νευρώνα 35

2.2.2 Το μαθηματικό μοντέλο 37

2.2.3 Συνάρτηση Ενεργοποίησης 39

2.2.4 Ταξινόμηση Νευρωνικών αλγορίθμων 42

2.2.5 Το δίκτυο Perceptron 44

2.2.5.1 Αλγόριθμος μάθησης 46

2.2.5.2 Παράδειγμα του αλγόριθμου μάθησης στον απλό Perceptron 48

2.2.6 Perceptron πολλών στρωμάτων 51

2.3 ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ 55

2.3.1 Εκτίμηση παραμέτρων με τη μέθοδο μέγιστης πιθανοφάνειας (maximum likelihood) .. 58

2.3.2 Άλλες μορφές στατιστικής συμπερασματολογίας για τις οποίες γίνεται χρήση λογιστικής παλινδρόμησης 61

2.3.3 Ιδιότητες της διασποράς των εκτιμητών μέγιστης πιθανοφάνειας στη λογιστική παλινδρόμηση. 63

2.3.4 Συμπερασματολογία με χρήση της μεθόδου Wald στη λογιστική παλινδρόμηση 65

2.3.5 Συμπερασματολογία με χρήση πιθανοφάνειας στη λογιστική παλινδρόμηση 68

2.3.6 Παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης 69

ΚΕΦΑΛΑΙΟ 3: ΕΦΑΡΜΟΓΗ ΣΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ 72

3.1 ΟΡΙΣΜΟΣ ΤΟΥ CREDIT SCORING 72

3.1.1 Ιστορική Αναδρομή 72

3.2 ΠΡΟΒΛΗΜΑ ΠΟΥ ΚΑΛΕΙΤΑΙ ΝΑ ΛΥΣΕΙ ΤΟ CREDIT SCORING 73

3.2.1 Τεχνικές που χρησιμποιούνται 74
3.3 ΣΥΜΠΕΡΑΣΜΑΤΙΚΑ 74
3.4 ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΦΑΡΜΟΓΗΣ 75
3.4.1 Εισαγωγή στο πρόβλημα 76
3.4.1.1 C&RT 85
3.4.1.2 Νευρωνικά Δίκτυα (χρήση της μεθόδου MLP) 91
3.4.1.3 Λογιστική Παλινδρόμηση (stepwise , forwards , backwards) 94
3.5 ΣΥΝΟΨΗ 102
3.6 ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ 102

ΚΕΦΑΛΑΙΟ 1: DATA MINING

1.1 Εισαγωγή

Η συνεχής ανάπτυξη στον τομέα της πληροφορικής σε συνδυασμό με τα σύγχρονα εργαλεία αυτοματοποιημένης συλλογής δεδομένων έχουν οδηγήσει στην δημιουργία μιας κοινωνίας ικανής να παρέχει διαρκώς νέες πληροφορίες. Το υλικό που συγκεντρώνεται καταγράφεται διαρκώς , με αποτέλεσμα τη δημιουργία τεράστιων βάσεων δεδομένων. Το ζήτημα λοιπόν που προκύπτει είναι εαν μπορούμε να διαχειριστούμε αυτές τις βάσεις δεδομένων.

Η ανάγκη εξόρυξης πληροφοριών από αυτό τον τεράστιο όγκο δεδομένων , όπου οι μέχρι πρότινος κλασσικές μέθοδοι της στατιστικής δεν αποδεικνύονταν επαρκείς και ικανοποιητικές οδήγησε στη διαδικασία της Εξόρυξης Δεδομένων (Data Mining). Πρόκειται για μια σειρά από τεχνικές που βασίζονται σε ανάπτυξη αλγορίθμων και είναι χρήσιμες σε πολλούς κλάδους όπως οι: η οικονομία , η βιοστατιστική, η δημογραφία και η μετεωρολογία.

Η λέξη “Data” είναι μια λατινική λέξη που σημαίνει <<τα πράγματα που έχουν δοθεί>>. Στην πληροφορική αναφέρεται ως μια συλλογή από αριθμούς ή σύμβολα σε τέτοια μορφή που είναι εύκολα επεξεργάσιμη από ηλεκτρονικούς υπολογιστές. Οι συλλογές αυτές από δεδομένα δεν έχουν καμία αξία αν δεν μπορούν να μετασχηματιστούν σε γνώση.

Γενικά υπάρχουν αντικρουόμενες απόψεις γύρω από το ποιος θα μπορούσε να είναι ένας σαφής και περιεκτικός ορισμός για την Εξόρυξη Δεδομένων (ΕΔ). Ωστόσο , αποδεχόμαστε σαν ορισμό του data mining τον εξής:

«Εξόρυξη Δεδομένων είναι η ανάλυση , συνήθως τεράστιων , παρατηρούμενων συνόλων δεδομένων , έτσι ώστε να βρεθούν μη παρατηρηθείσες σχέσεις και να συνοψιστούν τα δεδομένα με καινοφανείς τρόπους οι οποίοι να είναι κατανοητοί και χρήσιμοι στον κάτοχο δεδομένων».

Ο κύριος στόχος του data mining είναι η εξαγωγή νέων πληροφοριών από τα δεδομένα. Η ανακάλυψη της γνώσης γίνεται με τεχνικές οι οποίες διακρίνονται σε δυο βασικές κατηγορίες:

- **μέθοδοι με επίβλεψη (supervised methods)**. Αλγόριθμοι εκμάθησης με επίβλεψη είναι εκείνοι που χρησιμοποιούνται στην ταξινόμηση και στην πρόβλεψη. Ουσιαστικά μοντελοποιούν μια μεταβλητή απόκρισης βασιζόμενοι σε μια ή περισσότερες επεξηγηματικές μεταβλητές (input variable). Μερικές από αυτές τις supervised τεχνικές είναι και τα νευρωνικά δίκτυα (neural networks) , δέντρα αποφάσεων (decision trees) , λογιστική παλινδρόμηση (logistic regression) με τις οποίες θα ασχοληθούμε εκτεταμένα στη συνέχεια της παρούσας εργασίας.

- **μέθοδοι χωρίς επίβλεψη (unsupervised methods)**. Αλγόριθμοι εκμάθησης χωρίς επίβλεψη είναι εκείνοι που χρησιμοποιούνται όταν δεν υπάρχει μια μεταβλητή απόκρισης να προβλεφθεί ή να ταξινομηθεί. Ουσιαστικά, οι unsupervised τεχνικές χρησιμοποιούνται όταν δεν υπάρχει κάποιο πεδίο να προβλεφθεί αλλά οι σχέσεις των δεδομένων εξερευνούνται ώστε να ανακαλυφθεί η γενική δομή τους. Μερικές από τις τεχνικές αυτές είναι οι kohonen networks, two step , k-means.

Τα έξι βασικά αποτελέσματα που αναμένεται να λάβουμε ανάλογα με τους στόχους που έχουμε θέσει (tasks) είναι:

- Ταξινόμηση (classification): εξέταση των χαρακτηριστικών ενός νέου αντικειμένου και η ταξινόμησή του σε ήδη προκαθορισμένες κλάσεις.
- Εκτίμηση (estimation): εύρεση τιμών για μια άγνωστη μεταβλητή , με δεδομένα κάποια δεδομένα εισόδου.
- Πρόβλεψη (prediction): παρόμοια με την ταξινόμηση και την εκτίμηση αλλά οι εγγραφές ταξινομούνται με βάση κάποιες προβλεπόμενες μελλοντικές τάσεις ή εκτιμώμενες μελλοντικές τιμές.

- Ομαδοποίηση (grouping): καθορισμός των αντικειμένων που ανήκουν σε συγκεκριμένη ομάδα.
- Συσταδοποίηση (clustering): κατάτμηση ενός πληθυσμού σε ένα αριθμό υποομάδων ή συστάδων.
- Περιγραφή και οπτικοποίηση (description and visualization): διερευνητικό ή οπτικό data mining.

1.2 Διαδικασία KDD

Επεξεργαζόμενοι μια τεράστια βάση δεδομένων είναι πιθανό να ανακαλύψουμε την ύπαρξη «κρυμμένης γνώσης». Δηλαδή, μπορεί να εντοπίσουμε συσχετίσεις , αλληλεξαρτήσεις ή ομαδοποιήσεις μεταξύ των δεδομένων , πράγματα τα οποία να μην είναι άμεσα εμφανή. Το είδος αυτής της «γνώσης» θεωρείται ότι δεν είναι εκ των προτέρων διαθέσιμο αλλά μπορεί να αποδειχθεί πολύ χρήσιμο.

Την ανάγκη αυτή ανάκτησης γνώσης έρχεται να καλύψει η ΕΔ , η οποία αποτελεί τον πυρήνα της γενικότερης μεθοδολογίας της ανακάλυψης της γνώσης από βάσεις δεδομένων (Knowledge Discovery in Databases - KDD).

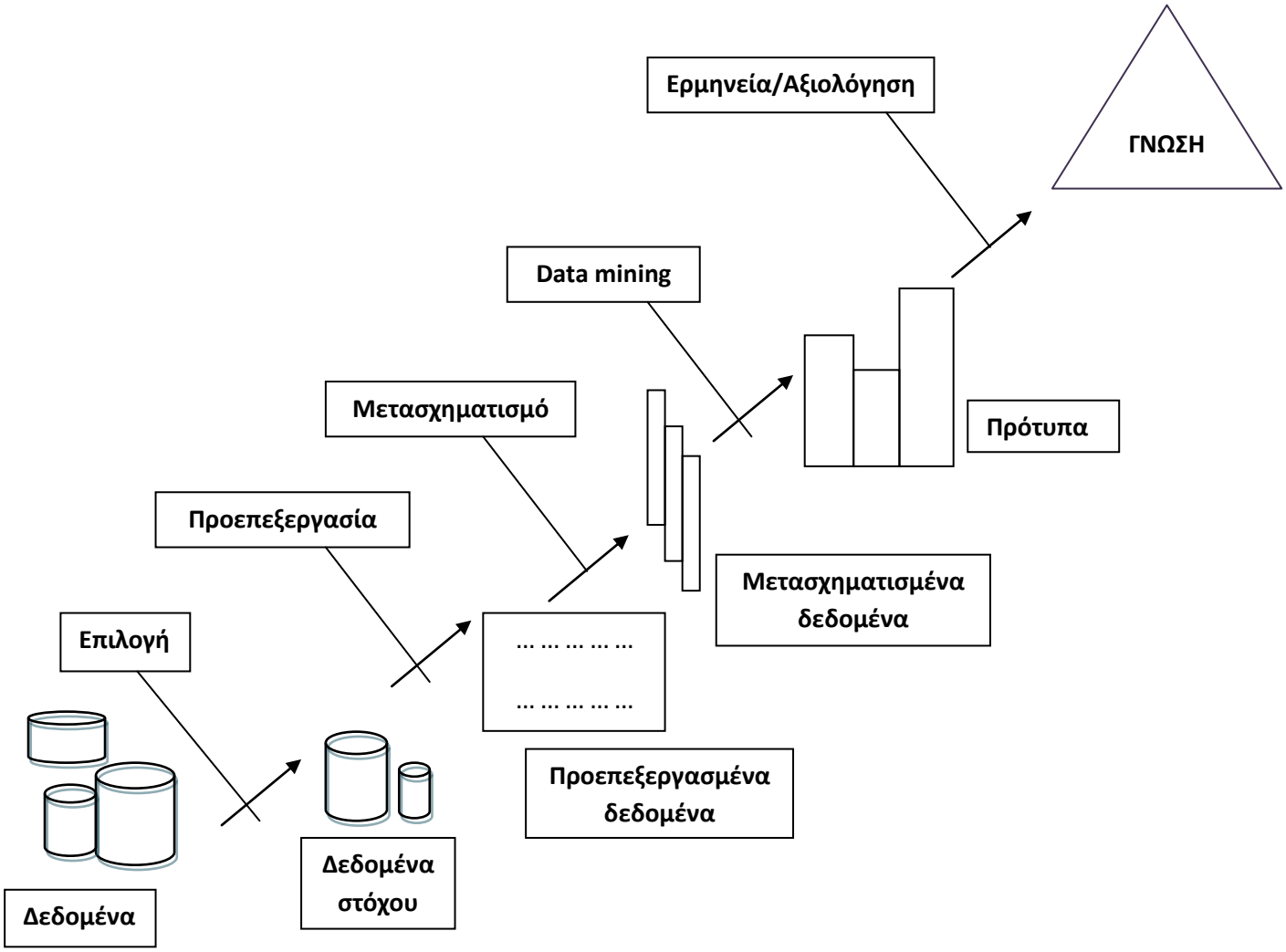
Η KDD είναι μια αυτοματοποιημένη διαδικασία ανάλυσης και μοντελοποίησης τεράστιων αποθηκών δεδομένων. Πρόκειται για μια συγκροτημένη μεθοδολογία αναγνώρισης έγκυρων και πρωτότυπων προτύπων μέσα από πολύ μεγάλους και περίπλοκους πίνακες δεδομένων , με στόχο τα πρότυπα που θα προκύψουν να είναι χρήσιμα και κατανοητά.

Τα βασικά βήματα της KDD διαδικασίας είναι τα ακόλουθα:

- **Ανάπτυξη και κατανόηση του πεδίου της εφαρμογής** συμπεριλαμβανόμενης οποιασδήποτε σχετικής προηγούμενης γνώσης για το πρόβλημα καθώς επίσης και των στόχων / προσδοκιών των τελικών χρηστών.
- **Δημιουργία του στοχευόμενου συνόλου δεδομένων (target data)** , το οποίο θα περιλαμβάνει τα δεδομένα από τα οποία πρόκειται να εξαχθεί η γνώση.

Το βήμα αυτό είναι εξαιρετικά κρίσιμο καθώς η ποιότητα των δεδομένων επηρεάζει την απόδοση του συστήματος αποκάλυψης γνώσης.

- **Καθαρισμός και επεξεργασία δεδομένων (data cleaning).** Το βήμα αυτό περιλαμβάνει βασικές λειτουγίες όπως η απομάκρυνση του θορύβου , η αντιμετώπιση του προβλήματος των δεδομένων με ελλιπείς τιμές κ.α
- **Μείωση της ποσότητας των δεδομένων (data reduction).** Το βήμα αυτό περολαμβάνει την εύρεση χρήσιμων χαρακτηριστικών για την αναπαράσταση των δεδομένων του προβλήματος ανάλογα με τους στόχους της ανακάλυψης γνώσης , τη μείωση του πλήθους αυτών των χαρακτηριστικών κ.α.
- **Επιλογή των εργασιών εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για τις ανάγκες του προβλήματος π.χ ταξινόμηση , πρόβλεψη , ομαδοποίηση κ.α
- **Επιλογή των αλγορίθμων εξόρυξης γνώσης (data mining)** που θα χρησιμοποιηθούν για την αναζήτηση προτύπων στα δεδομένα. Το βήμα αυτό περιλαμβάνει την επιλογή του κατάλληλου μοντέλου , την επιλογή των κατάλληλων παραμέτρων του μοντέλου κ.α
- **Data Mining**: αναζήτηση στα δεδομένα των προτύπων που μας ενδιαφέρουν.
- **Ερμηνεία των προτύπων** που ανακαλύφθηκαν από την KDD διαδικασία – πιθανόν να χρειαστεί να επιστρέψουμε και πάλι σε κάποια από τα παραπάνω βήματα.
- **Ενοποίηση της γνώσης που έχει εξαχθεί**: Σε αυτό το βήμα , η εξορυγμένη γνώση ενσωματώνεται στο σύστημα και χρησιμοποιούνται κάποιες τεχνικές αντιπροσώπευσης αυτής προκειμένου να παρουσιαστεί ευκρινώς στο χρήστη.



Σχήμα: Διαδικασία του Data mining

1.3 Εφαρμογές του Data mining

Το data mining χρησιμοποιείται σε μια πληθώρα πεδίων και εφαρμογών καθώς βοηθάει στη λήψη ολοκληρωμένων αποφάσεων. Για να επιτευχθεί αυτό όμως , πρέπει αρχικά τα δεδομένα να συγκεντρωθούν και να οργανωθούν με ένα συνεπή και χρήσιμο τρόπο (data warehousing). Οι data warehouses πρέπει να έχουν ακριβή ιστορικά δεδομένα , μια και η διαδικασία των data mining , γεννά μοντέλα από ιστορικά δεδομένα που χρησιμοποιούνται για προβλέψεις , ανίχνευση τάσεων κ.α.

Το data mining χρησιμοποιείται σήμερα σε πολλούς επιστημονικούς κλάδους όπως: η **ιατρική** , η **οικονομία** , οι **τηλεπικοινωνίες** , το **marketing**. Συνοπτικά παραθέτουμε μερικές εφαρμογές του data mining σε συνδυασμό με παραδείγματα:

1) Ανάλυση εταιριών και διαχείριση ρίσκου:
    i. Προβλέψεις
    ii. Διατήρηση πελατολογίου
    iii. Βελτιωμένη χρηματοδότηση

Π.χ 1) Κατασκευή δένδρων αποφάσεων από ιστορικά στοιχεία τραπεζικών δανείων για την παραγωγή αλγόριθμων , ώστε να αποφασίζεται αν πρέπει η όχι να δοθεί ένα δάνειο σε έναν υποψήφιο πελάτη.

Π.χ 2) Ιατρικά εργαστήρια θέλουν να συσχετίσουν ασθένειες με χαρακτηριστικά των ασθενών, όπως τόπος διαμονής, διατροφικές συνήθειες, παλαιότερες ασθένειες, κ.α., .ώστε να καταφέρουν να βγάλουν κάποια ιατρικά συμπεράσματα και καινούρια γνώση, με τη βοήθεια των συγκεκριμένων χαρακτηριστικών.

2) Ανάλυση αγοράς και διαχείριση:
    i. Target marketing
    ii. Customer relation Management
    iii. Market basket analysis (supermarket)
    iv. Cross selling

Π.χ 1) τράπεζες

Έλεγχος ποιότητας και Ανάλυση ανταγωνισιμότητας

Π.χ 2) Η περίπτωση «Diapers and beer». Η παρατήρηση ότι πελάτες που αγοράζουν πάνες αγοράζουν και μπύρα επιτρέπουν στα καταστήματα να τοποθετούν αυτά τα είδη σχετικά κοντά , γνωρίζοντας ότι οι πελάτες θα κάνουν τη διαδρομή μεταξύ των ραφιών με τις πάνες και αυτών με τις μπύρες. Τοποθετώντας ανάμεσά τους και πατατάκια αυξάνουν τις πωλήσεις και στα τρία είδη.

Π.χ 3) Εταιρία πωλήσεως ηλεκτρονικών συσκευών θέλει να μελετήσει τις αγοραστικές συνήθειες των πελατών της, ώστε να προγραμματίσει ανάλογα την επόμενη διαφημιστική καμπάνια.

3) Εντοπισμός απάτης και διαχείριση ρίσκου:

Άλλες εφαρμογές που χρησιμοποιούν Εξόρυξη Δεδομένων:

i. Εξόρυξη κειμένου (newsgroup, Email, documents) και Web analysis
ii. Ευφυείς απαντήσεις σε ερωτήματα

Π.χ 1) Άτομα που σκηνοθετούν ατυχήματα για να εισπράξουν από τις ασφαλιστικές εταιρίες, ή κάποιοι που κάνουν ξέπλυμα «βρώμικου χρήματος» εντοπίζοντας ύποπτες μεταφορές χρημάτων ή κάποιοι που κλέβουν τους πάροχους τηλεπικοινωνιών και κάνουν τηλεφωνήματα που έχουν κάποια επαναλαμβανόμενα σχέδια είτε προς μια κλειστή ομάδα ατόμων (κινητά) είτε κάποια συγκεκριμένη ώρα της ημέρας κλπ.
Π.χ 2) Εντοπισμός ακατάλληλων ιατρικών μεθόδων και θεραπειών.

1.4 Ταξινόμηση

1.4.1 Εισαγωγή ταξινόμησης

Η ταξινόμηση αποτελεί μία από τις βασικές τεχνικές εξόρυξης δεδομένων. Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (μη κατηγοριοποιημένου) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων. Η διαδικασία της κατηγοριοποίησης χαρακτηρίζεται από ένα σαφή καθορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προκαθορισμένα παραδείγματα. Η ταξινόμηση δεδομένων είναι μια διαδικασία η οποία βρίσκει τις κοινές ιδιότητες μεταξύ ενός συνόλου αντικειμένων σε μια βάση δεδομένων και ταξινομεί τα αντικείμενα αυτά σε διαφορετικές κλάσεις(τάξεις) σύμφωνα με ένα μοντέλο ταξινόμησης.

1.4.2 Μαθηματική περιγραφή του προβλήματος ταξινόμησης

Έστω μια βάση δεδομένων $\mathrm{D} = \{\mathrm{t}_1, \mathrm{t}_2, \dots, \mathrm{t}_\mathrm{n}\}$, όπου t_i είναι πλειάδες της μορφής $< \mathrm{t}_{\mathrm{i}1}, \mathrm{t}_{\mathrm{i}2}, \dots, \mathrm{t}_{\mathrm{ip}} >$ (που καλούνται στοιχεία ή εγγραφές ή παραδείγματα) , και ένα

σύνολο κλάσεων $\mathrm{C} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{m}\}$. Το πρόβλημα της κατηγοριοποίησης συνίσταται στον προσδιορισμό της απεικόνισης

$$f: D \to C$$

όπου κάθε t_i αντιστοιχεί σε μια κλάση C_j. Η απεικόνιση αυτή ονομάζεται και *μοντέλο*.

Έτσι μια κλάση C_j ορίζεται ως το σύνολο των παραδειγμάτων που κατατάσσονται σ' αυτήν:

$$C_j = \{ \frac{t_i}{f(t_i)} = C_j, 1 \leq i \leq n, t_i \in D \}$$

Όπου κάθε παράδειγμα t_i θεωρείται ως ένα διάνυσμα . Τα t_{ik} , $k = 1$, p είναι τιμές (διακριτές ή αριθμητικές) , που αναφέρονται σε αντίστοιχα φυσικά χαρακτηριστικά (features) $X_1, X_2, \ldots, X_p$. Γι ' αυτό και ένα τέτοιο διάνυσμα ονομάζεται διάνυσμα χαρακτηριστικών (feature vector). Κάθε χαρακτηριστικό X_k μπορεί να πάρει κάποιες τιμές $D_{xk} = \{ x_{ki}, i = 1, r \}$. Επομένως σ' ένα παράδειγμα κάθε t_{ik} είναι μια από τις x_{ki} , δηλ. $t_{ik} \in D_{xk}$.

Οι κλάσεις αναφέρονται και αυτές σ' ένα χαρακτηριστικό X_f , που ονομάζεται χαρακτηριστικό στόχου (target feature). Πιο συγκεκριμένα , οι κλάσεις αντιστοιχούν στις διαφορετικές τιμές που μπορεί να πάρει το χαρακτηριστικό στόχου.

ΚΕΦΑΛΑΙΟ 2: ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Για την επιτυχή διεκπεραίωση των διαφόρων εργασιών data mining έχουν αναπτυχθεί πολλές τεχνικές. Κάποιες από τις πιο σημαντικές τεχνικές είναι οι ακόλουθες:

- Τα δέντρα απόφασης (decision trees)
- Τα νευρωνικά δίκτυα (neural networks)
- Λογιστική παλινδρόμηση (logistic regression)

Οι παραπάνω τεχνικές διαφέρουν ως προς την ακρίβεια και τη δυνατότητα κατανόησής τους. Στη συνέχεια αναλύουμε κάθε επιμέρους τεχνική.

2.1 Δέντρα Αποφάσεων

Τα δέντρα απόφασης είναι πολύ ισχυρά εργαλεία που χρησιμοποιούνται ευρέως για τις περιπτώσεις της ταξινόμησης και της πρόβλεψης. Ένα δέντρο απόφασης αντιπροσωπεύει μια σειρά από IF THEN κανόνες ξεκινώντας από τη ρίζα του δέντρου και καταλήγοντας στα φύλλα του.

Οι εσωτερικοί κόμβοι ενός δέντρου απόφασης περιέχουν τα γνωρίσματα του προβλήματος, οι ακμές περιέχουν τις δυνατές τιμές των γνωρισμάτων και τα φύλλα περιέχουν τις πιθανές κλάσεις του προβλήματος. Απαραίτητο για την κατασκευή ενός δέντρου απόφασης είναι ένα σύνολο από στιγμιότυπα εκπαίδευσης , τα οποία περιγράφονται από κάποια γνωρίσματα και την κλάση του προβλήματος στην οποία ανήκουν.

Η διαδικασία που ακολουθούν οι αλγόριθμοι κατασκευής ενός δέντρου απόφασης συνοψίζεται στα ακόλουθα: Ξεκινώντας από τη ρίζα του δέντρου ο αλγόριθμος διασπά το σύνολο των στιγμιότυπων εκπαίδευσης σε υποσύνολα με βάση τη βέλτιστη ιδιότητα (best attribute) του κόμβου – η βέλτιστη ιδιότητα ενός κόμβου καθορίζεται από κάποιο κριτήριο όπως το information gain , το gain ratio , δείκτη Gini (Index Gini).Επομένως , μπορούμε να πούμε συμπερασματικά ότι ως ρίζα

επιλέγουμε εκείνο το χαρακτηριστικό που δίνει το μέγιστο κέρδος πληροφορίας και για να το ποσοτικοποιήσουμε θα χρησιμοποιήσουμε την έννοια της εντροπίας. Έτσι προκύπτει ένα πλήθος υποσυνόλων που το καθένα περιέχει λιγότερα παραδείγματα από το αρχικό σύνολο. Για καθένα από αυτά τα επιμέρους υποσύνολα εφαρμόζεται επαναληπτικά η παραπάνω διαδικασία χρησιμοποιώντας τα εναπομείναντα γνωρίσματα, οπότε η διάσπαση των στιγμιότυπων προχωρά και σταματά όταν όλα τα στιγμιότυπα του υποσυνόλου ανήκουν στην ίδια κλάση ή έχουν εξαντληθεί όλα τα γνωρίσματα. Στην ουσία πρόκειται για εφαρμογή της μεθόδου «Διαίρει και βασίλευε».

Εκτός από το σύνολο των στιγμιότυπων εκπαίδευσης υπάρχει και το σύνολο ελέγχου με βάση τα οποία ελέγχεται η απόδοση του δέντρου , δηλαδή η ακρίβεια με την οποία το κατασκευασμένο δέντρο απαντά στο πρόβλημα της ταξινόμησης. Στην περίπτωση αυτή δίνουμε ως είσοδο στο δέντρο τις τιμές των γνωρισμάτων του στιγμιότυπου ελέγχου και περιμένουμε ως απάντηση την τάξη του στιγμιότυπου. Το πλήθος των λανθασμένων απαντήσεων (δηλαδή τα στιγμιότυπα στα οποία το δέντρο απάντησε διαφορετική κλάση από την πραγματική) καθορίζει την ακρίβεια του δέντρου.

Τα δέντρα απόφασης χρησιμοποιούνται ευρέως τόσο από την επιστημονική κοινότητα όσο και από τη βιομηχανία και αρκετοί αλγόριθμοι έχουν αναπτυχθεί για το σκοπό αυτό. Οι γνωστότεροι αλγόριθμοι εκπαίδευσης (ID3 , C4.5 , C&RT) χρησιμοποιούν μια top-down , εξαντλητική αναζήτηση στο χώρο των πιθανών δέντρων απόφασης. Αρχίζουν με ένα κενό δέντρο και προοδευτικά θέτουν πιο περίπλοκες προθέσεις με στόχο την εύρεση ενός δέντρου που ταξινομεί σωστά τα δεδομένα εκπαίδευσης.

Έτσι η διαδικασία κατασκευής δέντρου απόφασης είναι η εξής:

- Επιλογή χαρακτηριστικού για τη θέση του αρχικού κόμβου (ρίζας) και δημιουργία κλάδων για κάθε πιθανή τιμή του χαρακτηριστικού.
- Διάσπαση υποδειγμάτων σε υποσύνολα , ένα για κάθε κλάδο που εκτείνεται από τη ρίζα.

- Επανάληψη των παραπάνω για κάθε κλάδο με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου.
- Ολοκλήρωση της διαδικασίας όταν όλα τα υποδείγματα σε ένα κόμβο ανήκουν στην ίδια τάξη.

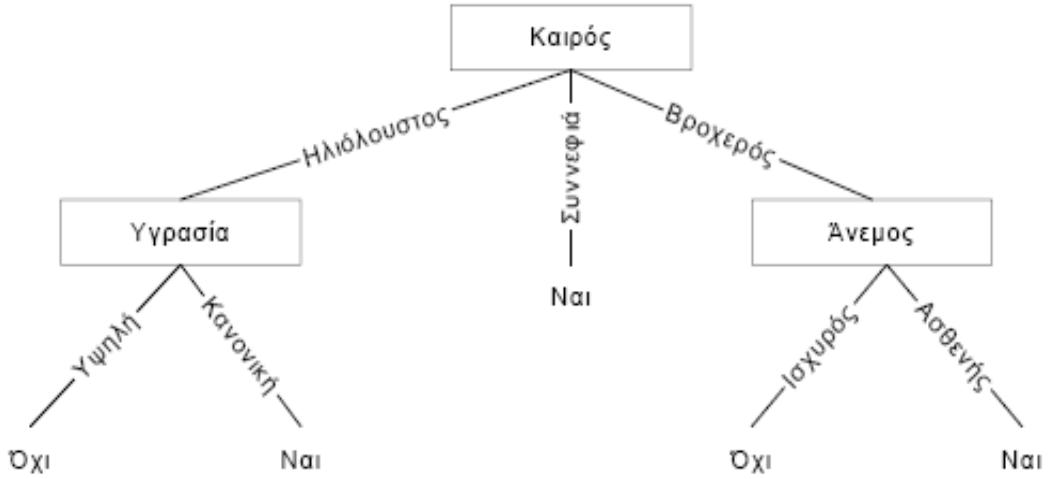
Όταν έχει ολοκληρωθεί η διαδικασία ανακάλυψης γνώσης με χρήση του αλγορίθμου , τότε το δέντρο μπορεί να αναπαρασταθεί ως σύνολο κανόνων της μορφής:

« If <ΣΥΝΟΛΟ ΣΥΝΘΗΚΩΝ> then <ΣΥΜΠΕΡΑΣΜΑ>» .

Η ανακάλυψη γνώσης με χρήση αλγορίθμων δέντρων απόφασης αποτελεί μια από τις πλέον δημοφιλείς τεχνικές επαγωγικής εκμάθησης και έχει μεγάλη εφαρμογή στη διάγνωση ιατρικών περιπτώσεων , στην εκτίμηση πιθανού ρίσκου από πιστοληπτικές τραπεζικές εργασίες κ.α

Έστω για παράδειγμα το κλασικό πρόβλημα που προσπαθεί να απαντήσει στο ερώτημα «Παίζεις τέννις» και το οποίο έχει δυο κλάσεις: «Ναι» και «Όχι» (Σχήμα 2.1). Η απάντηση στο πρόβλημα εξαρτάται από τους εξής παράγοντες: τον Καιρό (με πιθανές τιμές: ήλιος , βροχή , συννεφιά) , την Υγρασία (με πιθανές τιμές: υψηλή , κανονική) και τον Αέρα (με πιθανές τιμές: δυνατός , αδύνατος).

Στο Σχήμα 2.2 φαίνεται το δέντρο απόφασης του προβλήματος. Περιέχει τρεις εσωτερικούς κόμβους , σε κάθε κόμβο γίνεται έλεγχος ως προς κάποιο από τα γνωρίσματα του προβλήματος , ενώ στα φύλλα του περιέχονται οι κλάσεις του προβλήματος.



<u>Σχήμα</u>: Δέντρο απόφασης για το πρόβλημα «παίζεις τέννις»

2.1.1 Αλγόριθμος CART

Οι CART αλγόριθμοι βασίζονται στη θεωρία των δέντρων ταξινόμησης και παλινδρόμησης που διατυπώθηκε από τον Breiman et al (1984). Σ' ένα αλγόριθμο CART τα δεδομένα χωρίζονται σε δύο υποσύνολα , ξεκινώντας βέβαια με τον αρχικό κόμβο-ρίζα ο οποίος περιέχει ολόκληρο το δείγμα εκπαίδευσης , με τρόπο ώστε κάθε υποσύνολο να εξασφαλίζει περισσότερη ομοιογένεια απ'ό,τι το προηγούμενο. Η διαδικασία αυτή επαναλαμβάνεται έως ότου να επιτευχθεί το κριτήριο ομοιογένειας ή κάποιο άλλο κριτήριο διακοπής. Το πεδίο πρόβλεψης μπορεί να χρησιμοποιηθεί αρκετές φορές σε διαφορετικά επίπεδα στο δέντρο.

Πλεονεκτήματα αλγορίθμου CART:

- Είναι αρκετά ευέλικτος
- Καλύτερη διαχείριση δεδομένων με ελλείπουσες τιμές χρησιμοποιώντας υποκατάστατο διαχωρισμό
- Δίνει στο χρήστη τη δυνατότητα να καθορίσει την προηγούμενη (prior) κατανομή πιθανότητας σε ένα πρόβλημα ταξινόμησης.
- Ο χρήστης μπορεί να εφαρμόσει ένα αυτόματο κλάδεμα , που θα παρουσιάζει πολυπλοκότητα ως προς τα κόστη με στόχο να αποκομίσει ένα πιο γενικευμένο δέντρο.

2.1.1.1 Πεδία συχνότητας και πεδία βάρους

Για την κατασκευή του μοντέλου είναι απαραίτητο να γίνουν κάποιοι υπολογισμοί. Για παράδειγμα , για τη μείωση του μεγέθους του συνόλου δεδομένων χρειάζεται υπολογισμός των:

- Πεδίων συχνότητας
- Πεδίων βάρους

Είναι πολύ σημαντικό να γίνει ο σωστός διαχωρισμός μεταξύ των πεδίων βάρους και συχνότητας γιατί διαφορετικά θα προκύψουν λανθασμένα αποτελέσματα. Στην περίπτωση όπου τα πεδία συχνότητας ή βάρους δεν ορίζονται τότε η συχνότητα και τα βάρη για όλες τις καταχωρήσεις παίρνουν τις τιμές 1,0.

2.1.1.1.1 Πεδία συχνότητας

Ένα πεδίο συχνότητας αναπαριστά το συνολικό αριθμό των παρατηρήσεων που αντιπροσωπεύονται από κάθε καταχώρηση. Στην ανάλυση των συνολικών δεδομένων είναι σημαντικό να γνωρίζουμε σε ποιο πεδίο μια καταχώρηση (συνδυασμός περιπτώσεων) αναπαριστά περισσότερες από μια παρατηρήσεις. Ο συνολικός αριθμός των παρατηρήσεων μέσα στο δείγμα πρέπει πάντοτε να είναι ίσος με το άθροισμα των τιμών στο πεδίο συχνότητας. Το αποτέλεσμα που προκύπτει αν χρησιμοποιήσουμε πεδίο συχνότητας είναι ίδιο με αυτό που λαμβάνουμε χρησιμοποιώντας δεδομένα case by case.

Στον παρακάτω πίνακα παρατηρούμε ένα υποθετικό παράδειγμα , με τα πεδία πρόβλεψης «φύλλο» , «απασχόληση» και το πεδίο στόχος -κάπνισμα- «απόκριση» . Το πεδίο συχνότητας μας λέει , για παράδειγμα ότι 11 εργαζόμενοι άντρες ανταποκρίθηκαν με «ναι» στην ερώτηση αν καπνίζουν και 18 άνεργες γυναίκες ανταποκρίθηκαν με «όχι»».

ΦΥΛΛΟ	ΕΡΓΑΣΙΑ	ΑΠΟΚΡΙΣΗ	ΣΥΧΝΟΤΗΤΑ
Άντρας	ΝΑΙ	ΝΑΙ	11
Άντρας	ΝΑΙ	ΟΧΙ	17
Άντρας	ΟΧΙ	ΝΑΙ	12
Άντρας	ΟΧΙ	ΟΧΙ	21
Γυναίκα	ΝΑΙ	ΝΑΙ	11
Γυναίκα	ΝΑΙ	ΟΧΙ	15
Γυναίκα	ΟΧΙ	ΝΑΙ	15
Γυναίκα	ΟΧΙ	ΟΧΙ	18

Πίνακας: Πίνακας με Πεδίο Συχνότητας

Στο συγκεκριμένο παράδειγμα χρησιμοποιώντας το πεδίο συχνότητας επεξεργαζόμαστε ένα πίνακα 8 καταχωρήσεων ενώ, εάν χρησιμοποιούσαμε δεδομένα case by case θα ήταν απαραίτητες 120 καταχωρήσεις.

2.1.1.1.2 Πεδία βάρους

Κάνοντας χρήση ενός πεδίου βάρους οδηγούμαστε σε μια άνιση μεταχείριση στις καταχωρήσεις, σε ολόκληρο το σύνολο δεδομένων. Έτσι, η συνεισφορά μιας καταχώρησης στην ανάλυση είναι σταθμισμένη (weighted) σε αναλογία με τον πληθυσμό των μονάδων που η καταχώρηση αναπαριστά μέσα στο δείγμα. Για παράδειγμα , στην ερώτηση μιας έρευνας σε δείγμα 100.000 καταναλωτών , για λογαριασμό μιας επώνυμης βιομηχανίας παραγωγής καπνού , κατά πόσο καπνίζουν ή όχι , 20.000 ερωτηθέντες απάντησαν θετικά και 80.000 αρνητικά. Σε μια προσπάθεια να μειώσουμε το μέγεθος των δεδομένων, πιθανότατα θα συμπεριλάβουμε όλους όσους είναι καπνιστές και μόνο 25% του δείγματος (20.000) που δεν είναι καπνιστές. Κάτι τέτοιο μπορούμε να το κάνουμε αν ορίσουμε μια περίπτωση βάρους ίση με 1 γι' αυτούς που καπνίζουν και 4 γι' αυτούς που δεν καπνίζουν.

2.1.1.2 Κατασκευή ενός CART δέντρου

Η βασική ιδέα κατασκευής ενός δέντρου είναι να επιλέξουμε έναν διαχωρισμό (split) μεταξύ όλων των πιθανών διαχωρισμών σε κάθε κόμβο έτσι ώστε οι

θυγατρικοί κόμβοι που θα προκύψουν ως αποτέλεσμα να είναι οι καθαρότεροι. Με τον όρο καθαρότητα αναφερόμαστε στην ομοιότητα των τιμών του πεδίου στόχος. Σ' ένα εντελώς καθαρό κόμβο , όλες οι καταχωρήσεις έχουν την ίδια τιμή στο πεδίο στόχος. Ο CART αλγόριθμος μετρά την καθαρότητα ενός διαχωρισμού σε ένα κόμβο ορίζοντας ένα μέτρο καθαρότητας.

Τα βήματα που χρησιμοποιούνται για την κατασκευή ενός CART δέντρου είναι τα ακόλουθα (ξεκινώντας βέβαια με τον αρχικό κόμβο-ρίζα ο οποίος περιέχει όλες τις καταχωρήσεις):

1) Για κάθε πεδίο πρόβλεψης (predictor field) , βρίσκουμε τον καλύτερο δυνατό διαχωρισμό γι'αυτό ως ακολούθως:
   - *Αριθμητικά πεδία (range):* Ταξινομούμε τις τιμές των πεδίων στον κόμβο από την μικρότερη στη μεγαλύτερη. Επιλέγουμε κάθε σημείο με τη σειρά σαν σημείο διαχωρισμού και υπολογίζουμε το στατιστικό μη καθαρότητας για τους θυγατρικούς κόμβους που προκύπτουν σαν αποτέλεσμα του διαχωρισμού. Έπειτα διαλέγουμε σαν σημείο διαχωρισμού για το πεδίο , αυτό το οποίο αποδίδει τη μεγαλύτερη μείωση στη μη καθαρότητα σε σύγκριση με τη μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.
   - *Κατηγορικά πεδία (συμβολικά):* Εξετάζουμε τον κάθε πιθανό συνδυασμό των τιμών σαν δυο υποσύνολα. Για κάθε συνδυασμό , υπολογίζουμε τη μη καθαρότητα των θυγατρικών κόμβων για το διαχωρισμό που βασίζεται σε αυτό το συνδυασμό. Επιλέγουμε σαν καλύτερο σημείο διαχωρισμού για το πεδίο , αυτό το οποίο αποδίδει τη μεγαλύτερη μείωση στη μη καθαρότητα σε σύγκριση με τη μη καθαρότητα του κόμβου ο οποίος διαχωρίζεται.
2) Βρίσκουμε τον καλύτερο διαχωρισμό για τον κόμβο και προσδιορίζουμε το πεδίο του οποίου ο καλύτερος διαχωρισμός δίνει την μεγαλύτερη μείωση στην μη καθαρότητα για τον κόμβο. Στη συνέχεια επιλέγουμε αυτό τον καλύτερο διαχωρισμό του πεδίου ως το βέλτιστο συνολικό διαχωρισμό για τον κόμβο.

3) Ελέγχουμε εάν ικανοποιούνται οι κανόνες διακοπής , και επαναλαμβάνουμε. Εάν οι κανόνες διακοπής δεν ικανοποιούνται από το διαχωρισμό ή από τον γεννήτορα κόμβο , εφαρμόζουμε το διαχωρισμό για να δημιουργήσουμε δυο θυγατρικούς κόμβους. Επαναλαμβάνουμε όλη τη διαδικασία σε κάθε θυγατρικό κόμβο.

2.1.1.3 Διαχείριση κενών-ελλειπουσών τιμών

Πολλές φορές καλούμαστε να διαχειριστούμε κάποια κενά που τυχόν υπάρχουν σε πεδία πρόβλεψης. Για το σκοπό αυτό χρησιμοποιείται ο υποκατάστατος διαχωρισμός (surrogate splitting) . Στην περίπτωση που το βέλτιστο πεδίο πρόβλεψης το οποίο χρησιμοποιείται για ένα διαχωρισμό έχει μια ελλείπουσα τιμή ή ένα κενό σε κάποιο συγκεκριμένο κόμβο , ένα άλλο πεδίο το οποίο αποδίδει ένα παρόμοιο διαχωρισμό με αυτό του πεδίου πρόβλεψης στο ευρύτερο πλαίσιο αυτού του κόμβου , χρησιμοποιείται σαν υποκατάστατο για το πεδίο πρόβλεψης και η τιμή του χρησιμοποιείται για να εκχωρήσει την εγγραφή σε έναν από τους θυγατρικούς κόμβους.

Για παράδειγμα έστω ότι το X^* είναι το πεδίο πρόβλεψης το οποίο καθορίζει το βέλτιστο διαχωρισμό s^* στον κόμβο t. Η διαδικασία υποκατάστατου διαχωρισμού βρίσκει ένα άλλο διαχωρισμό s δηλαδή τον υποκατάστατο , βασισμένο σ' ένα άλλο πεδίο πρόβλεψης X τέτοιο ώστε αυτός ο διαχωρισμός να είναι ο πιο κοντινός με τον s^* στον κόμβο t. (Για καταχωρήσεις με έγκυρες τιμές και για τα πεδία πρόβλεψης). Εάν μια νέα καταχώρηση προορίζεται για πρόβλεψη και έχει μια ελλείπουσα τιμή στο πεδίο X^* στον κόμβο t τότε ο υποκατάστατος διαχωρισμός s εφαρμόζεται τελικά. Στην περίπτωση όμως που η καταχώρηση έχει επίσης μια ελλείπουσα τιμή στο X τότε χρησιμοποιείται το αμέσως καλύτερο υποκατάστατο και με τον τρόπο αυτό συνεχίζεται η διαδικασία μέχρις ότου να φτάσουμε στο καθοριμένο περιορισμένο όριο του αριθμού υποκατάστατων το οποίο φυσικά σχετίζεται με την ταχύτητα και την μνήμη.

Όταν ενδεχομένως μια καταχώρηση έχει ελλείπουσες τιμές στο πεδίο διαχωρισμού και σε όλα τα υποκατάστατα πεδία , μεταβιβάζεται στο θυγατρικό κόμβο με τη μεγαλύτερη συνάρτηση πυκνότητας πιθανότητας (weighted probability) , η οποία υπολογίζεται ως

$$\frac{N_{f,j}(t)}{N_f(t)}$$

Όπου:

$N_{f,j}(t)$ είναι το άθροισμα το οποίο προκύπτει από τα βάρη συχνότητας (frequency weights) για τις καταχωρήσεις στην κατηγορία j και για τον κόμβο t , και

$N_f(t)$ είναι το άθροισμα το οποίο προκύπτει από τα βάρη συχνότητας για όλες τις καταχωρήσεις στον κόμβο t.

Στην περίπτωση που το μοντέλο κατασκευάστηκε χρησιμοποιώντας ίσα ή καθορισμένα από το χρήστη priors , τα priors ενσωματώνονται στον υπολογισμό ως εξής:

$$\frac{\pi(j)}{p_f(t)} \times \frac{N_{f,j}(t)}{N_f(t)}$$

Όπου:

$\pi(j)$ είναι η prior πιθανότητα για την κατηγορία j

$P_f(t)$είναι η σταθμισμένη πιθανότητα μιας εγγραφής που εκχωρείται στον κόμβο με

$$p_f(t) = \sum_j \frac{\pi(j) N_{f,j}(t)}{N_{f,j}}$$

$N_{f,j}(t)$ είναι το άθροισμα από τα βάρη συχνότητας ή το πλήθος των καταχωρήσεων αν δεν γνωρίζονται τα βάρη συχνότητας στον κόμβο t που ανήκουν στην κατηγορία j.

$N_{f,j}$ είναι το άθροισμα από τα βάρη συχνότητας των καταχωρήσεων που ανήκουν σε κατηγορία μέσα σε ολόκληρο το δείγμα εκπαίδευσης.

Στο πεδίο στόχο οι καταχωρήσεις με ελλείπουσες τιμές δε λαμβάνονται υπόψην στην κατασκευή του μοντέλου του δέντρου. Όταν έχουμε να ταξινομήσουμε νέες καταχωρήσεις διαχειριζόμαστε τα κενά με τον ίδιο ακριβώς τρόπο όπως και κατά τη διάρκεια ανάπτυξης του δέντρου δηλαδή χρησιμοποιώντας υποκατάστατα όταν υπάρχει δυνατότητα , και όπου χρειάζεται χρησιμοποιούμε διαχωρισμό βασισμένο στις σταθμισμένες πιθανότητες.

2.1.1.4 Μέτρα μή καθαρότητας

Για την εύρεση διαχωρισμών στα CART μοντέλα υπάρχουν τρία γνωστά διαφορετικά μέτρα μη καθαρότητας τα οποία όμως εξαρτώνται από τον τύπο του πεδίου στόχου. Οι τύποι *Gini* και *Towing* χρησιμοποιούνται για συμβολικά πεδία στόχου ενώ η LSD (Least Squared deviation) ή μέθοδος απόκλισης ελάχιστων τετραγώνων χρησιμοποιείται για συνεχείς στόχους.

Gini:

Ο δείκτης ακαθαρσίας Gini $g(t)$ σε έναν κόμβο t ενός CART δέντρου , ορίζεται ως:

$$g(t) = \sum_{j \neq i} p(j/t)\, p(i/t)$$

Όπου i και j είναι κατηγορίες στο πεδίο στόχου

$$p(j/t) = \frac{p(j,t)}{p(t)}$$

$$p(j,t) = \frac{\pi(j) N_j(t)}{N_j}$$

$$p(t) = \sum_{j} p(j,t)$$

Όπου $\pi(j)$ είναι η τιμή της prior πιθανότητας για την κατηγορία j

$N_j(t)$ είναι το πλήθος των καταχωρήσεων στην κατηγορία j του κόμβου t

N_j είναι το πλήθος των των καταχωρήσεων στην κατηγορία j στον αρχικό κόμβο-ρίζα.

Σημειώνουμε ότι όταν χρησιμοποιείται ο δείκτης Gini για την εύρεση της βελτίωσης για έναν διαχωρισμό κατά την διάρκεια της ανάπτυξης του δέντρου , για να υπολογιστούν το N_j και το $N_j(t)$ χρησιμοποιούνται οι καταχωρήσεις στον αρχικό κόμβο-ρίζα και στον κόμβο t αντίστοιχα που έχουν έγκυρες τιμές για το πεδίο διαχωρισμού (split-predictor).

Μια άλλη μορφή του δείκτη μη καθαρότητας Gini είναι:

$$g(t) = 1 - \sum_j p^2(j/t)$$

Έτσι όταν οι καταχωρήσεις σε ένα κόμβο διανέμονται ομαλά δια μέσου των κατηγοριών , ο δείκτης Gini λαμβάνει την μεγαλυτερη τιμή του $1 - \frac{1}{k}$, όπου k έιναι το πλήθος των κατηγοριών για το πεδίο στόχος. Ο δείκτης Gini ισούται με 0 όταν όλες οι καταχωρήσεις σε ένα κόμβο ανήκουν στην ίδια κατηγορία.

Για το διαχωρισμό s στον κόμβο t η συνάρτηση του κριτηρίου Gini $\Phi(s,t)$ορίζεται:

$$\Phi(s,t) = g(t) - p_L g(t_L) - p_R g(t_R)$$

Όπου

p_Lείναι η μερίδα των καταχωρήσεων στον κόμβο t οι οποίες στέλνονται στον αριστερό θυγατρικό κόμβο

p_R είναι η μερίδα των καταχωρήσεων στον κόμβο t οι οποίες στέλνονται στον δεξιό θυγατρικό κόμβο.

Οι λόγοι p_Lκαι p_R ορίζονται ως εξής:

$$p_L = \frac{p(t_L)}{p(t)}$$

και

$$p_R = \frac{p(t_R)}{p(t)}$$

Επιλέγεται ο κατάλληλος διαχωρισμός s ούτως ώστε να μεγιστοποιηθεί η τιμή της $\Phi(s,t)$ συνάρτησης.

Twoing:

Ο δείκτης Twoing είναι βασισμένος στο διαχωρισμό των κατηγοριών στόχου σε δύο υπερκλάσεις, και ακολούθως στην εύρεση του βέλτιστου διαχωρισμού στο πεδίο πρόβλεψης και στηρίζεται στις δύο υπερκλάσεις. Οι υπερκλάσεις C_1και C_2 ορίζονται ως εξής:

$$C_1 = \{j: p(j/t_L) \geq p(j/t_R)\}$$

και

$$C_2 = C - C_1$$

Όπου:

C είναι το σύνολο των κατηγοριών του πεδίου στόχος

$p(j/t_R)$, $p(j/t_L)$ είναι τα $p(j/t)$ όπως ορίζονται στο κριτήριο Gini για τους δεξιούς και αριστερούς θυγατρικούς κόμβους αντίστοιχα.

Η συνάρτηση του κριτηρίου του Twoing για το διαχωρισμό s στον κόμβο t ορίζεται ως:

$$\Phi(s,t) = p_L p_R \left[\sum_j |p(j/t_L) - p(j/t_R)|\right]^2$$

Όπου t_Lκαι t_R είναι οι κόμβοι που δημιουργούνται από το διαχωρισμό s.

Ο διαχωρισμός που επιλέγεται είναι αυτός ο οποίος μεγιστοποιεί το Twoing κριτήριο.

LSD (least squared deviation)

Το LSD μέτρο μη καθαρότητας χρησιμοποιείται για τα συνεχή πεδία στόχου. Είναι η σταθμισμένη, μέσα στον κόμβο t διακύμανση και συμβολίζεται με $R(t)$. Ορίζεται ως

$$R(t) = \frac{1}{N_W(t)} \sum_{i \in t} w_i f_i \left(y_i - \overline{y(t)}\right)^2$$

Όπου:

$N_W(t)$ είναι ο σταθμισμένος αριθμός των καταχωρήσεων στον κόμβο t,

w_i είναι η τιμή του αντισταθμισμένου πεδίου για οποιαδήποτε καταχώρηση i,

f_i είναι η τιμή οποιουδήποτε πεδίου συχνότητας,

y_i είναι η τιμή του πεδίου στόχος,

$\overline{y(t)}$ είναι ο σταθμισμένος μέσος όρος για τον κόμβο t.

Η συνάρτηση του LSD, κριτηρίου για το διαχωρισμό στον κόμβο t ορίζεται:

$$\Phi(s,t) = R(t) - p_L R(t_L) - p_R R(t_R)$$

Ο διαχωρισμός s επιλέγεται έτσι ώστε να μεγιστοποιείται η τιμή της συνάρτησης $\Phi(s,t)$.

2.1.1.5 Κανόνες διακοπής – ολοκλήρωσης της διαδικασίας

Οι κανόνες διακοπής-ολοκλήρωσης της διαδικασίας ελέγχουν αν η διαδικασία κατασκευής δέντρου πρέπει να σταματήσει ή όχι. Χρησιμοποιούνται οι εξής κανόνες διακοπής:

- Αν ο κόμβος γίνει καθαρός: δηλαδή αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές της εξαρτημένης μεταβλητής τότε ο κόμβος δε θα διαχωριστεί.
- Αν όλες οι περιπτώσεις μέσα σε ένα κόμβο έχουν πανομοιότυπες τιμές για κάθε μεταβλητή πρόβλεψης τότε ο κόμβος δε θα διαχωριστεί.

- Αν το βάθος του πρόσφατου δέντρου πλησιάζει την τιμή του μέγιστου ορίου βάθους το οποίο καθορίζεται από το χρήστη, η διαδικασία κατασκευής δέντρου θα σταματήσει.
- Αν το μέγεθος ενός κόμβου είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από τον χρήστη τότε ο κόμβος δεν θα διαχωριστεί.
- Αν ο διαχωρισμός ενός κόμβου έχει σαν αποτέλεσμα ένα θυγατρικό κόμβο του οποίου το μέγεθος είναι μικρότερο από την ελάχιστη τιμή μεγέθους του κόμβου που ορίζεται από το χρήστη τότε ο κόμβος δε θα διαχωριστεί.
- Ο καλύτερος διαχωρισμός για ένα κόμβο αποδίδει μια μείωση στη μη καθαρότητα η οποία είναι μικρότερη από την ελάχιστη αλλαγή στη μη καθαρότητα που ορίζεται από το χρήστη.

2.1.1.6 *Κέρδη-κόστη*

Κέρδη (profits):

Τα κέρδη είναι αριθμητικές τιμές οι οποίες σχετίζονται με τις κατηγορίες ενός συμβολικού πεδίου στόχου τα οποία μπορούν να χρησιμοποιηθούν για να εκτιμήσουν το κέρδος ή τη ζημιά που σχετίζεται με ένα τμήμα. Καθορίζουν τη σχετική τιμή κάθε καταχώρισης του πεδίου στόχου. Οι τιμές χρησιμοποιούνται στον υπολογισμό των κερδών αλλά όχι κατά την διάρκεια ανάπτυξης του δέντρου. Το κέρδος για κάθε κόμβο στο δέντρο υπολογίζεται ως:

$$\sum_j f_j(t)\, P_j$$

Όπου

j είναι η κατηγορία του πεδίου στόχος ,

$f_j(t)$ είναι το άθροισμα των τιμών των πεδίων συχνότητας για όλες τις καταχωρήσεις στον κόμβο t με κατηγορία j για το πεδίο στόχος ,

P_j είναι η τιμή κέρδους για την κατηγορία j (καθορίζεται από το χρήστη).

Κόστη:

- Gini: Αν τα κόστη καθορίζονται , ο δείκτης Gini υπολογίζεται ως:

$$g(t) = \sum_{j \neq i} C(i/j)p(j/t)p(i/t)$$

Όπου

$C(i/j)$ είναι το κόστος της λανθασμένης ταξινόμησης μιας κατηγορίας j σαν κατηγορία i.

- Twoing: Αν τα κόστη, δεν λαμβάνονται υπόψη στον διαχωρισμό κόμβων χρησιμοποιώντας το towing κριτήριο. Παρόλ'αυτά , τα κόστη θα ενσωματωθούν στην εκχώρηση κόμβου και στην εκτίμηση του ρίσκου (η διαδικασία αυτή περιγράφεται πιο κάτω)
- LSD: Τα κόστη δεν εφαρμόζονται στα δέντρα παλινδρόμησης.

2.1.1.7 *Priors πιθανότητες*

Οι priors (πιθανότητες ορισμένες εκ των προτέρων) είναι αριθμητικές τιμές οι οποίες επηρεάζουν τα ποσοστά λανθασμένης ταξινόμησης για τις κατηγορίες του πεδίου στόχος. Καθορίζουν την αναλογία των καταχωρήσεων που αναμένονται να ανήκουν σε κάθε κατηγορία του πεδίου στόχος πριν από την ανάλυση. Οι τιμές των priors εμπλέκονται στην ανάπτυξη του δέντρου καθώς και στην εκτίμηση του ρίσκου.

Υπάρχουν τρεις τρόποι υπολογισμού για τις prior πιθανότητες:

- **Εμπειρικές priors** που υπολογίζονται με βάση τα δεδομένα εκπαίδευσης.

$$\pi(j) = \frac{N_{w,j}}{N_w}$$

- **Ίσες priors:** Η επιλογή των ίσων priors ορίζει την prior πιθανότητα για κάθε μια από τις J κατηγορίες στην ίδια τιμή.

$$\pi(j) = \frac{1}{J}$$

- **Priors καθορισμένες από το χρήστη** των οποίων οι καθορισμένες-εξειδικευμένες τιμές χρησιμοποιούνται στους υπολογισμούς οι οποίοι περιέχουν priors.

2.1.1.8 *Διαδικασία κλαδέματος*

Το κλάδεμα (pruning) αναφέρεται στη διαδικασία του ελέγχου ενός πλήρους αναπτυσσόμενου δέντρου και της αφαίρεσης των διαχωρισμών των κάτω επιπέδων που δεν έχουν σημαντική συνεισφορά στην ακρίβεια του δέντρου. Το λογισμικό στο κλάδεμα του δέντρου προσπαθεί να δημιουργήσει το μικρότερο δέντρο του οποίου το ρίσκο λανθασμένης ταξινόμησης δεν είναι πολύ μεγαλύτερο από το ρίσκο λανθασμένης ταξινόμησης του μεγαλύτερου πιθανού δέντρου. Η διαδικασία αφαιρεί ένα κλαδί, δέντρου αν το κόστος στο οποίο σχετίζεται με τη μεγαλύτερη πολυπλοκότητα του δέντρου είναι μεγαλύτερο από το κέρδος το οποίο σχετίζεται με το εάν έχουμε ένα άλλο επίπεδο κόμβων (κλαδί). Χρησιμοποιεί ένα δείκτη ο οποίος μετρά το ρίσκο λανθασμένης ταξινόμησης και την πολυπλοκότητα του δέντρου αφού στόχος μας είναι να ελαχιστοποιήσουμε και τα δυο.

Το μέτρο κόστους πολυπλοκότητας (cost complexity) ορίζεται ως εξής:

$$R_a(T) = R(T) + a\left|\tilde{T}\right|$$

Όπου:

$R(T)$είναι το ρίσκο λανθασμένης ταξινόμησης του δέντρου T ,

$\left|\tilde{T}\right|$είναι το πλήθος των τερματικών κόμβων για το δέντρο T ,

a είναι το κόστος πολυπλοκότητας ανά τερματικό κόμβο για το δέντρο

Η τιμή a υπολογίζεται από τον αλγόριθμο κατά τη διάρκεια του κλαδέματος.

Κάθε δέντρο που μπορούμε να παράγουμε έχει ένα μέγιστο μέγεθος (T_{max}), όπου σε κάθε τερματικό κόμβο περιέχεται μόνο μια καταχώρηση. Στην περίπτωση που το κόστος πολυπλοκότητας είναι μηδενικό $(\alpha = 0)$, το μέγιστο δέντρο έχει το

χαμηλότερο ρίσκο , αφού κάθε εγγραφή προβλέπεται τέλεια. Επομένως , όσο μεγαλύτερη είναι η τιμή του α , τόσο μικρότερος είναι ο αριθμός των τερματικών κόμβων στο $T(a)$, δηλαδή το δέντρο με το μικρότερο κόστος πολυπλοκότητας για το δοσμένο α. Όταν το α αυξάνεται από το 0 τότε παράγει μια πεπερασμένη ακολουθία από υποδέντρα (T_1, T_2, T_3), το καθένα με λιγότερους τερματικούς κόμβους από το προηγούμενο. Το κλάδεμα κόστους πολυπλοκότητας δουλεύει αφαιρώντας τον πιο αδύναμο διαχωρισμό. Οι εξισώσεις που ακολουθούν εκφράζουν το κόστος πολυπλοκότητας για τον κόμβο $\{t\}$, που είναι ένας οποιοσδήποτε ξεχωριστός-μόνος κόμβος , και για T_t , τον υπο-κλάδο του $\{t\}$:

$$R_a(\{t\}) = R(t) + a$$

και

$$R_a(T_t) = R(T_t) + a\left|\widetilde{T}_t\right|$$

Στην περίπτωση που το $R_\alpha(T_t)$ είναι μικρότερο από το $R_\alpha(\{t\})$, το κλαδί T_t έχει μικρότερο κόστος πολυπλοκότητας από αυτό του ξεχωριστού κόμβου $\{t\}$.

Η διαδικασία ανάπτυξης του δέντρου εξασφαλίζει ότι για $(\alpha = 0)$ ισχύει

$$R_a(\{t\}) \geq R_a(T_t) \qquad (\mathbf{1})$$

Καθώς το α αυξάνεται από το 0, τα $R_\alpha(\{t\})$ και $R_\alpha(T_t)$ αυξάνονται γραμμικά με το $R_\alpha(T_t)$ να αυξάνεται με ταχύτερο ρυθμό. Τελικά , βρίσκουμε ένα κάτω φράγμα α' τέτοιο ώστε $R_a(\{t\}) < R_a(T_t)$ για όλα τα $\alpha > \alpha'$. Συμπεραίνουμε ότι όταν το α γίνεται μεγαλύτερο από το α' , το κόστος πολυπλοκότητας του δέντρου μειώνεται αν κόψουμε το υποκλάδι (sub branch) T_t κάτω από το $\{t\}$.

Μπορούμε εύκολα να υπολογίσουμε το κάτω όριο λύνοντας την $(\mathbf{1})$ ούτως ώστε να βρούμε τη μεγαλύτερη τιμή του α για την οποία ισχύει η ανισότητα , η οποία συμβολίζεται και ως $g(t)$. Συνεπώς προκύπτει:

$$\alpha \leq g(t) = \frac{R(t) - R(T_t)}{\left|\widetilde{T}_t\right| - 1}$$

Μπορούμε να ορίσουμε σαν το πιο αδύναμο (link) σύνδεσμο (t) στο δέντρο T τον κόμβο ο οποίος παίρνει την μικότερη τιμή του $g(t)$:

$$g(\bar{t}) = \min_{t \in T} g(t)$$

Κατά συνέπεια, καθώς το α αυξάνεται , ο $\bar{t}$ είναι ο πρώτος κόμβος για τον οποίο ισχύει

$R_a(\{t\}) = R_a(T_t)$. Στο σημείο αυτό , το $\{t\}$ προτιμάται από το $T_{\bar{t}}$, και το υποκλάδι κλαδεύεται.

Συνοπτικά ο αλγόριθμος κλαδέματος βασίζεται στα εξής βήματα:

- Ορίζουμε το $\alpha_1 = 0$ και ξεκινούμε με το δέντρο για το οποίο $T_1 = T(0)$ δηλαδή το πλήρως αναπτυσσόμενο δέντρο.
- Αυξάνουμε το α μέχρι το κλάδεμα ενός κλαδιού. Έπειτα κλαδεύουμε το κλαδί από το δέντρο και υπολογίζουμε την εκτίμηση του ρίσκου του δέντρου το οποίο έχουμε κλαδέψει.
- Επαναλαμβάνουμε το προηγούμενο βήμα μέχρι να απομείνει μόνο ο αρχικός κόμβος ρίζα , αποδίδοντας μια σειρά από υποδέντρα $T_1, T_2, \dots, T_k$.
- Στην περίπτωση που επιλέξουμε τον κανόνα του τυπικού σφάλματος , τότε διαλέγουμε το μικρότερο δέντρο T_{opt} για το οποίο
$$R\big(T_{opt}\big) \leq \min_k R(T_k) + m \times SE\big(R(T)\big)$$
- Στην περίπτωση που δεν επιλέγουμε τον κανόνα τυπικού σφάλματος τότε διαλέγουμε το δέντρο με τη μικρότερη τιμή της συνάρτησης ρίσκου $R(T)$.

2.2 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένα εργαλείο που έχει πολλές και ποικίλες εφαρμογές στον κλάδο του data mining λόγω της δυναμικής τους , της ευελιξίας τους και της ευκολίας στη χρήση τους. Ο όρος νευρωνικά δίκτυα χρησιμοποιείται για μια αόριστη οικογένεια μοντέλων , που χαρακτηρίζονται από ένα μεγάλο χώρο παραμέτρων και ευέλικτη δομή. Η ιδέα για τη μελέτη και την ανάπτυξη των νευρνικών δικτύων , προήλθε από τη λειτουργία και τη δομή του ανθρώπινου εγκεφάλου και των διαδικασιών του σχετικά με τη μάθηση , τη μνήμη , τη γενίκευση , την ομαδοποίηση προτύπων κ.λ.π. Οι ορισμοί για τα νευρωνικά δίκτυα, ποικίλουν όσο και οι τομείς στους οποίους χρησιμοποιούνται. Εφόσον δεν υπάρχει ένας συγκεκριμένος ορισμός που να καλύπτει απόλυτα ολόκληρη την οικογένεια των μοντέλων , (αποδεχόμαστε) την ακόλουθη περιγραφή (Haykin , 1998):

Νευρωνικό δίκτυο είναι ένας μαζικός παράλληλος διανεμημένος επεξεργαστής ο οποίος εκ φύσεως αποθηκεύει εμπειρική γνώση και την καθιστά διαθέσιμη για χρήση.Προσομοιάζει στον ανθρώπινο εγκέφαλο σε δυο τομείς:

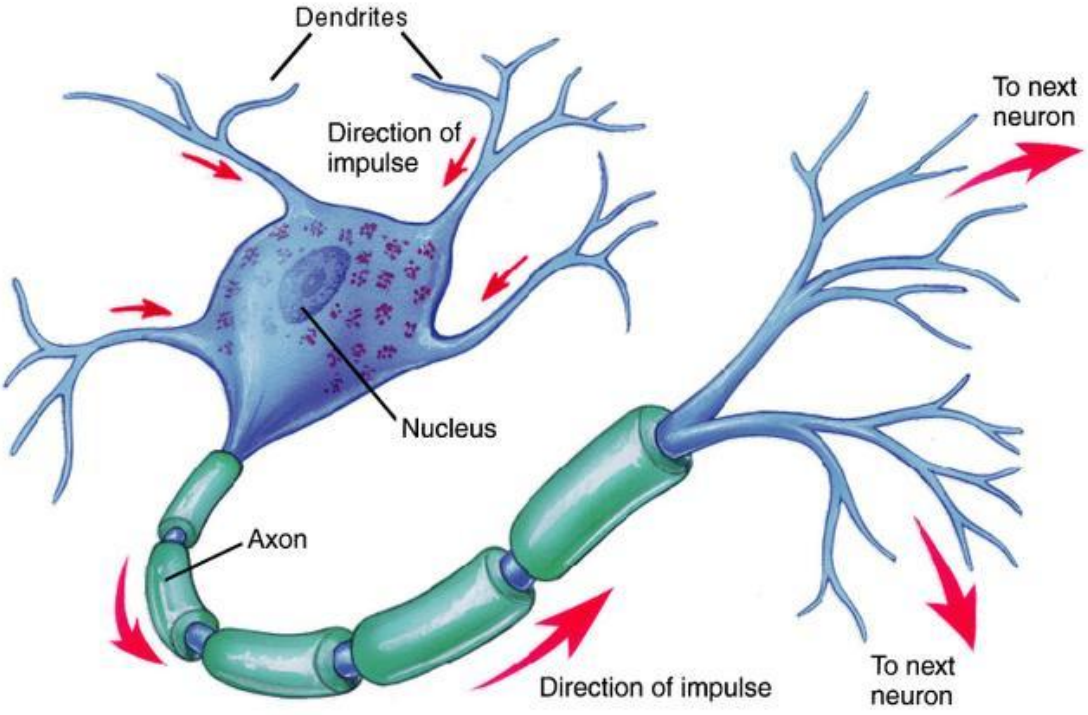
- Η γνώση αποκτάται από το δίκτυο μέσω μιας διαδικασίας μάθησης
- Οι ενδονευρωνικές συνδέσεις , γνωστές και ως συναπτικά βάρη , χρησιμοποιούνται για τη φύλαξη γνώσης.

2.2.1 Νευρώνας-Λειτουργία του βιολογικού νευρώνα

Ο ανθρώπινος εγκέφαλος είναι ένας ιδιαίτερα πολύπλοκος , μη γραμμικός και παράλληλος ηλεκτρονικός υπολογιστής (σύστημα επεξεργασίας πληροφοριών). Έχει την ικανότητα να οργανώνει τους νευρώνες με τέτοιο τρόπο ώστε να κάνει συγκεκριμένους υπολογισμούς , όπως για παράδειγμα την αναγνώριση προτύπων (pattern recognition) , την αντίληψη (perception) και την κίνηση , πολύ πιο γρήγορα από τον γρηγορότερο ηλεκτρονικό υπολογιστή που υπάρχει.

Ο ανθρώπινος εγκέφαλος αποτελείται από περίπου 100 δισεκατομμύρια νευρικά κύτταρα ή νευρώνες. Ο νευρώνας είναι ένα μεγάλο κύτταρο του οποίου η δομή περιλαμβάνει τέσσερα κύρια τμήματα που λειτουργικά παίζουν διαφορετικούς ρόλους:

- Το **σώμα**, που περιέχει τον πυρήνα και αποτελεί την καρδιά του κυττάρου.
- Οι **δενδρίτες**, που είναι οι πύλες εισόδου του νευρώνα και δέχονται ηλεκτρικά σήματα από άλλους νευρώνες.
- Ο **άξονας**, που είναι η πύλη εξόδου του νευρώνα. Μοιάζει με μακρόστενη κλωστή και ο σκοπός του είναι να μεταδώσει τα ηλεκτρικά σήματα, που δημιουργούνται στο νευρώνα, στους άλλους νευρώνες.
- Οι **συνάψεις** οι οποίες αποτελούν την περιοχή της σύνδεσης μεταξύ δύο νευρώνων. Είναι τα σημεία ένωσης των διακλαδώσεων του άξονα ενός νευρώνα-αποστολέα, και των δενδριτών των νευρώνων-παραληπτών.



<u>Εικόνα</u>: Ο Βιολογικός νευρώνας

Στους βιολογικούς νευρώνες , φορείς πληροφορίας είναι ηλεκτρικοί παλμοί που ταξιδεύουν στον άξονα κάθε νευρώνα και μέσω των συνάψεων διαδίδονται στους δενδρίτες των παραληπτών νευρώνων. Κάθε νευρώνας A συλλέγει όλο το ηλεκτρικό φορτίο που δέχεται από κάθε σύναψη στους δενδρίτες του, ζυγίζοντας το εισερχόμενο φορτίο με το αντίστοιχο συνοπτικό βάρος. Έτσι , όσο πιο ισχυρή είναι η συναπτική ζεύξη τόσο πιο πολύ έντονα συμμετέχει το συγκεκγριμένο φορτίο εισόδου στο συνολικό άθροισμα. Αν το άθροισμα του φορτίου ξεπερνάει κάποιο κατώφλι τότε ο άξονας του A αρχίζει να παράγει ηλεκτρικούς παλμούς με μεγάλη συχνότητα οπότε λέμε ότι ο νευρώνας *πυροβολεί (fires).* Αν όμως το φορτίο δεν περνάει το συγκεκριμένο αυτό όριο τότε ο νευρώνας παράγει πολύ αραιά παλμούς σε τυχαίες στιγμές οπότε λέμε ότι ο νευρώνας είναι *αδρανής.* Κάθε παλμός έχει συγκεκριμένο χρονικό πλάτος t_p και μετά από κάθε παλμό ο νευρώνας χρειάζεται ένα ελάχιστο χρόνο ανάπαυσης t_r. Έτσι ο μέγιστος αριθμός των παλμών δεν ξεπερνάει το όριο

$$Firing\ frequency\ < 1/(t_p + t_r)$$

Τελικά οι παλμοί που παράγονται ταξιδεύουν κατά μήκος του άξονα και τροφοδοτούν τους άλλους νευρώνες με τους οποίος συνδέεται ο A.

2.2.2 Το μαθηματικό μοντέλο

Κατ'αρχάς είναι αδύνατο για τα τεχνητά νευρωνικά δίκτυα να προσομοιάσουν πλήρως τη πολυπλοκότητα του ανθρώπινου εγκεφάλου. Τα τεχνητά νευρωνικά δίκτυα αποτελούνται το πολύ από μερικές εκατοντάδες (ή χιλιάδες) νευρώνες και περιορισμένο αριθμό συνδέσεων μεταξύ τους. Παρόλα αυτά κάποια δίκτυα έχουν χρησιμοποιηθεί για την επίλυση αρκετά περίπλοκων υπολογιστικών προβλημάτων.

Για τη μοντελοποίηση ενός βιολογικού νευρώνα σε ένα μαθηματικό μοντέλο, πρέπει να ληφθούν υπόψη τρεις βασικές συνιστώσες. Αρχικά, οι συνάψεις των βιολογικών νευρώνων μοντελοποιούνται σαν **συναπτικά βάρη (synaptic weights)**.

Ας θυμηθούμε πως οι συνάψεις των βιολογικών νευρώνων είναι υπεύθυνες για τη διασύνδεση του νευρωνικού δικτύου και δίνουν τη δύναμη των συνδέσεων. Για ένα τεχνητό νευρώνα, τα βάρη είναι πραγματικοί αριθμοί και αντιπροσωπεύουν τις συνάψεις. Ένα αρνητικό βάρος εκφράζει μια ανασταλτική σύνδεση, ενώ ένα θετικό μια διεγερτική σύνδεση. Πολλά μοντέλα νευρώνων περιλαμβάνουν επίσης και ένα εξωτερικό βάρος, που ονομάζεται **μεροληψία (bias)**. Σκοπός της μεροληψίας είναι η αύξηση ή η μείωση της τιμής που δίνει το δίκτυο στη συνάρτηση ενεργοποίησης ανάλογα με το αν είναι αρνητικό ή θετικό.

Οι υπόλοιπες συνιστώσες του μοντέλου αντιπροσωπεύουν τη δραστηριότητα του νευρώνα. Οι **είσοδοι (inputs)** του νευρώνα αθροίζονται και τροποποιούνται από τα συναπτικά βάρη. Τέλος, μια **συνάρτηση ενεργοποίησης (activation function)** ελέγχει το εύρος του εξερχόμενου φορτίου.

Θα δούμε τώρα πιο αναλυτικά τη διαδικασία μοντελοποίησης ενός βιολογικού νευρώνα σε ένα μαθηματικό μοντέλο. Έστω ,

$w_{k1}, w_{k2}, \dots, w_{kp}$ είναι τα συναπτικά βάρη,

$x_1, \dots, x_p$ οι είσοδοι του νευρώνα k και

b_k η μεροληψία.

Τότε το άθροισμα v_k του φορτίου που δέχεται ο νευρώνας εκφράζεται ως (Ρίζος, 1996):

$$v_k = \sum_{j=1}^{p} w_{kj} x_j$$

Η έξοδος του νευρώνα y_k, θα είναι το αποτέλεσμα της εφαρμογής μιας συνάρτησης ενεργοποίησης $\varphi(.)$ στη τιμή του v_k:

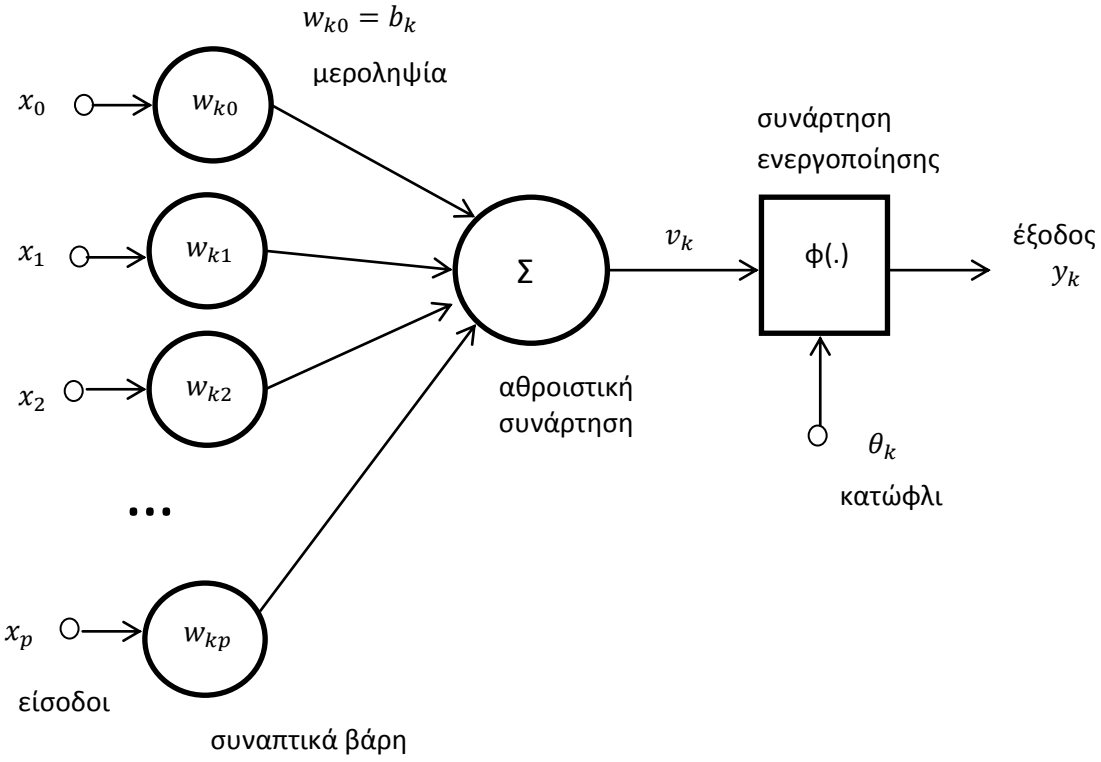
$$y_k = \varphi(v_k - b_k).$$

Η μεροληψία b_k είναι μια εξωτερική παράμετρος του νευρώνα που δεν εξαρτάται από καμιά τιμή εισόδου, και μπορούμε να την εντάξουμε στο μοντέλο του νευρώνα ως μια νέα σύναψη που έχει σαν είσοδο $x_0 = \pm 1$ (ανάλογα αν αυξάνει ή μειώνει τη

τιμή εισόδου στο δίκτυο) και βάρος $w_{k0} = b_k$. Έτσι θέτοντας: $u_k = v_k - b_k$ οι εξισώσεις που περιγράφουν το νευρώνα γίνονται τελικά:

$$u_k = \sum_{j=0}^{p} w_{kj} x_j$$

$$y_k = \varphi(u_k)$$



<u>Σχήμα</u>: Ο τεχνητός νευρώνας

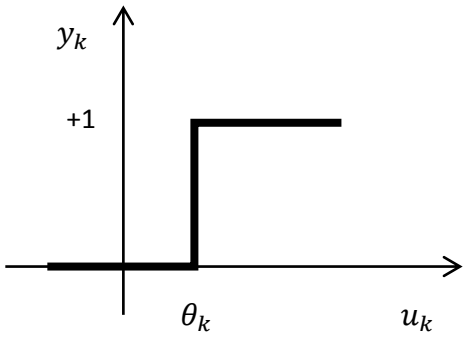
2.2.3 Συνάρτηση Ενεργοποίησης

Η συνάρτηση ενεργοποίησης λειτουργεί σαν συμπιεστική συνάρτηση, ώστε η έξοδος του νευρώνα σε ένα νευρωνικό δίκτυο να είναι μεταξύ συγκεκριμένων τιμών (συνήθως μεταξύ 0 και 1, ή -1 και 1). Γενικά, υπάρχουν τρεις τύποι συναρτήσεων ενεργοποίησης $\boldsymbol{\varphi}(.)$:

- Η **συνάρτηση κατώφλι (threshold function)** η οποία έχει σαν έξοδο 0 αν το εισερχόμενο άθροισμα είναι μικρότερο από μια καθορισμένη τιμή-κατώφλι θ_k, και 1 αν είναι μεγαλύτερο ή ίσο με αυτό.

$$v = u_k - \theta_k$$

$$y_k = \varphi(v) = \begin{cases} 0, & \alpha\nu\ v < 0 \\ 1, & \alpha\nu\ v \geq 0 \end{cases}$$

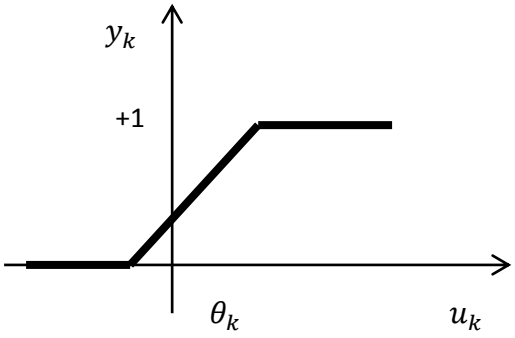


<u>Γράφημα</u>: Συνάρτηση κατώφλι

- Η **τμηματική γραμμική συνάρτηση (pricewise-linear function).** Όπως και η συνάρτηση κατώφλι έχει σαν έξοδο 0 ή 1, καθώς επίσης και τιμές μεταξύ αυτών που εξαρτώνται από τον παράγοντα ενίσχυσης μέσα στη γραμμική περιοχή της συνάρτησης.

$$v = u_k - \theta_k$$

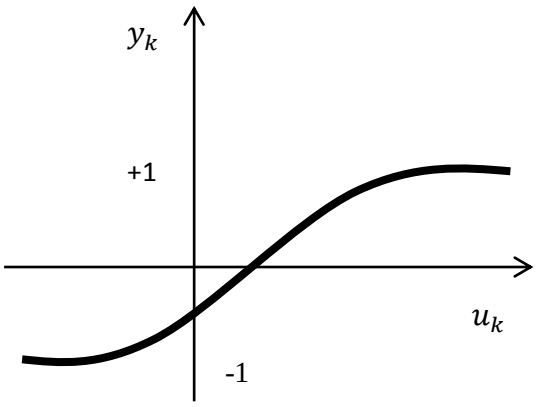
$$\varphi(v) = \begin{cases} 1, & \alpha\nu \quad v \geq \frac{1}{2} \\ v, & \alpha\nu \ -\frac{1}{2} < v < \frac{1}{2} \\ 0, & \alpha\nu \quad v \leq -\frac{1}{2} \end{cases}$$



Γράφημα: Τμηματική γραμμική συνάρτηση

- Η **σιγμοειδής συνάρτηση (sigmoid function)**. Είναι μια γνησίως αύξουσα συνάρτηση που είναι ομαλή και ασυμπτωτική. Αντίθετα με τη συνάρτηση κατώφλι είναι διαφορίσιμη και είναι η πλέον χρησιμοποιούμενη συνάρτηση ενεργοποίησης για τα νευρωνικά δίκτυα. Ένα παράδειγμα της σιγμοειδούς συνάρτησης είναι η υπερβολική εφαπτόμενη:

$$\varphi(\upsilon) = \tanh(\upsilon) = \frac{1 - e^{-\upsilon}}{1 + e^{-\upsilon}}$$

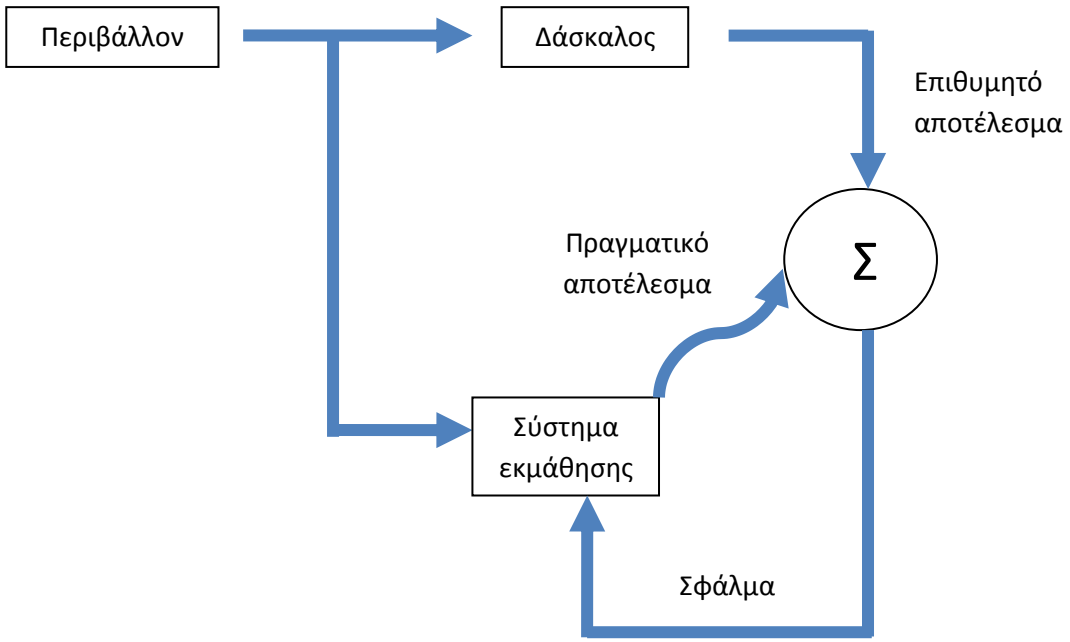


Γράφημα: Σιγμοειδής συνάρτηση

2.2.4 Ταξινόμηση Νευρωνικών αλγορίθμων

Ένα νευρωνικό δίκτυο είναι απαραίτητο να ρυθμιστεί έτσι ώστε όταν ένα σετ δεδομένων δοθεί σαν είσοδος, να παράγει το επιθυμητό σετ δεδομένων εξόδου. Υπάρχουν διάφορες μέθοδοι για την προσαρμογή των δυνάμεων των συνδέσεων σε ένα δίκτυο. Ένας τρόπος για να επιτευχθεί η μάθηση του δικτύου, είναι η ρύθμιση των συναπτικών βαρών, χρησιμοποιώντας μια *a priori* γνώση. Μια άλλη μέθοδος είναι η εκπαίδευση του νευρωνικού δικτύου τροφοδοτώντας το, με πρότυπα διδασκαλίας και επιτρέποντάς του να αλλάξει τα συναπτικά βάρη ανάλογα με κάποιο διδακτικό κανόνα. Μπορούμε να κατηγοριοποιήσουμε τις διαδικασίες μάθησης σε τρεις κλάσεις:

- **Μάθηση με επίβλεψη (supervised learning).** Στη μάθηση με επίβλεψη είναι απαραίτητη η παρουσία ενός εξωτερικού , ως προς το δίκτυο , παράγοντα που μπορούμε να ονομάσουμε «δάσκαλο». Στο σχήμα που ακολουθεί παρουσιάζεται το πώς επιδρά ο δάσκαλος στο δίκτυο και το περιβάλλον κατά τη διαδικασία της μάθησης. Ο δάσκαλος έχει την απαραίτητη γνώση για το περιβάλλον , που πρακτικά είναι ένα σύνολο από παραδείγματα εισόδου και την αντίστοιχη επιθυμητή έξοδο. Το Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) δεν έχει καμιά γνώση για το περιβάλλον. Αν υποθέσουμε ότι παρουσιάζουμε στο δάσκαλο και το δίκτυο ένα πρότυπο από το περιβάλλον , τότε λόγω της προηγούμενης γνώσης του δασκάλου για το περιβάλλον , θα είναι σε θέση να παρέχει στο δίκτυο την επιθυμητή απάντηση – έξοδο. Στη συνέχεια οι παράμετροι του δικτύου προσαρμόζονται ανάλογα με το πρότυπο που χρησιμοποιείται για την εκπαίδευση και το σφάλμα του δικτύου (δηλαδή τη διαφορά μεταξύ της επιθυμητής εξόδου και της εξόδου που στην πράξη δίνει το δίκτυο). Η προσαρμογή αυτών των παραμέτρων , γίνεται επαναληπτικά , βήμα προς βήμα , με στόχο το δίκτυο να μπορεί να προσομοιάσει το δάσκαλο. Αν αυτό γίνει εφικτό , τότε μπορούμε να επιτρέψουμε στο δίκτυο να αλληλεπιδράσει με το περιβάλλον χωρίς την παρουσία του δασκάλου.

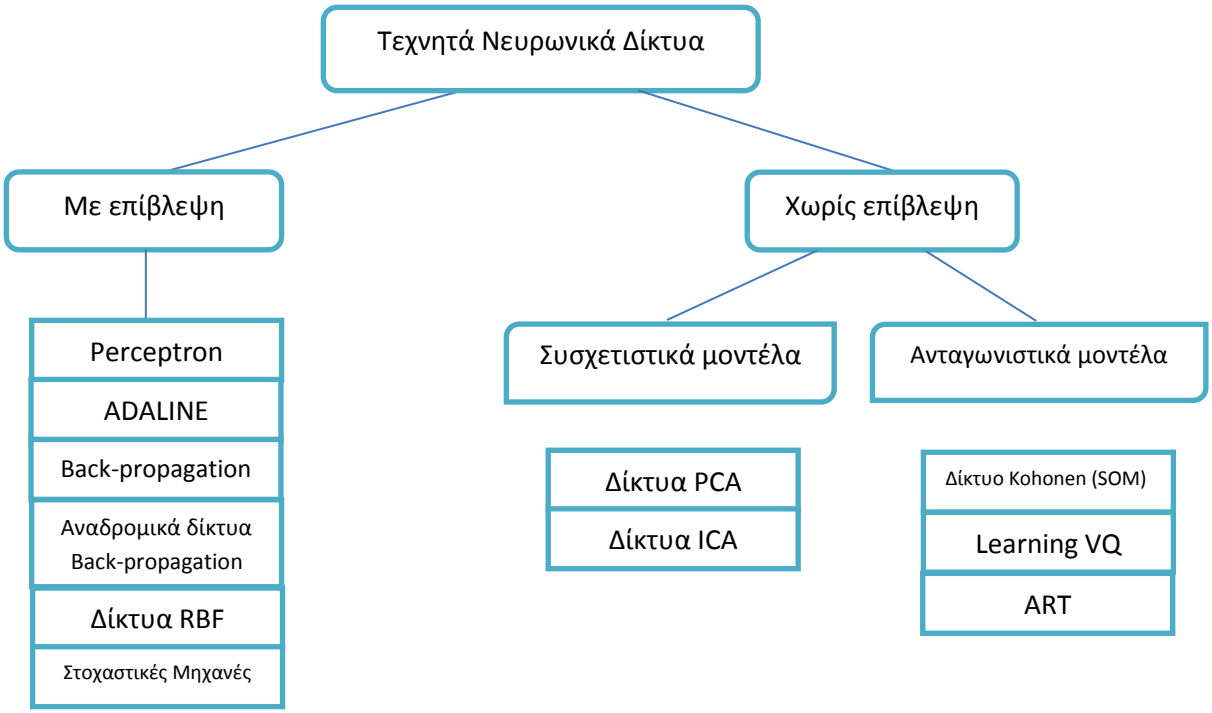


<u>Σχήμα</u>: Εκπαίδευση με επίβλεψη

- **Μάθηση χωρίς επίβλεψη (unsupervised learning) ή Μάθηση με αυτο-οργάνωση (self-organization).** Στην περίπτωση της μάθησης χωρίς επίβλεψη δεν υπάρχει κάποιος εξωτερικός παράγοντας που επιβλέπει τη διαδικασία μάθησης. Αυτό σημαίνει ότι δεν υπάρχουν παραδείγματα της συνάρτησης που πρέπει να μάθει το δίκτυο. Υπάρχει όμως ένα μέτρο , ανεξάρτητο από το εκάστοτε έργο που πρέπει να φέρει εις πέρας το ΤΝΔ , που μετράει την ποιότητα της αναπαράστασης που πρέπει να μάθει το δίκτυο. Οι ελεύθερες παράμετροι του δικτύου βελτιστοποιούνται ως προς αυτό το μέτρο. Όταν το δίκτυο «μάθει» τις στατιστικές ιδιότητες των προτύπων που του δίνονται σαν είσοδος , αναπτύσσει την ικανότητα να δημιουργεί εσωτερικές αναπαραστάσεις για την κωδικοποίηση των χαρακτηριστικών των προτύπων. Αποκτά δηλαδή την ικανότητα να δημιουργεί νέες κλάσεις αυτόματα.

- **Ενισχυτική μάθηση (reinforcement learning)**. Αυτός ο τύπος μάθησης θεωρείται μια ενδιάμεση μορφή των δύο προηγούμενων τύπων. Εδώ, το σύστημα μάθησης αξιολογεί τις ενέργειές του ως καλές (επιβράβευση) ή

κακές (αξιόποινη) βασισμένο σε κάποιες αντιδράσεις από το περιβάλλον και ανάλογα προσαρμόζει τις παραμέτρους του δικτύου. Γενικά, η προσαρμογή των παραμέτρων συνεχίζεται μέχρι να επιτευχθεί μια κατάσταση ισορροπίας, στην οποία, αν εφαρμοστεί ο μηχανισμός μάθησης, δε θα υπάρξουν περαιτέρω αλλαγές στις παραμέτρους.



<u>Διάγραμμα</u>: Διάγραμμα Τεχνητών Νευρωνικών Δικτύων

2.2.5 Το δίκτυο Perceptron

Ο όρος "Perceptrons" επινοήθηκε από τον Frank RosenBlatt το 1962, και χρησιμοποιείται για να περιγράψει τη σύνδεση των απλών νευρώνων σε ένα δίκτυο. Αυτά τα δίκτυα είναι απλοποιημένες μορφές του βιολογικού νευρικού συστήματος, όπου όμως κάποιες ιδιότητες του μεγαλοποιούνται και κάποιες αγνοούνται. Προς το παρόν θα επικεντρωθούμε στο Perceptron ενός στρώματος (Single Layer Perceptrons), δηλαδή το δίκτυο Perceptron όπου δεν υπάρχουν κρυφά επίπεδα νευρώνων. Η λέξη δίκτυο εδώ χρησιμοποιείται καταχρηστικά, αφού δεν υπάρχουν περισσότεροι από ένα νευρώνες για να συνδεθούν μεταξύ τους, παρά μόνο υπάρχουν οι συνδέσεις μεταξύ των εισόδων του νευρώνα.

Γενικά τα δεδομένα εισάγονται στο στρώμα εισόδου και το δίκτυο στη συνέχεια τα επεξεργάζεται πολλαπλασιάζοντάς τα με τα συναπτικά βάρη. Το αποτέλεσμα αυτού του πολλαπλασιασμού, μεταβάλλεται από το τελικό στρώμα, το στρώμα εξόδου, χρησιμοποιώντας μια συνάρτηση που καθορίζει καταπόσον ο κόμβος εξόδου «πυροβολά» ή όχι.

Η διαδικασία κατά την οποία το δίκτυο «εκπαιδεύεται», ο κανόνας εκπαίδευσης δηλαδή, περιλαμβάνει την εύρεση των σωστών τιμών των συναπτικών βαρών. Πρώτα όμως, ο πίνακας των βαρών αρχικοποιείται με τυχαίους αριθμούς μεταξύ -1 και +1. Έπειτα όσο το δίκτυο «μαθαίνει», αυτές οι τιμές μεταβάλλονται μέχρι να αποφασιστεί ότι το δίκτυο έχει επιλύσει το πρόβλημα. Για να εκπαιδευτεί το δίκτυο χρησιμοποιούνται set δεδομένων ως πρότυπα εισόδου για τα οποία οι σωστές έξοδοι είναι γνωστές. Ξεκινώντας από τυχαία συναπτικά βάρη, ένα πρότυπο εισόδου παρουσιάζεται στο δίκτυο, το οποίο κάνει μια αρχική υπόθεση για το ποια πρέπει να είναι η σωστή έξοδος.

Κατά τη διάρκεια της φάσης εκπαίδευσης, η διαφορά μεταξύ της υπόθεσης που κάνει το δίκτυο και της σωστής τιμής της εξόδου αξιολογείται, και τα συναπτικά βάρη αλλάζουν έτσι ώστε το λάθος να ελαχιστοποιηθεί.

Το απλό perceptron υλοποιείται όπως το βασικό μοντέλο που περιγράφεται πιο πάνω και έχει σαν συνάρτηση ενεργοποίησης μια απλή συνάρτηση κατώφλι:

$$f(x) = \begin{cases} 1, & x > t \\ 0, & διαφορετικά \end{cases}$$

, όπου x είναι η έξοδος του νευρώνα και t μια σταθερά-κατώφλι (threshold).

Αν συμβολίσουμε τα συναπτικά βάρη με τον πίνακα W_{ij} , όπου i είναι ο αριθμός των εισόδων, και j ο αριθμός των εξόδων, και το διάνυσμα εισόδου με $\boldsymbol{I}$ τότε η έξοδος $\boldsymbol{O}$ του νευρώνα υπολογίζεται ως εξής:

$$O = f(IW_{ij})$$

Ο κανόνας εκπαίδευσης του απλού perceptron είναι σχετικά απλός. Ξεκινώντας από έναν πίνακα από τυχαία συναπτικά βάρη, παρουσιάζουμε στο δίκτυο ένα πρότυπο

εκπαίδευσης, και υπολογίζουμε την έξοδο του δικτύου όπως πιο πάνω. Καθορίζουμε έτσι μια **συνάρτηση λάθους E**:

$$E(O) = (T - O)$$

Όπου σ' αυτή την περίπτωση, το **T** είναι το επιθυμητό διάνυσμα εξόδου για την είσοδο του προτύπου εκπαίδευσης. Για να καθορίσουμε λοιπόν το πώς πρέπει να προσαρμοστούν τα συναπτικά βάρη ώστε το δίκτυο να παράγει την επιθυμητή έξοδο για τη συγκεκριμένη είσοδο, θα πρέπει η συνάρτηση λάθους να ελαχιστοποιηθεί.

Στα νευρωνικά δίκτυα, ο στόχος είναι να εκτιμηθεί η επίδραση των συναπτικών βαρών στην ολική συνάρτηση λάθους. Συνδυάζοντας τα πιο πάνω διαπιστώνουμε ότι η συνάρτηση λάθους εκφράζεται:

$$E(O) = (T - O) = T - f(IW_{ij})$$

Στη συνέχεια παραγωγίζουμε τη συνάρτηση αυτή και εκτιμούμε έτσι τον πίνακα συναπτικών βαρών που την ελαχιστοποιεί. Η διαδικασία αυτή θα περιγραφεί εκτενέστερα στα επόμενα κεφάλαια και συγκεκριμένα στα δίκτυα Perceptron πολλαπλών στρωμάτων (MLP). Η συνάρτηση που ελαχιστοποιεί το λάθος στον αλγόριθμο του Perceptron ενός στρώματος είναι ιδιαίτερα απλή. Σε κάθε κόμβο εξόδου, υπολογίζεται το λάθος και προστίθεται στο συναπτικό βάρος που τροφοδοτεί αυτό τον κόμβο.

2.2.5.1 *Αλγόριθμος μάθησης*

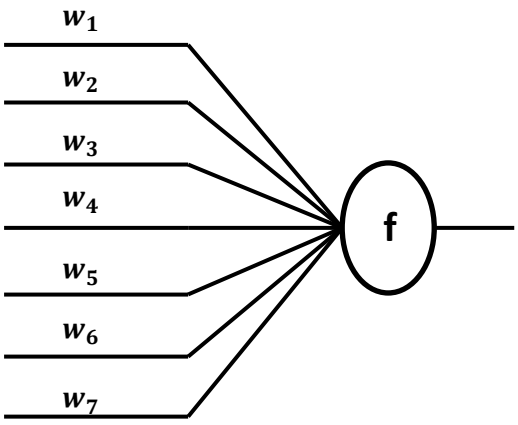
Για να περιγραφεί ο αλγόριθμος μάθησης του απλού Perceptron, πρέπει να οριστούν κάποιες μεταβλητές:

- $y = f(z)$ είναι η έξοδος του δικτύου για ένα διάνυσμα εισόδου z
- b είναι η μεροληψία, η οποία στο παράδειγμα παρακάτω θεωρείται 0
- $D = \{(x_1, d_1), \dots, (x_s, d_s)\}$ είναι το σετ δεδομένων εκπαίδευσης που αποτελείται από s δυάδες προτύπων:

- x_j είναι ένα n-διάστατο διάνυσμα εισόδου
- d_j είναι η επιθυμητή έξοδος του δικτύου για αυτή την είσοδο

- $x_{j,i}$ είναι η τιμή του i-οστού κόμβου του j-οστού διανύσματος εισόδου εκπαίδευσης ($x_{j,0} = 1$)
- w_i είναι η i-οστή τιμή του διανύσματος συναπτικών βαρών, που θα πολλαπλασιαστεί με την i-οστή τιμή του διανύσματος εισόδου
- Για τη χρονική εξάρτηση του w θεωρούμε $w_i(t)$ το i-συναπτικό βάρος τη χρονική στιγμή t
- α είναι μια σταθερά που ονομάζεται ρυθμός μάθησης

Μπορούμε επίσης να προσθέσουμε μια επιπλέον διάσταση, με δείκτη $n + 1$, τέτοια ώστε $x_{j,n+1} = 1$ και $w_{n+1} = b$ και με τον τρόπο αυτό εισάγεται η μεροληψία στο δίκτυο.



<u>Σχήμα</u>: Απλό Perceptron

Ο αλγόριθμος εκτελεί τα πιο κάτω βήματα:

1) *Αρχικοποίησε τα συναπτικά βάρη και το κατώφλι στη συνάρτηση ενεργοποίησης. Τα βάρη μπορούν να αρχικοποιηθούν θέτοντας τα* $w_i(0) = 0$ *ή κάποια μικρή τυχαία τιμή.*
2) *Για κάθε πρότυπο* j *στο σετ εκπαίδευσης* D *, εκτέλεσε τα πιο κάτω βήματα για τη είσοδο* x_j *και την επιθυμητή έξοδο* d_j*:*
   - *Υπολόγισε την έξοδο:*

$$y_j(t) = f\left[w(t) \cdot x_j\right] = f\left[w_0(t) + w_1(t)x_{j,1} + w_2(t)x_{j,2} + \cdots + wntxj,n\right.$$

- *Προσάρμοσε τα συναπτικά βάρη ως εξής:*

  $w_i(t+1) = w_i(t) + a\left(d_j - y_i(t)\right)x_{j,i}$, *για όλους τους κόμβους* $0 \leq i \leq n$

- *Το βήμα 2 επαναλαμβάνεται μέχρι το λάθος* $d_j - y_j(t)$ *να είναι μικρότερο από μια προκαθορισμένη τιμή* γ *, ή όταν συμπληρωθεί ένας προκαθορισμένος αριθμός επαναλήψεων.*

Αυτός είναι ο γενικός αλγόριθμος του perceptron. Μπορεί να δειχθεί ότι αυτή η τεχνική ελαχιστοποιεί τη συνάρτηση λάθους. Ο χρόνος που χρειάζεται για να καταλήξουμε σε λύση (ο χρόνος δηλαδή για να βρεθεί η ελάχιστη τιμή του λάθους) μπορεί να είναι απρόβλεπτος. Αυτό συμβαίνει γιατί, αν η συνάρτηση λάθους προσεγγίζεται με μεγάλα βήματα, η ελάχιστη τιμή ενδέχεται να βρεθεί πιο αργά. Αν γίνονται μικρότερα βήματα είναι πιο πιθανόν να χτυπήσουμε την ελάχιστη τιμή της συνάρτησης. Έτσι λοιπόν, για να ελέγξουμε το ρυθμό σύγκλισης, και να μειώσουμε τον αριθμό των βημάτων που γίνονται, χρησιμοποιούμε την παράμετρο που ονομάζεται **ρυθμός μάθησης (learning rate)** $\boldsymbol{\alpha}$. Αυτή η παράμετρος ορίζεται στο διάστημα [0,1] και έτσι τα συναπτικά βάρη αλλάζουν με μικρότερα βήματα.

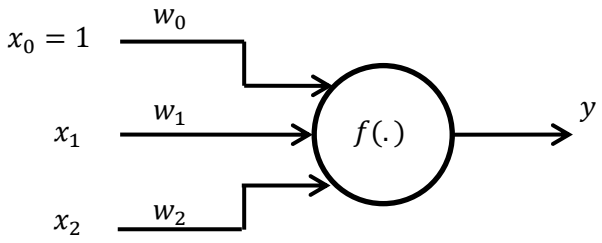
2.2.5.2 *Παράδειγμα του αλγόριθμου μάθησης στον απλό Perceptron*

Στο παράδειγμα αυτό, θα δούμε πώς ένας απλός perceptron με μια διαδικασία μάθησης μπορεί να χρησιμοποιηθεί ώστε να μιμείται τη λογική πύλη AND. Στον πιο κάτω πίνακα φαίνονται όλες οι πιθανές είσοδοι και η αντίστοιχη επιθυμητή έξοδος που θέλουμε να παράγει το δίκτυο.

x_1	x_2	z
0	0	0
0	1	0
1	0	0
1	1	1

Πίνακας: Η λογική πύλη AND

Τα δεδομένα που φαίνονται στον πίνακα αποτελούν το set εκπαίδευσης, και δείχνουν την τοπολογία που πρέπει να έχει το δίκτυο. Αυτό θα περιέχει τρεις κόμβους εισόδου (δύο για τα x_1, x_2 και ένα για το $x_0 = 1$) και έναν κόμβο εξόδο.



Σχήμα: Διάταξη νευρωνικού δικτύου που υλοποιεί τη λογική πύλη AND

Είσοδοι: x_0, x_1, x_2 όπου η είσοδος x_0 παραμένει σταθερή και ίση με 1

Κατώφλι συνάρτησης ενεργοποίησης: $t = 0.5$

Μεροληψία: $b = 0$

Ρυθμός μάθησης: $\alpha = 0.1$

Η αθροιστική συνάρτηση υπολογίζεται ως: $s = x_0 w_0 + x_1 w_1 + x_2 w_2$

Η έξοδος του δικτύου: $n = \begin{cases} 1, & s > t \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$

Η συνάρτηση λάθους: $e = z - n$

Η τελική διόρθωση στα συναπτικά βάρη: $d = \alpha * e$

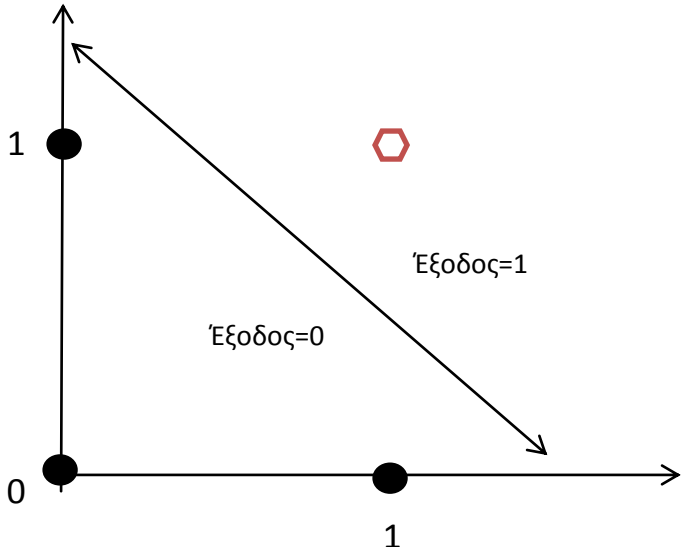
είσοδος				αρχικά βάρη			έξοδος		Λάθος	διόρθωση	τελικά βάρη		
x_0	x_1	x_2	z	w_0	w_1	w_2	s	n	e	d	w_0	w_1	w_2
1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	1	+0.1	0.1	0.1	0.1
1	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0.1	0.1	0.1
1	0	1	0	0.1	0.1	0.1	0.2	0	0	0	0.1	0.1	0.1
1	1	0	0	0.1	0.1	0.1	0.2	0	0	0	0.1	0.1	0.1
1	1	1	1	0.1	0.1	0.1	0.3	0	1	+0.1	0.2	0.2	0.2
1	0	0	0	0.2	0.2	0.2	0.2	0	0	0	0.2	0.2	0.2
1	0	1	0	0.2	0.2	0.2	0.4	0	0	0	0.2	0.2	0.2
1	1	0	0	0.2	0.2	0.2	0.4	0	0	0	0.2	0.2	0.2
1	1	1	1	0.2	0.2	0.2	0.6	1	0	0	**0.2**	**0.2**	**0.2**

Πίνακας: Βήματα αλγορίθμου μάθησης του δικτύου για τη λογική πύλη AND

Ο αλγόριθμος τερματίζεται, αφού μετά την τρίτη είσοδο των δεδομένων εκπαίδευσης, το λάθος μηδενίζεται. Τελικά τα συναπτικά βάρη με τα οποία το δίκτυο μας προσομοιάζει τη λογική πύλη AND είναι: $w_0, w_1, w_2 = 0.2$.

Μπορούμε να δούμε και γραφικά το πιο πάνω πρόβλημα. Στο Γράφημα 1.2.8.1 φαίνεται η χωρική διάταξη των δεδομένων εισόδου. Παρατηρούμε ότι είναι δυνατόν να σχεδιάσουμε μια ευθεία γραμμή μεταξύ των συντεταγμένων των τιμών εισόδου που έχουν σαν έξοδο τη τιμή 1 και αυτών των οποίων απαιτείται έξοδος

ίση με 0. Τα προβλήματα που έχουν την ιδιότητα αυτή ονομάζονται **γραμμικά διαχωρίσιμα**.



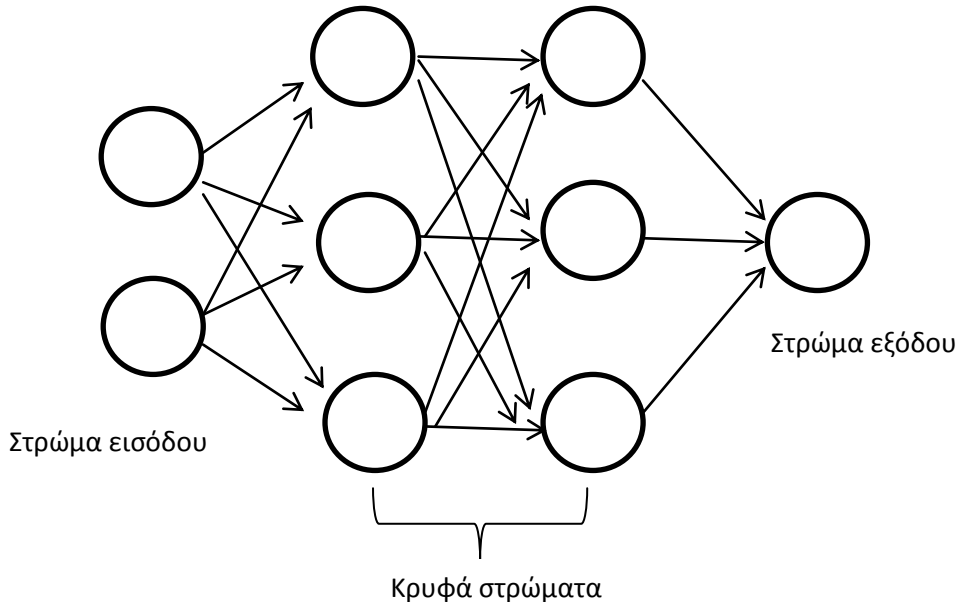
Γράφημα: Χωρική διάταξη δεδομένων εισόδου για τη λογική πύλη AND

Το απλό δίκτυο perceptron με ένα νευρώνα και μια συνάρτηση ενεργοποίησης/κατώφλι, μπορεί μόνο να λύσει προβλήματα που είναι γραμμικά διαχωρίσιμα. Τα περισσότερα όμως προβλήματα που καλούνται να λύσουν τα νευρωνικά δίκτυα δεν είναι της κλάσης αυτής, και γι' αυτό το απλό δίκτυο perceptron δεν μπορεί φυσικά να αντεπεξέλθει στις απαιτήσεις για τις οποίες η τεχνολογία των νευρωνικών δικτύων αναπτύχθηκε.

2.2.6 Perceptron πολλών στρωμάτων

Τα περισσότερα προβλήματα που καλούνται να επιλύσουν τα τεχνητά νευρωνικά δίκτυα, δεν ανήκουν στην κλάση των γραμμικά διαχωρίσιμων, τα οποία το απλό perceptron ενός νευρώνα μπορεί να επιλύσει. Έτσι γύρω στο 1986 αναπτύχθηκαν τα **perceptron πολλών στρωμάτων** ή αλλιώς **multilayer perceptrons (MLP)**. Όπως υποδηλώνει η ονομασία των δικτύων αυτών, ένα MLP είναι ένα δίκτυο αποτελούμενο από πολλούς νευρώνες, κατανεμημένους σε στρώματα (layers). Τα στρώματα αυτά χωρίζονται ως εξής:

- Το **στρώμα εισόδου (input layer)**, από όπου εισέρχονται τα δεδομένα στο δίκτυο. Ο αριθμός των νευρώνων από τους οποίους αποτελείται το στρώμα αυτό, εξαρτάται από τον αριθμό των εισόδων που θέλουμε να πάρει το δίκτυο.
- Ένα ή περισσότερα **κρυφά στρώματα (hidden layers)**. Αυτά τα στρώματα, παρεμβάλλονται μεταξύ του στρώματος εισόδου και εξόδου και ο αριθμός τους ποικίλει. Η λειτουργία των κρυφών στρωμάτων είναι να κωδικοποιούν τις εισόδους και να καθορίζουν τις εξόδους του δικτύου. Έχει αποδειχθεί ότι ένα MLP δίκτυο μπορεί να προσεγγίσει οποιαδήποτε συνάρτηση που συνδέει τις εισόδους της με τις εξόδους της, δεδομένου ότι μια τέτοια συνάρτηση υπάρχει.
- Το **στρώμα εξόδου (output layer)** στο οποίο παρουσιάζεται η έξοδος του δικτύου. Ο αριθμός των νευρώνων στο στρώμα αυτό, εξαρτάται από το πρόβλημα που θέλουμε να μάθει το κάθε δίκτυο.

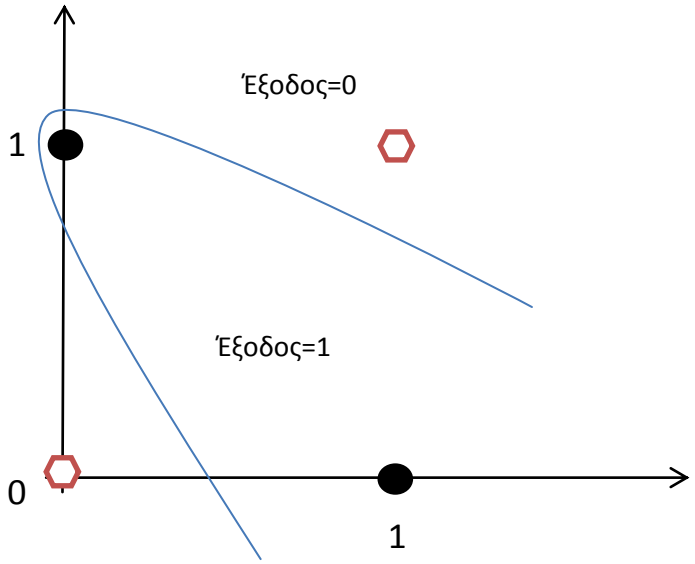


<u>Σχήμα</u>: MLP δίκτυο

Έστω ότι ενδιαφερόμαστε να υλοποιήσουμε τη λογική πύλη XOR, ένα πρόβλημα το οποίο δεν είναι γραμμικά διαχωρίσιμο όπως φαίνεται και στο Γράφημα 2.1.1.1.

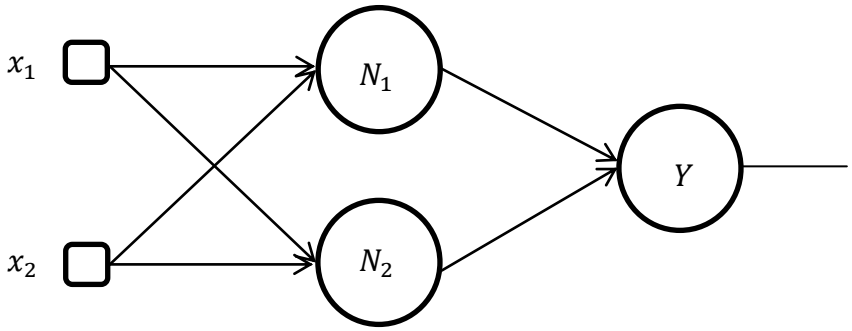
x_1	x_2	z
0	0	0
0	1	1
1	0	1
1	1	0

Πίνακας: Λογική πύλη XOR



Γράφημα: Χωρική διάταξη δεδομένων εισόδου για τη λογική πύλη XOR

Το γράφημα, που δείχνει τη διάταξη των προτύπων, υποδεικνύει πως για το διαχωρισμό αυτών είναι απαραίτητη μια καμπύλη και όχι ευθεία γραμμή. Γι' αυτό και ο αλγόριθμος του απλού perceptron δεν μπορεί να εφαρμοστεί αποτελεσματικά στο συγκεκριμένο πρόβλημα. Θεωρούμε λοιπόν τη διάταξη ενός MLP δικτύου όπως φαίνεται στο σχήμα.



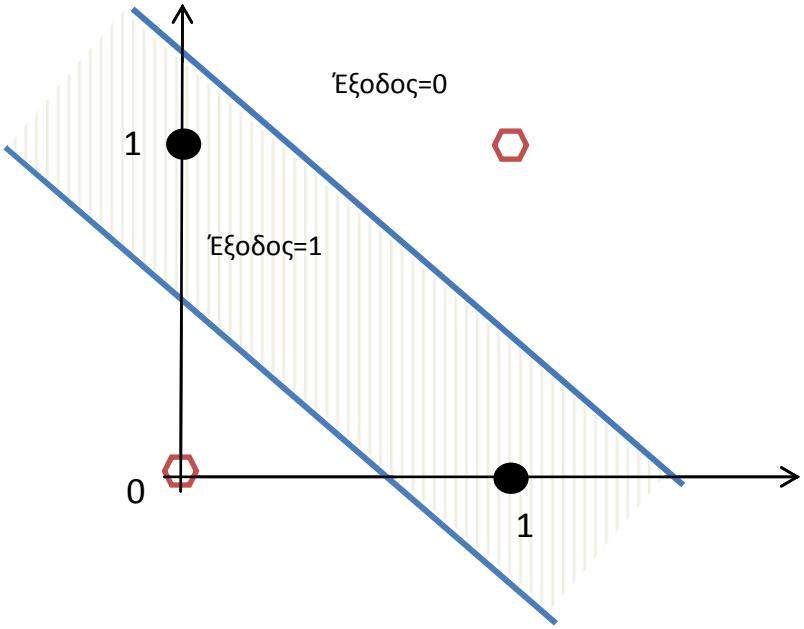
<u>Σχήμα</u>: Διάταξη MLP δικτύου που υλοποιεί τη λογική πύλη XOR

Οπότε, εφαρμόζοντας τον αλγόριθμο του perceptron χρησιμοποιώντας μια κλασική βηματική συνάρτηση μπορούμε να επιτύχουμε τις εξόδους που φαίνονται στον Πίνακα 2.1.1.2 για τους τρεις νευρώνες του δικτύου.

Είσοδοι		$\boldsymbol{N_1}$	$\boldsymbol{N_2}$	$\boldsymbol{Y}$
$\boldsymbol{x_1}$	$\boldsymbol{x_2}$			
0	0	0	0	0
0	1	1	0	1
1	0	1	0	1
1	1	1	1	0

<u>Πίνακας</u>: Έξοδοι των νευρώνων

Από το πίνακα μπορούμε να δούμε, ότι οι κρυφοί νευρώνες N_1, N_2 υλοποιούν τις λογικές πύλες OR και AND αντίστοιχα και ο νευρώνας εξόδου Y συνδυάζει τις εξόδους τους δίνοντας το επιθυμητό αποτέλεσμα του δικτύου. Σχηματικά, το πιο πάνω φαίνεται με το Γράφημα 2.1.1.2.



<u>Γράφημα</u>: Χωρική διάταξη των εισόδων για τη λογική πύλη XOR και οι έξοδοι του MLP δικτύου

Στα προβλήματα που απαιτούνται περισσότερες διαχωριστικές γραμμές, χρησιμοποιούνται και περισσότεροι κρυφοί νευρώνες. Συνδυάζοντας τις ευθείες, μπορούμε να πάρουμε μια μεγάλη ποικιλία περιοχών τις οποίες τα δίκτυα MLP μπορούν να διαχωρίσουν.

2.3 Λογιστική Παλινδρόμηση

Το μοντέλο της Λογιστικής παλινδρόμησης (logistic regression) αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων. Άρχισε να χρησιμοποιείται ευρέως κατά την δεκαετία του $50'$, κυρίως με εφαρμογές στη βιοστατιστική. Είναι μια μέθοδος στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξάρτητων μεταβλητών για την διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.

Η Λογιστική παλινδρόμηση είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός

συνόλου ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης.

Σε πολλές εφαρμογές η εξαρτημένη μεταβλητή παίρνει δυο μόνο τιμές , οι οποίες αντιστοιχούν σε δυο ενδεχόμενα. Για παράδειγμα , το αν ο ασθενής ζει ή απεβίωσε , το αν ο άνεργος βρίσκει εργασία ή όχι , το αν ραγίζει ή αντέχει το δοκάρι. Οι τιμές της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δυο ενδεχομένων , συνήθως 0 και 1.

Εάν ορίσουμε την τιμή $y = 1$ σαν «επιτυχία» και την τιμή $y = 0$ σαν «αποτυχία» , τότε η y είναι τ.μ της κατανομής Bernoulli , δηλαδή $y \sim B(p)$, με μέση τιμή $E(y) = p$ και διασπορά $V(y) = p(1-p)$.

Γενικεύοντας σε μια σειρά από n επαναλήψεις (δηλαδή πραγματοποιήσεων των ενδεχομένων), ορίζουμε την τ.μ

$$y = αριθμός\ επιτυχιών\ σε\ n\ δοκιμές$$

Υπό την υπόθεση ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους , τότε ισχύει η Διωνυμική (binomial) κατανομή

$$y \sim b(n, p)$$

Με συνάρτηση πυκνότητας

$$f(y) = \binom{n}{p} p^y (1-p)^{n-y} \ , y = 0,1,2, \dots , n$$

Όπου

p η πιθανότητα επιτυχίας η οποία είναι παράμετρος της κατανομής.

Η Διωνυμική κατανομή αποτελεί τη βασική κατανομή για την περιγραφή και ανάλυση μιας μεταβλητής αυτής της φύσης. Η μέση τιμή της y είναι ίση με $E(y) = np$ και η διασπορά με $V(y) = np(1-p)$. Στην ειδική περίπτωση που $n = 1$ μιλάμε για *δυαδικά δεδομένα ,* αλλιώς για *διωνυμικά δεδομένα.*

Σε πολλές περιπτώσεις η τ.μ y ενδέχεται να εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της y από τις επεξηγηματικές μεταβλητές x (ανεξάρτητες μεταβλητές ή συμμεταβλητές) εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις x (π.χ η πιθανότητα να μείνει κάποιος άνεργος εξαρτάται από το φύλο , την ηλικία , το μορφωτικό επίπεδο κ.α). Πιο συγκεκριμένα , κατασκευάζεται το αποκαλούμενο *μοντέλο λογιστικής παλινδρόμησης* , το οποίο είναι ένα γενικευμένο γραμμικό μοντέλο και εκφράζεται μέσω της σχέσης:

$$n_x = g\big(E(y_x)\big) = g(\mu_x) = \boldsymbol{x'\beta}$$

με την ακόλουθη δομή:

1. $y_x \sim b(n_x, \mu_x)$ $(n_x > 1\,,\ διωνυμικά\ δεδομένα)$

ή $y_x \sim B(n_x, \mu_x)$ $(n_x = 1\,, δυαδικά\ δεδομένα)$

2. $n_x = g(\mu_x) = \ln\frac{\mu_x}{n_{x-\mu_x}} = \ln\frac{p_x}{1-p_x} = logit(p_x) = \boldsymbol{x'\beta}$ $(συνάρτηση\ Logit)$
3. Ανεξαρτησία μεταξύ των παρατηρήσεων y_x **,**

Όπου ,

n_x είναι ο αριθμός των επαναλήψεων της τιμής του διανύσματος $\boldsymbol{x}$ των επεξηγηματικών μεταβλητών.

Αντιστρέφοντας τη συνάρτηση σύνδεσης προκύπτει:

$$p_x = e^{n_x}/(1 + e^{n_x})$$

για την οποία ισχύει ο περιορισμός $0 < p_x < 1$.

Για κάθε παρατήρηση i το μοντέλο γράφεται ως:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\,, i = 1, \ldots, n$$

όπου

$$p_i = p_{x_i} = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} = \frac{1}{1 + e^{-\boldsymbol{x_i'\beta}}} \qquad (1)$$

η πιθανότητα «επιτυχίας»

$$x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$$

είναι ο linear predictor.

$$E(y_i) = n_i p_i = n_i \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}$$

2.3.1 Εκτίμηση παραμέτρων με τη μέθοδο μέγιστης πιθανοφάνειας (maximum likelihood)

Ας υποθέσουμε ότι τα δεδομένα μας είναι χωρισμένα σε κατηγορίες. Δηλαδή , έχουμε n_i στο πλήθος πειραματικές μονάδες στο i-οστό σημείο δεδομένων (για παράδειγμα μπορούμε να θεωρήσουμε ότι το n_i είναι το πλήθος των πειραματόζωων στα οποία έχουμε δώσει μια συγκεκριμένη δοσολογία φαρμάκου). Το μοντέλο μας βάση της εξίσωσης $(\mathbf{1})$, γράφεται στη μορφή:

$$E(y_i) = n_i P(x_i) = n_i \frac{1}{1 + e^{-\boldsymbol{x_i'\beta}}} \text{ , } i = 1,2, \dots m$$

Με $y_1, y_2, \dots, y_m$ να είναι οι παρατηρούμενες τιμές των ανεξάρτητων διωνυμικών τυχαίων μεταβλητών. Σε αυτήν την περίπτωση ισχύει

$$var(y_i) = n_i P(x_i)[1 - P(x_i)]$$

και το άθροισμα

$$\sum_{i=1}^{m} n_i = n$$

είναι το συνολικό πλήθος του δείγματός μας.

Η συνάρτηση πιθανότητας μιας απλής διωνυμικής τυχαίας μεταβλητής y με παραμέτρους n , P δίνεται από τον τύπο:

$$\binom{n}{y} P^y (1 - P)^{n-y}$$

Ωστόσο , ο όρος $\binom{n}{y}$ δεν περιλαμβάνει το $\boldsymbol{\beta}$, οπότε δεν μπορεί να χρησιμοποιηθεί. Επομένως , η log πιθανοφάνεια για το λογιστικό μοντέλο παλινδρόμησης δίνεται από τον τύπο:

$$\ln[\mathcal{L}(\boldsymbol{P}; y)] = \sum_{i=1}^{m} \left\{ y_i \ln\left[\frac{P(x_i)}{1-P(x_i)}\right] + n_i \ln[1-P(x_i)] \right\} \qquad (2)$$

Είναι εφικτό τώρα να εισάγουμε τη μορφή του λογιστικού μοντέλου στην εξίσωση (1).

Ο όρος $\ln\left[\frac{P(x_i)}{1-P(x_i)}\right]$ ονομάζεται $logit$ και γράφεται ως:

$$\ln\left[\frac{\mathrm{P(x_i)}}{1-\mathrm{P(x_i)}}\right] = \mathrm{x_i^{'}\beta} = \beta_0 + \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{x_{ij}}\,\beta_{\mathrm{j}} \ , \qquad \mathrm{i} = 1,2,\dots,\mathrm{m}, \ \ \mathrm{m} \geq \mathrm{k}+1$$

Σαν αποτέλεσμα η log πιθανοφάνεια της εξίσωσης (2) γράφεται ως:

$$\ln[\mathcal{L}(\beta; \mathrm{y})] = \sum_{\mathrm{i=1}}^{\mathrm{m}} \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{y_i x_{ij}} \beta_{\mathrm{j}} - \sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{n_i} \ln\left(1 + \exp \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{x_{ij}} \beta_{\mathrm{j}}\right) \qquad (3)$$

Στη συνέχεια η εξίσωση (3) πρέπει να μεγιστοποιηθεί ως προς τον όρο β_{j}. Σε μορφή πινάκων , η εξίσωση (3) γράφεται ως:

$$\ln[\mathcal{L}(\beta; \mathrm{y})] = \beta' \mathrm{X y} - \sum_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{n_i} \ln\left(1 + \exp(\mathrm{x_i}'\beta)\right) \qquad (4)$$

Όπου:

X είναι ο κλασσικός πίνακας του μοντέλου που συναντάμε και στην γραμμική παλινδρόμηση και

y το διάνυσμα της απόκρισης.

Παραγωγίζουμε τώρα την εξίσωση (4) ως προς β:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\beta_i}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}' \boldsymbol{y} - \sum_{i=1}^{m} \left[\frac{n_i}{1 + e^{\boldsymbol{x_i}'\boldsymbol{\beta}}}\right] e^{\boldsymbol{x_i}'\boldsymbol{\beta}} \boldsymbol{x_i}$$

Γνωρίζοντας ότι ισχύει

$$\frac{e^{\boldsymbol{x_i'\beta}}}{1+e^{\boldsymbol{x_i'\beta}}} = \frac{1}{1+e^{-\boldsymbol{x_i'\beta}}} = P(x_i)\,,$$

Προκύπτει ότι:

$$\frac{\partial \ln \mathcal{L}(\boldsymbol{\beta_i}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \boldsymbol{X'y} - \sum_{i=1}^{m} n_i P(x_i) x_i$$

Εφόσον ο όρος $\mathrm{n_i P(\mathbf{x_i})}$ αποτελεί τον μέσο της διωνυμικής τυχαίας μεταβλητής το δεξί μέλος , της παραπάνω σχέσης γράφεται σε μορφή πινάκων ως $\mathrm{X'(y-\mu)}$ όπου:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_m \end{bmatrix}$$

και $\mathrm{\mu_i = n_i P(\mathbf{x_i})}$. Σαν αποτέλεσμα ο εκτιμητής μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator-MLE) είναι η λύση της εξίσωσης (score equation):

$$\boldsymbol{X'(y-\mu)} = \mathbf{0} \qquad (5)$$

Για την λύση της εξίσωσης (5) μπορούμε να χρησιμοποιήσουμε μια επαναληπτική διαδικασία για να παράγουμε τις εκτιμήσεις $\mathrm{b_0, b_1, \dots, b_k}$ των όρων $\mathrm{\beta_0, \beta_1, \dots, \beta_k}$ για τις $\mathrm{p = k+1}$ παραμέτρους του μοντέλου. Μια τέτοια επαναληπτική μέθοδος είναι αυτή των σταθμισμένων ελαχίστων τετραγώνων (weighted least squares).

Ο τύπος για το σταθμισμένο άθροισμα ελαχίστων τετραγώνων των υπολοίπων είναι:

$$S = \sum_{i=1}^{m} \left[\frac{(y_i - \mu_i)^2}{{\sigma_i}^2} \right]$$

Όπου $\mu_i = n_i P(\boldsymbol{x_i})$ και ${\sigma_i}^2$είναι η διωνυμική διακύμανση στο i-οστό σημείο δεδομένων με

$${\sigma_i}^2 = n_i P(\boldsymbol{x_i})[1 - P(\boldsymbol{x_i})] = n_i \frac{e^{-\boldsymbol{x_i'\beta}}}{(1+e^{-\boldsymbol{x_i'\beta}})^2}$$

Ελαχιστοποιούμε το S:

$$\min S = \min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \left[\frac{(y_i - \mu_i)^2}{{\sigma_i}^2} \right]$$

Η διακύμανση ${\sigma_i}^2$είναι σταθερή, επομένως παραγωγίζουμε μόνο τον αριθμητή του S και παίρνουμε:

$$2 \left[\frac{\sum_{i=1}^{m} (y_i - \mu_i)}{{\sigma_i}^2} \right] \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)$$

Ισχύει ότι:

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = n_i P(\boldsymbol{x_i})[1 - P(\boldsymbol{x_i})]\boldsymbol{x_i} = {\sigma_i}^2 \boldsymbol{x_i}$$

Επομένως , η λύση που παίρνουμε από την ελαχιστοποίηση του σταθμισμένου αθροίσματος τετραγώνων των υπολοίπων με σταθερό ${\sigma_i}^2$είναι:

$$\sum_{i=1}^{m} (y_i - \mu_i)\boldsymbol{x_i} = \mathbf{0}$$

η οποία είναι παρόμοια με την εξίσωση $\mathrm{X}'(\mathrm{y} - \mu) = \mathbf{0}$, εξίσωση (5). Άρα μια επαναληπτική μέθοδος όπως η παραπάνω μπορεί να χρησιμοποιηθεί για να προσδιοριστούν οι αριθμητικές τιμές των $\mathrm{b}_0, \mathrm{b}_1, \ldots, \mathrm{b}_\mathrm{k}$δηλαδή των εκτιμητών μέγιστης πιθανοφάνειας.

2.3.2 Άλλες μορφές στατιστικής συμπερασματολογίας για τις οποίες γίνεται χρήση λογιστικής παλινδρόμησης

Όπως έχουμε δει η λογιστική παλινδρόμηση χρησιμοποιείται σε πολλές διαφορετικές περιπτώσεις για την εξαγωγή συμπερασμάτων , όπως για παράδειγμα σε κλινικές δοκιμές όπου πρέπει να συγκρίνουμε τα αποτελέσματα διαφορετικών θεραπειών των οποίων το αποτέλεσμα έχει δυαδική μορφή. Για την βελτίωση του μοντέλου εξετάζεται η σημασία της κάθε μεταβλητής.Σε πολλές περιπτώσεις τα δεδομένα που έχουμε δεν είναι ομαδοποιημένα δηλαδή $n_i = 1$. Όταν όμως οι πειραματικές μονάδες του δείγματος είναι σχετικά ομοιογενείς , τότε η λογιστική παλινδρόμηση μπορεί να πάρει τη μορφή μιας καμπύλης «δόσης-απόκρισης» ,

όπου μετράει την ανταπόκριση ενός ασθενή ανάλογα με την δοσολογία που του χορηγείται. Σε μια τέτοια περίπτωση , ισχύει ότι $k = 1$ και $p = 2$ και το μοντέλο παίρνει την μορφή:

$$P(x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

Σε αρκετές περιπτώσεις τα όρια εμπιστοσύνης για τα β_0 και β_1καθώς επίσης και τα όρια εμπιστοσύνης για τους συντελεστές της $P(x_i)$ είναι σημαντικά για τους ερευνητές. Το ενδιαφέρον για την μελέτη του κάθε συντελεστή παλινδόμησης ξεχωριστά πηγάζει από την ανάγκη να προσδιορίσουμε τους λόγους πιθανοτήτων (odd ratios).

Για παράδειγμα αρκετά συχνά η ανεξάρτητη μεταβλητή x είναι κατηγορική και μια ομάδα από τα πειραματικά μας υποκείμενα χωρίζεται σε αυτά που τους χορηγήθηκε μεγάλη ποσότητα βιταμίνης C $(x = 0)$ και σε αυτά που δεν τους χορηγήθηκε τίποτα $x = 1$. Έτσι η απόκριση μπορεί να είναι η μόλυνση του αναπνευστικού συστήματος ή όχι , παίρνει τις τιμές $y = 1$ και $y = 0$ αντίστοιχα.

Η ιδέα προσδιορισμού του λόγου πιθανοτήτων είναι αποτέλεσμα της χρήσης του $logit(P)$ που δίνεται από τον τύπο:

$$Log\left[\frac{P}{(1-P)}\right]$$

Στη γενική σχέση (1) του λογιστικού μοντέλου παλινδρόμησης , το $logit[P(\boldsymbol{x_i})]$ δίνεται από τη σχέση:

$$\ln\left[\frac{P(\boldsymbol{x_i})}{1 - P(\boldsymbol{x_i})}\right] = \boldsymbol{x_i'\beta}$$

και μέσω του μετασχηματισμού του P γραμμικοποιείται η λογιστική συνάρτηση.Αν χρησιμοποιήσουμε την εξίσωση (7) για το μοντέλο της εξίσωσης (6) , τότε στο παραπάνω παράδειγμα προκύπτει η σχέση:

$$\ln\left[\frac{P(\boldsymbol{x_i})}{1 - P(\boldsymbol{x_i})}\right] = \beta_0 + \beta_1 \boldsymbol{x_i}$$

Αν τώρα θεωρήσουμε ένα υποκείμενο στο οποίο χορηγείται η βιταμίνη C , δηλαδή $x = 0$, τότε η ποσότητα $\exp(\beta_0)$ μπορεί να μεταφραστεί σαν ο λόγος συχνοτήτων για τα υποκείμενα που μολύνθηκαν προς αυτά που δεν μολύνθηκαν , για όλο τον

πληθυσμό που μελετάμε. Όσον αφορά την ομάδα υποκειμένων στα οποία δεν χορηγήθηκε βιταμίνη $(x = 1)$, τότε έχουμε:

$$\ln\left[\frac{P(\boldsymbol{x_i})}{1 - P(\boldsymbol{x_i})}\right] = \beta_0 + \beta_1$$

Χρησιμοποιώντας την παραπάνω ερμηνεία του β_0 μπορούμε να βρούμε την αντίστοιχη *odd ratio* ερμηνεία του β_1. Για την ομάδα υποκειμένων που δεν δέχτηκε θεραπεία ισχύει:

$$\ln\left[\frac{Pr(Y = 1|x = 1)}{\Pr(Y = 0|x = 1)}\right] = \ln\left[\frac{\Pr(Y = 1|x = 0)}{\Pr(Y = 0|x = 0)}\right] + \beta_1$$

Άρα, η ποσότητα $\exp(\beta_1)$ μπορεί να ερμηνευτεί ως ο λόγος συχνοτήτων της ομάδας που δεν δέχτηκε θεραπεία, σε σχέση με αυτή που δέχτηκε. Προφανώς ένας ερευνητής ερμηνεύει μια τιμή $\beta_0 \ll 0$, όπως επίσης και μια τιμή $\beta_1 \gg 0$, να είναι ευνοϊκή προς την θεραπεία.

2.3.3 Ιδιότητες της διασποράς των εκτιμητών μέγιστης πιθανοφάνειας στη λογιστική παλινδρόμηση.

Είναι ευρέως γνωστό ότι οι εκτιμητές μέγιστης πιθανοφάνειας παρουσιάζουν ασυμπτωτικές ιδιότητες στη διακύμανση και τη συνδιακύμανση , στον πίνακα πληροφορίας. Στην περίπτωση ενός γραμμικού μοντέλου με κανονικά , ανεξάρτητα και ομοιόμορφα κατανεμημένα (iid) σφάλματα , ο πίνακας πληροφορίας για τους εκτιμώμενους συντελεστές παλινδρόμησης εκφράζεται από τον τύπο:

$$\boldsymbol{I}(\boldsymbol{b}) = \frac{\boldsymbol{X'X}}{\sigma^2}$$

Όπου σ^2 είναι η διακύμανση του σφάλματος. Σαν αποτέλεσμα , σε αυτήν την περίπτωση , ο πίνακας διακύμανσης – συνδιακύμανσης (variance-covariance matrix) των εκτιμώμενων συντελεστών είναι:

$$\boldsymbol{I^{-1}}(\boldsymbol{b}) = (\boldsymbol{X'X})^{-\mathbf{1}}\sigma^2$$

Ο πίνακας πληροφορίας παρουσιάζει , κατά μία έννοια , την ποιότητα των πληροφοριών των παραμέτρων που διατίθενται από τα δεδομένα μας. Ένας σχετικά μεγάλος, πίνακας πληροφορίας σημαίνει μικρότερες διακυμάνσεις στους εκτιμώμενους συντελεστές του μοντέλου. Μπορούμε να υπολογίσουμε τον πίνακα πληροφορίας με διάφορες μεθόδους , όπως με τη βοήθεια της εξίσωσης (5):

$$\boldsymbol{I}(\boldsymbol{b}) = var(score)$$

$$= var[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu})]$$

Όπου var συμβολίζουμε τον πίνακα διακύμανσης-συνδιακύμανσης (variance-covariance matrix).

Χρησιμοποιώντας το συντελεστή τυπικής διακύμανσης , η παραπάνω εξίσωση αποκτά τη μορφή:

$$var[\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{\mu})] = \boldsymbol{X}'var[(\boldsymbol{y} - \boldsymbol{\mu})]\boldsymbol{X}$$

Για το μοντέλο της λογιστικής παλινδρόμησης , έχουμε υποθέσει για τις $y_1, y_2, \dots, y_m$ ανεξάρτητες παρατηρήσεις , ότι κάθε y_i παρατήρηση είναι μια διωνυμική τυχαία μεταβλητή με μέσο $n_i P(\boldsymbol{x_i})$ και διασπορά ${\sigma_i}^2 = n_i[P(\boldsymbol{x_i})][1 - P(\boldsymbol{x_i})]$. Επομένως , ισχύει ότι:

$$\boldsymbol{V} = diag\{{\sigma_1}^2, {\sigma_2}^2, \dots, {\sigma_m}^2\}$$

Και

$$\boldsymbol{I}(\boldsymbol{b}) = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}$$

Ο ασυμπτωτικός πίνακας διακύμανσης-συνδιακύμανσης του b δίνεται , λοιπόν , από τον τύπο:

$$var(\boldsymbol{b}) = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}$$

Συνεπώς , τα εκτιμώμενα τυπικά σφάλματα βρίσκονται στα διαγώνια στοιχεία του $\boldsymbol{V}$, το οποίο αντικαθιστά ο $\widehat{\boldsymbol{V}}$ από τη στιγμή που τα β της $P(\boldsymbol{x_i})$ έχουν αντικατασταθεί από τα εκτιμώμενα b.

2.3.4 Συμπερασματολογία με χρήση της μεθόδου Wald στη λογιστική παλινδρόμηση

Η πρώτη εφαρμογή της μεθόδου Wald χρησιμοποιείται στον έλεγχο υποθέσεων για κάθε ξεχωριστό συντελεστή του μοντέλου της λογιστικής παλινδρόμησης. Πιο συγκεκριμένα , θέλουμε να ελέγξουμε:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

με το β_j να εμφανίζεται στον linear predictor $\boldsymbol{\chi_i}'\boldsymbol{\beta}$ του λογιστικού μοντέλου στην εξίσωση (1).

Για έναν εκτιμητή μέγιστης πιθανοφάνειας b_j ισχύει ότι:

$$z_j = \frac{b_j - \beta_j}{\sigma b_j}$$

Αυτός ακολουθεί την τυπική κανονική κατανομή $N(0,1)$ και έτσι ισχύει ότι το

$${z_j}^2 = \left(\frac{b_j}{\sigma b_j}\right)^2$$

Ακολουθεί ασυμπτωτικά την ${x_1}^2$ κατανομή , υπό την H_0 υπόθεση , όπου σb_jείναι το κατάλληλο διαγώνιο στοιχείο του ασυμπτωτικού πίνακα variance-covariance των b.

Στην πράξη , αντικαθιστούμε τα σb_j με τα $\hat{\sigma} b_j$. Ο έλεγχος που διαξάγουμε , είναι ο συνηθισμένος μονομερής ή διμερής έλεγχος (one or two sided test). Για τον υπολογισμό των τιμών της ${\chi_1}^2$ και της $p-τιμής$ για κάθε συντελεστή του απαιτούμενου μοντέλου γίνεται χρήση διαφόρων στατιστικών πακέτων.

Μια δεύτερη μορφή της Wald συμπερασματολογίας έχει να κάνει με τον υπολογισμό του διαστήματος εμπιστοσύνης της διωνυμικής πιθανότητας για κάποια δοσμένα ή αυθαίρετα δεδομένα. Θα μπορούσε να χρησιμοποιηθεί η μέθοδος Δέλτα για το σκοπό αυτό αλλά λόγω ύπαρξης του linear predictor $\boldsymbol{\chi_i}'\boldsymbol{\beta}$ στο

λογιστικό μοντέλο ακολουθείται μια εναλακτική διαδικασία υπολογισμού των διαστημάτων εμπιστοσύνης.

Στη λογιστική παλινδρόμηση πρέπει να έχουμε υπόψην ότι η μέση απόκριση στο $\boldsymbol{x = x_i}$ δίνεται από τον τύπο $\frac{1}{1+e^{-x_i'\beta}}$ και άρα είναι πιθανότητα. Για παράδειγμα ένας μηχανικός πιθανόν να απαιτεί ένα 95% διάστημα εμπιστοσύνης για την πιθανότητα «ελαττωματικού» προϊόντος σε μια βιομηχανία όπου οι συνθήκες παραγωγής ορίζονται ως $\boldsymbol{x = x_i}$**.**

Η σημειακή εκτίμηση της πιθανότητας δίνεται από το $\widehat{y_i} = \hat{P}(\boldsymbol{x_i})$

Στο λογιστικό μοντέλο $P = \frac{1}{1+e^{-x_i\beta}}$ το P είναι μια μονότονη εξίσωση του $\boldsymbol{x'\beta}$. Μπορούμε να ορίσουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης στο P , χρησιμοποιώντας ένα διάστημα εμπιστοσύνης στο $\boldsymbol{x'\beta}$**.** Προφανώς ο linear predictor περιέχει όρους που είναι γραμμικοί στο $\boldsymbol{\beta}$ και μπορούμε να εκμεταλλευτούμε το γεγονός ότι ο $\boldsymbol{b}$ (εκτιμητής μέγιστης πιθανοφάνειας του **β)** είναι ασυμπτωτικά κανονικός. Συνεπώς , ένα άνω διάστημα εμπιστοσύνης για το $\boldsymbol{x'\beta}$, παράγει ένα άνω διάστημα για το P.

Ασυμπτωτικά ισχύει ότι

$$\boldsymbol{x'b} \sim N[\boldsymbol{x'\beta}, \boldsymbol{x'(X'VX)^{-1}x}]$$

Έτσι το διάστημα εμπιστοσύνης για το $\boldsymbol{x'\beta}$ δίνεται από τον τύπο:

$$\boldsymbol{x'b} \pm z_{\frac{\alpha}{2}}\sqrt{\boldsymbol{x'(X'VX)^{-1}x}}$$

Σε διάφορα παραδείγματα από τον τομέα της βιολογίας και της χημείας , όπου τα δεδομένα είναι ομαδοποιημένα και η $i-$ οστή παρατηρούμενη απόκριση y_i είναι διωνυμική με παραμέτρους $P(\boldsymbol{x_i})$ και n_i είναι πολύ ενδιαφέρον να υπολογίσουμε το διάστημα πρόβλεψης για το y_i. Για το σκοπό αυτό είναι απαραίτητη μια έκφραση για τη διακύμανσή του

$$\hat{P}(\boldsymbol{x_i}) = \frac{1}{1+e^{-\boldsymbol{x_i'b}}} \qquad i = 1,2,\dots,m$$

Μπορούμε επίσης να χρησιμοποιήσουμε τη μέθοδο Δέλτα , όπου προκύπτει η σχέση:

$$var\left[\hat{P}(\boldsymbol{x_i})\right] = \left(\frac{\partial \hat{P}(\boldsymbol{x_i})}{\partial \boldsymbol{b}}\right)' (\boldsymbol{X'VX})^{-1} \left(\frac{\partial \hat{P}(\boldsymbol{x_i})}{\partial \boldsymbol{b}}\right)$$

Μια πολύ σημαντική ιδιότητα της λογιστικής παλινδρόμησης θεωρείται η παρακάτω:

$$\frac{\partial P(\boldsymbol{x_i})}{\partial \boldsymbol{\beta}} = n_i[P(\boldsymbol{x_i})][1 - P(\boldsymbol{x_i})]\boldsymbol{x_i}$$

ή γενικότερα

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = [var(y_i)]\boldsymbol{x_i}$$

Από την παραπάνω σχέση προκύπτει:

$$var\left[\hat{P}(\boldsymbol{x_i})\right] = [var(y_i)]^2 \boldsymbol{x_i}'(\boldsymbol{X'VX})^{-1}\boldsymbol{x_i}$$

Έτσι το διάστημα πρόβλεψης μπορεί να βρεθεί όπως και σε όλα τα γραμμικά μοντέλα.

Κατ' αρχήν

$$\frac{y_i - \hat{P}(\boldsymbol{x_i})}{n_i[P(\boldsymbol{x_i})][1 - P(\boldsymbol{x_i})]\sqrt{1 + \boldsymbol{x_i}'(\boldsymbol{X'VX})^{-1}\boldsymbol{x_i}}} \sim N(0,1)$$

ασυμπτωτικά.

Επομένως ένα κατάλληλο $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το y_i , μπορεί να βρεθεί από τη σχέση:

$$\hat{P}(\boldsymbol{x_i}) \pm z_{\frac{\alpha}{2}}\{n_i[P(\boldsymbol{x_i})][1 - P(\boldsymbol{x_i})]\}\sqrt{1 + \boldsymbol{x_i}'(\boldsymbol{X'VX})^{-1}\boldsymbol{x_i}} \quad , \qquad i = 1,2,\dots,m$$

Φυσικά , στην πράξη πρέπει να αντικαταστήσουμε το $\hat{P}(\boldsymbol{x_i})$ στον πίνακα **V**.

2.3.5 Συμπερασματολογία με χρήση πιθανοφάνειας στη λογιστική παλινδρόμηση

Με τη συμπερασματολογία πιθανοφάνειας , μπορούμε να ενισχύσουμε τον έλεγχο υποθέσεων , χρησιμοποιώντας την $\log\ likelihood$. Η χρήση της μοιάζει αρκετά με τη χρήση της αρχής του επιπλέον αθροίσματος τετραγώνων (extra sum of squares principles) των γραμμικών μοντέλων. Για παράδειγμα , στα γραμμικά μοντέλα μπορούμε να χρησιμοποιήσουμε κάτω από τη μηδενική υπόθεση ένα μοντέλο ελαττωμένο (reduced model) , δηλαδή η μηδενική υπόθεση θέτει σε ένα υποσύνολο συντελεστών παλινδρόμησης την τιμή μηδέν. Ο έλεγχος χρησιμοποιεί τη διαφορά στο άθροισμα τετραγώνων του σφάλματος:

$$SS_E(reduced) - SS_E(full)$$

Η διαφορά στο άθροισμα τετραγώνων του σφάλματος αντικαθίσταται , στη λογιστική παλινδρόμηση , από τη διαφορά της log πιθανοφάνειας.

Ασυμπτωτικά ισχύει:

$$-2\ln\left[\frac{\mathcal{L}(reduced)}{\mathcal{L}(full)}\right] \sim {\chi_{\Delta}}^2$$

όπου:

το $\mathcal{L}(.)$ είναι η πιθανοφάνεια και στην περίπτωσή μας , ζητούμε την πιθανοφάνεια για το πλήρες και για το ελαττωμένο μοντέλο.

η παράμετρος Δ είναι η διαφορά στον αριθμό των παραμέτρων ανάμεσα στο πλήρες και το ελαττωμένο μοντέλο.

Υποθέτουμε ότι ο linear predictor είναι $\beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3$ και ενδιαφερόμαστε να εξετάσουμε την αρχική υπόθεση $H_0 : \beta_1 = \beta_2 = 0$

Το στατιστικό ελέγχου για το λόγο πιθανοφάνειας (likelihood ratio test statistic) δίνεται από τον τύπο:

$$2[\ln \mathcal{L}[b_0, b_1, b_2, b_3] - \ln \mathcal{L}[b_0^*, b_3^*]]$$

Όπου

$\mathcal{L}(b_0^*, b_3^*)$ είναι η πιθανοφάνεια για το λογιστικό μοντέλο στο οποίο έχουμε επικαλεστεί την μηδενική υπόθεση (δηλαδή $\beta_1 = \beta_2 = 0$)

Σαν αποτέλεσμα , η υπόθεση απορρίπτεται στην πρίπτωση που η log πιθανοφάνεια αυξηθεί σημαντικά εισάγοντας τα β_1, β_2 στο μοντέλο μαζί με τα β_0, β_3. Στην περίπτωσή μας , η κατανομή που χρησιμοποιείται για την άνω-ουρά (upper-tail) ενός μονόπλευρου ελέγχου είναι η χ_2^2. Άρα , η συμπερασματολογία με χρήση πιθανοφάνειας , χρησιμοποιείται για ελέγχους ενός σετ υποθέσεων.

2.3.6 Παράδειγμα εφαρμογής της λογιστικής παλινδρόμησης

Ο σκοπός του παραδείγματος είναι να μελετήσουμε τη χρήση της λογιστικής παλινδρόμησης για την ανάλυση της επίδρασης μιας ουσίας σε ένα πείραμα τοξικότητας. Ο παρακάτω πίνακας δείχνει την επίδραση διαφορετικών δόσεων νικοτίνης στην κοινή μύγα των φρούτων.

Συγκέντρωση x (g/100cc)	**Αριθμός εντόμων N**	**Αριθμός εντόμων που απεβίωσαν y**	**Ποσοστό**
0.10	47	8	17.0
0.15	53	14	26.4
0.20	55	24	43.6
0.30	52	32	61.5
0.50	46	38	82.6
0.70	54	50	92.6
0.95	52	50	96.2

Με χρήση της λογιστικής παλινδρόμησης θα καταλήξουμε σε ένα κατάλληλο μοντέλο και θα εκτιμήσουμε τις αποτελεσματικές δόσεις (ED) , τις τιμές δηλαδή της νικοτίνης που οδηγούν σε μια συγκεκριμένη τιμή πιθανότητας P. Τέτοιες ποσότητες χρησιμοποιούνται συχνά για να χαρακτηρίσουν τα αποτελέσματα μιας

πειραματικής διαδικασίας. Θα εκτιμήσουμε την ED_{50} , όπου ED_p είναι η τιμή του x για την οποία η πιθανότητα του θανάτου μιας μύγας των φρούτων παίρνει την τιμή P.

Το στατιστικό πακέτο (PROC LOGIST from SAS) δίνει τα ακόλουθα αποτελέσματα:

Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Wald chi-square	Pr>Chi-square	Standardized Estimate
INTERCPT	*1*	*-1.7361*	*0.2420*	*51.4482*	*0.0001*	
X	*1*	*6.2954*	*0.7422*	*71.9399*	*0.0001*	*1.024917*
INTERCPT	*1*	*3.1236*	*0.3349*	*86.9818*	*0.0001*	
LOGX	*1*	*2.1279*	*0.2214*	*92.3628*	*0.0001*	*0.898802*

Χρησιμοποιήθηκαν δυο λογιστικά μοντέλα με διαφορετική μορφή το καθένα για τον Linear predictor. Αρχικά χρησιμοποιήθηκε το τυπικό μοντέλο της εξίσωσης (1) με το τυπικό linear predictor $\beta_0 + \beta_1 x$. Ακόμη , χρησιμοποιήθηκε το μοντέλο

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \ln x)}}$$

Σε τέτοιου είδους πειράματα συχνά αντικαθιστούμε το x με το $\ln x$. Αυτό είναι ιδιαίτερα χρήσιμο όταν το x έχει μεγάλο εύρος τιμών. Οι $p - τιμές$ των παραμέτρων που παράχθηκαν από τα στατιστικά χ^2 του Wald ελέγχου είναι αρκετά σημαντικές και για τα δύο μοντέλα , έτσι έχουμε δύο υποψήφια μοντέλα. Μια μέθοδος για να συγκρίνουμε τα δύο μοντέλα είναι να συγκρίνουμε το εύρος των διαστημάτων εμπιστοσύνης γύρω από το $\hat{y}$ (Lewis , Montgomery and Myers 2001). Μια άλλη σχετική μέθοδος είναι να παρατηρήσουμε το τυπικό σφάλμα του εκτιμώμενου predictor $\boldsymbol{x'b}$ για τα δύο μοντέλα. Στον παρακάτω πίνακα παρουσιάζονται τα τυπικά σφάλματα των linear predictors.

$b_0 + b_1 x$	$b_0' + b_i' \ln x$
0.1844	0.2440
0.1607	0.1763
0.1428	0.1439
0.1336	0.1408
0.2139	0.2041
0.3432	0.2646
0.5194	0.3246

Στην περίπτωσή μας είναι δύσκολο να επιλέξουμε ανάμεσα στα δύο μοντέλα χρησιμοποιώντας τις πιο πάνω πληροφορίες , παρόλο που τα τυπικά σφάλματα είναι αρκετά μικρότερα για το log μοντέλο στις υψηλές δόσεις. Η χρήση των υπολοίπων (residuals) για την εξέταση αυτών των μοντέλων με τον ίδιο τρόπο που χρησιμοποιούνται στα συνηθισμένα γραμμικά μοντέλα θέλει προσοχή , καθώς τα υπόλοιπα δεν έχουν κοινή διακύμανση.

Υπολογίζουμε το ED_{50} χρησιμοποιώντας και τα δυο μοντέλα για το Linear predictor:

- Για το μοντελο $b_0 + b_1 x$, ο $\widehat{ED}_{50}$ δίνεται από την εξίσωση:

$$\widehat{ED}_{50} = \frac{b_0}{b_1}$$

Στο παράδειγμα ισούται με 0.277g/100cc.

- Για το μοντέλο $b_0' + b_1' \ln x$, το $\widehat{ED}_{50}$ δίνεται από την εξίσωση

$$\widehat{ED}_{50} = e^{-1.42} = 0.242g/100cc$$

ΚΕΦΑΛΑΙΟ 3: ΕΦΑΡΜΟΓΗ ΣΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ

3.1 Ορισμός του Credit scoring

Είναι ένα εξελιγμένο στατιστικά εργαλείο ή σύστημα που χρησιμοποιείται ευρέως από χρηματοοικονομικά ιδρύματα (κυρίως από Τράπεζες) στην προσπάθεια τους να κρίνουν αν θα εγκρίνουν ή να απορρίψουν ένα προσωπικό ή εταιρικό δάνειο. Καθορίζεται έτσι το ρίσκο που υπάρχει κατά τη διάρκεια της χορήγησης ενός δανείου.

Για να κατασκευαστεί ένα τέτοιο μοντέλο, θα πρέπει να αναλυθούν εις βάθος ιστορικά δεδομένα από πιο παλιά δάνεια που τυχόν να έκανε ο πελάτης αλλά και άλλες χρήσιμες πληροφορίες για τον πελάτη-δανειζόμενο οι οποίες μπορούν να ληφθούν από την αίτηση δανείου την οποία θα συμπληρώσει. Στοιχεία όπως το μηνιαίο εισόδημα του πελάτη, τα συνολικά έξοδα του, πόσο καιρό ο πελάτης εργάζεται στην ίδια δουλειά, αν ο πελάτης είναι ιδιοκτήτης σπιτιού ή ενοικιάζει, αν κατέχει δικό του αυτοκίνητο, αν οι λογαριασμοί που διατηρεί σε Τράπεζες δεν είναι χρεωστικοί είναι μερικά από τα στοιχεία που οι περισσότερες Τράπεζες ζητούν να μάθουν έτσι ώστε να δημιουργήσουν ένα «προφίλ» για το δανειζόμενο.

Στα περισσότερα μοντέλα (scoring systems), ένα ψηλό σκορ συνεπάγεται χαμηλό ρίσκο ενώ η Τράπεζα θέτει ένα όριο (το λεγόμενο cut off) βασιζόμενη στο βαθμό ρίσκου που είναι διαθετημένη να ανεχτεί. Είναι βέβαια στη δική της αρμοδιότητα και θέληση να αποδεχτεί πελάτες με σκορ πιο ψηλό από το όριο που έθεσε και να απορρίψει πελάτες με σκορ μικρότερο από το όριο.

3.1.1 Ιστορική Αναδρομή

Το σύστημα του Credit Scoring αρχικά γεννήθηκε σαν Credit Reporting και χρησιμοποιείτο από mail order organizations κυρίως για υποστήριξη πωλήσεων και

από εκδότες πιστωτικών και επαγγελματικών καρτών αφού οι υπηρεσίες και τα προϊόντα τους είχαν τα εξής χαρακτηριστικά:

- Μεγάλος όγκος λειτουργιών
- Ελάχιστη επαφή με τον πελάτη
- Συγκεντρωμένες διαδικασίες
- Χαμηλής αξίας συναλλαγές

Ακολούθως, χρησιμοποιείτο ευρέως από τους λιανικούς έμπορους (retail merchants) της εποχής οι οποίοι, στην προσπάθεια τους να δυναμώσουν τη σχέση και τη συνεργασία τους, αντάλλαζαν πληροφορίες για τους πελάτες τους. Έτσι είχαμε και τη δημιουργία των πρώτων credit bureaus.

Μετά το τέλος της Βιομηχανικής Επανάστασης, η ιδέα του Credit Reporting άρχισε να παίρνει την\ σημερινή του μορφή (Credit Scoring) και να εισάγεται δειλά-δειλά, στις Τράπεζες αφού επικρατούσαν τα εξής χαρακτηριστικά:

- Μικρός όγκος λειτουργιών
- Πολλές συναλλαγές με μεγάλα ποσά
- Σημαντική η επαφή με τον πελάτη

Τα τελευταία 15 χρόνια όμως οι σύγχρονες τάσεις της αγοράς που κάνει τις ανάγκες των ανθρώπων για δανεισμό από Χρηματοοικονομικά Ιδρύματα τεράστιες, έχει καταστήσει την εξάπλωση του credit scoring ραγδαία.

3.2 Πρόβλημα που καλείται να λύσει το credit scoring

Εδώ και αρκετά χρόνια διάφοροι οργανισμοί και εταιρείες, στην πλειοψηφία τους τράπεζες, προσπαθούν να υιοθετήσουν διάφορες αυτοματοποιημένες τεχνικές, οι οποίες θα τους βοηθήσουν τόσο στις αποφάσεις τους σε θέματα έγκρισης δανείων όσο και στην εξοικονόμηση χρόνου και κόστους. Το μεγαλύτερο πρόβλημα που

υπάρχει εδώ και χρόνια, είναι τα χρέη που δημιουργούνται στους χρηματοοικονομικούς οργανισμούς, λόγω της αδυναμίας διάφορων πελατών να ξεπληρώσουν τα δάνεια που τους παραχωρήθηκαν. Τρανό παράδειγμα αποτελεί το πρόσφατο θέμα χρεοκοπίας δύο μεγάλων τραπεζών της Αμερικής, που οφειλόταν στη συνεχή και απερίσκεπτη παραχώρηση δανείων, χωρίς την απαιτούμενη μελέτη, όσον αφορά την αξιοπιστία και δυνατότητα των πελατών να ξεπληρώσουν το δανειζόμενο ποσό. Για το σκοπό αυτό έχουν αναπτυχθεί διάφορα συστήματα τα οποία διαχωρίζουν τους πελάτες ως κακοπληρωτές ή καλοπληρωτές, βασισμένα σε παλαιότερες συναλλαγές τους. Με τον τρόπο αυτό, τα συστήματα παρέχουν περισσότερες πληροφορίες στους αρμόδιους για τα θέματα δανεισμού, βελτιώνοντας έτσι την ορθότητα των αποφάσεών τους.

3.2.1 Τεχνικές που χρησιμποιούνται

Πολλές τεχνικές χρησιμοποιούνται ευρέως για την υλοποίηση συστημάτων credit scoring. Από τις πιο διάσημες, είναι αυτές των νευρωνικών δικτύων και των δέντρων αποφάσεων και ταξινόμησης (decision/classification trees) οι οποίες παρουσιάζουν αρκετά καλά αποτελέσματα. Άλλες κοινές τεχνικές οι οποίες χρησιμοποιούνται ήδη, είναι οι στατιστικές μέθοδοι Discriminate analysis και Logistic regression, οι οποίες είναι αρκετά αξιόπιστες.

Τεχνικές, λιγότερο διαδεδομένες και με πιο μικρά ποσοστά επιτυχίας είναι οι γενετικοί αλγόριθμοι, γενετικός προγραμματισμός καθώς και data mining.
Δειλά-δειλά κάνουν την εμφάνιση τους και συστήματα βασισμένα στη Θεωρία της Ασαφούς Λογικής τα οποία όμως βρίσκονται ακόμη σε πιλοτικό στάδιο.

3.3 Συμπερασματικά

Ένα καλό scoring system δε θα προβλέψει με ακρίβεια τη συμπεριφορά του δανειζόμενου κατά τη διάρκεια της αποπληρωμής του δανείου του αλλά θα δώσει στην Τράπεζα μια πολύ καλή εικόνα για αυτή τη συμπεριφορά. Τουλάχιστον όμως,

όσο είναι δυνατόν, το σύστημα θα μειώσει το ρίσκο που παίρνει η Τράπεζα για να δανείσει χρήματα.

Λαμβάνοντας υπόψη και τις σύγχρονες τάσεις της αγοράς που αναφέρουν ότι ο πλανήτης μπήκε ήδη σε κλοιό παγκόσμιας κρίσης, οι Τράπεζες και γενικότερα οι Χρηματοοικονομικοί οργανισμοί, θα πρέπει να είναι πολύ προσεκτικοί πού και σε ποιο δανείζουν χρήματα. Ως εκ τούτου, η χρήση του συστήματος του credit scoring κρίνεται επιτακτική.

3.4 Περιγραφή της εφαρμογής

Η βάση δεδομένων που χρησιμοποιείται στην εφαρμογή αυτή έχει ληφθεί από το UCI Machine Learning Repository [Blake and Merz, 1998]. Τα δεδομένα , αφορούν στον πιστωτικό έλεγχο για δάνειο 1000 αιτητών σε μια γερμανική τράπεζα. Υπάρχουν 20 μεταβλητές-χαρακτηριστικά (7 αριθμητικές και 13 κατηγορηματικές) καθώς και ένα δυαδικό (binary) αποτέλεσμα (0 ή 1) . Από τις 1.000 παρατηρήσεις, 700 (ή 70,0%) έχουν μικρό πιστωτικό κίνδυνο (καλοπληρωτές=1) και 300 (ή 30,0%) έχουν μεγάλο πιστωτικό κίνδυνο (κακοπληρωτές=0). Τα 20 χαρακτηριστικά που είναι διαθέσιμα για την κατασκευή μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας, περιλαμβάνουν δημογραφικά χαρακτηριστικά όπως π.χ. το φύλο και η ηλικία καθώς και μερικά λογιστικά στοιχεία (π.χ. ιστορικό , το ποσό πίστωσης που ζητά ο αιτητής). Ας υποθέσουμε ότι η γερμανική τράπεζα επιθυμεί να αναπτύξει ένα μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας ώστε να προβλέψει τον πιστωτικό κίνδυνο των αιτούντων δανείου. Η τράπεζα προτίθεται να αναπτύξει το μοντέλο κατά το χρόνο επεξεργασίας των αιτήσεων για παραχώρηση δανείων. Η κατασκευή του μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας απαιτεί την κατασκευή μοντέλων πρόβλεψης. Για το σκοπό αυτό χρησιμοποιούνται, τρεις τεχνικές εξόρυξης δεδομένων: η λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα, και τα δέντρα απόφασης. Προκειμένου να εφαρμόσουμε τις μεθόδους αυτές στην παρούσα εργασία γίνεται χρήση του Clementine SPSS (ενός λογισμικού εξόρυξης δεδομένων).

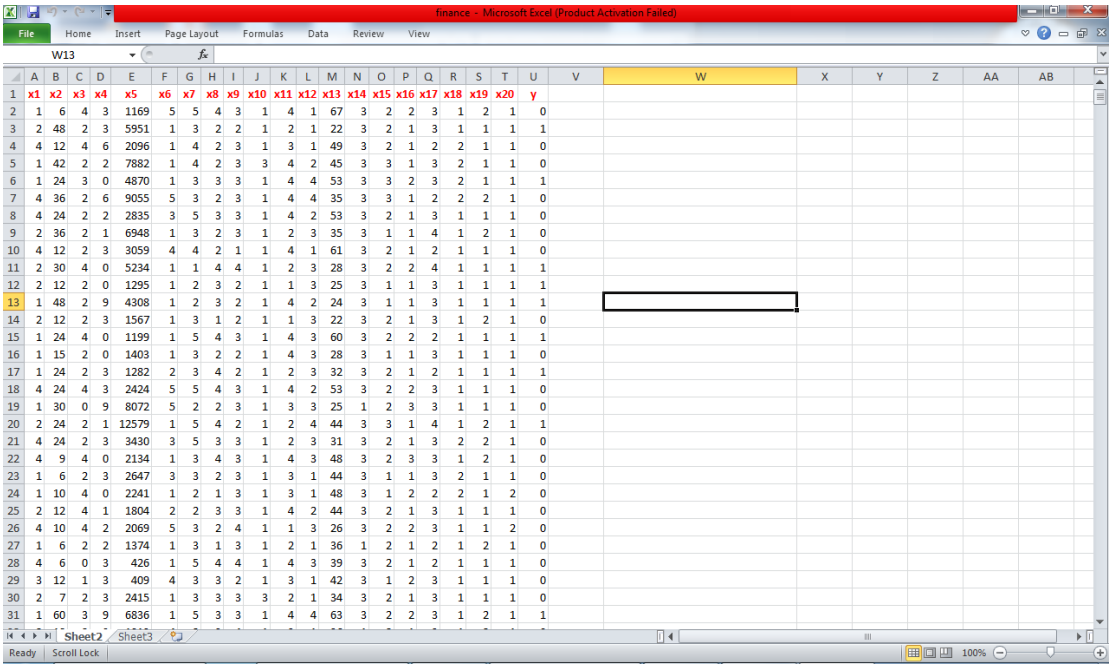
3.4.1 Εισαγωγή στο πρόβλημα

Όπως έχουμε ήδη προαναφέρει πρόκειται για μια βάση δεδομένων με 1000 εγγραφές και 20 επεξηγηματικές μεταβλητές οι οποίες δηλώνονται στον παρακάτω πίνακα:

X_1	**Ισοζύγιο τρεχουσών συναλλαγών** σε γερμανικά μάρκα (DM)
X_2	**Διάρκεια σε μήνες**
X_3	**Ιστορικό πελάτη**
X_4	**Σκοπός δανείου**
X_5	**Ποσό πίστωσης**
X_6	**Λογαριασμοί καταθέσεων / ταμιευτηρίου**
X_7	**Διάρκεια εργασίας** σε χρόνια
X_8	**Δόση δανείου σε ποσοστό του διαθέσιμου εισοδήματος**
X_9	**Συζυγική κατάσταση**
X_{10}	**Άλλοι πιστωτές / εγγυητές**
X_{11}	**Διάρκεια ζωής σε μόνιμη κατοικία** σε χρόνια
X_{12}	**Περιουσιακά στοιχεία που διαθέτει ο αιτητής**
X_{13}	**Ηλικία**
X_{14}	**Σχέδια αποπληρωμής δανείου**
X_{15}	**Στέγαση**
X_{16}	**Αριθμός προηγούμενων δανείων σε αυτή την τράπεζα**
X_{17}	**Απασχόληση**
X_{18}	**Πλήθος ατόμων που καλείται να συντηρήσει**
X_{19}	**Τηλέφωνο**
X_{20}	**Αλλοδαπός εργαζόμενος**

ΒΗΜΑΤΑ στο Clementine:

Αρχικά εισάγουμε τα δεδομένα στο πρόγραμμα μέσω ενός xls (excel) αρχείου.



x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19	x20	y
1	6	4	3	1169	5	5	4	3	1	4	1	67	3	2	2	3	1	2	1	0
2	48	2	3	5951	1	3	2	2	1	2	1	22	3	2	1	3	1	1	1	1
4	12	4	6	2096	1	4	2	3	1	3	1	49	3	2	1	2	2	1	1	0
1	42	2	2	7882	1	4	2	3	3	4	2	45	3	3	1	3	2	1	1	0
1	24	3	0	4870	1	3	3	3	1	4	4	53	3	3	2	3	2	1	1	1
4	36	2	6	9055	5	3	2	3	1	4	4	35	3	3	1	2	2	2	1	0
4	24	2	2	2835	3	5	3	3	1	4	2	53	3	2	1	3	1	1	1	0
2	36	2	1	6948	1	3	2	3	1	2	3	35	3	1	1	4	1	2	1	0
4	12	2	3	3059	4	4	2	1	1	4	1	61	3	2	1	2	1	1	1	0
2	30	4	0	5234	1	1	4	4	1	2	3	28	3	2	2	4	1	1	1	1
2	12	2	0	1295	1	2	3	2	1	1	3	25	3	1	1	3	1	1	1	1
1	48	2	9	4308	1	2	3	2	1	4	2	24	3	1	1	3	1	1	1	1
2	12	2	3	1567	1	3	1	2	1	1	3	22	3	2	1	3	1	2	1	0
1	24	4	0	1199	1	5	4	3	1	4	3	60	3	2	2	2	1	1	1	1
1	15	2	0	1403	1	3	2	2	1	4	3	28	3	1	1	3	1	1	1	0
1	24	2	3	1282	2	3	4	2	1	2	3	32	3	2	1	2	1	1	1	1
4	24	4	3	2424	5	5	4	3	1	4	2	53	3	2	2	3	1	1	1	0
1	30	0	9	8072	5	2	2	3	1	3	3	25	1	2	3	3	1	1	1	0
2	24	2	1	12579	1	5	4	2	1	2	4	44	3	3	1	4	1	2	1	1
4	24	2	3	3430	3	5	3	3	1	2	3	31	3	2	1	3	2	2	1	0
4	9	4	0	2134	1	3	4	3	1	4	3	48	3	2	3	3	1	2	1	0
1	6	2	3	2647	3	3	2	3	1	3	1	44	3	1	1	3	2	1	1	0
1	10	4	0	2241	1	2	1	3	1	3	1	48	3	1	2	2	2	1	2	0
2	12	4	1	1804	2	2	3	3	1	4	2	44	3	2	1	3	1	1	1	0
4	10	4	2	2069	5	3	2	4	1	1	3	26	3	2	2	3	1	1	2	0
1	6	2	2	1374	1	3	1	3	1	2	1	36	1	2	1	2	1	2	1	0
4	6	0	3	426	1	5	4	4	1	4	3	39	3	2	1	2	1	1	1	0
3	12	1	3	409	4	3	3	2	1	3	1	42	3	1	2	3	1	1	1	0
2	7	2	3	2415	1	3	3	3	2	1	1	34	3	2	1	3	1	1	1	0
1	60	3	9	6836	1	5	3	3	1	4	4	63	3	2	2	3	1	2	1	1

Στη συνέχεια τοποθετούμε ένα type node στο stream canvas με σκοπό να διαβαστούν οι τύποι των τιμών των πεδίων. Άρα καθορίζεται ο τύπος των δεδομένων (type) για κάθε πεδίο και επιπρόσθετα καθορίζεται η κατεύθυνση(direction) που επιδεικνύει το ρόλο που παίζει κάθε πεδίο στη μοντελοποίηση.

Αφού συγκεκριμενοποιήσαμε την πηγή των δεδομένων μας, το επόμενο βήμα πριν το data mining είναι να ορίσουμε τον τύπο της πληροφορίας για κάθε πεδίο των δεδομένων. Ο τύπος πληροφορίας για κάθε πεδίο πρέπει να τεθεί πριν τα πεδία χρησιμοποιηθούν στα διάφορα modeling nodes. Το Clementine διακρίνει τους εξής τύπους δεδομένων :

- *Εύρος (range)*: Το range χρησιμοποιείται για να περιγράψει συνεχείς αριθμητικές τιμές, ένα σύνολο ή μία κλίμακα 0-100 ή 0.75-1.25. ή μία τιμή range μπορεί να είναι ακέραιος, πραγματικός αριθμός ή ημερομηνία/ώρα.

- *Διακριτοποίηση (discrete)*: Το discrete χρησιμοποιείται για να περιγράψει αλφαριθμητικές τιμές όταν ένας ακριβής αριθμός διαφορετικών τιμών είναι άγνωστος (π.χ 1,5,8)
- *Δίτιμη* παράμετρος-λογική παράμετρος τύπου *Boolean (flag)*: Το flag χρησιμοποιείται από δεδομένα με δύο μόνο τιμές yes/no ή 0/1 ή 1/2.
- *Σύνολο (set)*: Το set χρησιμοποιείται για να περιγράψει δεδομένα με πολλαπλές διακεκριμένες τιμές όπου η καθεμιά αντιμετωπίζεται ως μονάδα ενός συνόλου, ή διακεκριμένες κατηγορίες όπως small/medium/large.
- *Ανένταχτος τύπος (typeless)*: Το typeless χρησιμοποιείται για δεδομένα που δεν εντάσσονται σε καμία από τις παραπάνω κατηγορίες ή για δεδομένα τύπου set με πάρα πολλές διακεκριμένες τιμές. Η επιλογή του τύπου typeless ορίζεται αυτόματα το πεδίο του direction se none, δηλαδή το πεδίο που δεν μπορεί να χρησιμοποιηθεί σε μοντέλα.

Τα δεδομένα, εντάσσονται, αρχικά, σε μία από τις παραπάνω κατηγορίες με την είσοδό τους στο σύστημα. Για παράδειγμα, ο discrete τύπος δίνεται προσωρινά σε κατηγορικές μεταβλητές μέχρι να μπορεί να προσδιορισθεί αν πρόκειται για set ή flag τύπο και ο τύπος range δίνεται σε όλες τις αριθμητικές μεταβλητές.

Η τιμή του direction ενός πεδίου σχετίζεται μόνο με τη μοντελοποίηση. Υπάρχουν τέσσερις δυνατές κατευθύνσεις :

- *IN* : το πεδίο χρησιμοποιείται σαν input, δηλαδή είναι μία τιμή που θα βοηθήσει στην πρόβλεψη
- *OUT:* το πεδίο χρησιμοποιείται σαν output-στόχος της τεχνικής μοντελοποίησης. Δηλαδή είναι το πεδίο που θα προβλέψουμε.
- *BOTH*: το πεδίο επιτρέπεται να είναι και input και output σε κανόνα συσχέτισης (association rule). Όλες οι άλλες τεχνικές μοντελοποίησης αγνοούν αυτό το πεδίο.
- *NONE*: το πεδίο δε χρησιμοποιείται στη μοντελοποίηση.

Επόμενο βήμα είναι η τοποθέτηση στο stream canvas ενός Partition node με σκοπό να χωριστούν τα δεδομένα σε δεδομένα εκπαίδευσης (training set), δεδομένα

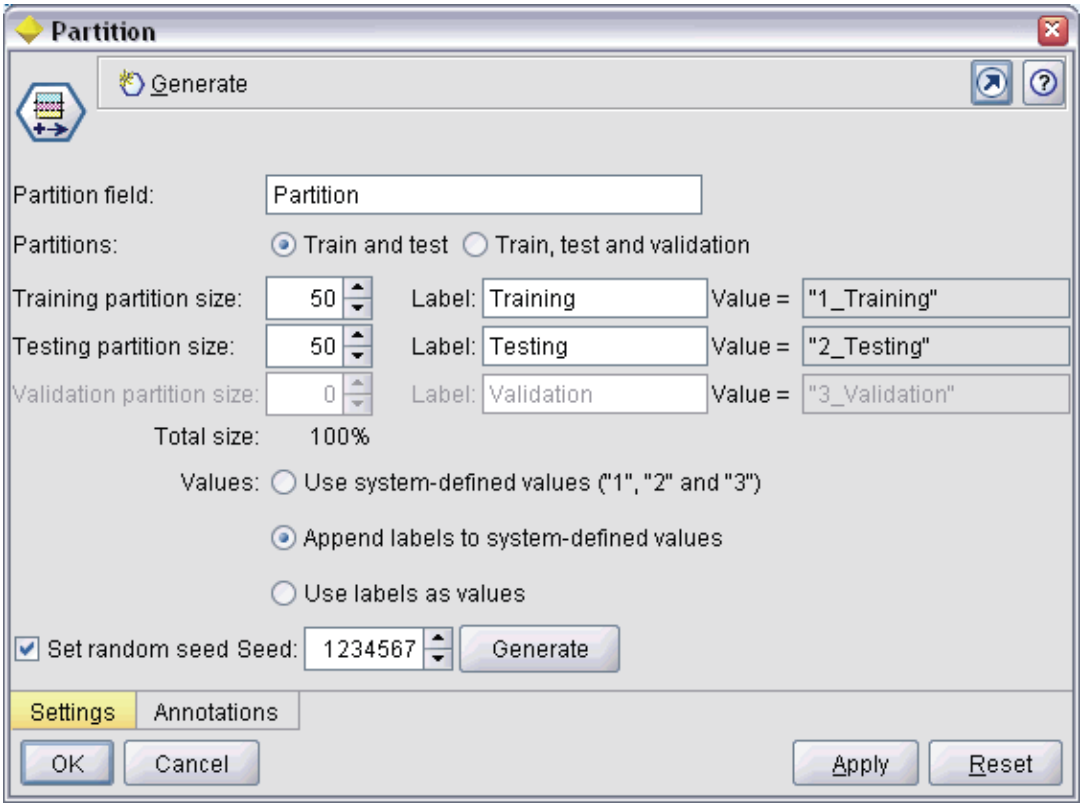
επαλήθευσης-επικύρωσης (quiz set-validation set) και δεδομένα ελέγχου-εξέτασης (test dataset).

Πιο αναλυτικά :

Σε ένα τυπικό πρόβλημα του data mining, έχουμε ένα σύνολο δεδομένων εκπαίδευσης (training set) στο οποίο γνωρίζουμε την τιμή του αποτελέσματος και τις τιμές των χαρακτηριστικών που μας ενδιαφέρουν, και προσπαθούμε με βάση αυτά τα δεδομένα να κατασκευάσουμε ένα μοντέλο πρόβλεψης. Το μοντέλο αυτό θα το χρησιμοποιήσουμε στη συνέχεια για να προβλέψουμε το αποτέλεσμα νέων συνόλων δεδομένων εξέτασης (test set), στα οποία σύνολα είναι γνωστές οι τιμές των χαρακτηριστικών αλλά δεν είναι γνωστή η τιμή του αποτελέσματος, δηλαδή η τιμή της τάξης. Στην περίπτωση τώρα όπου ο αλγόριθμος που εφαρμόζουμε στηρίζεται σε κατασκευή και εκτίμηση μοντέλου, τα δεδομένα διαχωρίζονται σε τρία υποσύνολα: 1) τα δεδομένα εκπαίδευσης(training data) τα οποία χρησιμοποιούνται για την προσαρμογή του μοντέλου, 2) τα δεδομένα επικύρωσης(validation data) που χρησιμοποιούνται για την εκτίμηση του σφάλματος πρόβλεψης για την επιλογή του μοντέλου και 3) τα δεδομένα ελέγχου (test data) που χρησιμοποιούνται για τον υπολογισμό της γενικευμένης τιμής σφάλματος του τελικά επιλεγμένου μοντέλου. Καθένα από αυτά τα σύνολα θα πρέπει να επιλεγεί ανεξάρτητα.

Εφαρμόζουμε λοιπόν την παραπάνω διάκριση στις δικές μας 1000 εγγραφές και προκύπτουν:

- Training set: $\approx 80\% \times 1000 = 800$
- Test set: $\approx 20\% \times 1000 = 200$



Εικόνα: Creating a partition in the data

Μετά το τέλος της διαδικασίας διαχωρισμού (partition) του συνόλου των δεδομένων προχωρούμε στο επόμενο βήμα που είναι η τοποθέτηση στο stream canvas ενός Feature selection node τον οποίο συνδέουμε στον Partition node με σκοπό να επιλεγούν , για επίπεδο σημαντικότητας α=0.05 , οι σημαντικές μεταβλητές

Πιο αναλυτικά :

Στο δικό μας πρόβλημα εξόρυξης δεδομένων, όπως συμβαίνει και στην πλειοψηφία των προβλημάτων εξόρυξης δεδομένων, εμπεριέχονται πολλά πεδία-μεταβλητές τα οποία είναι πιθανόν να χρησιμοποιηθούν με σκοπό την πρόβλεψη. Σαν αποτέλεσμα, χρειάζεται να ξοδευτεί αρκετός χρόνος και προσπάθεια για να εξεταστεί ποια από αυτά τα πεδία πρέπει να συμπεριληφθούν στο μοντέλο. Για να μειώσουμε στο ελάχιστο τις πιθανές επιλογές, ο αλγόριθμος της επιλογής των χαρακτηριστικών (Feature Selection Algorithm) μπορεί να χρησιμοποιηθεί για να προσδιορίσει τα πεδία εκείνα τα οποία είναι πιο σημαντικά για την δεδομένη

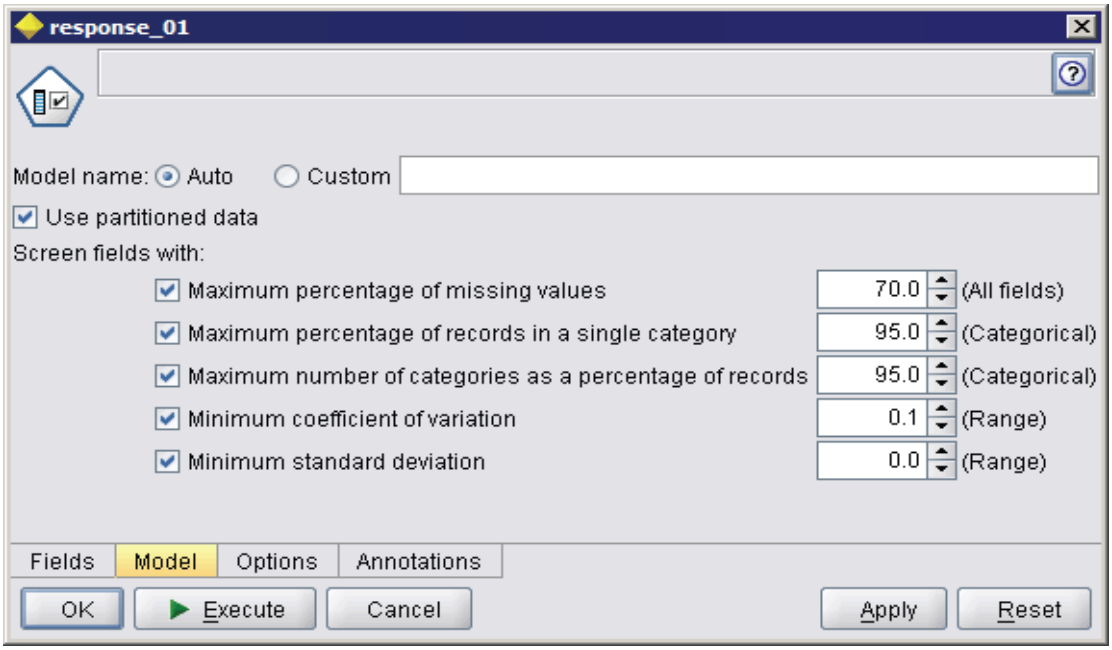
ανάλυση. Στην δική μας εφαρμογή προσπαθούμε να προβλέψουμε την ικανότητα αποπληρωμής δανείων από τους διάφορους πελάτες-αιτητές ούτως ώστε να γίνει δεκτή , (εαν ο πελάτης κριθεί ως «καλοπληρωτής») ή αντίστοιχα να απορριφθεί (εαν ο πελάτης κριθεί ως «κακοπληρωτής») , η χορηγεία δανείου. Επομένως πρέπει να βρούμε τους παράγοντες-μεταβλητές που δείχνουν πιο σημαντικοί από τους υπόλοιπους.

Η επιλογή χαρακτηριστικών αποτελείται από τρία βήματα :

- *Screening* (κρισάρισμα): Σε αυτό το βήμα απομακρύνονται οι μη σημαντικές και προβληματικές μεταβλητές πρόβλεψης καθώς και εγγραφές, όπως στην περίπτωση που έχουμε μεταβλητές με πολλές ελλείπουσες τιμές ή μεταβλητές με πολύ μεγάλη ή πολύ μικρή διακύμανση για να τις καθιστά χρήσιμες.
- *Ranking* (Στοίχιση): Σε αυτό το βήμα ξεχωρίζονται οι εναπομείνασες μεταβλητές πρόβλεψης και καθορίζονται ranks βασισμένα στην σημαντικότητα.
- *Επιλογή*: Σε αυτό το βήμα αναγνωρίζεται το υποσύνολο των χαρακτηριστικών που θα χρησιμοποιηθεί στα μοντέλα που ακολουθούν κρατώντας μόνο τις πιο σημαντικές μεταβλητές πρόβλεψης και φιλτράροντας ή αποκλείοντας όλες τις υπόλοιπες.

Τα πλεονεκτήματα από την επιλογή χαρακτηριστικών είναι ότι η διαδικασία της μοντελοποίησης απλοποιείται και φυσικά γίνεται ταχύτερη. Μειώνοντας τον αριθμό των πεδίων που χρησιμοποιούνται στο μοντέλο, μειώνεται ο χρόνος αξιολόγησης του μοντέλου και επιπρόσθετα αποκτούμε απλούστερα, ακριβέστερα μοντέλα τα οποία μπορούν πολύ πιο εύκολα να εξηγηθούν .

Model tab-Options tab: Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα model tab το οποίο περιλαμβάνει βασικές επιλογές για το μοντέλο καθώς και ρυθμίσεις που επιτρέπουν την εύρεση κριτηρίων για το κρισάρισμα των μεταβλητών πρόβλεψης.



Εικόνα: Model tab

Model name : (auto) το όνομα του μοντέλου παράγεται αυτόματα.

Use partitioned data : (ν) το σημειώνουμε διότι αυτή η επιλογή μας εξασφαλίζει ότι χρησιμοποιούνται τα δεδομένα μόνο από το το training set για την κατασκευή του μοντέλου.

Τα πεδία κρισάρονται με την βοήθεια των παρακάτω κριτηρίων :

- Μέγιστο ποσοστό ελλειπουσών τιμών: Κρισάρει τα πεδία με μεγάλο αριθμό ελλειπουσών τιμών που προσφέρουν ελάχιστη πληροφορία πρόβλεψης
- Μέγιστο ποσοστό εγγραφών σε μια απλή κατηγορία: Κρισάρει τα πεδία τα οποία έχουν πάρα πολλές εγγραφές να ανήκουν στην ίδια κατηγορία , για παράδειγμα το 95% των πελατών-αιτητών στην βάση δεδομένων να είναι άντρες παντρεμένοι, μιας και το να συμπεριληφθεί αυτή η πληροφορία δεν είναι χρήσιμη για να ξεχωρίσουμε τον ένα πελάτη από τον άλλο.
- Μέγιστος αριθμός των κατηγοριών ως ποσοστό των εγγραφών: Κρισάρει τα πεδία με πολλές κατηγορίες συγκριτικά με τον συνολικό αριθμό των εγγραφών δηλαδή εάν ένα μεγάλο ποσοστό των κατηγοριών περιέχει μόνο μία περίπτωση, το πεδίο δεν μπορεί παρά να χρησιμοποιηθεί ελάχιστα.

- Ελάχιστος συντελεστής διακύμανσης: Κρισάρει τα πεδία με συντελεστή βάρους μικρότερο ή ίσο από το καθορισμένο ελάχιστο όριο. Εάν η τιμή είναι κοντά στο 0 , δεν υπάρχει μεγάλη μεταβλητότητα στις τιμές της μεταβλητής.
- Ελάχιστη τυπική απόκλιση: Κρισάρει τα πεδία με τυπική απόκλιση μικρότερη ή ίση από το καθορισμένο ελάχιστο όριο.

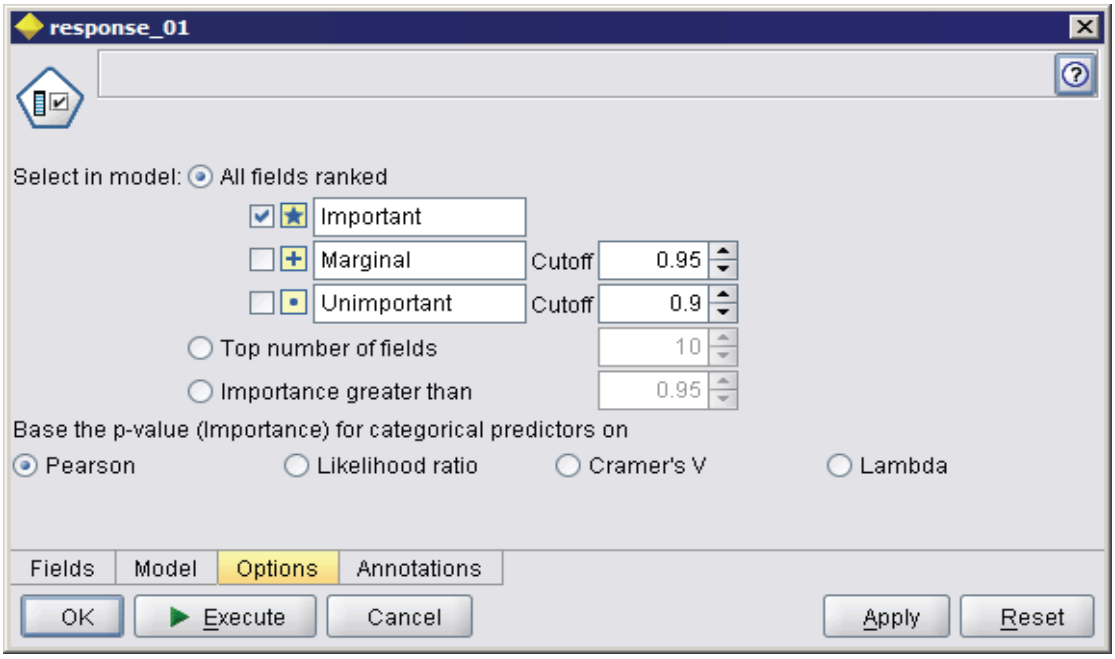
Οι εγγραφές οι οποίες έχουν ελλείπουσες τιμές για το πεδίο στόχου ή ελλείπουσες τιμές για όλες τις μεταβλητές πρόβλεψης , αποκλείονται αυτόματα από όλους τους υπολογισμούς μέσα στα rankings.

Στον κόμβο της επιλογής χαρακτηριστικών υπάρχει ένα options tab το οποίο σου επιτρέπει να καθορίσεις τις default (εξ'ορισμού) ρυθμίσεις για την επιλογή ή τον αποκλεισμό των πεδίων πρόβλεψης του μοντέλου.

Σε αυτό το βήμα θεωρείται μία μεταβλητή πρόβλεψης τη φορά για να εξεταστεί, πόσο καλά προβλέπει κάθε μεταβλητή πρόβλεψης ξεχωριστά τη μεταβλητή στόχο. Οι μεταβλητές πρόβλεψης ιεραρχούνται σύμφωνα με το κριτήριο που καθορίζεται από τον πειραματιστή.

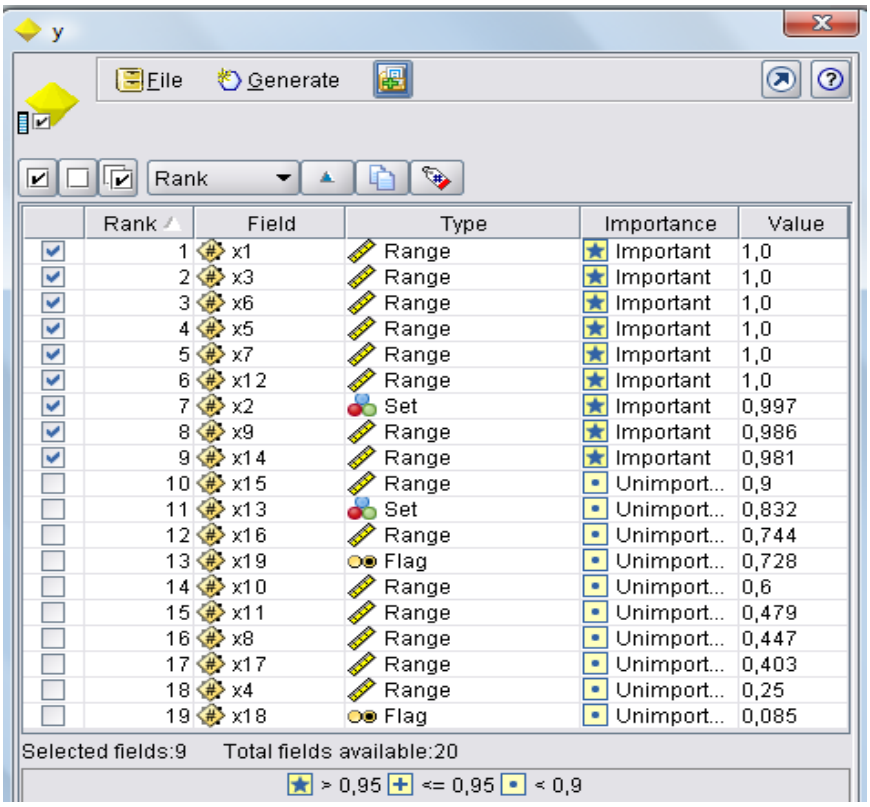
Η τιμή σημαντικότητας κάθε μεταβλητής ή διαφορετικά ένα μέτρο το οποίο χρησιμοποιείται για να βάλει σε σειρά τα πεδία ή τα αποτελέσματα σε ποσοστιαία κλίμακα ορίζεται ως $(1-p)$ όπου p είναι η τιμή $p-value$ του κατάλληλου στστιστικού τεστ της σχέσης μεταξύ της υποψήφιας μεταβλητής πρόβλεψης και της μεταβλητής στόχου .

Στην\ δική μας εφαρμογή χρησιμοποιήσαμε τιμή p value βασισμένη στο στατιστικό του Pearson , το Pearson chi-square το οποίο εξετάζει την ανεξαρτησία του στόχου και της μεταβλητής πρόβλεψης χωρίς να δείχνει τη δύναμη ή την κατεύθυνση οποιασδήποτε υπάρχουσας σχέσης .



Εικόνα: Feature Selection Options tab

Μετά από όλη αυτή την διαδικασία επιλογής μεταβλητών παρατηρούμε ότι οι επεξηγηματικές μας μεταβλητές από 20 που ήταν αρχικά μειώνονται στις 9. Για επίπεδο σημαντικότητας α=0.05 αυτές οι 9 σημαντικές μεταβλητές με p value =1,0 παρουσιάζονται παρακάτω :



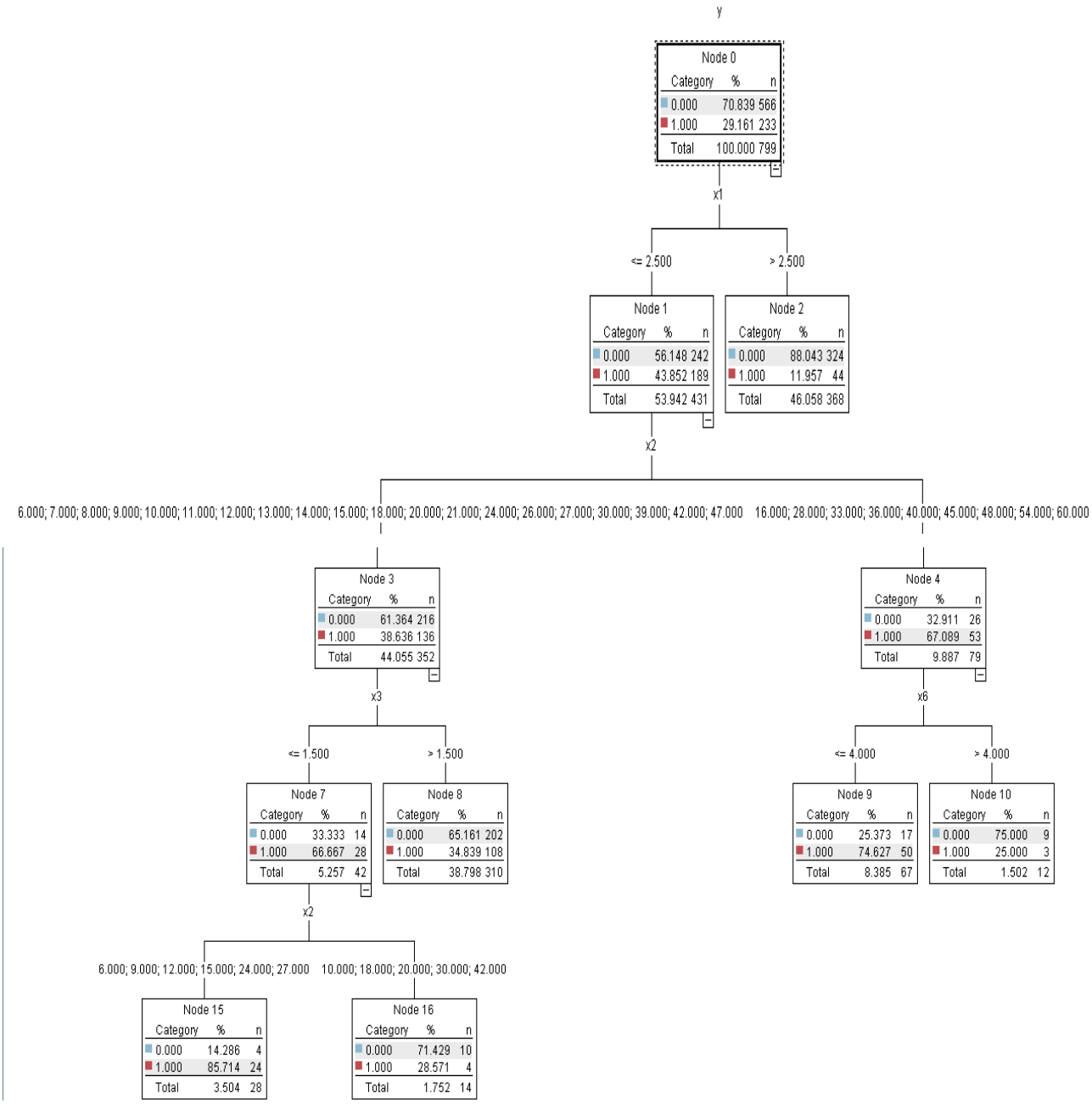
Εικόνα: Feature Selection model results

3.4.1.1 C&RT

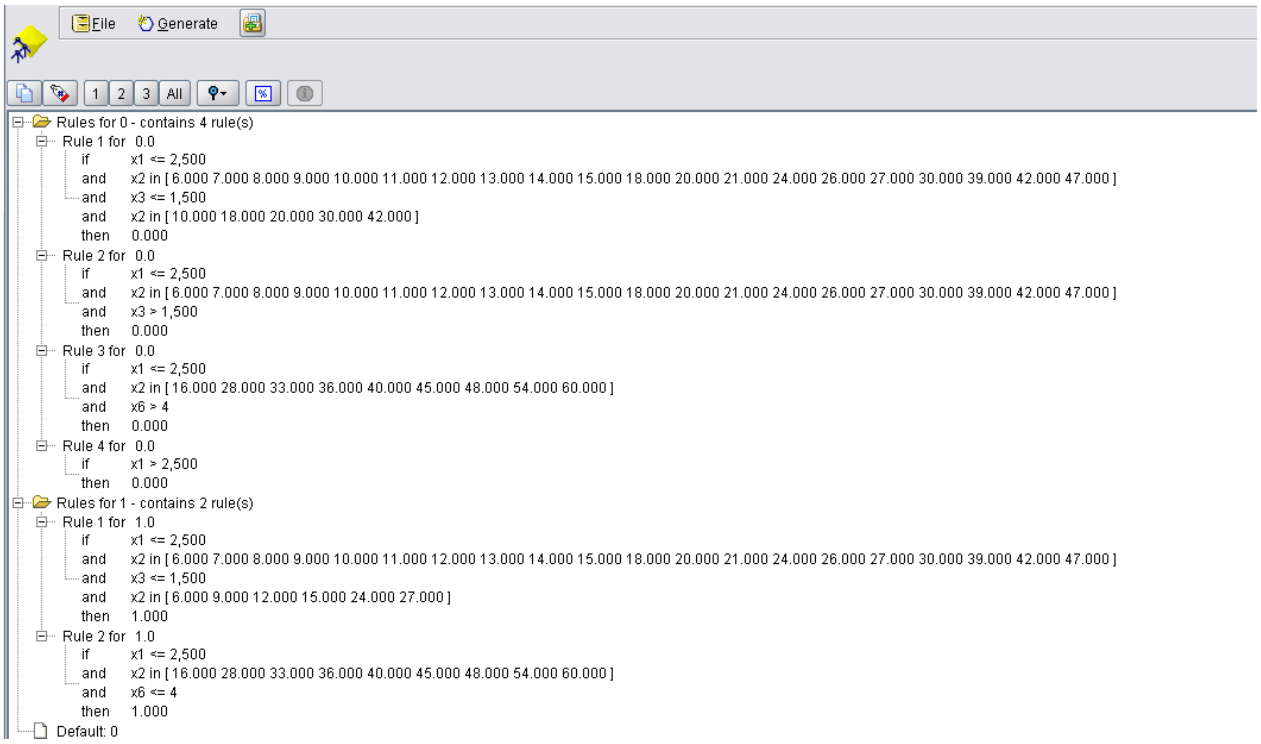
Προχωρώντας , το επόμενο βήμα είναι η τοποθέτηση στο stream canvas ενός C&R Tree node.

Συνδέουμε ένα C&R Tree node στο Feature selection model . Ο Classification and Regression (C&R) Tree node παράγει ένα δέντρο απόφασης το οποίο επιτρέπει στον πειραματιστή να προβλέψει ή να ταξινομήσει τις μελλοντικές παρατηρήσεις. Η μέθοδος χρησιμοποιεί επαναληπτικό διαμερισμό για να διασπάσει τις εγγραφές από το training set σε τμήματα ελαχιστοποιώντας τη μη καθαρότητα σε κάθε βήμα. Ένας κόμβος θεωρείται καθαρός εάν το 100% των περιπτώσεων στον κόμβο βρίσκονται μέσα σε μία συγκεκριμένη κατηγορία του πεδίου στόχου.

Χρησιμοποιώντας το κριτήριο Gini προκύπτει το παρακάτω δέντρο:



Προκειμένου τα παραπάνω αποτελέσματα να γίνουν πιο κατανοητά παραθέτουμε τους κανόνες ταξινόμησης του συγκεκριμένου δέντρου ,



όπου

- X_1, X_2, X_3, X_6 επεξηγηματικές μεταβλητές που έχουν δηλωθεί παραπάνω
- με 0.000 συμβολίζονται οι καλοπληρωτές
- με 1.000 συμβολίζονται οι κακοπληρωτές

Για την αξιολόγηση του αλγόριθμου (με χρήση του Gini κριτηρίου) έχουμε τα παρακάτω αποτελέσματα:

Results for output field y

Comparing $R-y with y

'Partition'	1_Training		2_Testing	
Correct	619	77,47%	145	72,14%
Wrong	180	22,53%	56	27,86%
Total	799		201	

Coincidence Matrix for $R-y (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	545	21
1.000000	159	74

'Partition' = 2_Testing	0.000000	1.000000
0.000000	126	8
1.000000	47	19
$null$	1	0

Performance Evaluation

'Partition' = 1_Training	
0.000000	0,089
1.000000	0,983

'Partition' = 2_Testing	
0.000000	0,083
1.000000	0,762

Για την εκτίμηση της λειτουργίας μιας διαγνωστικής διαδικασίας χρησιμοποιούνται τα μέτρα ευαισθησία, ειδικότητα, θετική προγνωστική αξία (Θ.Π.Α) και αρνητική προγνωστική αξία (Α.Π.Α). Ωστόσο και σε πολλές εφαρμογές εξόρυξης δεδομένων το κριτήριο αξιολόγησης είναι η ακρίβεια (accuracy) δηλαδή το ποσοστό των σωστά ταξινομημένων περιπτώσεων που προκύπτουν από τον αλγόριθμο, όπως αναφερθήκαμε και παραπάνω.

Μέτρα αξιολόγησης διαγνωστικών τεστ:

Διαγνωστικό τεστ	**Αποτέλεσμα**	**Αποτέλεσμα**	
	Κακοπληρωτές(+)	Καλοπληρωτές(-)	Σύνολο
Θετικό (+)	a (πραγματικά θετικά)	b (λανθασμένα θετικά)	a + b
Αρνητικό (-)	c (λανθασμένα αρνητικά)	d (πραγματικά αρνητικά)	c+d
Σύνολο	a + c	b + d	a+b+c+d

Όπου :

- ευαισθησία = $\frac{a}{a+c}$
- ειδικότητα = $\frac{d}{b+d}$
- θετική προγνωστική αξία (Θ.Π.Α) = $\frac{a}{a+b}$
- αρνητική προγνωστική αξία (Α.Π.Α) = $\frac{d}{d+c}$
- ακρίβεια = $\frac{a+d}{a+b+c+d}$

Θα υπολογίσουμε τα μέτρα αυτά αξιολόγησης στα σύνολα εκπαίδευσης, και ελέγχου ξεχωριστά και θα τα απεικονίσουμε αναλυτικά σε μορφή ποσοστών .
Για το σύνολο εκπαίδευσης έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	545	21
1	Κακοπληρωτές (+)	159	74

Άρα a=(+ +)= πραγματικά θετικά = 74

b = (+ -) =λανθασμένα θετικά = 159

c =(- +) =λανθασμένα αρνητικά= 21

d = (- -) = πραγματικά αρνητικά= 545

και τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 77,89 (%)
- ειδικότητα =77,41 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 31,76 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 96,29 (%)
- ακρίβεια = 74,47 (%)

Για το *σύνολο ελέγχου* έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	126	8
1	Κακοπληρωτές (+)	47	19

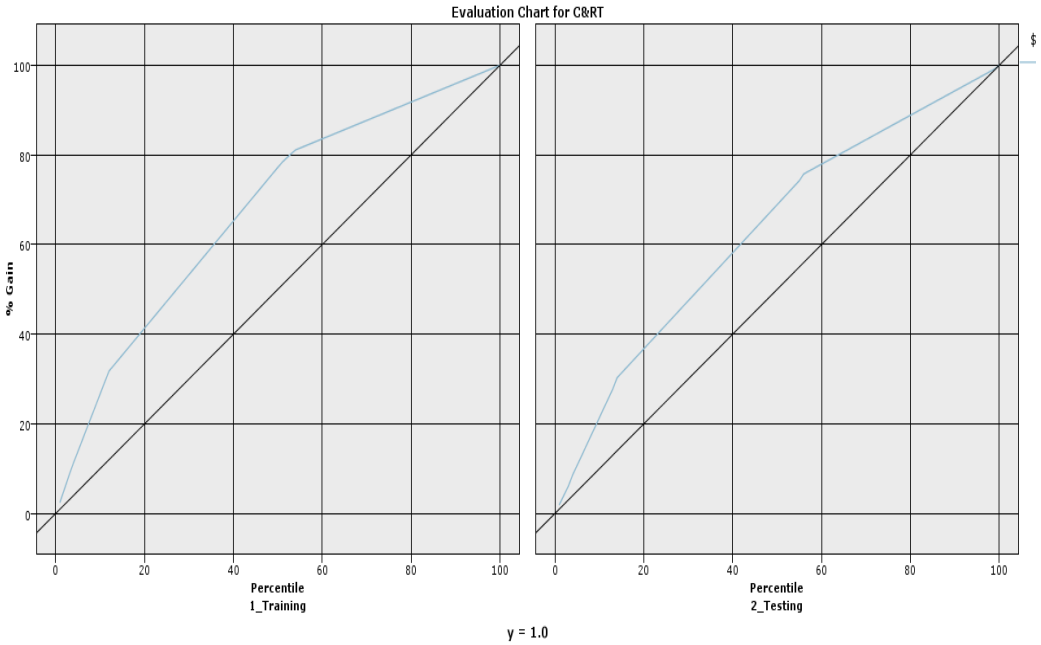
Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 70,37 (%)
- ειδικότητα = 72,83 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 28,79 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 94,03 (%)
- ακρίβεια = 72,14 (%)

Επόμενο μας βήμα είναι η κατασκευή των διαγραμμάτων αξιολόγησης (evaluation charts)τα οποία μπορούν εύκολα να συγκρίνουν τις επιδόσεις των διαφόρων μοντέλων. Στόχος είναι να δημιουργήσουμε ένα γράφημα για τα κέρδημε βάση τις προβλέψεις από το δέντρο απόφασης. Η κατηγορία στόχος από προεπιλογή θα έχει

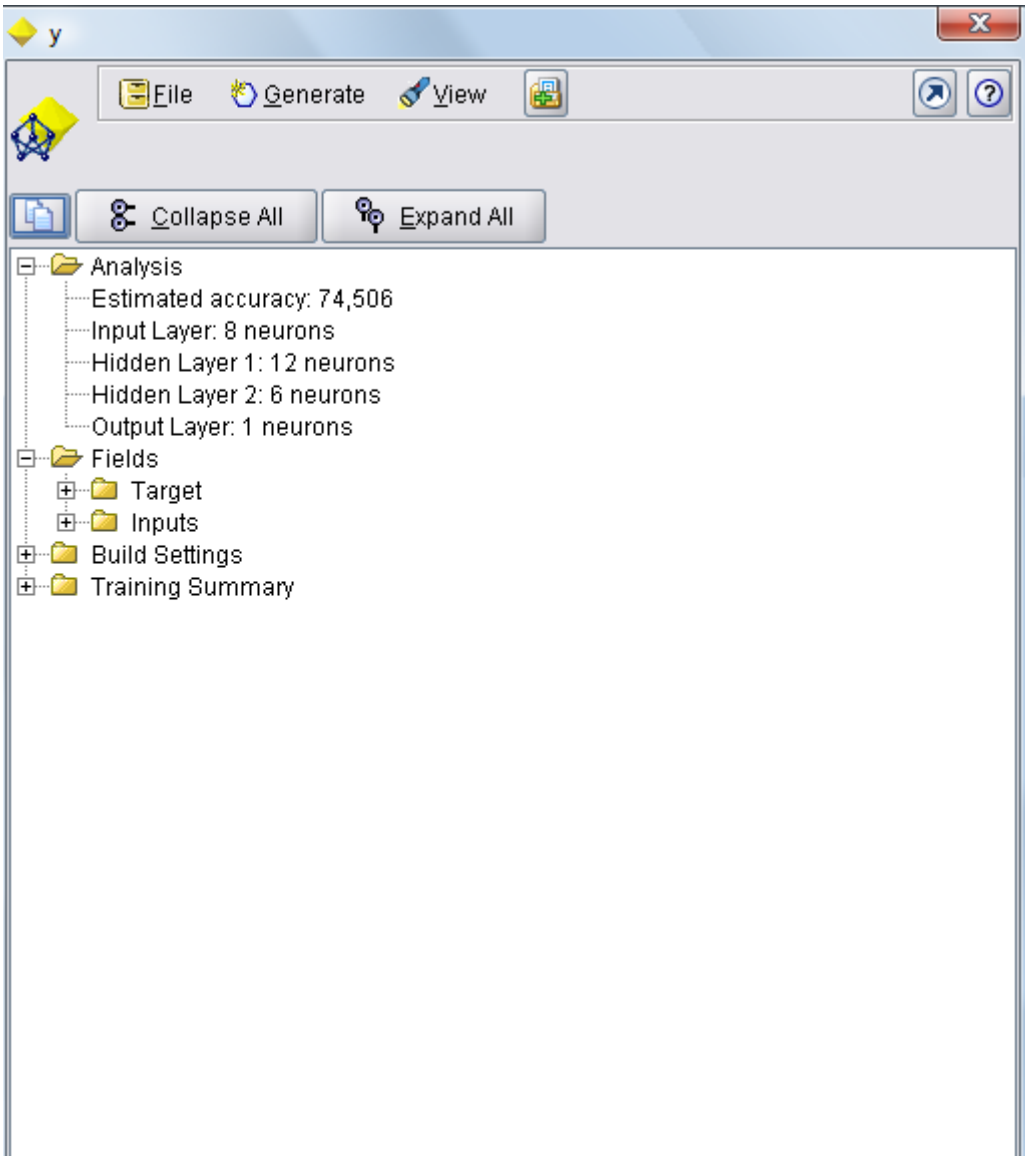
την τιμή 1 (κακοπληρωτές). Για κάθε κομμάτι του συνόλου δεδομένων (partition) θα πρέπει να γίνεται ξεχωριστό διάγραμμα αξιολόγησης.

Θα δούμε παρακάτω ένα evaluation chart για τα δεδομένα μας όπου εμφανίζονται σε κάθε διάγραμμα ξεχωριστά μια βασική γραμμή (baseline) και μια καμπύλη.



3.4.1.2 Νευρωνικά Δίκτυα (χρήση της μεθόδου MLP)

Με τον ίδιο ακριβώς τρόπο που έχουμε περιγράψει , αρχικά την κατασκευή και στη συνέχεια τα αποτελέσματα της ανάλυσης και αξιολόγησης ενός δέντρου απόφασης, θα προχωρήσουμε τώρα στην προσαρμογή ενός νευρωνικού δικτύου με την μέθοδο MLP στο Feature selection model.



Για την αξιολόγηση του αλγόριθμου (με χρήση του Gini κριτηρίου) έχουμε τα παρακάτω αποτελέσματα:

Results for output field y

Comparing $N-y with y

'Partition'	1_Training		2_Testing	
Correct	610	76,35%	150	74,63%
Wrong	189	23,65%	51	25,37%
Total	799		201	

Coincidence Matrix for $N-y (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	496	70
1.000000	119	114

'Partition' = 2_Testing	0.000000	1.000000
0.000000	120	14
1.000000	36	30
$null$	0	1

Performance Evaluation

'Partition' = 1_Training	
0.000000	0,13
1.000000	0,754

'Partition' = 2_Testing	
0.000000	0,143
1.000000	0,708

Για το σύνολο εκπαίδευσης έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	496	70
1	Κακοπληρωτές (+)	119	114

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 61,96 (%)
- ειδικότητα = 80,65 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 48,93 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 87,63 (%)
- ακρίβεια = 76,35 (%)

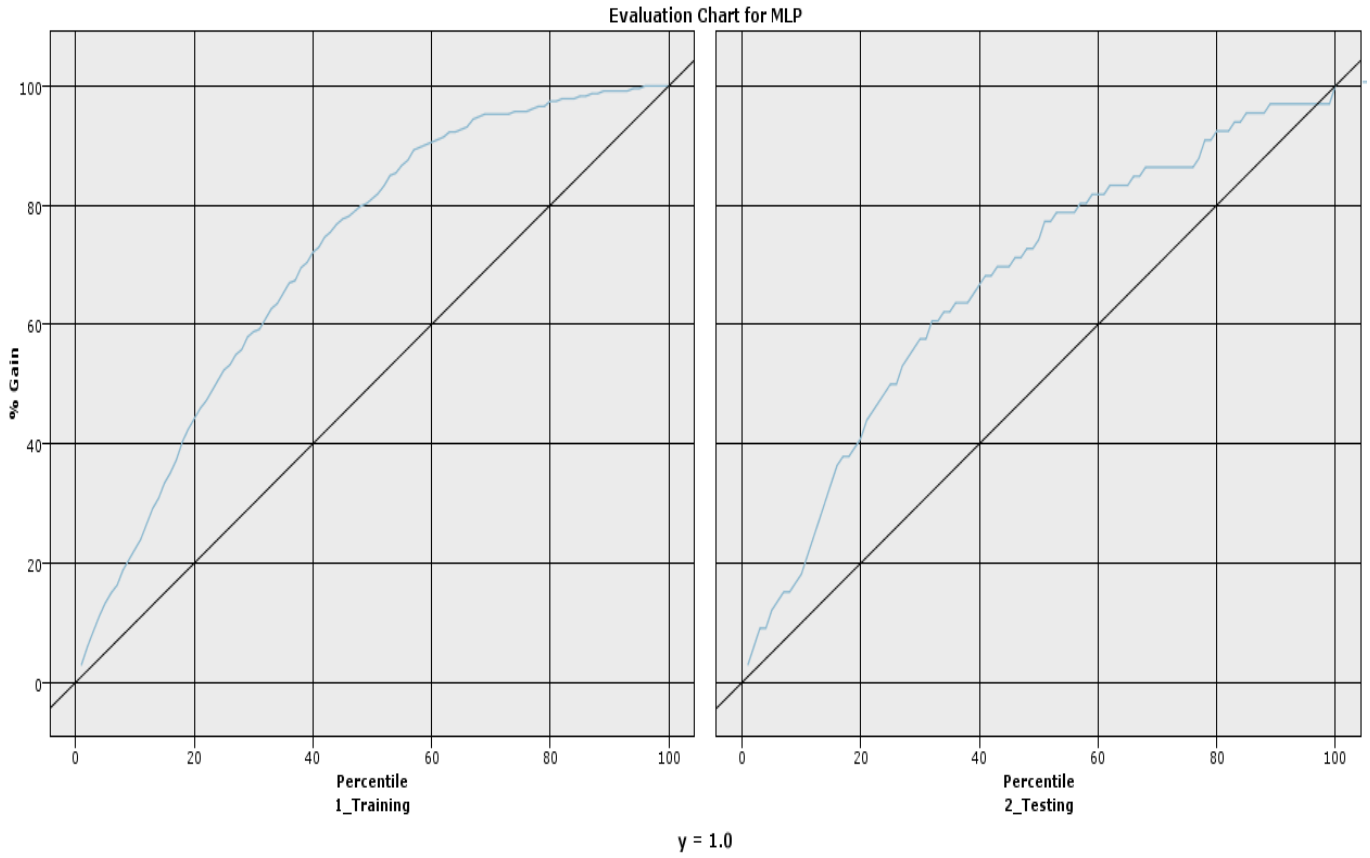
Για το σύνολο ελέγχου έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	120	14
1	Κακοπληρωτές (+)	36	30

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 68,18 (%)
- ειδικότητα = 76,92 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 45,45 (%)
- αρνητική προγνωστική αξία (Α.Π.Α)= 89,55 (%)
- ακρίβεια = 76,63 (%)

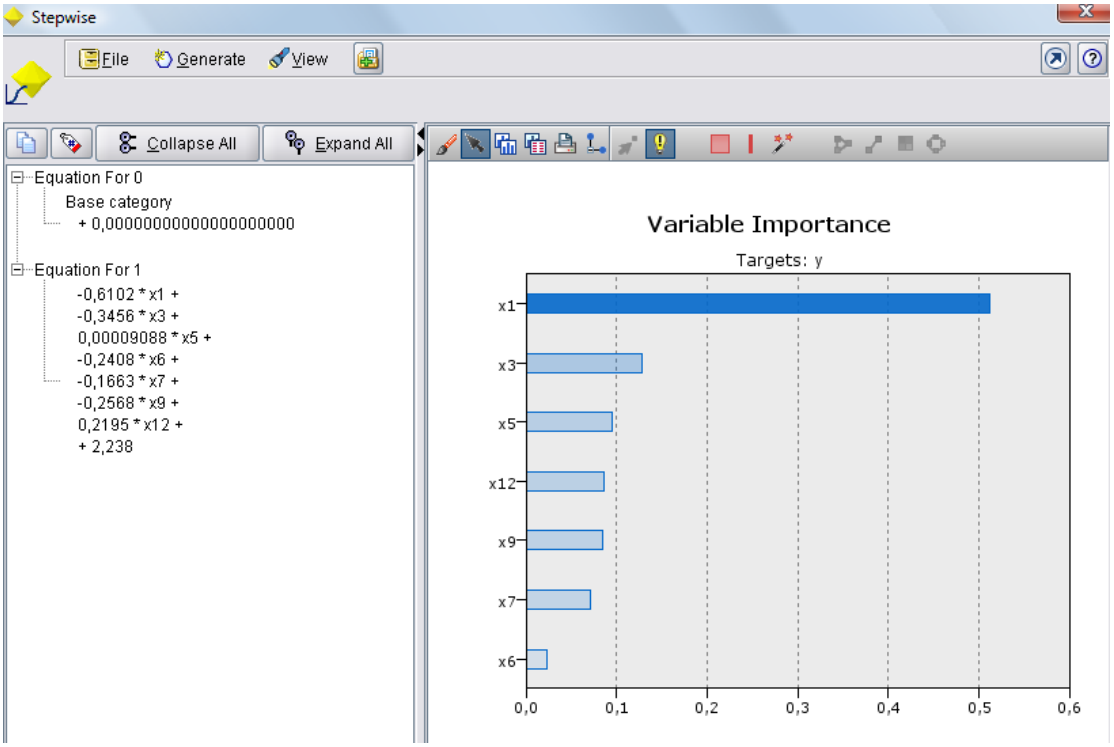
Ακολουθεί τώρα το evaluation chart για τον MLP αλγόριθμο:



3.4.1.3 Λογιστική Παλινδρόμηση (stepwise , forwards , backwards)

Stepwise:

Ένας τρίτος και τελευταίος αλγόριθμος που χρησιμοποιούμε στη συγκεκριμένη εφαρμογή είναι αυτός της Λογιστικής παλινδρόμησης (Logistic Regression).



Η παραπάνω εικόνα είναι το output που προκύπτει αφού εφαρμόσουμε τη μέθοδο stepwise στην επιλογή των σημαντικών μεταβλητών του μοντέλου. Στην αριστερή στήλη εμφανίζονται οι εξισώσεις του μοντέλου ενώ στη δεξιά στήλη απεικονίζεται η σημαντικότητα-συμβολή των διαφόρων μεταβλητών στο μοντέλο.

Όπως έχουμε αναφέρει υπάρχουν διάφορες μέθοδοι για την επιλογή των ανεξάρτητων μεταβλητών. Υπάρχουν δύο μέθοδοι σταδιακής (stepwise) επιλογής:

Η μέθοδος προς τα εμπρός και η μέθοδος προς τα πίσω. Οι stepwise μέθοδοι μπορούν να χρησιμοποιήσουν είτε το στατιστικό Wald, το λόγο πιθανοφάνειας, ή έναν υπό συνθήκη αλγόριθμο αφαίρεσης μεταβλητών. Για τις δύο μεθόδους ,

λαμβάνεται υπόψη η P-τιμή στην επιλογή μεταβλητών για την είσοδο τους στο μοντέλο.

Παρουσιάζουμε τώρα την αξιολόγηση (coincidence matrices και evaluation charts) των μεθόδων.

Results for output field y

Comparing $L-y with y

'Partition'	1_Training		2_Testing	
Correct	611	76,47%	145	72,14%
Wrong	188	23,53%	56	27,86%
Total	799		201	

Coincidence Matrix for $L-y (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	507	59
1.000000	129	104

'Partition' = 2_Testing	0.000000	1.000000
0.000000	121	13
1.000000	42	24
$null$	1	0

Performance Evaluation

'Partition' = 1_Training	
0.000000	0,118
1.000000	0,783

'Partition' = 2_Testing	
0.000000	0,101
1.000000	0,681

Για το σύνολο εκπαίδευσης έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	507	59
1	Κακοπληρωτές (+)	129	104

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 63,80 (%)
- ειδικότητα = 79,72 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 44,64 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 89,58 (%)
- ακρίβεια = 76,47 (%)

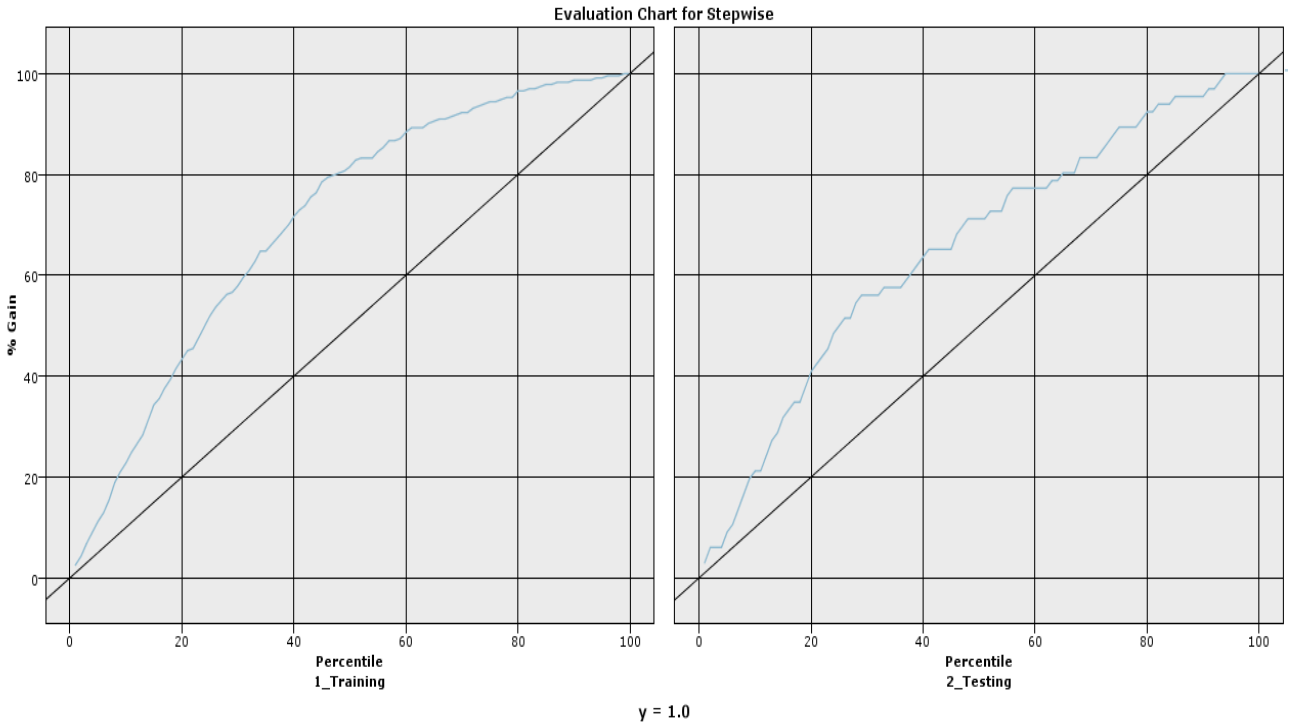
Για το σύνολο ελέγχου έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	121	13
1	Κακοπληρωτές (+)	42	24

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 64,86 (%)
- ειδικότητα = 74,23 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 36,36 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 90,3 (%)
- ακρίβεια = 72,14 (%)

Ακολουθεί το διάγραμμα αξιολόγησης (evaluation chart) της μεθόδου stepwise



Backwards:

Results for output field y

Comparing $L-y with y

'Partition'	1_Training		2_Testing	
Correct	610	76,35%	148	73,63%
Wrong	189	23,65%	53	26,37%
Total	799		201	

Coincidence Matrix for $L-y (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	504	62
1.000000	127	106

'Partition' = 2_Testing	0.000000	1.000000	$null$
0.000000	120	14	0
1.000000	37	28	1
$null$	1	0	0

Performance Evaluation

'Partition' = 1_Training	
0.000000	0,12
1.000000	0,772
'Partition' = 2_Testing	
0.000000	0,13
1.000000	0,708
	-0,005

Για το σύνολο εκπαίδευσης έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	504	62
1	Κακοπληρωτές (+)	127	106

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 63,1 (%)
- ειδικότητα = 79,87 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 45,49 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 89,05 (%)
- ακρίβεια = 76,35 (%)

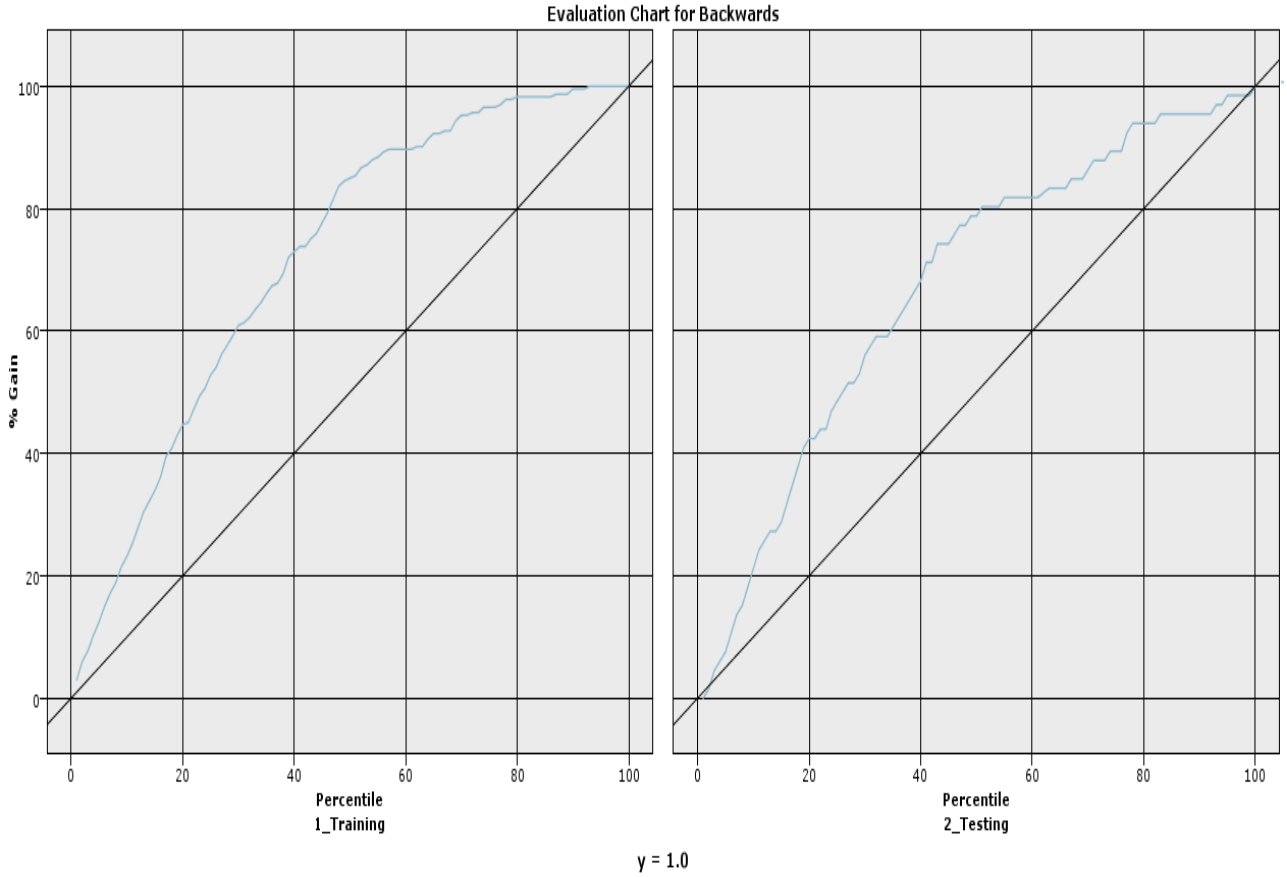
Για το σύνολο ελέγχου έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	120	14
1	Κακοπληρωτές (+)	37	28

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 66,67 (%)
- ειδικότητα = 76,43 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 43,08 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 89,55 (%)
- ακρίβεια = 73,63 (%)

Ακολουθεί το διάγραμμα αξιολόγησης (evaluation chart) της μεθόδου backwards



Forwards:

Results for output field y

Comparing $L-y with y

'Partition'	1_Training		2_Testing	
Correct	611	76,47%	145	72,14%
Wrong	188	23,53%	56	27,86%
Total	799		201	

Coincidence Matrix for $L-y (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	507	59
1.000000	129	104

'Partition' = 2_Testing	0.000000	1.000000
0.000000	121	13
1.000000	42	24
$null$	1	0

Performance Evaluation

'Partition' = 1_Training	
0.000000	0,118
1.000000	0,783

'Partition' = 2_Testing	
0.000000	0,101
1.000000	0,681

Για το σύνολο εκπαίδευσης έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	507	59
1	Κακοπληρωτές (+)	129	104

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 63,80 (%)
- ειδικότητα = 79,72 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 44,64 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 89,58 (%)
- ακρίβεια = 76,47 (%)

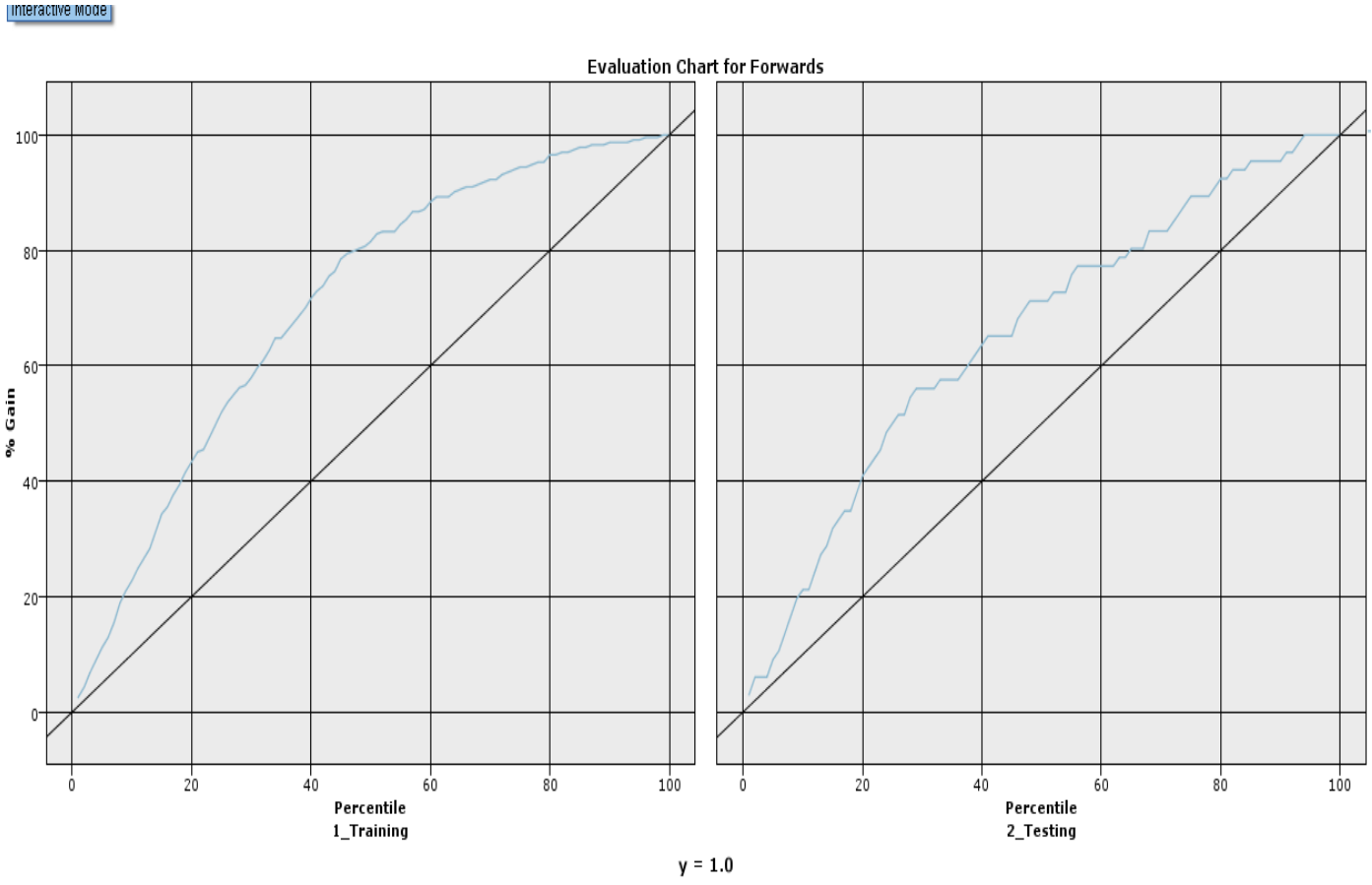
Για το σύνολο ελέγχου έχω :

		0 (-)	1 (+)
0	Καλοπληρωτές (-)	121	13
1	Κακοπληρωτές (+)	42	24

Τα ποσοστά των μέτρων % είναι αναλυτικά :

- ευαισθησία = 64,86 (%)
- ειδικότητα = 74,23 (%)
- θετική προγνωστική αξία (Θ.Π.Α) = 36,36 (%)
- αρνητική προγνωστική αξία (Α.Π.Α) = 90,3 (%)
- ακρίβεια = 72,14 (%)

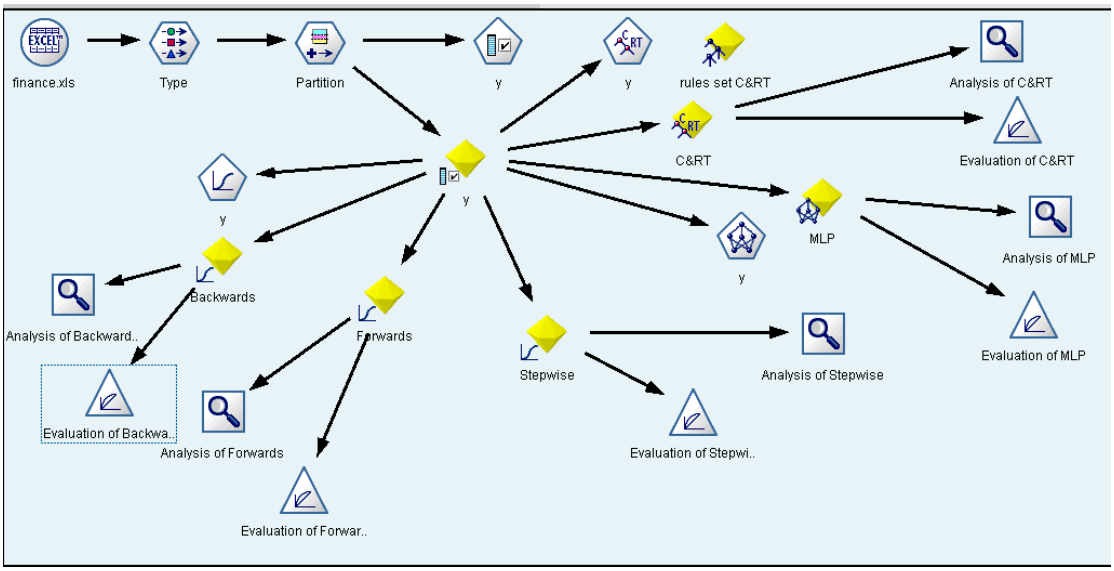
Ακολουθεί το διάγραμμα αξιολόγησης (evaluation chart) της μεθόδου forwards



3.5 ΣΥΝΟΨΗ

Για να συνοψίσουμε, αρχικά εισάγαμε στο πρόγραμμα το σύνολο των δεδομένων μας που αποτελείτο από 1000 εγγραφές-καταχωρήσεις και στη συνέχεια το χωρίσαμε σε δείγμα εκπαίδευσης και δείγματα ελέγχου (περίπου σε ποσοστό 80% έως 20%, αντίστοιχα). Στη συνέχεια, κατασκευάστηκαν μοντέλα βαθμολόγησης πιστοληπτικής ικανότητας με τα δεδομένα του δείγματος εκπαίδευσης χρησιμοποιώντας τις μεθόδους (λογιστικής παλινδρόμησης , νευρωνικών δικτύων και δέντρων απόφασης)και τα δεδομένα ελέγχου για έλεγχο των αποτελεσμάτων.

Η τελική εικόνα του stream canvas μετά την εφαρμογή των αλγορίθμων με την βοήθεια του Clementine είναι :



3.6 ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία εξετάζει και παρουσιάζει τη χρήση των τεχνικών εξόρυξης γνώσης για την κατασκευή μοντέλων βαθμολόγησης. Τα τελευταία χρόνια, η εξόρυξη γνώσης έχει αποκτήσει τεράστια σημασία και έγινε ευρέως γνωστή στον εμπορικό κόσμο. Εκτός από την πιστωτική βαθμολόγηση (credit scoring), υπάρχουν

και άλλες πιθανές εφαρμογές εξόρυξης γνώσης για τις επιχειρήσεις. Για παράδειγμα, μπορεί να χρησιμοποιηθεί ώστε: (1) να κατασκευάσει μοντέλα ανίχνευσης της απάτης τα οποία θα δίνουν έγκαιρη προειδοποίηση για δόλιες συναλλαγές, που πιθανών να γίνονται (2) να εντοπίσει τις προτιμήσεις των καταναλωτών και των πελατών (π.χ., μέσω ανάλυσης καλαθιού της αγοράς), (3) να κατηγοριοποιήσει τους διάφορους πελάτες (π.χ., μέσω ομαδοποίησης), ή (4) να κατασκευάσει μοντέλα τα οποία προβλέπουν την πιθανότητα να αγοράσουν ορισμένα προϊόντα ή υπηρεσίες για τη διευκόλυνση των πωλήσεων . Τα ευρήματα της ανάλυσης, μπορούν να χρησιμοποιηθούν, για παράδειγμα, στην προετοιμασία ηλεκτρονικών καταλόγων, στις διαφημίσεις, στις εκστρατείες προώθησης, κ.λπ.

Περιορισμοί στο Credit Scoring:

Παρόλο που το credit scoring έχει σημαντικά πλεονεκτήματα, οι περιορισμοί στη χρήση του πρέπει να σημειωθούν. Ένα από τα σημαντικότερα προβλήματα που μπορεί να προκύψουν κατά την κατασκευή ενός μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας είναι ότι το μοντέλο μπορεί να κατασκευαστεί χρησιμοποιώντας ένα δείγμα προκατειλημμένων καταναλωτών και πελατών στους οποίους έχει χορηγηθεί πίστωση [Hand, 2001]. Αυτό μπορεί να συμβεί, διότι οι αιτούντες (δηλαδή, οι πιθανοί πελάτες), οι οποίοι έχουν απορριφθεί δεν θα πρέπει να περιλαμβάνονται στα στοιχεία για την κατασκευή του μοντέλου. Ως εκ τούτου, το δείγμα θα είναι προκατειλημμένο (δηλαδή, θα διαφέρει από το γενικό πληθυσμό), καθώς οι καλοί πελάτες θα εκπροσωπούνται σε υψηλό βαθμό.. Το μοντέλο βαθμολόγησης πιστοληπτικής ικανότητας έχει κατασκευαστεί χρησιμοποιώντας αυτό το δείγμα το οποίο μπορεί να μην αποδίδει καλά στο σύνολο του πληθυσμού και τα δεδομένα που χρησιμοποιήθηκαν για την κατασκευή του μοντέλου είναι διαφορετικά από αυτά που θα εφαρμοστούν στο μοντέλο.

Το δεύτερο πρόβλημα που μπορεί να προκύψει κατά την κατασκευή μοντέλων βαθμολόγησης πιστοληπτικής ικανότητας είναι η αλλαγή των προτύπων κατά την πάροδο του χρόνου. Η βασική υπόθεση για κάθε προγνωστική μοντελοποίηση είναι ότι το παρελθόν μπορεί να προβλέψει το μέλλον [Berry και Linoff, 2000]. Αυτό

σημαίνει ότι τα χαρακτηριστικά των παλιών αιτητών, οι οποίοι έπειτα ταξινομούνται ως "καλοί" ή "κακοί" πιστωτές, μπορούν να χρησιμοποιηθούν για την πρόβλεψη της πιστοληπτικής ικανότητας των νέων υποψηφίων. Μερικές φορές, η τάση για την κατανομή των χαρακτηριστικών αλλάζει με την πάροδο του χρόνου και η αλλαγή είναι τόσο γρήγορη που απαιτεί μια σχετικά συνεχή ανανέωση του μοντέλου βαθμολόγησης πιστοληπτικής ικανότητας.

Παρά τους περιορισμούς που επισημαίνονται παραπάνω, δεν υπάρχει καμία αμφιβολία ότι η βαθμολόγηση πιστοληπτικής ικανότητας θα συνεχίσει να είναι ένα σημαντικό εργαλείο για την πρόβλεψη του πιστωτικού κινδύνου στον τομέα των καταναλωτών. Αναμένεται ότι οι οργανισμοί που χρησιμοποιούν σωστά το credit scoring θα αποκτήσουν σημαντικό στρατηγικό πλεονέκτημα και ανταγωνιστικό πλεονέκτημα έναντι των αντιπάλων τους.

Βιβλιογραφία

1. Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

2. Breiman, L. , Friedman, J.H. , Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

3. Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, Vol. 66, pp. 429-436.

4. Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, second edition, New York : Springer-Verlac

5. Collet, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall/CRC, Boca Raton.

6. Cundiff, K. (2004). Closing the loop: How credit scoring drives performance improvements along the financial value chain*, Business Credit,* 106(3), 38-42.

7. Diana, T. (2005). Credit risk analysis and credit scoring – *now and in the future, Business Credit,* 107(3), 12-16.

8. Faraggi, D. and Simon, R. (1995). *A Neural Network Model for Survival Data. Statistics in medicine*, Vol. 14, pp. 73-82.

9. Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for Data Mining researchers*, Intelligent Enterprise Technologies Laboratory*.

10. Fensterstock, A. (2005). Credit scoring and the next step*, Business Credit,* 107(3), 46-49.

11. Fieller, N. (2007). *Medical Statistics: Survival Data, Course Booklet*. Department of Probability & Statistics, University of Sheffield.

12. Goh, C.P. , Koh, H.C. and Tan, W.C. (*2006)*. A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques *, International Journal of Business and Information, Volume 1, No. 1, 96-118.*

13. Gurney, K. (1997). *An introduction to Neural Networks*, University of Sheffield.

14. Hand, D.J. (2001). Modelling consumer credit risk, *IMA Journal of Management Mathematics* 12(1) 139-155.

15. Hauck, W.W. and Donner, A. (1997). Wald's test applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, Vol. 72, pp. 851-853.

16. Haykin, S. (1999). *Neural Networks, a comprehensive foundation*. McMaster University Hamilton, Ontario, Canada, second edition.

17. Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*, John Wiley and Sons, New York, second edition.

18. Kamath, C. (2009). *Scientific Data Mining, a Practical Perspective*. Philadelphia: Society for Industrial and Applied Mathematics.

19. Kawaguchi, K. (2000). *A multithreaded software model for backpropagation neural network applications*. [Chp. 2.4.4].

20. Larose, D.T (2005). *Discovering Knowledge in Data. An Introduction to Data Mining*, John Wiley and Sons, Hoboken, New Jersey.

21. Liestol, K. , Andersen, P.K. and Andersen, U. (1994). Survival analysis and neural nets*, Statistics in Medecine*, Vol. 13, pp. 1189-1200.

22. Marubini, E. and Valecchi, G.M (2004). *Analysing Survival Data from Clinical Trials and Observational Studies*.

23. Menard, S.W. *Applied Logistic Regression Analysis*. Second Edition, Sage Publications, London.

24. Noriega, L. (2005). *Multilayer Perceptron Tutorial*. School of Computing, Staffordshire University.

25. Pardalos, P.M. , Boginski, V.L. and Vazacopoulos, A. (2007). *Data Mining in Biomedicine*, Springer, New York.

26. Pearson, R.L (1983). Karl Pearson and the chi-squared test, *International Statistical Review*, 51, 59-72.

27. Rojas, R. (1996). *Neural Networks*. Springer-Verlag, Berlin.

28. Suka, M. , Shinichi, O., Ichimura, T., Yoshida, K. and Takezawa J. (2004). Comparison of Proportional Hazard Model and Neural Network Models in a real data set of intensive care unit patients. *Studies in Health Technology and Informatics*, Vol. 107, pp. 741-745.

29. T. Hastie, T. , Tibshirani, R. and Friedman, J. (2001*), The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York.

30. Zhang, T. , Ramakrishnan, R. and Livny, M. (1996). An effcient data clustering method for very large databases, In Proc. of the ACM SIGMOD *International, Conference on Management of Data (SIGMOD)*, 103-114.

31. Zhou , X.H. , Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*. Chapter 2, Wiley, New York.

32. Καρώνη, Χ. (2005). Μοντέλα Αξιοπιστίας και Επιβίωσης. Ε.Μ.Π.

33. Ρίζος, Γ. (1996). *Τεχνητά νευρωνικά δίκτυα*. Αθήνα: Εκδόσεις Νέων Τεχνολογιών.