ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

ΠΡΟΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΥΡΙΑΚΗ Ι. ΙΩΑΝΝΙΔΟΥ

Επιβλέπων: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2011

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ

ΠΡΟΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΚΥΡΙΑΚΗ Ι. ΙΩΑΝΝΙΔΟΥ

Επιβλέπων: ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2011

ΠΡΟΛΟΓΟΣ

Στη διπλωματική αυτή εργασία περιγράφεται η μέθοδος της διαχωριστικής ανάλυσης. Εξετάζεται και συγκρίνεται με άλλες γνωστές μεθόδους ταξινόμησης.

Ολοκληρώνοντας τη διπλωματική εργασία, θα ήθελα να ευχαριστήσω θερμά τον κ. Χ. Κουκουβίνο, επιβλέποντα καθηγητή, για την ανάθεση της διπλωματικής εργασίας, την καθοδήγηση και την υποστήριξη του κατά τη διάρκεια της εκπόνησης της. Επίσης, θα ήθελα να ευχαριστήσω την υποψήφια διδάκτωρ κ. Χ. Παρπούλα για το πολύτιμο συμβουλευτικό έργο, την αμέριστη βοήθεια της, την πλούσια βιβλιογραφία που μου παρείχε και τις χρήσιμες παρατηρήσεις της, καθώς και την οικογένεια μου για την στήριξη τους.

Οκτώβριος 2011

Ιωαννίδου Κυριακή

ΠΕΡΙΛΗΨΗ

Η διαχωριστική ανάλυση είναι μια στατιστική τεχνική, η οποία έχει δύο βασικούς στόχους: Τη διάκριση ενός πληθυσμού σε ευδιάκριτα σύνολα και την ταξινόμηση παρατηρήσεων στα παραπάνω σύνολα (με την χρήση ενός κανόνα-σχέση).
Η παρούσα εργασία χωρίζεται σε δύο μέρη. Στο πρώτο μέρος (κεφάλαια 1-6) γίνεται μια αναφορά στις βασικές αρχές της Διαχωριστικής Ανάλυσης (εισαγωγή, κανόνες διαχωρισμού ομάδων, έννοιες, σύγκριση της Διαχωριστικής ανάλυσης με άλλες προσεγγίσεις για τον διαχωριστμό ομάδων, σύγκριση της απόδοσης της Διαχωριστικής ανάλυσης και των Νευρωνικών δικτύων για την ταξινόμηση).
Στο δεύτερο μέρος (κεφάλαια 7-8) εξετάζεται ειδικά η διαχωριστική ανάλυση στο SPSS Clementine, και ακολουθούν δύο εφαρμογές της διαχωριστικής ανάλυσης, η πρώτη στο SPSS Clementine και η δεύτερη στο SPSS.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ

Διαχωριστική ανάλυση, Διαχωριστική συνάρτηση, Διαχωριστική ταξινόμηση, Διαχωριστικός βαθμός.

ABSTRACT

The Discriminant analysis is a statistical technical, which has two basic aims:
The discrimination of population in distinct total and the classification of observations in these groups (using rule – relation).
The present work is separated in two parts: In the first part (capital 1-6) there exists a report in the basic rules of Discriminant Analysis (Import, rules of segmentation, rules, relation of Discriminant Analysis with other statistical techniques, relation of Discriminant Analysis with Neural Networks in classification).
In the second part (capital 7-8), Discriminant Analysis is specifically examined in SPSS Clementine, and it is followed by two applications of Discriminant analysis. The first application is conducted in SPSS Clementine and the second in SPSS.

WORDS KEYS

Discriminant Analysis, Discriminant function, Discriminant classification, Discriminant score.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

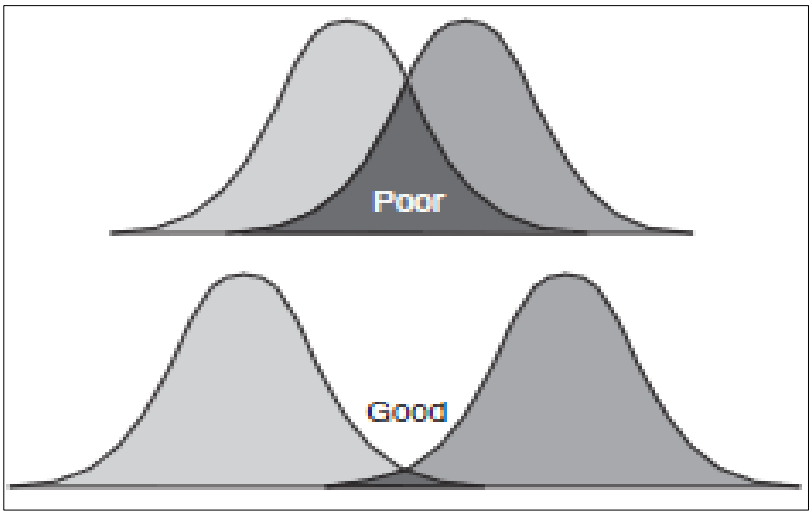
ΚΕΦΑΛΑΙΟ 1^{Ο}

1.1 ΕΙΣΑΓΩΓΗ

Η διαχωριστική ανάλυση (ή διακριτικής ανάλυσης, Discriminant analysis) είναι μία στατιστική τεχνική που αναπτύχθηκε από τον R. A. Fisher το 1936 και έχει δύο βασικούς στόχους:

i. τη διάκριση ενός πληθυσμού σε ευδιάκριτα σύνολα (ομάδες – υποπληθυσμούς) και
ii. την ταξινόμηση παρατηρήσεων στους προηγούμενους γνωστούς πληθυσμούς, με γνωστές κατανομές για κάθε πληθυσμό, με τη βοήθεια ενός κανόνα (διαχωριστική συνάρτηση).

Ας υποθέσουμε ότι έχουμε K πληθυσμούς (ομάδες) $\pi_1, \pi_2, ..., \pi_k$ με $k \geq 2$ και ότι για κάθε πληθυσμό γνωρίζουμε την κατανομή του. Έστω για τον πληθυσμό π_k η κατανομή του είναι $f_k(x)$, όπου x είναι το διάνυσμα στήλη p τυχαίων μεταβλητών
Σκοπός της διαχωριστικής ανάλυσης είναι να διαχωρίσει, δηλαδή να κατανείμει κάθε παρατήρηση στους k γνωστούς πληθυσμούς-ομάδες. Δηλαδή ψάχνουμε για έναν κανόνα, διαχωριστική συνάρτηση, που στόχο έχει να κατατάξει όσο το δυνατόν πιο σωστά περισσότερες παρατηρήσεις.
Στο τέλος της διαχωριστικής ανάλυσης, το επιθυμητό αποτέλεσμα είναι κάθε ομάδα να έχει κανονική κατανομή των διαχωριστικών βαθμών (discriminant scores). Ο βαθμός της κάλυψης των κατανομών των διαχωριστικών αποτελεσμάτων μπορεί να χρησιμοποιηθεί ως μέτρο αναφοράς για το αν η τεχνική είναι επιτυχημένη. Για παράδειγμα, στο πρώτο ζευγάρι της Εικόνας 1.1 υπάρχει μεγάλη υπερκάλυψη μεταξύ των δύο κατανομών και δεν υπάρχει σωστός διαχωρισμός μεταξύ των δύο ομάδων. Η λανθασμένη ταξινόμηση (misclassification) μειώνεται στο δεύτερο ζευγάρι, σε αντίθεση με το πρώτο ζευγάρι όπου πολλές από τις παρατηρήσεις είναι λάθος ταξινομημένες.



Εικονα 1.1: Κατανομές διαχωριστικής ανάλυσης

Σε άλλες επιστήμες η μέθοδος αναφέρεται και με άλλες ονομασίες, όπως για παράδειγμα στην πληροφορική αναφέρεται ως αναγνώριση προτύπων (pattern recognition). Οι εφαρμογές της μεθόδου είναι πάρα πολλές και μερικά από τα παραδείγματα εφαρμογών της μεθόδου είναι τα ακόλουθα:

- Στην Ιατρική, όταν θέλουμε να διαγνώσουμε την ασθένεια κάποιου ασθενή με βάση κάποια συμπτώματα που έχει. Δεδομένου ότι κάθε ασθένεια έχει και τα συμπτώματα της, μπορούμε να κατασκευάσουμε έναν κανόνα σύμφωνα με τον οποίον θα γίνεται η διάγνωση για κάθε

καινούριο ασθενή λαμβάνοντας υπόψη τα συμπτώματα του.

- Στα Χρηματοοικονομικά, όπου για παράδειγμα οι τράπεζες ενδιαφέρονται να εντοπίσουν "καλούς" και "κακούς" πελάτες πριν την χορήγηση ενός δανείου ή μιας πιστωτικής κάρτας. Ως "καλοί" πελάτες μπορούν να θεωρηθούν αυτοί που πληρώνουν κανονικά τις δόσεις τους και ως "κακοί" αυτοί που δεν πληρώνουν. Συνεπώς, η τράπεζα μπορεί να σχηματίσει κανόνες, ώστε να κατατάξει κάθε καινούριο πελάτη σε μια από τις κατηγορίες, και πιθανότητα να αρνηθεί την έγκριση ενός δανείου είτε να εγκρίνει το δάνειο με όρους σύμφωνα με το επίπεδο κινδύνου (risk) που έχει διαγνώσει για κάθε νέο πελάτη.

- Στις κοινωνικές επιστήμες όπου υπάρχει έντονο ενδιαφέρον να καταταχθούν ομάδες πληθυσμού σε συγκεκριμένες κοινωνικές ομάδες λαμβάνοντας υπόψη διάφορα χαρακτηριστικά, όπως για παράδειγμα προβλήματα, κοινωνικοοικονομικά χαρακτηριστικά κλπ.

- Στις προεκλογικές καμπάνιες και δημοσκοπήσεις συνήθως υπάρχει ένα έντονο πρόβλημα με τους αναποφάσιστους και με αυτούς που δεν δηλώνουν καθαρά την προτίμηση τους. Σε αυτή την περίπτωση δημιουργούνται κανόνες σύμφωνα με την διαχωριστική ανάλυση και ο αναποφάσιστος να εντάσσεται σε κάποια ομάδα ψήφου.

Τα παραδείγματα προφανώς δεν εξαντλούνται σε αυτά που μόλις αναφέρθηκαν, αλλά δείχνουν την ποικιλία εφαρμογών της μεθόδου. Ενδιαφέρον είναι να παρατηρήσει κανείς ότι η κατάταξη γίνεται είτε σε δύο ομάδες, όπως στο παράδειγμα της τράπεζας, είτε σε περισσότερες, όπως στο παράδειγμα της ιατρικής διάγνωσης.

Η διαχωριστική ανάλυση έχει αρκετές ομοιότητες με την ανάλυση παλινδρόμησης και την ανάλυση διασποράς (ANOVA) καθώς είναι μία μέθοδος με την οποία μπορούμε να καταλάβουμε την σχέση μεταξύ της εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η βασική διαφορά μεταξύ της διαχωριστικής ανάλυσης και των δύο άλλων μεθοδών αφορά την μορφή της εξαρτημένης μεταβλητής. Στην ανάλυση παλινδρόμησης και στην ανάλυση διασποράς, η εξαρτημένη μεταβλητή πρέπει να είναι συνεχής μεταβλητή, ενώ στην διαχωριστική ανάλυση η εξαρτημένη μεταβλητή πρέπει να είναι κατηγορική μεταβλητή. Οι κατηγορίες πρέπει να είναι αμοιβαία αποκλειστικές, δηλαδή ένα αντικείμενο μπορεί να ανήκει σε μία μόνο ομάδα σύμφωνα με την κατηγορική μεταβλητή.

Επίσης η διαχωριστική ανάλυση μοιάζει με την ανάλυση κατά συστάδες (cluster analysis) αλλά έχει και σημαντικές διαφορές από αυτή. Η σημαντικότερη διαφορά είναι ότι στην διαχωριστική ανάλυση οι ομάδες είναι γνωστές ενώ σκοπός της ανάλυσης κατά συστάδες είναι να βρει αυτές τις ομάδες. Γι'αυτό το λόγο, στην διαχωριστική ανάλυση προσπαθούμε να φτιάξουμε έναν κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις για το μέλλον, ενώ στην ανάλυση κατά συστάδες ο κύριος στόχος είναι να δημιουργήσουμε ομοειδείς ομάδες. Οι ομάδες αυτές έχουν ως στόχο την κατανόηση των ήδη υπαρχόντων στοιχείων και την μείωση της διασποράς σε επιμέρους ομάδες.

ΚΕΦΑΛΑΙΟ 2^{Ο}

Βασικοί όροι και έννοιες στη Διαχωριστική Ανάλυση

Για τη σωστή ερμηνεία της διαχωριστικής ανάλυσης απαιτείται η κατανόηση τεσσάρων εννοιών: τους διαχωριστικούς συντελεστές (discriminant coefficients), την διαχωριστική συνάρτηση (discriminant function) , τους διαχωριστικούς βαθμούς (discriminant scores) , και τους κεντροειδείς των ομάδων (group centroids).

2.1 Συντελεστές της διαχωριστικής ανάλυσης (Discriminant Coefficients)

Στη διαχωριστική ανάλυση υπολογίζονται τα μαθηματικά βάρη (weights) για κάθε βαθμό σε κάθε διαχωριστική μεταβλητή. Το βάρος κάθε μεταβλητής δείχνει τη διαφορά, μεταξύ των ομάδων, των βαθμών για τη συγκεκριμένη μεταβλητή που υπολογίστηκε το βάρος. Έτσι, οι διαχωριστικές μεταβλητές βάσει των οποίων διαφέρουν περισσότερο οι ομάδες ονομάζονται διαχωριστικοί συντελεστές.
Στα περισσότερα στατιστικά προγράμματα, ο ερευνητής έχει την δυνατότητα να υπολογίσει και τους τυποποιημένους (standardized) και τους μη-τυποποιημένους (unstandardized) διαχωριστικούς συντελεστές. Οι μη-τυποποιημένοι συντελεστές χρησιμοποιούνται κυρίως όταν ο ερευνητής θέλει να διασταυρώσει ή να αναπαράγει ξανά τα αποτελέσματα μιας διαχωριστικής ανάλυσης ή για να κατατάξει μη ταξινομημένες παρατήρησεις σε μια ομάδα. Ωστόσο, οι μη τυποποιημένοι συντελεστές δεν μπορούν να χρησιμοποιηθούν για να συγκριθούν οι μεταβλητές ή για να καθοριστούν ποιες μεταβλητές παίζουν τον σημαντικότερο ρόλο στον διαχωρισμό των ομάδων επειδή η τυποποίηση που γίνεται σε κάθε μεταβλητή είναι συχνά διαφορετική.
Οι τυποποιμένοι συντελεστές χρησιμοποιούνται για να καθοριστούν οι σχέσεις των διαχωριστικών μεταβλητών στις συναρτήσεις. Οι τυποποιημένοι συντελεστές μετατρέπονται σε *z* βαθμούς (δηλ. *m=0* και *sd=1*) για να εξαλειφθούν οι διαφορές λόγω τυποποίησης μεταξύ των διαχωριστικών μεταβλητών. Το αποτέλεσμα είναι να μπορεί ο ερευνητής να καθορίσει τον βαθμό στον οποίο κάθε διαχωριστική μεταβλητή σχετίζεται με τις διαφορές μεταξύ των ομάδων εξετάζοντας την απόλυτη τιμή των τυποποιημένων συντελεστών καθώς και άμα είναι σχετικά σημαντική κάθε μεταβλητή στον διαχωρισμό των ομάδων.

2.2 Συνάρτηση και βαθμοί της διαχωριστικής ανάλυσης (Discriminant Function and Discriminant Scores)

Η διαχωριστική ανάλυση περιλαμβάνει τον καθορισμό μιας γραμμικής εξίσωσης η οποία θα χρησιμοποιείται για τον προβλέψουμε σε ποια ομάδα θα ανήκει κάθε παρατήρηση. Η μορφή αυτή της εξίσωσης ή συνάρτησης είναι:

$$D = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

όπου

D = διαχωριστικός βαθμός (discriminant score)
α = σταθερά
b = συντελεστές διαχωριστικής συνάρτησης
x = βαθμός κάθε μεταβλητής

p = αριθμός των διαχωριστικών μεταβλητών

Η διαχωριστική συνάρτηση είναι όμοια με την συνάρτηση της ανάλυσης παλινδρόμησης. Οι συντελεστές b μεγιστοποιούν την απόσταση μεταξύ των μέσων από την εξαρτημένη μεταβλητή. Ένας διαχωριστικός βαθμός για κάθε συνάρτηση υπολογίζεται πολλαπλασιάζοντας κάθε μεταβλητή με το αντίστοιχο βάρος. Έπειτα, οι διαχωριστικοί βαθμοί χρησιμοποιούνται για να υπολογιστεί ο μέσος διαχωριστικός βαθμός των περιπτώσεων που ανήκουν στην ομάδα (δηλ. ο κεντροειδής της ομάδας) για κάθε διαχωριστική συνάρτηση.
Ο σκοπός της διαχωριστικής συνάρτησης είναι να μεγιστοποιήσει την απόσταση μεταξύ των κατηγοριών, δηλαδή η διαχωριστική συνάρτηση πρέπει να έχει μεγάλη διαχωριστική δύναμη μεταξύ των ομάδων.
Αφού χρησιμοποιήσουμε ένα σύνολο δεδομένων για να διεξάγουμε την διαχωριστική συνάρτηση και να ταξινομήσουμε αυτά τα δεδομένα, έπειτα οποιαδήποτε παρατήρηση θα μπορεί να ταξινομηθεί στις ήδη υπάρχουσες ομάδες. Ο αριθμός των διαχωριστικών συναρτήσεων ισούται με τον αριθμό των ομάδων μειωμένος κατά ένα. Δηλαδή υπάρχει μόνο μια διαχωριστική συνάρτηση για την ανάλυση δύο ομάδων.

2.3 Κεντροειδείς ομάδων (Group Centroids)

Στην διαχωριστική ανάλυση, ο κεντροειδής κάθε ομάδας αναπαριστάνει το μέσο διαχωριστικό βαθμό των μελών της ομάδας για την δεδομένη συνάρτηση. Για τους σκοπούς της πρόβλεψης και της ταξινόμησης, ο διαχωριστικός βαθμός σε κάθε περίπτωση συγκρίνεται με τον κεντροειδή κάθε ομάδας και υπολογίζεται η πιθανότητα να είναι μέλος στην συγκεκριμένη ομάδα η συγκεκριμένη περίπτωση. Όσο πιο κοντά είναι ο βαθμός στον κεντροειδή της ομάδας, τόσο μεγαλύτερη είναι η πιθανότητα η περίπτωση να ανήκει στη συγκεκριμένη ομάδα.
Οι κεντροειδείς των ομάδων αποκαλύπτουν το μέγεθος και τον τρόπο με τον οποίο διαφοροποιούνται οι ομάδες σε κάθε συνάρτηση. Η απόλυτη τιμή δείχνει τον βαθμό ως προς τον οποίο μια ομάδα διαφοροποιείται στην συνάρτηση, ενώ το πρόσημο δείχνει τη κατεύθυνση της διαφοροποίησης.

2.4 Υποθέσεις διαχωριστικής ανάλυσης

Όπως και σε όλες τις στατιστικές μεθόδους πολλών μεταβλητών, έτσι και στην διαχωριστική ανάλυση υπάρχουν κάποιες υποθέσεις που πρέπει να καλυφθούν. Οι περισσότερες από αυτές τις υποθέσεις ισχύουν και στην πολυμεταβλητή ανάλυση παλινδρόμησης και στην ανάλυση διασποράς.

Οι κύριες υποθέσεις της διαχωριστικής ανάλυσης είναι:

- Οι παρατηρήσεις είναι ένα τυχαίο δείγμα. Η διαχωριστική ανάλυση προυποθέτει ότι οι παρατηρήσεις είναι ανεξάρτητες η μια από την άλλη, δηλαδή ζευγάρια δεδομένων που είναι ίδια δεν επιτρέπονται.

- Κάθε μεταβλητή πρόβλεψης ακολουθεί κανονική κατανομή. Εφ'όσον τα μεγέθη των δειγμάτων είναι μεγάλα, τα μεγέθη των ομάδων είναι περίπου ίσα και οι ακραίες τιμές (outliers) ελάχιστες, η διαχωριστική ανάλυση μπορεί να παραλείψει κάποιες παραβιάσεις αυτής της υπόθεσης, ειδικά αν το πρόβλημα είναι η λοξότητα και όχι οι ακραίες τιμές. Καθώς η διαχωριστική ανάλυση είναι ευαίσθητη στο θέμα των ακραίων τιμών θα πρέπει να εξαταστεί η παρουσία τους πριν διεξαχθεί όλη η ανάλυση. Εάν υπάρχουν ακραίες τιμές, τότε αυτές θα πρέπει να μετασχηματιστούν ή να αφαιρεθούν. Εναλλακτικά, κάποιος μπορεί

να εφαρμόσει λογιστική παλινδρόμηση, όπου δεν υπάρχει καμία προυπόθεση για την κατανομή των ανεξάρτητων μεταβλητών.

- Οι εξαρτημένες μεταβλητές στην αρχική ταξινόμηση είναι σωστά ταξινομημένες.

- Οι ομάδες ή κατηγορίες πρέπει να οριστούν πριν συλλεχθούν τα δεδομένα.

- Πρέπει να υπάρχουν τουλάχιστον δύο ομάδες ή κατηγορίες, με κάθε περίπτωση να ανήκει σε μια μόνο ομάδα έτσι ώστε όλες οι ομάδες να είναι αποκλειστικές (όλες οι περιπτώσεις μπορούν να τοποθετηθούν σε μια ομάδα).

- Τα χαρακτηριστικά που χρησιμοποιούνται για να διαχωρίσουν τις ομάδες πρέπει να ευδιάκριτα έτσι ώστε η υπερκάλυψη μεταξύ των ομάδων να είναι ανύπαρκτη ή ελάχιστη.

- Τα μεγέθη των ομάδων δεν θα πρέπει να διαφέρουν πολύ μεταξύ τους και θα πρέπει να είναι τουλάχιστον πέντε φορές του αριθμού των ανεξάρτητων μεταβλητών.

- Ομοσκεδαστικότητα των υπολοίπων. Η διαχωριστική ανάλυση υποθέτει ότι η διακύμανση των υπολοίπων (πραγματικές-προβλέψιμες τιμές) είναι σταθερή σε όλες τις τιμές των ανεξάρτητων μεταβλητών.

- Ομοιογένεια διακύμανσης/συνδιακύμανσης. Η διαχωριστική ανάλυση υποθέτει ότι οι πίνακες συνδιακύμανσης για τις ανεξάρτητες μεταβλητές είναι όμοιοι για κάθε ομάδα. Η ομοιογένεια της διακύμανσης και της συνδιακύμανσης μπορεί να εξεταστεί γραφικά, για παράδειγμα με scatterplot. Υπάρχουν και άλλοι έλεγχοι οι οποίοι μπορούν να εφαρμοστούν για να αποδειχθεί αν παραβιάζεται η αρχική υπόθεση ή όχι. Ωστόσο, ο έλεγχος Box M για την ομοιογένεια της διακύμανσης/συνδιακύμανσης είναι ευαίσθητος σε αποκλίσεις από την κανονικότητα, και δεν θα πρέπει να λαμβάνεται σοβαρά υπ'όψιν. Τέλος, η διαχωριστική ανάλυση μπορεί να ανεχθεί παραβιάσεις αυτής της υπόθεσης, ειδικά αν τα μεγέθη των ομάδων είναι σχεδόν ίδια.

- Μη πολυσυγγραμικότητα. Αν και η διαχωριστική ανάλυση μπορεί να ανεχθεί κάποια συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών, δεν πρέπει να υπάρχει πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών. Όταν παρουσιάζεται μεγάλη πολυσυγγραμμικότητα μεταξύ δύο ή περισσότερων ανεξάρτητων μεταβλητών, οι συντελεστές της διαχωριστικής συνάρτησης δεν θα είναι στατιστικά σημαντικοί. Επιπλέον, εάν υπάρχει τέλεια πολυσυγγραμικότητα, δηλαδή μια από τις ανεξάρτητες μεταβλητές γράφεται συναρτήσει των υπολοίπων (για παράδειγμα, ως το άθροισμά τους) η ανάλυση αποτυγχάνει.

Όταν δεν ισχύει μια από τις παραπάνω υποθέσεις δεν σημαίνει ότι δεν μπορεί να εφαρμοστεί η διαχωριστική ανάλυση. Για παράδειγμα, μπορεί να εφαρμοστεί με εξαιρετικά αποτελέσματα σε δεδομένα από μη-κανονική κατανομή, αν και δεν θα είναι πια εύκολο να γίνουν οι έλεγχοι υποθέσεων για την σημαντικότητα ως προς τις διαφορές μεταξύ των ομάδων.

ΚΕΦΑΛΑΙΟ 3^{Ο}

Ερμηνεία των διαχωριστικών συναρτήσεων

Η ερμηνεία των αποτελεσμάτων μιας διαχωριστικής ανάλυσης εξαρτάται, σε μεγάλο βαθμό, από την ερμηνεία των διαχωριστικών συναρτήσεων. Η συνάρτηση ορίζεται από τους συντελεστές οι οποίοι χρησιμοποιούνται για να 'ζυγίσουν' τον βαθμό κάθε περίπτωσης στις διαχωριστικές μεταβλητές.
Πρώτα απ'όλα, ένας ερευνητής πρέπει να αναγνωρίσει τις διαχωριστικές μεταβλητές οι οποίες έχουν το υψηλότερο και το χαμηλότερο βάρος στην συνάρτηση. Το μέγεθος των συντελεστών δείχνει πόσο πολύ μία διαχωριστική μεταβλητή συνεισφέρει στον διαχωρισμό των ομάδων, ενώ το πρόσημο δείχνει την κατεύθυνση της σχέσης.
Επίσης, ο ερευνητής θα πρέπει να εξετάσει τον πίνακα δομής των συντελεστών, ο οποίος δείχνει τη σχέση των διαχωριστικών μεταβλητών με την συνάρτηση. Η απόλυτη τιμή δείχνει τη δύναμη της σχέσης μεταξύ κάθε μεταβλητής και συνάρτησης, ενώ το πρόσημο, όπως και πριν, δείχνει την κατεύθυνση της σχέσης. Η διαδικασία είναι αρκετά αποτελεσματική, καθώς η πληροφορία που προέρχεται όταν εξετάζονται πολλές μεταβλητές ταυτόχρονα είναι πολύ πιο χρήσιμη σε σύγκριση με την πληροφορία που παίρνει μελετώντας κάθε μεταβλητή χωριστά.
Τέλος, θα πρέπει πρέπει να εξεταστούν οι κεντροειδείς κάθε ομάδας για κάθε συνάρτηση και να βρεθούν οι ομάδες με το υψηλότερο και το χαμηλότερο βαθμό. Οι κεντροειδείς δίνουν πληροφορίες για τις ιδιότητες κάθε ομάδας.

3.1 Εκτίμηση σημαντικότητας των συναρτήσεων

3.1.1 Wilks's Lambda και Chi-Square

Το Wilk's Lambda είναι το κλάσμα της μεταβλητότητας ανάμεσα στις ομάδες (within-groups variability) προς την συνολική μεταβλητότητα (total variability) των διαχωριστικών μεταβλητών. Δηλαδή, $\frac{SS_{within-groups}}{SS_{total}}$ και είναι ένα αντίστροφο μέτρο σημαντικότητας των συναρτήσεων. Όσο μικρότερο είναι το Wilk's lambda για μια μεταβλητή, τόσο περισσότερο συμβάλλει η μεταβλητή στην διαχωριστική συνάρτηση. Οι τιμές κοντά στο 1 δείχνουν ότι σχεδόν όλη η μεταβλητότητα των μεταβλητών σχετίζεται με τις διαφορές μέσα στην ομάδα (διαφορές μεταξύ των περιπτώσεων σε κάθε ομάδα), ενώ οι τιμές κοντά στο 0 δείχνουν ότι η μεταβλητότητα οφείλεται στις διαφορές ανάμεσα στις ομάδες.
Το chi-square test που βασίζεται στο lambda δείχνει εάν η μεταβλητότητα η οποία σχετίζεται με τις διαφορές των ομάδων είναι στατιστικά σημαντική.

3.1.2 Η ιδιοτιμή (eigenvalue)

Οι ιδιοτιμές είναι το κλάσμα της μεταβλητότητας μεταξύ των ομάδων (between groups variability) προς την μεταβλητότητα ανάμεσα στις ομάδες (within-groups variability), δηλαδή $\frac{SS_{between-groups}}{SS_{within-groups}}$.
Η ιδιοτιμή αποτελεί, μέσω των διαχωριστικών βαθμών στις ομάδες, έναν δείκτη αποτελεσματικότητας της διαχωριστικής συνάρτησης καθώς και δείκτη επιλογής του αριθμού των διαχωριστικών συναρτήσεων. Όσο μεγαλύτερη είναι η τιμή της ιδιοτιμής, τόσο πιο «τέλεια» είναι η διαχωριστική συνάρτηση. Υπάρχει μία ιδιοτιμή για κάθε διαχωριστική συνάρτηση. Για την διαχωριστική ανάλυση δύο ομαδών, υπάρχει μια διαχωριστική συνάρτηση και επομένως μία

ιδιοτιμή. Εάν υπάρχουν περισσότερες από μία διαχωριστικές συναρτήσεις, η πρώτη θα είναι η μεγαλύτερη και η σημαντικότερη, η δεύτερη η αμέσως σημαντικότερη ως προς την επεξηγηματική δύναμη κοκ.

3.1.3 Το σχετικό ποσοστό (the relative percentage)

Το σχετικό ποσοστό ισούται με το κλάσμα της ιδιοτιμής μιας συνάρτησης προς το άθροισμα των ιδιοτιμών όλων των διαχωριστικών συναρτήσεων στο μοντέλο. Το σχετικό ποσοστό δείχνει την αναλογία της συνολικής διασποράς μεταξύ των ομάδων που συνδέεται με μια δεδομένη διαχωριστική συνάρτηση. Δηλαδή, το σχετικό ποσοστό δείχνει την σημαντικότητα κάθε συνάρτησης λαμβάνοντας υπ'όψιν τις διαφορές μεταξύ των ομάδων.

3.1.4 Το απόλυτο ποσοστό (the absolute percent)

Το απόλυτο ποσοστό υπολογίζεται διαιρώντας κάθε ιδιοτιμή προς τον αριθμό των διαχωριστικών μεταβλητών που χρησιμοποιούνται στην ανάλυση. Το απόλυτο ποσοστό δείχνει το μέγεθος της μεταβλητότητας μεταξύ των ομάδων όπως εξηγείται από κάθε συνάρτηση που συνδέεται με το μέγεθος της μεταβλητότητας μεταξύ των ομάδων.
Το απόλυτο ποσοστό είναι χρήσιμο για να μετρήσουμε πόσο αποτελεσματικές είναι οι διαχωριστικές μεταβλητές για τον διαχωρισμό των ομάδων.

3.1.5 Δείτης κανονικής συσχέτισης R (canonical correlation)

Πολλές φορές η μεταβλητότητα μιας συνάρτησης δεν σχετίζεται με τις διαφορές των ομάδων. Αυτή η μεταβλητότητα μπορεί να σχετίζεται με διαφορές εντός των ομάδων, ή με άλλα λάθη που συμβαίνουν κατά την συλλογή δεδομένων ή κατά την εισαγωγή δεδομένων. Ο δείκτης κανονικής συσχέτισης δείχνει την συσχέτιση μεταξύ των βαθμών των διαχωριστικών συναρτήσεων και των ομάδων. Παίρνει τιμές από μηδέν μέχρι και ένα και όσο πιο μεγάλος είναι ο δείκης τόσο μεγαλύτερος είναι ο βαθμός συσχέτισης μεταξύ των διαχωριστικών συναρτήσεων και των ομάδων.

3.1.6 Πίνακας δομών (structure matrix)

Ο πίνακας δομής (structure matrix) δίνει τους δείκτες συσχέτισης κάθε ανεξάρτητης μεταβλητής με τις διαχωριστικές συναρτήσεις. Οι δείκτες συσχέτισης μπορούν να χρησιμοποιηθούν για να αξιολογηθεί πόσο σημαντική είναι κάθε μεταβλητή για την κατασκευή της διαχωριστικής συνάρτησης.

3.1.7 Πίνακας ταξινόμησης (classification table)

Ο πίνακας ταξινόμησης (classification table) είναι χρήσιμος για τον υπολογισμό του ποσοστού των σωστά ταξινομημένων παρατηρήσεων. Στον πίνακα αυτόν, οι παρατηρούμενες κατηγορίες (ομάδες) καταγράφονται στις σειρές, ενώ στις στήλες οι προβλεπόμενες. Το ποσοστό στη διαγώνιο είναι το ποσοστό των σωστά ταξινομημένων παρατηρήσεων, το οποίο ονομάζεται hit ratio.

3.2 Εκτίμηση της ακρίβειας ταξινόμησης

3.2.1 Ακρίβεια ταξινόμησης

Η αξιολόγηση της ταξινόμησης είναι ένα μέσο για τον καθορισμό της στατιστικής και πρακτικής

χρησιμότητας των υπολογισμένων συναρτήσεων. Οι συναρτήσεις που χρησιμοποιούνται παράγονται στην διαχωριστική ανάλυση για να ταξινομηθούν περιπτώσεις οι οποίες δεν έχουν ακόμη ταξινομηθεί ή περιπτώσεις που έχουν ταξινομηθεί με κάποιον εναλλακτικό τρόπο.
Η διαδικασία περιλαμβάνει την ταξινόμηση των αρχικών περιπτώσεων στο δείγμα χρησιμοποιώντας τις συναρτήσεις και αξιολογώντας την ακρίβεια αυτών των ταξινομήσεων. Όταν εκτελούμε περιγραφική διαχωριστική ανάλυση, ξεκινάμε με ένα δείγμα περιπτώσεων του οποίου η ιδιότητα των μελών είναι γνωστή. Αφού εκτελεστεί η διαχωριστική ανάλυση και οι διαχωριστικές συναρτήσεις παραχθούν και αξιολογηθούν, υπολογίζονται σε κάθε συνάρτηση οι διαχωριστικοί βαθμοί και οι κεντροειδείς των ομάδων για κάθε συνάρτηση. Έπειτα, καθορίζεται η ομάδα στην οποία κάθε περίπτωση θα καταταχθεί σύμφωνα τον βαθμό στις διαχωριστικές μεταβλητές. Τελικά, η ομάδα στην οποία θα καταταχθεί κάθε περίπτωση συγκρίνεται με την ομάδα στην οποία η περίπτωση πραγματικά ανήκει, και υπολογίζεται το ποσοστό των σωστών κατατάξεων. Η διαδικασία έχει ως αποτέλεσμα το ποσοστό των σωστά καταταγμένων περιπτώσεων.

3.2.2 Επικύρωση (validation)

Είναι κρίσιμο να διασταυρώσουμε τα αποτελέσματα που παίρνουμε από μία διαχωριστική ανάλυση, ειδικά εάν πρόκειται να ταξινομήσουμε και άλλα δείγματα σ'αυτές ομάδες.
Αρχικά εξάγουμε από το δείγμα μας έναν αριθμό παρατηρήσεων από κάθε ομάδα, εφαρμόζουμε τη διαχωριστική ανάλυση στα νέα πλέον δεδομένα και στη συνέχεια παίρνουμε τις παρατηρήσεις που βγάλαμε από το αρχικό δείγμα και τις ταξινομούμε στις ομάδες. Τέλος, με βάση την αρχική ταξινόμηση των παρατηρήσεων που βγάλαμε από το δείγμα βλέπουμε αν τοποθετήθηκαν σωστά ή όχι οι παρατηρήσεις και το ποσοστό επιτυχίας. Έτσι, κρίνεται αν η ανάλυση που έγινε είναι πετυχημένη ή όχι.

ΚΕΦΑΛΑΙΟ 4^{Ο}

Κανόνες διαχωρισμού Δυο ομάδων

Σ'αυτό το κομμάτι, θα εξετάσουμε τους κανόνες διαχωρισμού ομάδων που βασίζονται στην απόσταση Mahalanobis, στον κανόνα μέγιστης πιθανοφάνειας και στον κανόνα Bayes. Αυτοί οι κανόνες βασίζονται σε πληθυσμούς που έχουν είτε γνωστούς μέσους και συνδιασπορές είτε γνωστές πυκνότητες.

4.1 Απόσταση Mahalanobis

Το τετράγωνο της απόστασης Mahalanobis δίνεται από τον τύπο:

$$D^2 = (x-\mu)'\Sigma^{-1}(x-\mu)$$

που χρησιμοποιείται για να δώσει την απόσταση ενός συγκεκριμένου διανύσματος από το κέντρο της κατανομής, όπου Σ είναι ο πίνακας συνδιασποράς. Στο πρόβλημα διαχωρισμού, έχουμε ένα τυχαίο διάνυσμα *x* και *t* είναι οι πιθανές κατανομές από τις οποίες μπορεί να προκύψει το διάνυσμα αυτό. Μια λογική διαδικασία διαχωρισμού είναι να ταξινομήσουμε την παρατήρηση *x* στο πληθυσμό για τον οποίο μειώνεται η απόσταση Mahalanobis. Δηλαδή:

- Εάν $(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) < (x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_1.
- Εάν $(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) < (x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_2.

4.2 Κανόνας Μέγιστης Πιθανοφάνειας

Δεδομένου ότι οι πυκνότητες πιθανοτήτων $f(x \mid i) = f_i(x)$ είναι γνωστές για κάθε πληθυσμό *i* και έστω ότι έχουμε μια παρατήρηση *x*, τότε η συνάρτηση πιθανότητας είναι

$$L(i) = f(x \mid i) \text{ για } i=1,...,t$$

Σύμφωνα με τον κανόνα μέγιστης πιθανοφάνειας η παρατήρηση *x* κατατάσεται στον πληθυσμό *r* για τον οποίο

$$L(r) = \max L(i)$$

ή ισοδύναμα

- Εάν $f_1(x) > f_2(x)$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_1
- Εάν $f_2(x) > f_1(x)$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_2

Εάν οι παρατηρήσεις ακολουθούν κανονική κατανομή, ο κανόνας μέγιστης πιθανοφάνειας μοιάζει

με τον κανόνα της απόστασης Mahalanobis. Οι πυκνότητες πιθανοτήτων βρίσκονται από την σχέση:

$$L(i) = f(x \mid i) = (2\pi)^{-q/2}\left|\Sigma_i\right|^2 \exp[-(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)/2] \text{ όπου } i=1,\ldots,t$$

Καθώς ο λογάριθμος είναι μια αύξουσα συνάρτηση, η ελαχιστοποίηση της λογαριθμικής πιθανότητας (log-likelihood) είναι ισοδύναμη με την μεγιστοποίηση της πιθανότητας.

$$l(i) = \log(L(i)) = -\frac{q}{2}\log(2\pi) - \frac{1}{2}\log(\left|\Sigma_i\right| - \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)$$

Εάν "πετάξουμε έξω" τον σταθερό όρο $-\frac{q}{2}\log(2\pi)$ και ελαχιστοποιώντας την αρνητική λογαριθμική πιθανότητα (και όχι μεγιστοποιώντας τη λογαριθμική πιθανότητα, παρατηρούμε ότι ο κανόνας μέγιστης πιθανοφάνειας για κανονικούς πληθυσμούς γίνεται:

- Εάν $\log(\left|\Sigma_1\right| + \frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1) < \log(\left|\Sigma_2\right| + \frac{1}{2}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1
- Εάν $\log(\left|\Sigma_2\right| + \frac{1}{2}(x-\mu_2)'\Sigma_2^{-1}(x-\mu_2) < \log(\left|\Sigma_1\right| + \frac{1}{2}(x-\mu_1)'\Sigma_1^{-1}(x-\mu_1)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

Ας σημειώσουμε ότι η μόνη διαφορά μεταξύ του κανόνα μέγιστης πιθανοφάνειας και του κανόνα Mahalanobis είναι ο όρος $\log(\left|\Sigma_i\right|)$. Και οι δύο κανόνες περιλαμβάνουν δευτεροβάθμιες εξισώσεις του y. Οι μέθοδοι που σχετίζονται με αυτούς τους κανόνες συχνά αναφέρονται και ως δευτεροβάθμιες διαχωρίστικες μέθοδοι.

4.3 Κανόνας του Bayes

Η μέθοδος του Bayes προϋποθέτει την ύπαρξη μιας πιθανότητας για κάθε πληθυσμό, έστω *P(i),* έτσι ώστε κάθε νέα παρατήρηση x να προέρχεται από αυτόν τον πληθυσμό.
Δεδομένων των εκ των προτέρων πιθανοτήτων και των δεδομένων x, οι εκ των υστέρων πιθανότητες ώστε η x να προέλθει από τον πληθυσμό π_1 δίνονται από τον τύπο:

$$P(\pi_1 \mid x) = \frac{f_1(x)p_1}{f_1(x)p_1 + f_2(x)p_2}$$

Ομοίως η εκ των υστέρων πιθανότητα να προέλθει η παρατήρηση x από τον πληθυσμό π_2 δίνεται από τον τύπο:

$$P(\pi_2 \mid x) = \frac{f_2(x)p_2}{f_1(x)p_1 + f_2(x)p_2}$$

Έπειτα ταξινομούμε την παρατήρηση x στον πληθυσμό με την μεγαλύτερη εκ των υστέρων πιθανότητα. Δηλαδή :

- Εάν $P(\pi_1 \mid x) > P(\pi_2 \mid x)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1
- Εάν $P(\pi_1 \mid x) < P(\pi_2 \mid x)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

Καθώς παρατηρούμε ότι οι παρανομαστές των εκ των υστέρων πιθανοτήτων δεν εξαρτώνται από το i=1 ή 2, ο κανόνας ταξινόμησης του Bayes είναι ισοδύναμος με τον εξής:

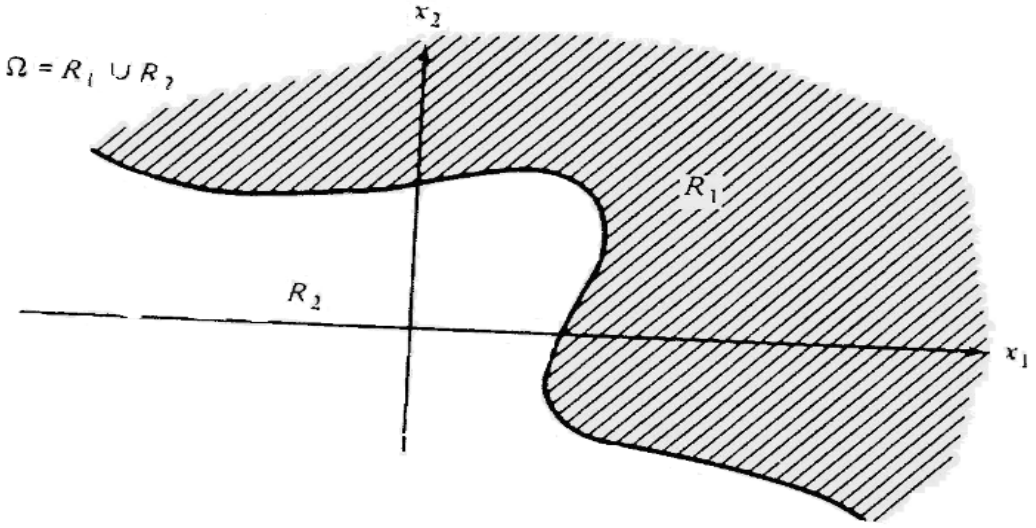
- Εάν $f_1(x)p_1 > f_2(x)p_2$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1
- Εάν $f_1(x)p_1 < f_2(x)p_2$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

Εάν και οι εκ των προτέρων πιθανότητες p_1 και p_2 είναι ίσες, τότε ο κανόνας ταξινόμησης του Bayes είναι ισοδύναμος με τον κανόνα μέγιστης πιθανοφάνειας. Δηλαδή

- Εάν $f_1(x) > f_2(x)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1
- Εάν $f_1(x) < f_2(x)$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

4.4 Ελαχιστοποιήση κόστους λανθασμένης ταξινόμησης

Ένας κανόνας διαχωρισμού ενδέχεται να κατατάξει λανθασμένα μερικές παρατηρήσεις μέσα στους δύο πληθυσμούς. Πρώτα ορίζουμε ως $f_1(x)$ και $f_2(x)$ τις συναρτήσεις πυκνότητας πιθανότητας για ένα τυχαίο διάνυσμα X για τους πληθυσμούς π_1 και π_2, αντίστοιχα. Ως R_1 ορίζουμε το σύνολο των τιμών x για τις οποίες οι παρατηρήσεις ταξινομούνται στον πληθυσμό π_1 και αντίστοιχα ως R_2. ορίζουμε το σύνολο των υπόλοιπων τιμών x για τις οποίες οι παρατηρήσεις ταξινομούνται στον πληθυσμό π_2. Δηλαδή η ένωση $R_1 \cup R_2$ ορίζει έναν χώρο Ω όπου ανήκουν οι τιμές x (Εικόνα 4.4.1).



Εικόνα 4.4.1: Περιοχές ταξινόμησης για δύο πληθυσμούς

Έστω p_1 και p_2 οι προηγούμενες πιθανότητες έτσι ώστε το x να ανήκει στους π_1 και π_2, αντίστοιχα,

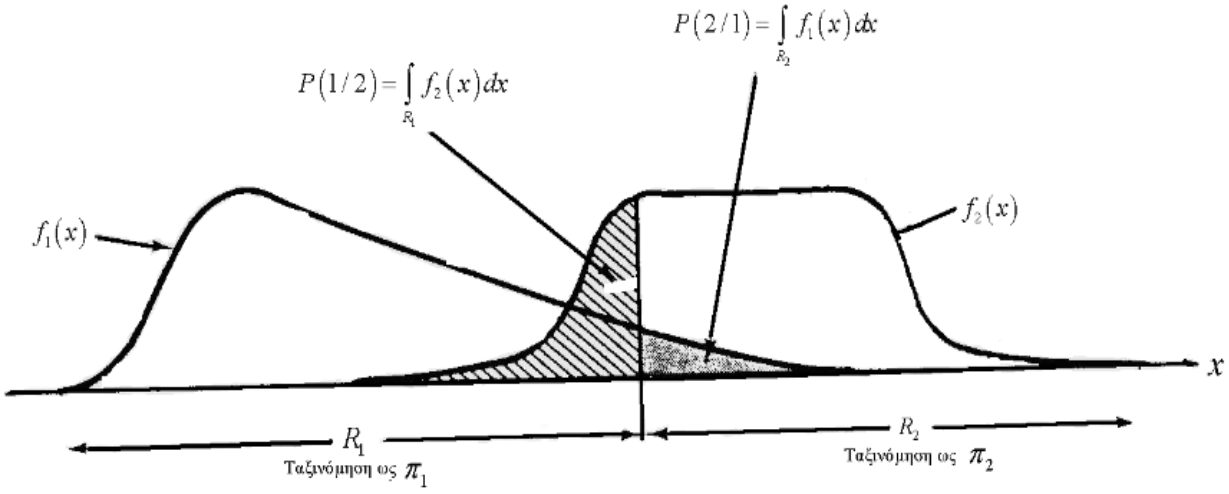
όπου p_1+p_2 =1. Η δεσμευμένη πιθανότητα *P(2|1),* της ταξινόμησης μιας παρατήρησης στον π_2 πληθυσμό, όταν στην πραγματικότητα αυτή προέρχεται από τον πληθυσμό π_1 είναι:

$$P(2\,|\,1) = P(x \in R_2\,|\,\pi_1) = \int_{R_2} f_1(x)dx \quad (1)$$

Ομοίως, η δεσμευμένη πιθανότητα *P(1|2),* της ταξινόμησης μιας παρατήρησης στον π_1 πληθυσμό, όταν στην πραγματικότητα αυτή προέρχεται από τον πληθυσμό π_2 είναι:

$$P(2\,|\,1) = P(x \in R_1\,|\,\pi_2) = \int_{R_1} f_2(x)dx \quad (2)$$

Το ολοκλήρωμα στην (1) παριστάνει τον όγκο που σχηματίζεται από την συνάρτηση πυκνότητας $f_1(x)$ πάνω στην περιοχή R_2. Ομοίως, το ολοκλήρωμα στην (2) παριστάνει τον όγκο που σχηματίζεται από την συνάρτηση πυκνότητας $f_2(x)$ πάνω στην περιοχή R_1 (Εικόνα 4.4.2).



Εικόνα 4.4.2: Πιθανότητες λανθασμένης ταξινόμησης για υποθετικές περιοχές ταξινόμησης όταν *p=1*

Θέτουμε c1=C(2|1) το κόστος λανθασμένης ταξινόμησης κατατάσσοντας μια παρατήρηση από το πληθυσμό π_2 στον π_1 και c2=C(1|2) το αντίστοιχο κόστος λανθασμένης ταξινόμησης από τον πληθυσμό π_1 στον π_2 και 0 για σωστή ταξινόμηση. Τα κόστη λανθασμένων ταξινομήσεων μπορούν να καθοριστούν από έναν πίνακα κόστους (Πίνακας 4.4.1).

	π_1	π_2
π_1	0	C(2|1)
π_2	C(1|2)	0

Πίνακας 4.4.1: Πίνακας κόστους

Επομένως, για κάθε κανόνα ταξινόμησης ο μέσος όρος, ή το αναμενόμενο κόστος λανθασμένης ταξινόμησης (expected cost of misclassification, ECM) δίνεται πολλαπλασιάζοντας τα διαγώνια μη

μηδενικά στοιχεία του πίνακα με τις αντίστοιχες πιθανότητες των σχέσεων (1) και (2). Δηλαδή,

$$ECM = p_1 P(2|1)C(2|1) + p_2 P(1|2)C(1|2)$$

Ένας κανόνας ταξινόμησης πρέπει να έχει όσο το δυνατόν πιο μικρό ECM.

Επομένως ο κανόνας ταξινόμησης που δίνει το μικρότερο ECM είναι ο εξής

- Εάν $\frac{f_1(x)}{f_2(x)} \geq (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_1
- Εάν $\frac{f_1(x)}{f_2(x)} < (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_2

Ειδικές περιπτώσεις

(α) Εάν $p_2 / p_1 = 1$, δηλαδή οι εκ των προτέρων πιθανότητες είναι ίσες, τότε μια παρατήρηση ταξινομείται στον πληθυσμό π_1 εάν $\frac{f_1(x)}{f_2(x)} \geq (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$ ενώ εάν $\frac{f_1(x)}{f_2(x)} < (\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$ ταξινομείται στον πληθυσμό π_2.

(β) Εάν $c(1|2)/c(2|1) = 1$, δηλαδή τα κόστη των λανθασμένων ταξινομήσεων είναι ίσα, τότε μια παρατήρηση ταξινομείται στον πληθυσμό π_1 εάν $\frac{f_1(x)}{f_2(x)} \geq (\frac{p_2}{p_1})$ ενώ εάν $\frac{f_1(x)}{f_2(x)} < (\frac{p_2}{p_1})$ ταξινομείται στον πληθυσμό π_2.

(γ) Εάν $p_2 / p_1 = c(1|2)/c(2|1) = 1$ ή $p_2 / p_1 = 1/c(1|2)/c(2|1)$, δηλαδή οι εκ των προτέρων πιθανότητες και τα κόστη των λανθασμένων ταξινομήσεων είναι ίσα, τότε μια παρατήρηση ταξινομείται στον πληθυσμό π_1 εάν $\frac{f_1(x)}{f_2(x)} \geq 1$ ενώ εάν $\frac{f_1(x)}{f_2(x)} < 1$ ταξινομείται στον πληθυσμό π_2.

Όταν οι εκ των προτέρων πιθανότητες είναι άγνωστες, αυτές λαμβάνονται να είναι ίσες και ο κανόνας για το ελάχιστο ECM περιλαμβάνει την σύγκριση του λόγου των πυκνότητων πιθανοτήτων προς τον λόγο των κόστων λανθασμένης ταξινόμησης. Εάν δεν μπορούμε να ορίσουμε τον λόγο των κόστων λανθασμένης ταξινόμησης, ο λόγος των πυκνοτήτων πιθανοτήτων συγκρίνεται με τον λόγο των προηγούμενων πιθανοτήτων. Ενώ τέλος, εάν και ο λόγος των εκ των προτέρων πιθανοτήτων και ο λόγος των κόστων λανθασμένης ταξινόμησης είναι μονάδα, οι βέλτιστες περιοχές ταξινόμησης καθορίζονται από την σύγκριση των τιμών των συναρτήσεων πυκνοτήτων πιθανότητων. Δηλαδή, εάν x_0 είναι μια καινούρια παρατήρηση και $f_1(x_0)/f_2(x_0) \geq 1$, τότε η παρατήρηση x_0 κατατάσσεται στον πληθυσμό π_1 και αντίστοιχα εάν $f_1(x_0)/f_2(x_0) < 1$ η παρατήρηση x_0 κατατάσσεται στον πληθυσμό π_2.

Παράδειγμα

Έστω ένας ερευνητής έχει αρκετά δεδομένα για να εκτιμήσει τις συναρτήσεις πυκνοτήτων πιθανοτήτων $f_1(x)$ και $f_2(x)$ που σχετίζονται με τους πληθυσμούς π_1 και π_2 αντίστοιχα. Υποθέτουμε

ότι $c\langle 2|1\rangle = 5$ μονάδες και $c\langle 1|2\rangle = 10$ μονάδες. Επιπλέον, είναι γνωστό ότι περίπου το 20% των παρατηρήσεων ανήκουν στον πληθυσμό π_2. Δηλαδή, οι εκ των προτέρων πιθανότητες είναι $p_1 = 0.8$ και $p_2 = 0.2$.

Δεδομένων των προηγούμενων πιθανοτήτων και των κόστων λανθασμένης ταξινόμησης, μία παρατήρηση κατατάσσεται στον πληθυσμό π_1 εάν $\frac{f_1(x)}{f_2(x)} \geq (\frac{10}{5})(\frac{0.2}{0.8}) = 0.5$ ενώ στον πληθυσμό π2 εάν $\frac{f_1(x)}{f_2(x)} < (\frac{10}{5})(\frac{0.2}{0.8}) = 0.5$.

Υποθέτουμε ότι οι συναρτήσεις πυκνοτήτων πιθανοτήτων που αξιολογούνται σε μια νέα παρατήρηση x_0 δίνουν $f_1(x_0)=0.3$ και $f_2(x_0)=0.4$. Για να αποφασίσουμε σε ποιον πληθυσμό θα κατατάξουμε την νέα παρατήρηση, βρίσκουμε τον λόγο

$$\frac{f_1(x_0)}{f_2(x_0)} = \frac{0.3}{0.4} = 0.75$$

και τον συγκρίνουμε με το 0.5. Συνεπώς, καθώς

$$\frac{f_1(x_0)}{f_2(x_0)} = 0.75 > (\frac{c\langle 1|2\rangle}{c\langle 2|1\rangle})(\frac{p_2}{p_1}) = 0.5$$

η παρατήρηση x_0 κατατάσσεται στον πληθυσμό π_1.

Έπειτα, υποθέτοντας ότι οι συναρτήσεις πυκνότητας πιθανότητας $f_1(x)$ και $f_2(x)$ είναι γνωστές και αγνοώντας τα κόστη λανθασμένης ταξινόμησης, η συνολική πιθανότητα λανθασμένης ταξινόμησης (total probability of misclassification, TPM) ισούται με το άθροισμα, p_1 φορές την πιθανότητα να προέρχεται μια παρατήρηση από τον πληθυσμό π_1 και να κατατάσσεται λανθασμένα, $P(2|1)$ συν p_2 φορές την πιθανότητα να προέρχεται μια παρατήρηση από τον πληθυσμό π_2 και να κατατάσσεται λανθασμένα, $P(1|2)$. Δηλαδή,

$$\begin{aligned} TRM &= p_1 P(2|1) + p_2 P(1|2) \\ &= p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \end{aligned}$$

Μαθηματικά, αυτό το πρόβλημα είναι ισοδύναμο με το να ελαχιστοποιήσουμε το εκτιμώμενο κόστος λανθασμένης ταξινόμησης, όταν τα κόστη λανθασμένης ταξινόμησης είναι ίσα. Τότε, ο βέλτιστος κανόνας ταξινόμησης είναι ισοδύναμος με την (β) ειδική περίπτωση που συναντήσαμε νωρίτερα.

ΚΕΦΑΛΑΙΟ 5^{Ο}

Ταξινόμηση κανονικών πληθυσμών σε δύο ομάδες

Οι κανόνες ταξινόμησης που βασίζονται σε κανονικούς πληθυσμούς υπερισχύουν πρακτικά στην στατιστική λόγω της απλότητας τους και λόγω της λογικά μεγάλης απόδοσης σε σύγκριση με άλλα μοντέλα. Έτσι υποθέτουμε ότι $f_1(x)$ και $f_2(x)$ είναι πολυμεταβλητές κανονικές πυκνότητες πιθάνοτητες, η πρώτη με μέσο μ_1 και πίνακα συνδιασποράς Σ_1 και η δεύτερη αντίστοιχα με μέσο μ_2 και πίνακα συνδιασποράς Σ_2.
Η ειδική περίπτωση όπου έχουμε ίσους πίνακες συνδιασποράς, οδηγεί σε μια ιδιαίτερα απλή γραμμική στατιστική ταξινόμησης.

5.1 Ταξινόμηση κανονικών πληθυσμών όταν $\Sigma_1=\Sigma_2=\Sigma$

Υποθέτουμε ότι οι πυκνότητες πιθανοτήτων $f_1(x)$ και $f_2(x)$ για το διάνυσμα $X'=[X_1, X_2, ..., X_p]$ για τους πληθυσμούς π_1 και π_2 δίνονται από τον τύπο

$$f_i(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)] \text{ για } i=1,2 \qquad (5.1)$$

Υποθέτουμε επίσης ότι οι παράμετροι του πληθυσμού μ_1, μ_2 και Σ είναι γνωστοί. Έπειτα, αφού παραλείψουμε τους όρους $(2\pi)^{p/2}|\Sigma|^{1/2}$ οι περιοχές με τον ελάχιστο εκτιμώμενο κόστος λανθασμένης ταξινόμησης (ECM) είναι ως εξής:

$$R_1 : \exp[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)\geq(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

και (5.2)

$$R_2 : \exp[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)<(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})$$

Δεδομένων των περιοχών R_1 και R_2, κατασκευάζουμε τον κανόνα ταξινόμησης. Δηλαδή:

- Εάν $(\mu_1-\mu_2)'\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)\geq\ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1

(5.3)

- Εάν $(\mu_1-\mu_2)'\Sigma^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)<\ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

Ωστόσο στις περισσότερες περιπτώσεις, οι ποσότητες μ_1, μ_2 και Σ είναι άγνωστες, γι'αυτό και ο κανόνας (5.3) πρέπει να τροποποιηθεί. Οι Wald και Anderson πρότειναν την αντικατάσταση των παραμέτρων από τις αντίστοιχες παραμέτρους των δειγμάτων.
Υποθέτωντας ότι έχουμε n_1 δείγμα παρατηρήσεων του πληθυσμού π_1 και αντίστοιχα n_2 δείγμα

παρατηρήσεων του πληθυσμού π_2 για ένα τυχαίο διάνυσμα $X^{'} = [X_1, X_2, ..., X_p]$, όπου $n_1 + n_2 - 2 \geq p$. Οι αντίστοιχοι πίνακες δεδομένων είναι οι εξής:

$$X_1 = \lfloor x_{11}, x_{12}, ..., x_{1n_1} \rfloor_{(p \text{ x } n_1)} \text{ για τον } \pi_1$$

και (5.4)

$$X_2 = \lfloor x_{21}, x_{22}, ..., x_{2n_2} \rfloor_{(p \text{ x } n_2)} \text{ για τον } \pi_2$$

Από τους πίνακες των δεδομένων, τα διανύσματα των δειγματικών μέσων και των πινάκων συνδιασποράς καθορίζονται ως εξής

$$\bar{x}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j} \text{ και } S_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)^{'} \text{ για τον } \pi_1$$

και (5.5)

$$\bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j} \text{ και } S_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)^{'} \text{ για τον } \pi_2$$

Καθώς ισχύει ότι $\Sigma_1 = \Sigma_2 = \Sigma$, οι πίνακες διασποράς του δείγματος S_1 και S_2 συδυάζονται (pooled) και δίνουν έναν αμερόληπτο εκτιμητή του Σ, το S_{pooled}. Συγκεκριμένα, το

$$S_{pooled} = [\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}]S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}]S_2 \quad (5.6)$$

είναι μια αμερόληπτη εκτίμηση του Σ εάν οι πίνακες δεδομένων X_1 και X_2 περιέχουν τυχαία δείγματα από τους πληθυσμούς π_1 και π_2.
Αντικαθιστώντας το μ_1 με $\bar{x}_1$, το μ_2 με $\bar{x}_2$, και το Σ με S_{pooled} στην (5.3) παίρνουμε τον εξής κανόνα ταξινομήσης για το "δείγμα"

- Εάν $(\bar{x}_1 - \bar{x}_2)^{'} S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^{'} S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατήρηση x ταξινομείται στον πληθυσμό π_1

(5.7)

- Εάν $(\bar{x}_1 - \bar{x}_2)^{'} S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^{'} S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) < \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατήρηση x ταξινομείται στον πληθυσμό π_2

Εάν υποθέσουμε ότι

$$(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1}) = 1$$

τότε ln(1)=0, δηλαδή

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq 0$$

$$\Rightarrow (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x \geq \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$$

Για τον κανόνα εκτίμησης ελάχιστου κόστους λανθασμένης ταξινόμησης, συγκρίνουμε την τυποποιημένη μεταβλητή

$$\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x = \hat{a}' x \tag{5.8}$$

με τον αριθμό

$$\begin{aligned} \hat{m} &= (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \\ &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \end{aligned} \tag{5.9}$$

όπου

$$\bar{y}_1 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1) = \hat{a}' \bar{x}_1$$

και

$$\bar{y}_2 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_2) = \hat{a}' \bar{x}_2$$

Δηλαδή, ο εκτιμώμενος ελάχιστος ECM κανόνας για δύο κανονικούς πληθυσμούς είναι ισοδύναμος με την δημιουργία δύο νέων μεταβλητών παίρνοντας τον γραμμικό συνδυασμό των παρατηρήσεων από τους πληθυσμούς π_1 και π_2 και ταξινομώντας μια παρατήρηση x_0 στον π_1 ή π_2.
Όταν οι εκτιμήσεις των παραμέτρων εισαχθούν για τις άγνωστες ποσότητες του πληθυσμού, ο κανόνας ταξινόμησης που θα προκύψει δεν είναι σίγουρο ότι θα μειώσει το κόστος της λανθασμένης ταξινόμησης. Αυτό συμβαίνει γιατί ο κανόνας (5.3) περιέχει την υπόθεση ότι οι πολυμεταβλητές κανονικές πυκνότητες $f_1(x)$ και $f_2(x)$ είναι γνωστές.
Συνοψίζοντας, εάν τα δεδομένα ακολουθούν κανονική κατανομή, το αριστερό μέλος της (5.7) μπορεί να υπολογιστεί για κάθε νέα παρατήρηση x_0. Αυτές οι παρατηρήσεις ταξινομούνται συγκρίνοντας τις τιμές που δίνει το αριστερό μέλος της (5.7) με την τιμή 0.

Τυποποίηση (Scaling)

Ο συντελεστής του διανύσματος $\hat{a} = S_{pooled}^{-1}(\hat{x}_1 - \hat{x}_2)$ δεν είναι μοναδικός και κάθε διάνυσμα της μορφής $c\hat{a}$ με $c \neq 0$ είναι συντελεστής διαχωρισμού.
Το διάνυσμα $\hat{a}$ συχνά "τυποποιείται" ή "κανονικοποιείται" για να είναι μοναδικό. Οι δύο πιο συχνοί τρόποι τυποποίησης είναι οι εξής:

1. Θέτουμε

$$\tilde{a} = \frac{\hat{a}}{\sqrt{\hat{a}'\hat{a}}} \tag{5.10}$$

έτσι ώστε το $\tilde{a}$ να έχει μοναδίαιο μήκος.

2. Θέτουμε

$$\tilde{a} = \frac{\hat{a}}{\hat{a}_1} \tag{5.11}$$

έτσι ώστε το πρώτο στοιχείο του διανύσματος $\tilde{a}$ να είναι 1.

Και στις δύο περιπτώσεις το $\tilde{a}$ είναι της μορφής $c\hat{a}$. Στην πρώτη περίπτωση $c = (\hat{a}'\hat{a})^{-1/2}$ και στην δεύτερη $c = \hat{a}_1^{-1}$.
Τα μεγέθη $\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_p$ στην (5.10) βρίσκονται στο διάστημα [-1,1]. Η διατήρηση των μεγεθών στο διάστημα [-1,1] χρησιμεύει στην οπτική σύγκριση των συντελεστών. Στην (5.11) το $\tilde{a}_1 = 1$ και τα $\tilde{a}_2, ..., \tilde{a}_p$ εκφράζονται ως πολλαπλάσια του $\tilde{a}_1$. Ομοίως, το να εκφράζουμε τους συντελεστές ως πολλαπλάσια του $\tilde{a}_i$ μας επιτρέπει να ελέγξουμε την σημαντικότητα των μεταβλητών $X_2, ..., X_p$ ως διαχωριστικούς συντελεστές.

5.2 Ταξινόμηση κανονικών πληθυσμών όταν $\Sigma_1 \neq \Sigma_2$

Όπως είναι αναμενόμενο, οι κανόνες ταξινόμησης είναι πιο πολύπλοκοι όταν οι πίνακες συνδιασποράς είναι άνισοι.
Έστω οι πυκνότητες πιθανοτήτων από την (5.1), όπου αντικαθιστούμε το Σ με Σ_i, i=1,2. Δηλαδή, οι πίνακες συνδιασποράς, όπου και οι μέσοι των διανυσμάτων είναι διαφορετικοί για τους δύο πληθυσμούς. Όπως έχουμε δει, οι περιοχές ελάχιστου ECM και ελάχιστης συνολικής πιθανότητας λανθασμένης ταξινόμησης (TPM) εξαρτώνται από το κλάσμα πυκνοτήτων, $f_1(x)/f_2(x)$ ή ισοδύναμα από τον φυσικό λογάριθμο της πυκνότητας πιθανότητας, $\ln[f_1(x)/f_2(x)] = \ln[f_1(x)] - \ln[f_2(x)]$. Όταν οι κανονικές πυκνότητες έχουν διαφορετικές συνδιασπορές, δεν μπορούμε να παραλείψουμε τους όρους $|\Sigma_i|^{1/2}$, όπως συμβαίνει συμβαίνει στην περίπτωση $\Sigma_1=\Sigma_2$.
Αντικαθιστώντας τις πολυμεταβλητές κανονικές πυκνότητες με διαφορετικούς πίνακες, παίρνοντας τον λογάριθμο και απλοποιώντας, οι περιοχές ταξινόμησης είναι οι εξής

$$\begin{aligned} R_1 &: -\frac{1}{2}x'(\Sigma_1' - \Sigma_2')x + (\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1})'x - k \geq \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})] \\ R_2 &: -\frac{1}{2}x'(\Sigma_1' - \Sigma_2')x + (\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1})'x - k < \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})] \end{aligned} \tag{5.12}$$

όπου

$$k = \frac{1}{2}\ln(\frac{|\Sigma_1|}{|\Sigma_2|}) + \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2) \tag{5.13}$$

Ο κανόνας ταξινόμησης βγαίνει άμεσα από τον κανόνα (5.12). Δηλαδή

- Εάν $-\frac{1}{2}x'(\Sigma_1' - \Sigma_2')x + (\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1})'x - k \geq \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατηρήση x ταξινομείται στον πληθυσμό π_1

(5.14)

- Εάν $-\frac{1}{2}x'(\Sigma_1' - \Sigma_2')x + (\mu_1\Sigma_1^{-1} - \mu_2\Sigma_2^{-1})'x - k < \ln[(\frac{c(1|2)}{c(2|1)})(\frac{p_2}{p_1})]$ η παρατηρήση x ταξινομείται στον πληθυσμό π_2

Στην πράξη, στον κανόνα ταξινόμησης αντικαθίστανται τα $\bar{x}_1, \bar{x}_2$ από τα μ_1, μ_2 αντίστοιχα και τα S_1, S_2 από τα Σ_1, Σ_2 αντίστοιχα.

5.3 Αξιολόγηση κανόνων ταξινόμησης

Εφ'οσον φτιάξουμε τις συναρτήσεις κατάταξης, το επόμενο βήμα είναι να τις αξιολογήσουμε, δηλαδή αν ελέγξουμε πόσο καλά αυτές οι συναρτήσεις διαχωρίζουν τους πληθυσμούς.

Ένα τρόπος αξιολόγησης των συναρτήσεων ταξινόμησης είναι ο υπολογισμός του ρυθμού σφάλματος

Στην περίπτωση που οι πληθυσμοί π_1 και π_2 είναι γνωστοί, μπορούμε να υπολογίσουμε την συνολική πιθανότητα λανθασμένης ταξινόμησης (TPM). Η ελάχιστη τιμή που μπορεί να πάρει το TPM είναι ο βέλτιστος βαθμός σφάλματος (optimum error rate, OER) όπου

$$\begin{aligned} OER &= \min TPM = \min(p_1 P(2|1) + p_2 P(1|2)) \\ &= \min p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx \end{aligned} \tag{5.15}$$

Όμως, επειδή οι πληθυσμοί δεν είναι εντελώς γνωστοί χρησιμοποιούμε τις δειγματικές συναρτήσεις ταξινόμησης, οι οποίες αξιολογούνται από τον πραγματικό βαθμό σφάλματος (actual error rate, AER) όπου

$$AER = p_1 \int_{\hat{R}_1} f_1(x)dx + p_2 \int_{\hat{R}_2} f_2(x)dx \tag{5.16}$$

όπου $\hat{R}_1$ και $\hat{R}_2$ παριστάνουν τις περιοχές κατάταξης όταν χρησιμοποιούμε δύο δείγματα μεγέθους n_1 και n_2 για να εκτιμήσουμε τους πληθυσμούς π_1 και π_2 αντίστοιχα.

Όμως επειδή οι συναρτήσεις πυκνότητας πιθανότητας $f_1(x)$ και $f_2(x)$ είναι άγνωστες δεν μπορούμε να υπολογίσουμε τον πραγματικό ρυθμό σφάλματος. Για αυτό τον λόγο χρησιμοποιούμε ένα άλλο μέτρο αξιολόγησης, που δεν εξαρτάται από καμία διαδικασία ταξινόμησης. Αυτό το μέτρο ονομάζεται φαινομενικός βαθμός σφάλματος (apparent error rate, APER) και ορίζεται ως το κλάσμα του αριθμού των λανθασμένων ταξινομημένων παρατηρήσεων προς τον συνολικό αριθμό παρατηρήσεων. Δηλαδή

$$APER = \frac{n_{1E} + n_{2E}}{n_1 + n_2} \tag{5.17}$$

Για τον υπολογισμό του APER χρησιμοποιείται χρησιμοποιείται ένας πίνακας ταξινόμησης (Πίνακας 5.3.1) που έχει ως εξής

Πραγματικός πληθυσμός	Προβλεπόμενος πληθυσμός π_1	Προβλεπόμενος πληθυσμός π_2	Μέγεθος δείγματος
π_1	n_{1c}	$n_{1E}=n_1-n_{1c}$	n_1
π_2	$n_{2E}=n_2-n_{2c}$	n_{2c}	n_2

Πίνακας 5.3.1: Πίνακας ταξινόμησης

Όπου:

n_1: αριθμός των παρατηρήσεων από τον πληθυσμό π_1

n_2: αριθμός των παρατηρήσεων από τον πληθυσμό π_2

n_{1c}: αριθμός των παρατηρήσεων του πληθυσμού π_1 που ταξινομούνται σωστά στο π_1

n_{2c}: αριθμός των παρατηρήσεων του πληθυσμού π_2 που ταξινομούνται σωστά στο π_2

n_{1E}: αριθμός των παρατηρήσεων του πληθυσμού π_1 που λανθασμένα ταξινομούνται στο π_2

n_{2E}: αριθμός των παρατηρήσεων του πληθυσμού π_2 που λανθασμένα ταξινομούνται στο π_1

Ο APER είναι μια εκτίμηση της πιθανότητας με την οποία ένας κανόνας ταξινόμησης που βασίζεται σε ένα δείγμα θα δώσει λανθασμένη ταξινόμηση μιας παρατήρησης. Τα μεγάθη των δειγμάτων n_1 και n_2 πρέπει να είναι αρκετά μεγάλα για να δώσει καλά αποτελέσματα ο APER. Αλλά επειδή τα ίδια δεδομένα χρησιμοποιούνται και για να κατασκευάσουν αλλά και για να αξιολογήσουν έναν κανόνα ταξινόμησης, ο APER τείνει να υποτιμήσει το AER.
Για να εξαλειφθεί η μεροληψία (bias) στον APER, μπορούμε να χωρίσουμε το δείγμα σε δύο μέρη. Το ένα θα είναι το δείγμα εκπαίδευσης (training sample) και το άλλο το δείγμα επικύρωσης (validation sample). Ο κανόνας ταξινόμησης δημιουργείται χρησιμοποιώντας το δοκιμαστικό δείγμα και η αξιολόγηση του γίνεται χρησιμοποιώντας το δείγμα επικύρωσης. Ο βαθμός σφάλματος είναι η αναλογία των λανθασμένων κατατάξεων στο δείγμα επικύρωσης.
Αυτή η διαδικασία έχει τα παρακάτω δυο σημαντικά μειονεκτήματα

- απαιτείται μεγάλο μέγεθος δείγματος και
- καθώς ο κανόνας ταξινόμησης βασίζεται σε ένα υποσύνολο του δείγματος, μπορεί να είναι ένας "αδύναμος" εκτιμητής της συνάρτησης ταξινόμησης του δειγμάτος, αναλόγως με τον διαχωρισμό.

Ένας εναλλακτικός μη παραμετρικός τρόπος αξιολόγησης μελετήθηκε από τους Lachenbruch και Mickey (1936). Η διαδικασία έχει ως εξής.

- Ξεκινώντας από τον πληθυσμό π_1, παραλείπουμε μια παρατήρηση και αναπτύσσουμε έναν κανόνα ταξινόμησης που βασίζεται στις παρατηρήσεις n_1-1 και n_2.
- Ταξινομούμε την "holdout" παρατήρηση χρησιμοποιώντας τον κανόνα που κατασκευάστηκε στον βήμα 1.

- Επαναλαμβάνουμε τα βήματα 1 και 2 μέχρι όλες οι παρατηρήσεις να ταξινομηθούν και ορίζουμε ως n_{1E}^{H} τον αριθμό των "holdout" παρατηρήσεων που λανθασμένα ταξινομήθηκαν στον πληθυσμό π_1.

- Έπειτα επαναλαμβάνουμε τα βήματα 1 εώς 3 για τον πληθυσμό π_2. Ορίζουμε ως n_{2E}^{H} τον αριθμό των "holdout" παρατηρήσεων που λανθασμένα ταξινομήθηκαν σ' αυτόν τον πληθυσμό.

Οι εκτιμώμενες πιθανότητες των δεσμευμένων πιθανοτήτων λανθασμένης ταξινόμησης ορίζονται ως:

$$\hat{p}(2 \mid 1) = n_{1E}^{(H)} / n_1$$

και (5.18)

$$\hat{p}(1 \mid 2) = n_{2E}^{(H)} / n_2$$

Η εκτίμηση του συνολικού ρυθμού σφάλματος $A\hat{P}AR$ ορίζεται ως:

$$A\hat{P}AR = \frac{n_{1E}^{H} + n_{2E}^{H}}{n_1 + n_2} \tag{5.19}$$

που είναι μια αμερόληπτη (unbiased) εκτίμηση του αναμενόμενου πραγματικού ρυθμού σφάλματος.

5.4 Διαχωριστική συνάρτηση του Fisher – Διαχωρισμός δύο πληθυσμών

Ο Fisher κατέληξε στον κανόνα ταξινόμησης που προκύπτει από την εξίσωση $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x = \hat{a}' x$ μετατρέποντας τις πολυμεταβλητές παρατηρήσεις x σε μονομεταβλητές παρατηρήσεις y έτσι ώστε οι παρατηρήσεις y που προέρχονται από τους πληθυσμούς π_1 και π_2 να διαχωρίζονται όσο το δυνατόν καλύτερα. Ο Fisher πρότεινε τους γραμμικούς συνδυασμούς του x για να δημιοργήσει τα y επειδή τα y είναι συναρτήσεις του x που αντιμετωπίζονται εύκολα. Η προσέγγιση του Fisher δεν προυποθέτει να είναι οι πληθυσμοί κανονικοί. Ωστόσο, προυποθέτει οι πίνακες συνδιασποράς να είναι ίσοι, επειδή χρησιμοποιείται συγκεντρωμένη εκτίμηση (pooled estimate) του κοινού πίνακα συνδιασποράς.
Η μετατροπή των x σε y (μονοδιάστατα σκορ) γίνεται μέσω της διαχωριστικής συνάρτησης. Τα σκορ των πληθυσμών θα πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένα, έτσι ώστε να μπορούμε να διαχωρίσουμε τους δύο πληθυσμούς και να κατατάσσουμε κάθε καινούρια παρατήρηση σε έναν από τους δύο πληθυσμούς.
Έστω ότι τα σκορ δίνονται ως $y_{11}, y_{12}, ..., y_{1n_1}$ για τις παρατηρήσεις του πρώτου πληθυσμού και $y_{21}, y_{22}, ..., y_{2n_2}$ για τις παρατηρήσεις του δεύτερου πληθυσμού. Ένα μέτρο απόστασης αυτών των σκορ δίνεται από την σχέση:

$$a_s = \frac{|\bar{y}_1 - \bar{y}_2|}{S}$$

όπου

$S^2 = \dfrac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$ είναι η συγκεντρωμένη εκτίμηση της διασποράς και $\bar{y}_i$ είναι η δειγματικός μέσος των παρατηρήσεων για κάθε ομάδα *i* (*i*=1,2).

Δηλαδή ο Fisher πήρε την απόσταση των μέσων των δυο πληθυσμών διαιρούμενο από την τυπική απόκλιση, έτσι ώστε να απαλλαγεί από τις μονάδες μέτρησης. Σκοπός είναι να μεγιστοποιήσουμε την απόσταση α_s, ή αντίστοιχα την απόσταση a_s^2, καθώς αυτό σημαίνει ότι τα σκορ των δυο πληθυσμών θα είναι όσο γίνεται πιο διαφορετικά μεταξύ τους. Δηλαδή έχουμε

$$\begin{aligned} a_s^2 &= \frac{|\,\bar{y}_1 - \bar{y}_2\,|^2}{S^2} \\ &= \frac{(\hat{a}'\bar{x}_1 - \hat{a}'\bar{x}_2)^2}{\hat{a}' S_{pooled}\hat{a}} \\ &= \frac{(\hat{a}'d)^2}{\hat{a}' S_{pooled}\hat{a}} \end{aligned} \qquad (5\text{-}20)$$

για όλα τα πιθανά διανύσματα συντελεστών $\hat{a}$ όπου $d = (\bar{x}_1 - \bar{x}_2)$.

Για να βρούμε το μέγιστο της (5-20) χρησιμοποιούμε την σχέση $\max = \dfrac{(x'd)^2}{x'Bx} = d'B^{-1}d$ για $x \neq 0$.

Οπότε έχουμε

$$\max \frac{(a'd)^2}{\hat{a}' S_{pooled}\hat{a}} = d' S_{pooled}^{-1} d = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 - \bar{x}_2) = D^2$$

Για να ολοκληρωθεί ο διαχωριστικός κανόνας ορίζουμε την κρίσιμη τιμή *m*, που είναι η μέση τιμή των $\bar{y}_1$ και $\bar{y}_2$. Δηλαδή έχουμε:

$$m = \frac{\bar{y}_1 + \bar{y}_2}{2} = \frac{\hat{a}'x_1 + \hat{a}'x_2}{2} = \frac{(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2)}{2}$$

- Εάν $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x \geq m$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_1

(5.21)

- Εάν $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x < m$ η παρατήρηση *x* ταξινομείται στον πληθυσμό π_2

Η διαχωριστική συνάρτηση του Fisher αναπτύχθηκε υπό την υπόθεση ότι οι δύο πληθυσμοί έχουν κοινό πίνακα συνδιασποράς. Ο πρώτος όρος, $\hat{y} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x$, στον κανόνα ταξινόμησης (5.7) είναι η γραμμική συνάρτηση που αποκτήθηκε από τον Fisher που μεγιστοποιεί την μεταβλητότητα "μεταξύ" των δειγμάτων σχετικά με την μεταβλητότητα "μέσα" στα δείγματα. Η σχέση

$$\hat{w} = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$$
$$= (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}[x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)] \quad (5.22)$$

ονομάζεται συνάρτηση ταξινόμησης του Anderson. Δεδομένου ότι οι κανονικοί πληθυσμοί έχουν τον ίδιο πίνακα συνδιασποράς, ο κανόνας ταξινόμησης του Fisher είναι ισοδύναμος με τον ελάχιστο κανόνα ECM με ίσες προηγούμενες πιθανότητες και ίσα κόστη λανθασμένης ταξινόμησης.
Συνοψίζοντας, ο μέγιστος διαχωρισμός που μπορεί να αποκτηθεί λαμβάνοντας υπ'όψιν γραμμικούς συνδυασμούς πολυμεταβλητών παρατηρήσεων είναι ίσος με την απόσταση D^2. Η απόσταση αυτή μπορεί να χρησιμοποιηθεί για να εξετασθεί εάν οι μέσοι μ_1 και μ_2 διαφέρουν σημαντικά. Συνεπώς, ο έλεγχος για την διαφορά των μέσων μπορεί να ληφθεί ως ένας έλεγχος για την "σημαντικότητα" του διαχωρισμού που μπορεί να επιτευχθεί.
Έστω λοιπόν, οι πληθυσμοί π_1 και π_2 είναι κανονικοί με κοινό πίνακα συνδιασποράς Σ.
Τότε, ο έλεγχος :

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

ισοδυναμεί με το αν

$$[\frac{n_1 + n_2 - p + 1}{(n_1 + n_2 - 2)p}][\frac{n_1 n_2}{n_1 + n_2}]D^2 \sim F_{(p, n1+n2-p-1)}$$

Αν απορριφθεί η H_0, συμπεραίνουμε ότι ο διαχωρισμός μεταξύ των πληθυσμών π_1 και π_2 είναι σημαντικός.
Ωστόσο, σημαντικός διαχωρισμός δεν σημαίνει απαραίτητα και καλή ταξινόμηση. Αλλά από την άλλη, εάν ο διαχωρισμός δεν είναι σημαντικός, η έρευνα για μια χρήσιμη ταξινόμηση δεν θα επιφέρει αποτελέσματα.

5.5 Ταξινόμηση για περισσότερους από δύο πληθυσμούς

Θεωρητικά, η γενίκευση των κανόνων ταξινόμησης για περισσότερους από δύο πληθυσμούς είναι άμεση. Ωστόσο, δεν είναι γνωστά τα πάντα για τις συναρτήσεις ταξινόμησης και συγκεκριμένα, οι ρυθμοί σφάλματος δεν έχουν ερευνηθεί τελείως.

Εκτιμώμενο ελάχιστο κόστος για την μέθοδο λανθασμένης ταξινόμησης

Έστω $f_i(x)$ οι πυκνότητες πιθανοτήτων για τους πληθυσμούς π_i για $i=1,2,...,g$. Ορίζουμε

p_i = εκ των προτέρων πιθανότητες για τους πληθυσμούς π_i για $i=1,2,...,g$

$c = (k \mid i)$ = κόστος ταξινόμησης μιας παρατήρησης στον πληθυσμό π_k
ενώ ανήκει στον πληθυσμό π_i για k, $i=1,2,...,g$

Για $k=i$, c(i|i)=0. Τέλος, R_k ορίζουμε το σύνολο των x που ταξινομούνται ως π_k, και

$$P(k \mid i) = P(x \in R_k \mid \pi_i) = \int_{R_k} f_i(x)dx$$

για $k, i=1,2,...,g$ όπου $P(i \mid i) = 1 - \sum_{k=1,k\neq i}^{g} P(k \mid i)$

Το εκτιμώμενο κόστος λανθασμένης ταξινόμησης μιας παρατήρησης x από τον πληθυσμό π_1 στον π_2, ή π_3,...., ή π_g είναι

$$\begin{aligned} ECM(1) &= P(2 \mid 1)c(2 \mid 1) + P(3 \mid 1)c(3 \mid 1) + ... + P(g \mid 1)c(g \mid 1) \\ &= \sum_{k=2}^{g} P(k \mid 1)c(k \mid 1) \end{aligned}$$

Με όμοιο τρόπο βρίσκουμε τα εκτιμώμενα κόστη λανθασμένης ταξινόμησης *ECM(2),...,ECM(g)*. Πολλαπλασιάζοντας κάθε ECM με την εκ των προτέρων πιθανότητα και αθροίζοντας τα βρίσκουμε το ολικό ECM:

$$\begin{aligned} ECM &= p_1 ECM(1) + p_2 ECM(2) + ... + p_k ECM(k) \\ &= p_1 (\sum_{g=2}^{k} P(g \mid 1)c(g \mid 1)) + p_1 (\sum_{g=1,g\neq 2}^{k} P(g \mid 2)c(g \mid 2)) + ... + p_1 (\sum_{g=1}^{k-1} P(g \mid k)c(g \mid k)) \\ &= \sum_{i=1}^{k} p_i (\sum_{g=1,g\neq i}^{k} P(g \mid i)c(g \mid i)) \end{aligned} \qquad (5.23)$$

Ο καθορισμός μιας βέλτιστης διαδικασία ταξινόμησης είναι ισοδύναμος με τον καθορισμό των περιοχών ταξινόμησης $R_1, R_2, ..., R_g$ τέτοιες ώστε η (5.23) να είναι ελάχιστη.

Οι περιοχές ταξινόμησης που ελαχιστοποιούν το ECM (5.23) ορίζονται κατατάσσοντας την παρατήρηση x στον πληθυσμό π_k, $\kappa=1,2,...,g$ για τον οποίο η ποσότητα

$$\sum_{i=1,i\neq k}^{g} p_i f_i(x)c(k \mid i) \qquad (5.24)$$

είναι ελάχιστη.

Υποθέτοντας ότι τα κόστη λανθασμένης ταξινόμησης είναι ίσα, κατατάσσουμε την παρατήρηση x στον πληθυσμό π_k, $k=1,2,...,g$ για τον οποίο η ποσότητα

$$\sum_{i=1,i\neq k}^{g} p_i f_i(x) \qquad (5.25)$$

είναι ελάχιστη.

Επομένως, όταν τα κόστη λανθασμένης ταξινόμησης είναι ίσα ο κανόνας ταξινόμησης γίνεται:

- Εάν $p_k f_k(x) > p_i f_i(x)$, για κάθε $i \neq k$, η παρατήρηση x κατατάσσεται στον πληθυσμό π_k (5.26)

Πρέπει να έχουμε υπ'όψιν ότι οι κανόνες ελαχιστοποιήσης λανθασμένης ταξινόμησης προυποθέτουν την εκτίμηση των εκ των προτέρων πιθανοτήτων, των κοστών λανθασμένης ταξινόμησης και των πυκνοτήτων πιθανοτήτων πριν εφαρμοστούν οι κανόνες.

Παράδειγμα

Έστω ότι θέλουμε να κατατάξουμε μια παρατήρηση x_0 σε έναν από τους πληθυσμούς π_1,π_2 ή π_3. Οι εκ των προτέρων πιθανότητες, τα κόστη λανθασμένης ταξινόμησης και οι πυκνότητες είναι οι εξής:

	π_1	π_2	π_3
π_1	c(1|1)=0	c(1|2)=500	c(1|3)=100
π_2	c(2|1)=10	c(2|2)=0	c(2|1)=50
π_3	c(3|1)=50	c(3|2)=200	c(3|3)=0
Προηγούμενες πιθανότητες	$p_1 = 0.05$	$p_2 = 0.6$	$p_3 = 0.35$
Πυκνότητες στο x_0	$f_1(x_0) = 0.01$	$f_2(x_0) = 0.85$	$f_3(x_0) = 2$

Οι τιμές $\sum_{i=1, i\neq g}^{3} p_i f_i(x_0) c(g \mid i)$ για κάθε g είναι οι εξής:

g=1: $p_2 f_2(x_0) c(1 \mid 2) + p_3 f_3(x_0) c(1 \mid 3)$
$= (0.6)(0.85)(500) + (0.35)(2)(100) = 325$

g=2: $p_1 f_1(x_0) c(2 \mid 1) + p_3 f_3(x_0) c(2 \mid 3)$
$= (0.05)(0.01)(10) + (0.35)(2)(50) = 35.055$

g=3: $p_1 f_1(x_0) c(3 \mid 1) + p_2 f_2(x_0) c(3 \mid 2)$
$= (0.05)(0.01)(50) + (0.6)(0.85)(200) = 102.025$

Καθώς το $\sum_{i=1, i\neq k}^{3} p_i f_i(x_0) c(k \mid i)$ παίρνει ελάχιστη τιμή για k=2, θα κατατάξουμε την παρατήρηση x_0 στον π_2.

Εάν τα κόστη λανθασμένης ταξινόμησης είναι ίσα, θα κατατάξουμε την παρατήρηση x_0 σύμφωνα με την σχέση $\sum_{i=1, i\neq k}^{g} p_i f_i(x)$. Επομένως έχουμε

$p_1 f_1(x_0) = (0.05)(0.01) = 0.0005$

$p_2 f_2(x_0) = (0.60)(0.85) = 0.510$

$p_3 f_3(x_0) = (0.35)(2) = 0.700$

Καθώς $p_3 f_3(x_0) = 0.700 \geq p_i f_i(x_0)$ για i=1,2
η παρατήρηση x_0 κατατάσσεται στον πληθυσμό π_3.

5.6 Διαχωριστική συνάρτηση Fisher για *g* ομάδες

Ο Fisher επέκτεινε την μέθοδο διαχωρισμού και για *g* ομάδες. Και σ'αυτή την περίπτωση δεν είναι απαραίτητη η προυπόθεση να ακολουθούν οι πληθυσμοί κανονική κατανομή, αλλά θα πρέπει οι πίνακες συνδιασποράς κάθε πληθυσμού να είναι ίσοι, δηλαδή $\Sigma_1=\Sigma_2=...=\Sigma_g=\Sigma$.

Έστω ότι $\bar{\mu}$ είναι το διάνυσμα της μέσης τιμής όλων των πληθυσμών, δηλαδή $\bar{\mu} = \frac{1}{g}\sum_{i=1}^{g} \mu_i$ και B_μ τα αθροίσματα μεταξύ των ομάδων $B_\mu = \sum_{i=1}^{g} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$. (5.27)

Έστω Y οι μεταβλητές που προκύπτουν από τον γραμμικό μετασχηματισμό των x παρατηρήσεων, τότε

$$Y = a'X$$

με μέση τιμή

$$E(Y) = a'E(X \mid \pi_i) = a'\mu_i \text{ για τον πληθυσμό } \pi_i$$

και διασπορά

$$Var(Y) = a'Cov(X)a = a'\Sigma a \text{ για όλους τους πληθυσμούς}$$

Συνεπώς, η τιμή $\mu_{iY} = a'\mu_i$ μεταβάλλεται καθώς αλλάζει ο πληθυσμός που επιλέγεται. Ο ολικός μέσος ορίζεται ως εξής

$$\bar{\mu}_Y = \frac{1}{g}\sum_{i=1}^{g} \mu_{iY} = \frac{1}{g}\sum_{i=1}^{g} a'\mu_i = a'\left(\frac{1}{g}\sum_{i=1}^{g} \mu_i\right) = a'\bar{\mu}$$

και σχηματίζουμε τον λόγο

$$\frac{\sum_{i=1}^{g} (\mu_{iY} - \bar{\mu})^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{g} (a'\mu_i - a'\bar{\mu})^2}{a'\Sigma a} = \frac{a'\left(\sum_{i=1}^{g} (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'\right)a}{a'\Sigma a} = \frac{a'B_\mu a}{a'\Sigma a} \quad (5.28)$$

ο οποίος αποτελεί ένα μέτρο μεταβλητότητας μεταξύ των ομάδων (variability between groups) των Y τιμών σε σχέση με την κοινή μεταβλητότητα ανάμεσα στις ομάδες (variability within groups). Τότε μπορούμε να επιλέξουμε το a που μεγιστοποιεί αυτόν τον λόγο, ώστε να πετύχουμε τον καλύτερο διαχωρισμό.

Καθώς στην πράξη τα Σ και μ_i είναι άγνωστα τα εκτιμούμε από τα S_{pooled} και $\bar{x}_i$ αντίστοιχα. Υποθέτουμε ότι έχουμε ένα δοκιμαστικό σύνολο δεδομένων που αποτελείται από ένα τυχαίο δείγμα μεγέθους n_i από τον πληθυσμό π_i $i=1,2,...,g$. Η δειγματική μέση τιμή για τον πληθυσμό π_i είναι $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ ενώ η ολική μέση τιμή είναι:

$$\bar{x} = \frac{\sum_{i=1}^{g} n_i \bar{x}_i}{\sum_{i=1}^{n} n_i} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{g} n_i}$$

το οποίο διάνυσμα είναι *p x 1*διάστασης.
Έπειτα, ορίζουμε τον πίνακα ανάμεσα στις ομάδες του δείγματος

$$B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{5.29}$$

Επίσης, μια εκτίμηση του Σ βασίζεται στον πίνακα εντός των ομάδων του δείγματος

$$W = \sum_{i=1}^{g} (n_i - 1) S_i = \sum_{i=1}^{g}\sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_i)(\bar{x}_{ij} - \bar{x}_i)' \tag{5.30}$$

Συνεπώς, μία εκτίμηση του Σ είναι $W/(n_1 + n_2 + ... + n_g - g) = S_{pooled}$. Επομένως το ίδιο $\hat{a}$ που μεγιστοποιεί το $\hat{a}'B\hat{a}/\hat{a}'S_{pooled}a$ μεγιστοποιεί και το $\hat{a}'B\hat{a}/\hat{a}'Wa$. Εξάλλου μπορούμε να παρουσιάσουμε το βέλτιστο $\hat{a}$ στην περισσότερο συνηθισμένη μορφή ως ιδιοδιανύσματα $\hat{e}_i$ του πίνακα $W^{-1}B$, επειδή εάν το $W^{-1}Be = \hat{\lambda}\hat{e}$ τότε $S_{pooled}^{-1}B\hat{e} = \hat{\lambda}(n_1 + n_2 + ... + n_g - g)\hat{e}.$

Έστω $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_s > 0$, όπου $s \le \min(g-1, p)$ είναι μη μηδενικές ιδιοτιμές του $W^{-1}B$ και $\hat{e}_1, \hat{e}_2, ..., \hat{e}_s$ είναι τα αντίστοιχα ιδιοδιανύσματα, έτσι ώστε $\hat{e}'S_{pooled}\hat{e} = 1$. Τότε, το διάνυσμα των συντελεστών $\hat{a}$ το οποίο μεγιστοποιεί τον λόγο

$$\frac{\hat{a}'B\hat{a}}{\hat{a}'W\hat{a}} = \frac{\hat{a}'(\sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})\hat{a}}{\hat{a}'(\sum_{i=1}^{g}\sum_{j=1}^{n_i} (\bar{x}_{ij} - \bar{x}_i)(\bar{x}_{ij} - \bar{x}_i)'\hat{a}} \tag{5.31}$$

δίνεται από $\hat{a}_1 = \hat{e}_1$. Ο γραμμικός συνδυασμός των $\hat{a}_1'x$ ονομάζεται απλή διαχωριστική συνάρτηση. Η επιλογή $\hat{a}_2 = \hat{e}_2$ δίνει τον δεύτερο δειγματικό διαχωριστή, $\hat{a}_2'x$, και συνεχίζοντας, παίρνουμε $\hat{a}_k'x = \hat{e}_k'x.$ τον $k_{\text{οστό}}$ διαχωριστή του δείγματος, $k \le s$.

Χρήση των διαχωριστών του Fisher για την ταξινόμηση των παρατηρήσεων

Οι διαχωριστές του Fisher δημιουργήθηκαν για να αποκτηθεί χαμηλοδιάστατη αναπαράσταση των δεδομένων η οποία διαχωρίζει τους πληθυσμούς όσο το δυνατόν καλύτερα.
Θέτοντας

$$Y_k = a_k^{'} X \;\; k_{\text{οστός}}\text{διαχωριστής, με } k \le s \tag{5.32}$$

συμπεραίνουμε ότι

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} \text{ έχει μέσο } \mu_{iY} = \begin{bmatrix} \mu_{iY1} \\ \vdots \\ \mu_{iYs} \end{bmatrix} = \begin{bmatrix} a_1^{'}\mu_i \\ \vdots \\ a_s^{'}\mu_i \end{bmatrix}$$

υπό τον πληθυσμό π_i και πίνακα συνδιασποράς I, για όλους τους πληθυσμούς.
Καθώς τα συστατικά του Y έχουν διασπορές μονάδες και μηδενική συνδιασπορά, το κατάλληλο μέτρο της απόστασης τετραγώνων από το $Y=y$ μέχρι το μ_{iy} είναι

$$(y-\mu_{iY})^{'}(y-\mu_{iY}) = \sum_{j=1}^{s}(y_j - \mu_{iY_j})^2$$

Ένας κανόνας ταξινόμησης είναι ο εξής: η παρατήρηση x ταξινομείται στον πληθυσμό π_k εάν το τετράγωνο της απόστασης του y μέχρι το μ_{kY} είναι μικρότερο από το τετράγωνο της απόστασης από το y μέχρι το μ_{iY} για $i \neq k$.

Εάν χρησιμοποιηθούν r διαχωριστές για την ταξινόμηση, ο κανόνας είναι ο εξής:
Ταξινομείται η παρατήρηση x στον πληθυσμό π_k εάν

$$\sum_{j=1}^{r}(y_j - \mu_{kY_j})^2 = \sum_{j=1}^{r}[a_j^{'}(x-\mu_k)]^2 \le \sum_{j=1}^{r}a_j^{'}(x-\mu_i)]^2 \text{ για όλα τα } i \neq k \tag{5.33}$$

Σε αυτό το σημείο θα ελέγχξουμε τον περιορισμό ότι έοχυμε s διαχωριστές, δηλαδή ότι έχουμε s μη μηδενικές ιδιοτιμές του πίνακα $\Sigma^{-1}B_\mu$ ή του $\Sigma^{-1/2}B_\mu\Sigma^{-1/2}$, τέτοιες ώστε $s \le min(g-1,p)$.
Καθώς ο $\Sigma^{-1}B_\mu$ είναι $p \times p$ πίνακας, πρέπει $s \le p$. Ακόμα, έχουμε g πληθυσμούς που θέλουμε να διαχωρίσουμε. Αν σχηματίσουμε τα g διανύσματα διαφορών τότε

$$\mu_1 - \overline{\mu}, \mu_2 - \overline{\mu}, ..., \mu_g - \overline{\mu} \tag{5.34}$$

και παρατηρούμε ότι $(\mu_1 - \overline{\mu}) + (\mu_2 - \overline{\mu}) + ... + (\mu_g - \overline{\mu}) = g\overline{\mu} - g\overline{\mu} = 0$. Δηλαδή, η πρώτη διαφορά $\mu_1 - \overline{\mu}$ μπορεί να γραφτεί ως γραμμικός συνδυασμός των τελευταίων $g-1$ διαφορών. Οι γραμμικοί συνδυασμοί των g διανυσμάτων στην (5.34) καθορίζουν ένα υπερεπίπεδο με διάσταση $q \le g-1$. Αν πάρουμε οποιοδήποτε διάνυσμα e κάθετο σε κάθε $\mu_1 - \overline{\mu}$, τότε το υπερεπίπεδο δίνει:

$$B_\mu e = \sum_{i=1}^{g}(\mu_i - \overline{\mu})(\mu_i - \overline{\mu})^{'} e = \sum_{i=1}^{g}(\mu_i - \overline{\mu})0 = 0$$

επομένως

$$\Sigma^{-1}B_\mu e = 0e$$

Υπάρχουν $p-q$ ορθογώνια ιδιοδιανύσματα που αντιστοιχούν στην μηδενική ιδιοτιμή. Αυτό σημαίνει ότι υπάρχουν q ή λιγότερες μη μηδενικές ιδιοτιμές. Καθώς ισχύει πάντα ότι $q \le g-1$, ο αριθμός των μη μηδενικών ιδιοτιμών s πρέπει να ικανοποιεί την $s \le min(g-1,p)$.
Δηλαδή, δεν υπάρχει απώλεια πληροφορίας, αν σχεδιάσουμε τους διαχωριστές σε δύο διαστάσεις,

εφ'όσον ισχύουν οι ακόλουθες συνθήκες:

Αριθμός μεταβλητών	Αριθμός πληθυσμών	Μέγιστος αριθμός διαχωριστικών
p	*g=2*	1
p	*g=3*	2
p=2	*G*	2

Σ'αυτό το σημείο παρουσιάζουμε μια σημαντική σχέση μεταξύ του κανόνα ταξινόμησης και του γραμμικού διαχωριστικού βαθμού,

$$d_i(x) = \mu_i^{'}\Sigma^{-1}x - \frac{1}{2}\mu_i^{'}\Sigma^{-1}\mu_i + \ln p_i$$

Ισοδύναμα,

$$d_i(x) - \frac{1}{2}x^{'}\Sigma^{-1}x = -\frac{1}{2}(x-\mu_i)^{'}\Sigma^{-1}(x-\mu_i) + \ln p_i$$

Έστω, $y_j = a_j^{'}x$ όπου $a_j = \Sigma^{-1/2}e_j$ και e_j είναι ιδιοδιάνυσμα του $\Sigma^{-1/2}B_\mu\Sigma^{-1/2}$. Τότε,

$$\sum_{j=1}^{p}(y_j - \mu_{iYj})^2 = \sum_{j=1}^{p}[a_j^{'}(x-\mu_i)]^2 = (x-\mu_i)^{'}\Sigma^{-1}(x-\mu_i) = -2d_i(x) + x^{'}\Sigma^{-1}x + \ln p_i$$

Εάν $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_s > 0 = \lambda_{s+1} = \ldots = \lambda_p$, η ποσότητα $\sum_{j=1}^{p}(y_j - \mu_{iYj})^2$ είναι σταθερή για όλους τους πληθυσμούς $i=1,2,...,g$, άρα μόνο οι πρώτοι s διαχωριστές y_j, ή το $\sum_{j=1}^{s}(y_j - \mu_{iYj})^2$ συνεισφέρει στην ταξινόμηση.

Επομένως, ο κανόνας ταξινόμησης για τους πρώτους $r \leq s$ διαχωριστές είναι ο εξής:
Ταξινομείται η παρατήρηση x στον πληθυσμό π_k εάν

$$\sum_{j=1}^{r}(\hat{y}_j - \bar{y}_{kj})^2 = \sum_{j=1}^{r}[a_j^{'}(x-\bar{x}_k)]^2 \leq \sum_{j=1}^{r}[a_j^{'}(x-\bar{x}_i)]^2 \quad \text{για όλα τα } i \neq k \qquad (5.35)$$

Όταν για τις εκ των προτέρων πιθανότητες ισχύει $p_1=p_2=\ldots=p_g=1/g$ και $r=s$, ο κανόνας (5.35) είναι ισοδύναμος με τον κανόνα 11-52, ο οποίος βασίζεται στο μεγαλύτερο γραμμικό διαχωριστικό βαθμό. Επιπλέον, εάν χρησιμοποιούνται $r<s$ διαχωριστές για την ταξινόμηση, υπάρχει απώλεια της απόστασης του τετραγώνου του $\sum_{j=r+1}^{p}[\hat{a}_j^{'}(x-\bar{x}_i)]^2$ για κάθε πληθυσμό π_i, όπου το κομμάτι $\sum_{j=r+1}^{s}[\hat{a}_j^{'}(x-\bar{x}_i)]^2$ είναι χρήσιμο για την ταξινόμηση.

Ουσιαστικά, ο κανόνας του Fisher στηρίζεται στις ευκλείδειες αποστάσεις. Συγκρίνει τις τετραγωνικές αποστάσεις των παρατηρήσεων που θέλουμε να κατατάξουμε από το μέσο όρο της κάθε ομάδας και κατατάσσει την παρατήρηση σε εκείνο τον πληθυσμό που η απόσταση είναι η

μικρότερη.

Σε αυτό το σημείο θα αναφέρουμε γιατί οι πρώτοι διαχωριστές είναι περισσότερο σημαντκοί από τους τελευταίους χρησιμοποιώντας ένα άλλο διαχωριστικό μέτρο το

$$\Delta_s^2 = \sum_{i=1}^{g} (\mu_i - \overline{\mu})' \Sigma^{-1} (\mu_i - \overline{\mu}) \qquad (5.36)$$

όπου

$$\overline{\mu} = \frac{1}{g} \sum_{i=1}^{g} \mu_i$$

και $(\mu_i - \overline{\mu})' \Sigma^{-1} (\mu_i - \overline{\mu})$ είναι η τετράγωνη στατιστική απόσταση του πληθυσμιακού μέσου μ_i από τον κεντροειδή $\overline{\mu}$. Θα δειχθεί ότι $\Delta_s^2 = \lambda_1 + \lambda_2 + ... + \lambda_p$ όπου τα $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_s$ είναι μη μηδενικές ιδιοτιμές του Σ^{-1}B και τα $\lambda_{s+1}, ..., \lambda_p$ είναι μηδενικές ιδιοτιμές.

Ο πρώτος διαχωριστής $Y_1 = e_1' \Sigma^{-1/2} X$ έχει μέσο $\mu_{iY1} = e_1' \Sigma^{-1/2} \mu_i$ και η απόσταση τετραγώνων $\sum_{i=1}^{g} (\mu_{iY1} - \overline{\mu}_{Y1})^2$ του μ_{iY1} από την κεντρική τιμή $\overline{\mu}_{iY1} = e_1' \Sigma^{-1/2} \mu_i$ είναι λ_1. Καθώς το Δ_s^2 μπορεί επίσης να γραφτεί και ως

$$\begin{aligned}
\Delta_s^2 &= \lambda_1 + \lambda_2 + ... + \lambda_p \\
&= \sum_{\iota=1}^{g} (\mu_{\iota Y} - \overline{\mu}_{Y})' (\mu_{\iota Y} - \overline{\mu}_{Y}) \\
&= \sum_{i'=1}^{g} (\mu_{\iota Y1} - \overline{\mu}_{Y1})^2 + \sum_{i'=1}^{g} (\mu_{\iota Y2} - \overline{\mu}_{Y2})^2 + ... + \sum_{i'=1}^{g} (\mu_{\iota Yp} - \overline{\mu}_{Yp})^2
\end{aligned}$$

συνεπάγεται ότι ο πρώτος διαχωριστής έχει την μεγαλύτερη ξεχωριστή συνεισφορά λ_1 στο διαχωριστικό μέτρο Δ_s^2. Γενικά, ο r διαχωριστής συνεισφέρει λ_r στο Δ_s^2. Εάν οι επόμενες s-r ιδιοτιμές είναι τέτοιες ώστε το $\lambda_{r+1} + \lambda_{r+2} + ... + \lambda_s$ να είναι μικρό σε σύγκριση με το $\lambda_1 + \lambda_2 + ... + \lambda_r$ τότε οι τελευταίοι διαχωριστές $Y_{r+1}, Y_{r+2}, ..., Y_s$ μπορούν να θεωρηθούν αμελητέοι χωρίς να μειωθεί η σημαντικότητα του διαχωρισμού.

ΚΕΦΑΛΑΙΟ 6^{Ο}

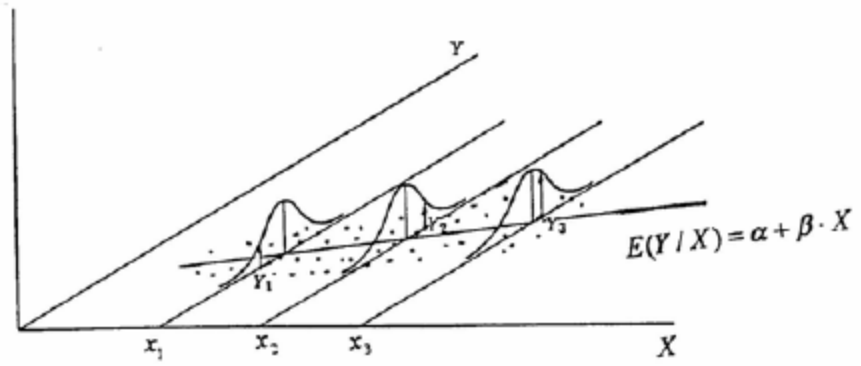
Άλλες προσεγγίσεις για τον διαχωρισμό ομάδων

6.1 Ανάλυση Παλινδρόμησης

Με την ανάλυση παλινδρόμησης (regression analysis) εξετάζουμε τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Στο απλό γραμμικό μοντέλο, η μία μεταβλητή είναι η συνεχής ανεξάρτητη μεταβλητή ενώ η άλλη συνεχής εξαρτημένη μεταβλητή απόκρισης.
Οι παρατηρήσεις y_i δίνονται από την σχέση:

$$y_i = E(y_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

όπου ε_i είναι το τυχαίο σφάλμα. Υπόλοιπο (residual) ονομάζεται η διαφορά $y_i - \hat{y}_i = e_i$ όπου $i=1,\ldots,n$ και είναι η κατακόρυφη απόκλιση της πραγματικής τιμής y_i από την ευθεία της εκτιμώμενης εξίσωσης παλινδρόμησης.



Με τη μέθοδο των ελαχίστων τετραγώνων προσδιορίζεται μια εκτίμηση $\hat{y} = \hat{b}_0 + \hat{b}_1 x$ της ευθείας $y = b_0 + b_1 x$. Η σταθερά b_0 εκφράζει την μέση τιμή της Y όταν X=0. Η σταθερά b_1 εκφράζει το πόσο αναμένεται να μεταβληθεί η τιμή της Y, αν η X αυξηθεί κατά μία μονάδα.
Η μέθοδος ελαχίστων τετραγώνων βασίζεται σε κάποιες υποθέσεις σχετικά με τα τυχαία σφάλματα ε_i, και αν κάποια από αυτές παραβιάζεται τότε το μοντέλο που προσαρμόζεται με τη μέθοδο αυτή δεν είναι κατάλληλο για να εξηγήσει τη συμπεριφορά των παρατηρήσεων. Οι υποθέσεις είναι οι εξής:

- Η μέση τιμή του ε είναι μηδέν, $E(\varepsilon_i)=0$, για κάθε i
- Η διασπορά του ε είναι άγνωστη, $Var(\varepsilon_i)=\sigma^2$ για κάθε i
- Τα ε_i είναι ασυσχέτιστα, $cov(\varepsilon_i,\varepsilon_j)$ για $i \neq j$.

Το άθροισμα

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

λέγεται ολικό άθροισμα τετραγώνων (total sum of squares) ή ολική μεταβλητότητα (total variation) των y_i το οποίο αναλύεται σε δύο συνιστώσες: στο άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

και στο άθροισμα τετραγώνων των σφαλμάτων (error sum of squares) ή υπόλοιπο μεταβλητότητας (residual variation)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

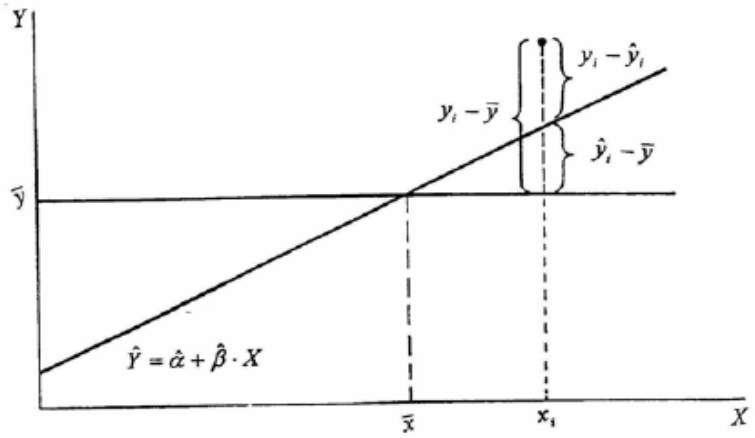
Δηλαδή,

$$SST = SSR + SSE$$

Το SST μετράει την συνολική μεταβλητότητα των παρατηρήσεων y_i, δηλαδή εκφράζει την αβεβαιότητα στην πρόβλεψη του Y όταν δεν χρησιμοποιείται το X. Το SSR εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στο X και το SSE εκφράζει την υπόλοιπη μεταβλητότητα που δεν εξηγείται από την παλινδρόμηση. Ο λόγος

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i, που εξηγείται από την παλινδρόμηση. Το r^2 λέγεται συντελεστής προσδιορισμού (coefficient of determination) και παίρνει τιμές στο κλειστό διάστημα [0,1]. Όσο πλησιέστερα βρίσκεται η τιμή του r^2 στο 1, τόσο καλύτερη είναι η ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης.



Εικόνα 6.1.1. Ευθεία ελαχίστων τετραγώνων $\hat{Y} = \hat{\alpha} + \hat{\beta} X$ και η απόσταση των υπολοίπων $y_i - \hat{y}_i$ από αυτή.

Οι προυποθέσεις του απλού γραμμμικού μοντέλου είναι:

- Γραμμικότητα (Linearity)
- Ομοσκεδαστικότητα (Homoscedasticity)
- Ανεξαρτησία σφαλμάτων (Error independence)
- Κανονικότητα σφαλμάτων (Error normality)

6.2 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυμεταβλητής στατιστικής ανάλυσης (multivariate statistical analysis) που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών (independent variables) για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής (dependent variable).
Η λογιστική παλινδρόμηση (Logistic Regression) είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της 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 (set) ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης (predictor variables). Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής.
Στη λογιστική παλινδρόμηση, σε αντίθεση με την πολλαπλή παλινδρόμηση (multiple regression) είναι δυνατό να χρησιμοποιηθούν ως εξαρτημένες μεταβλητές εκτός από αριθμητικές μεταβλητές (ratio scale) και κατηγορικές μεταβλητές (nominal scale).
Η πιο διαδεδομένη, έκφραση της εξίσωσης της Λογιστικής Παλινδρόμησης είναι:

$$\ln(odds) = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

Το δεξί μέρος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο παλινδρόμησης. Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με τη μορφή του λογαρίθμου των odds, δηλαδή του λογαρίθμου της σχέσης odds=Prob/(1-Prob). Το odds εναλλακτικά ονομάζεται logit και ο όρος Prob εκφράζει την πιθανότητα του συμβάντος του γεγονότος. Οι συντελεστές των ανεξάρτητων μεταβλητών στην εξίσωση παλινδρόμησης εκτιμούνται με βάση τη μέθοδου Μεγίστης Πιθανοφάνειας. Σύμφωνα με τη μέθοδο αυτή η τιμή των συντελεστών των ανεξάρτητων μεταβλητών είναι αυτή που κάνει τις παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάσει του συνόλου (set) των ανεξαρτήτων μεταβλητών.

Η λογιστική παλινδρόμηση, απαιτεί μεγάλο δείγμα για την σωστή εφαρμογή της προκειμένου να παράγει αξιόπιστο αποτέλεσμα. Σύμγωνα με έναν εμπειρικό κανόνα το δείγμα θα πρέπει να είναι 30 φορές μεγαλύτερο από το αριθμό των παραμέτρων που εκτιμά το μοντέλο.

Τα βήματα κατασκευής του μοντέλου της Λογιστικής Παλινδρόμησης είναι ανάλογα αυτών της γραμμικής παλινδρόμησης.

- Προσδιορισμός της εξαρτημένης μεταβλητής και του σετ των ανεξάρτητων μεταβλητών που θα συμμετέχουν στην παλινδρόμηση.
- Διερεύνηση των δεδομένων για τυχόν ύπαρξη ασυνήθιστων κινήσεων όπως, ακραίες τιμές, ελλείπουσες τιμές κ. λ. π.
- Ελέγχος για την ικανοποίηση των υποθέσεων για την σωστή εφαρμογή της Λογιστικής Παλινδρόμησης.
- Σχηματισμός της εξίσωσης παλινδρόμησης.
- Μελέτη της επίδρασης κάθε ανεξάρτητης μεταβλητής στο μοντέλο.

6.3 Δέντρα αποφάσεων

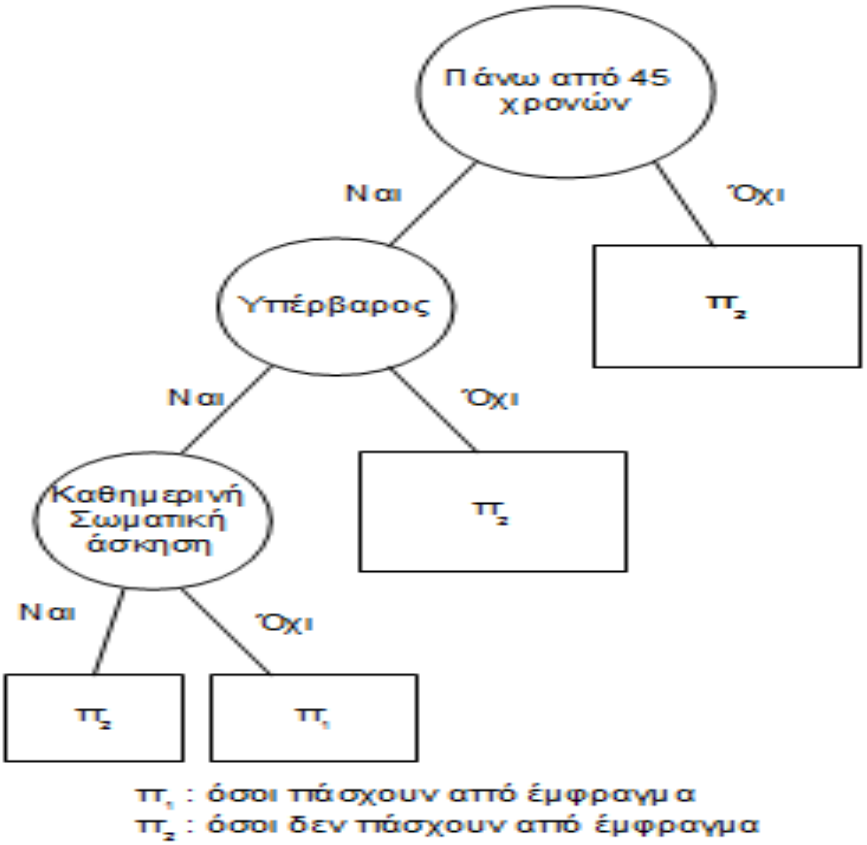
Η μέθοδος ονομάζεται δέντρα ταξινόμησης και παλινδρόμησης (classification and regression trees, CART) και μοιάζει περισσότερο με την μέθοδο κατά συστάδες παρά με την διαχωριστική ανάλυση. Αρχικά, όλα τα αντικείμενα βρίσκονται σε μία ομάδα. Αυτή η ομάδα χωρίζεται σε δύο υπο-ομάδες χρησιμοποιώντας, για παράδειγμα, υψηλές τιμές μιας μεταβλητής για μία ομάδα και χαμηλές τιμές για την άλλη. Έπειτα, οι δύο υπο-ομάδες χωρίζονται χρησιμοποιώντας τις τιμές μιας δεύτερης μεταβλητής. Αυτή η διαδικασία σταματάει όταν ένας κανόνας παύσης ικανοποιηθεί.
Η μέθοδος ταξινόμησης απαιτεί εκατοντάδες αντικείμενα και συχνά πολλές μεταβλητές. Το δέντρο που προκύπτει είναι πολύ πολύπλοκο. Πρέπει να χρησιμοποιηθούν αντικειμενικά κριτήρια για να δημιουργηθεί το δέντρο το οποίο να έχει ομάδες με πολλά αντικείμενα και έπειτα κάθε ομάδα ταξινομείται στον πληθυσμό με την μεγαλύτερη πιθανότητα. Ένα νέο αντικείμενο έπειτα ταξινομείται σύμφωνα με την τελική ομάδα.

Για παράδειγμα, έστω τα αντικείμενα μπορούν να ταξινομηθούν στις ομάδες

π_1: όσοι πάσχουν από έμφραγμα

π_2: όσοι δεν πάσχουν από έμφραγμα

σύμφωνα με την ηλικία, το βάρος, και τον ρυθμό εξάσκησης. Σ'αυτή την περίπτωση, η διαδικασία CART μπορεί να αναπαρισταθεί όπως φαίνεται στην Εικόνα 6.3.1. Η περιοχή R1, που ορίζεται από τα άτομα είναι πάνω από 45, υπέρβαρα, και δεν εξασκούνται καθόλου, χρησιμοποιείται για να ταξινομήσουν ένα αντικείμενο στον πληθυσμό π_1.



Εικόνα 6.3.1: Δέντρο ταξινόμησης

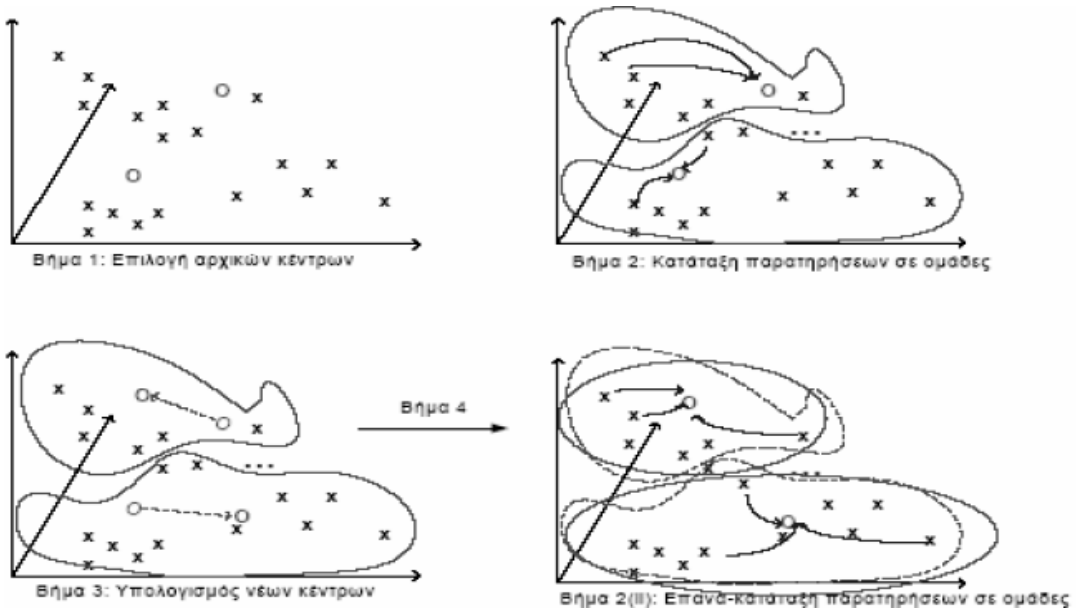
Συνήθως στα δένδρα ταξινόμησης χρειάζεται να κατηγοριοποιήσουμε τις συνεχείς μεταβλητές, ώστε να 'λειτουργούν' πιο αποδοτικά οι αλγόριθμοι και αυτό μπορεί να οδηγήσει σε απώλεια κάποιας πληροφορίας.

6.4 Ανάλυση κατά συστάδες (Cluster Analysis)

Η ανάλυση κατά συστάδες έχει σκοπό να διαχωρίσει το σύνολο των παρατηρήσεων σε φυσικές ομάδες, έτσι ώστε τα μέλη κάθε ομάδας να είναι όμοια μεταξύ τους στο μεγαλύτερο δυνατό βαθμό. Γεωμετρικά αυτό σημαίνει ότι δύο όμοιες παρατηρήσεις θα βρίσκονται σε γειτονικά σημεία, ενώ δύο ανόμοιες σε απομακρυσμένα σημεία.
Όπως αναφέρθηκε και στην εισαγωγή η βασική διαφορά μεταξύ της διαχωριστικής ανάλυσης και της ανάλυσης κατά συστάδες είναι ότι στη διαχωριστική ανάλυση οι ομάδες (συστάδες) καθορίζονται εκ των προτέρων και σκοπός είναι να καθοριστεί ο γραμμικός συνδυασμός ανεξάρτητων μεταβλητών που κάνει διακρίσεις καλύτερα μεταξύ των ομάδων. Ενώ στην ανάλυση συστάδων οι ομάδες (συστάδες) δεν προκαθορίζονται και στην πραγματικότητα, σκοπός της μεθόδου είναι να καθοριστεί ο καλύτερος τρόπος με τον οποίο οι παρατηρήσεις (περιπτώσεις) μπορούν να χωριστούν και να συγκεντρωθούν σε ομάδες.
Η πιο συνηθισμένη μέθοδος σχηματισμού των ομάδων είναι η ιεραρχική μέθοδος, όπου ξεκινάμε με κάθε παρατήρηση να είναι μόνη της μια ομάδα. Σε κάθε βήμα ενώνουμε τις δύο παρατηρήσεις που έχουν πιο μικρή απόσταση. Αν δύο παρατηρήσεις έχουν ενωθεί σε προηγούμενο βήμα ενώνουμε μια προυπάρχουσα ομάδα με μία παρατήρηση. Αυτή η διαδικασία συνεχίζεται μέχρι να καταταχθούν όλες οι παρατηρήσεις σε ομάδες.
Μια άλλη ευρέως γνωστή μέθοδος σχηματισμού των ομάδων είναι η k-means, όπου ο αριθμός των ομάδων είναι γνωστός εκ των προτέρων. Με έναν επαναληπτικό αλγόριθμο μοιράζουμε τις παρατηρήσεις στις ομάδες ανάλογα την απόσταση κάθε παρατήρησης από τις ομάδες. Κάθε παρατηρήσε κατατάσσεται στην ομάδα από την οποία έχει την μικρότερη απόσταση.
Γενικά, οι ιεραρχικές μέθοδοι δεν είναι καλή ιδέα να χρησιμοποιούνται για μεγάλο πλήθος δεδομένων καθώς απαιτούν πολύ χρόνο και υπολογιστική ισχύ. Επίσης, μπορεί να δημιουργηθούν ομάδες με ανομοιογενές μέγεθος. Από την άλλη, η μέθοδος k-means ενώ αποφεύγει αυτά τα προβλήματα, δουλεύει αρκετά καλά με μεγάλα δείγματα και δημιουργεί ομάδες παραπλήσιου μεγέθους, εξαρτάται πολύ από τις αρχικές τιμές που θα χρησιμοποιήσουμε.

Ο αλγόριθμος του k-means έχει ως εξής:

- Βήμα 1: Βρίσκουμε τα αρχικά σημεία.
- Βήμα 2: Κατατάσσουμε κάθε παρατήρηση στην ομάδα της οποίας το κέντρο έχει τη μικρότερη απόσταση από την παρατήρηση.
- Βήμα 3: Υπολογίζουμε τα νέα κέντρα από τις παρατηρήσεις που είναι μέσα στην ομάδα.
- Βήμα 4: Αν τα κέντρα δεν διαφέρουν από τα προηγούμενα σταματάμε, διαφορετικά πηγαίνουμε στο βήμα 2.



Εικόνα 6.4.1: Αλγόριθμος k-means

Στην αρχική ομαδοποιήση, ο αριθμός των ομάδων δεν είναι γνωστός από πριν.
Ο αλγόριθμος έχει ως εξής:

- Βήμα 1: Δημιουργούμε έναν πίνακα αποστάσεων για όλες τις παρατηρήσεις.

- Βήμα 2: Ενώνουμε τις παρατηρήσεις που έχουν την μικρότερη απόσταση. Δηλαδή, δημιουργούμε μια ομάδα με τις παρατηρήσεις που είναι πιο κοντά. Αν η μικρότερη απόσταση αφορά μια ήδη δημιουργηθήσα ομάδα και μια παρατήρηση απλά βάζουμε αυτή την παρατήρηση στην ομάδα ή αν αφορά δύο ομάδες που ήδη υπάρχουν τις ενώνουμε.

- Βήμα 3: Αν δεν έχουν μπει όλες οι ομάδες σε μια ομάδα πηγαίνουμε στο βήμα 1 αλλιώς σταματάμε.

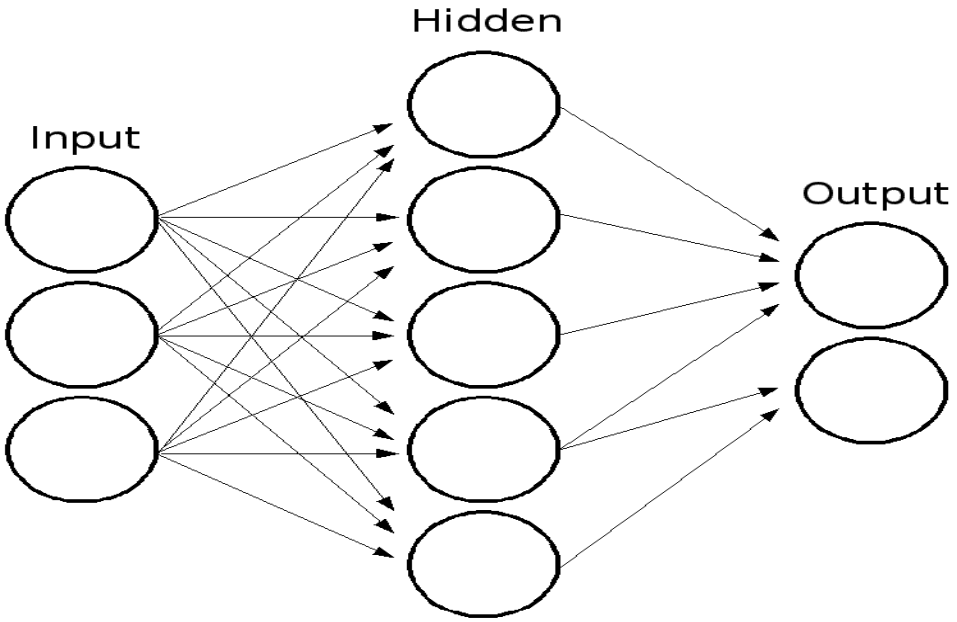
Για να υπολογιστεί η απόσταση της ομάδας που δημιουργήθηκε υπάρχουν πολλοί μέθοδοι, όπως:

- Η μέθοδος του κοντινότερου γείτονα (nearest neighbor)
- Η μέθοδος του μακρυνότερου γείτονα (furthest neighbor)
- Η μέθοδος του μέσου μεταξύ των ομάδων (average between groups)
- Η μέθοδος του μέσου ανάμεσα στις ομάδες (average within groups)
- Η μέθοδος του Ward και άλλες

Από τις μεθόδους αυτές η πιο απλή είναι αυτή του ‘κοντινότερου γείτονα’ η οποία όμως έχει το μειονέκτημα ότι δίνει ομάδες με μεγάλες διαφορές ως προς το μέγεθος τους. Η μέθοδος του Ward έχει το πλεονέκτημα ότι μας δίνει περίπου ισοπληθικές ομάδες.

6.5 Νευρωνικά Δίκτυα

Ένα νευρωνικό δίκτυο (Neural Network-NN), αποτελεί μια υπολογιστική προσέγγιση, αλγοριθμική διαδικασία η οποία μετατρέπει εισερχόμενη πληροφορία σε εξερχόμενη πληροφορία χρησιμοποιώντας δίκτυα που αποτελούνται από μικρές μονάδες, νευρώνες ή κόμβους (neurons or nodes).
Τα τρία απαραίτητα χαρακτηριστικά ενός νευρωνικού δικτύου είναι οι βασικές μονάδες υπολογισμού (νευρώνες ή κόμβοι), η αρχιτεκτονική του δικτύου που περιγράφει τις συνδέσεις μεταξύ των μονάδων υπολογισμού, και ο δοκιμαστικός αλγόριθμος που χρησιμοποιείται για τις τιμές των παραμέτρων που απαιτούνται για να εκτελεστεί κάποια εργασία.
Οι υπολογιστικές μονάδες συνδέονται η μια με την άλλη έτσι ώστε η εξερχόμενη πληροφορία μιας μονάδας να μπορεί να χρησιμοποιηθεί ως εισερχόμενη πληροφορία για κάποια άλλη μονάδα. Κάθε υπολογιστική μονάδα μετατρέπει την εισερχόμενη πληροφορία σε εξερχόμενη πληροφορία χρησιμοποιώντας κάποια ειδική μονότονη συνάρτηση. Αυτή η συνάρτηση εξαρτάται από τις σταθερές (παραμέτρους) των οποίων οι μεταβλητές μπορούν να καθοριστούν από ένα δοκιμαστικό σύνολο εισερχόμενης και εξερχόμενης πληροφορίας.
Η αρχιτεκτονική του δικτύου είναι ένας οργανισμός υπολογιστικών μονάδων. Στις στατιστικές μεθόδους, οι υπολογιστικές μονάδες είναι διατεταγμένες σε διαφορετικά επίπεδα, και οι συνδέσεις γίνονται μεταξύ κόμβων από διαφορετικά επίπεδα και όχι μεταξύ κόμβων από το ίδιο επίπεδο. Το επίπεδο που λαμβάνει τα αρχικά δεδομένα ονομάζεται επίπεδο εισερχόμενης πληροφορίας, ενώ το τελευταίο επίπεδο ονομάζεται επίπεδο εξερχόμενης πληροφορίας. Τα επίπεδα που βρίσκονται μεταξύ του αρχικού και του τελευταίου ονομάζονται "κρυφά" (hidden) επίπεδα. Μια απλή σχηματική αναπαράσταση ενός νευρωνικού δικτύου φαίνεται στην Εικόνα 6.5.1.
Τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν τόσο για τον διαχωρισμό όσο και για την ταξινόμηση. Οι μεταβλητές εισερχόμενης πληροφορίας είναι τα μετρήσιμα χαρακτηριστικά των ομάδων και οι μεταβλητές εξερχόμενης πληροφορίας είναι κατηγορικές μεταβλητές γνωστοποιώντας τις ιδιότητες των μελών των ομάδων. Τα νευρωνικά δίκτυα μπορούν να έχουν καλές αποδόσεις στον διαχωρισμό και στην ταξινόμηση, αντίστοιχες με τις αποδόσεις που έχει η λογιστική παλινδρόμηση και η διαχωριστική ανάλυση.



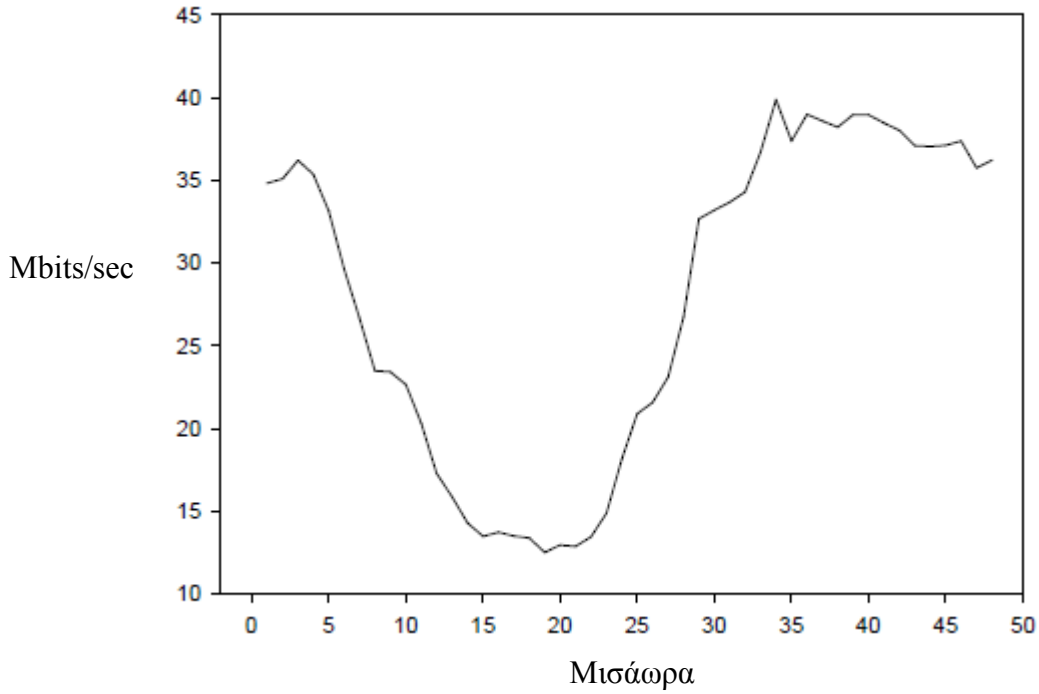
Εικόνα 6.5.1: Νευρωνικό δίκτυο με ένα κρυφό επίπεδο

6.6 Σύγκριση απόδοσης της Διαχωριστικής ανάλυσης και των Νευρωνικών δικτύων για την ταξινόμηση

Η σύγκριση απόδοσης της διαχωριστικής ανάλυσης και των νευρωνικών δικτύων για την ταξινόμηση, θα γίνει μέσω ενός παραδείγματος όπου ταξινομούνται χρήστες του διαδικτύου σε ομάδες σύμφωνα με τον μέσο ρυθμό λήψης δεδομένων σε περιόδους μισής ώρας. Για να παραχθεί η εκπαίδευση του νευρωνικού δικτύου, υπολογίζουμε τις γραμμικές διαχωριστικές συναρτήσεις της διαχωριστικής ανάλυσης, και για την εκτίμηση της απόδοσης και της διαχωριστικής ανάλυσης και του νευρωνικού δικτύου ως προς την ταξινόμηση των χρηστών του διαδικτύου, προαπαιτείται ταξινόμηση των χρηστών σύμφωνα με την μέθοδο κατά συστάδες (Cluster analysis). Και τελικά, για να αξιολογηθούν οι μέθοδοι ταξινόμησης, συγκρίνονται τα αποτελέσματα με τα αποτελέσματα της ταξινόμησης σύμφωνα με την ανάλυση κατά συστάδες που εφαρμόστηκε σε ολόκληρο το σύνολο δεδομένων,

Τα δεδομένα που χρησιμοποιήθηκαν μετρήθηκαν σε ένα Πορτογαλικό ISP που χρησιμοποιεί ένα δίκτυο CATV και παρέχει πολλούς τύπους υπηρεσιών, που χαρακτηρίζονται από το μέγιστο επιτρεπτό ρυθμό μεταφοράς/λήψης δεδομένων (μετρημένο σε Kbit/s): 128/64, 256/128, 512/256. Οι μετρήσεις των δεδομένων έγιναν το Σάββατο, 9 Νοεμβρίου, 2002. Τα δεδομένα είναι λεπτομερή πακέτα μετρήσεων, όπου καταγράφονται η στιγμή της άφιξης και τα πρώτα 57 byte κάθε πακέτου. Περιλαμβάνει πληροφορίες για το μέγεθος του πακέτου, τις διευθύνσεις IP αναχώρησης και άφιξης και τον τύπο του πρωτοκόλου IP.

Οι χρήστες αναγνωρίζονται συνδυάζοντας τις IP διευθύνσεις με τις πληροφορίες του λογαριασμού. Το σύνολο δεδομένων περιλαμβάνει 3432 χρήστες. Οι χρήστες ταξινομούνται σύμφωνα με τον ρυθμό λήψης δεδομένων σε διαστήματα μισής ώρας. Ο ρυθμός μεταφοράς δεδομένων ενός χρήστη (σε Kbits/s) στο *k*-οστο μισάωρο διάστημα συμβολίζεται ως $X_k, k = 1,2,...,48$. Η Εικόνα 6.6.1 δείχνει τους ρυθμούς μεταφοράς της συνολικής λαμβανόμενης κίνησης συναρτήσει του χρόνου.



Εικόνα 6.6.1: Ρυθμός μεταφοράς της συνολικής λαμβανόμενης κίνησης

Ο συνολικός ρυθμός μεταφοράς δείχνει ότι η χαμηλότερη χρήση γίνεται κατά τις πρωινές ώρες ενώ η υψηλότερη κατά τις απογευματινές ώρες. Ωστόσο, το γράφημα δεν δίνει ομάδες των χρηστών με συγκεκριμένα χρονικά προφίλ.

Χρήση της μεθόδου ανάλυση κατά συστάδες

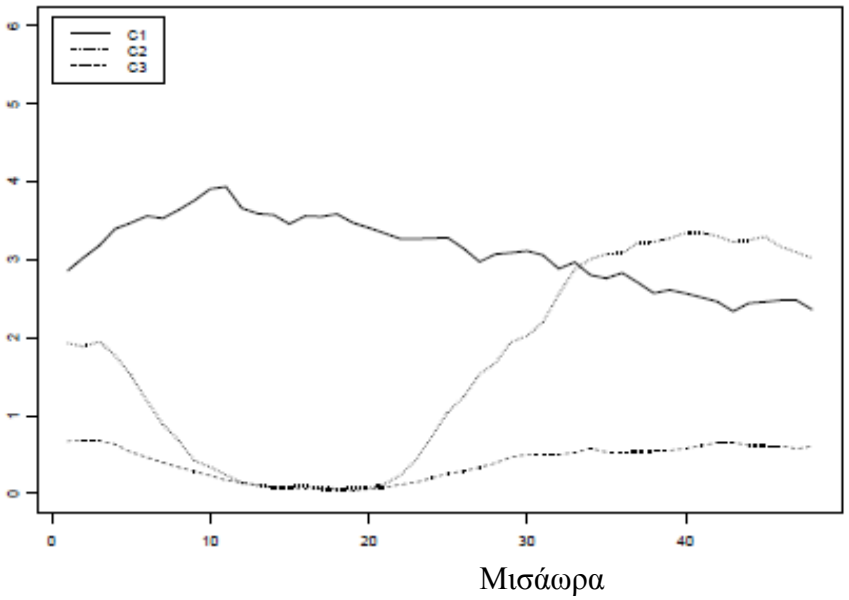
Οι τεχνικές ταξινόμησης, διαχωριστική ανάλυση και νευρωνικά δίκτυα, βασίζονται σε ήδη γνωστές ομάδες, οι οποίες καθορίζονται χρησιμοποιώντας την ανάλυση κατά συστάδες.
Στην συγκεκριμένη περίπτωση, η ανάλυση κατά συστάδες βασίζεται στην μέθοδο medoid. Σ'αυτή τη μέθοδο, πρέπει να οριστεί εκ των προτέρων ο αριθμός των επιπέδων που θα χρησιμοποιηθούν. Έστω ότι θα χρησιμοποιηθούν K επίπεδα και τα medoid συμβολίζονται ως $m_1, m_2, ..., m_K$. Αυτά είναι αντιπροσωπευτικά αντικείμενα τα οποία επιλέγονται έτσι ώστε η συνολική Ευκλείδια απόσταση όλων των αντικειμένων από το κοντινότερο medoid να είναι ελάχιστη. Δηλαδή, βρίσκεται ένα υποσύνολο $\{m_1, m_2, ..., m_k\} \subset \{1, ..., n\}$ (όπου n είναι ο αριθμός των αντικειμένων που θα ταξινομηθούν) το οποίο ελαχιστοποιεί την συνάρτηση

$$\sum_{i=1}^{n} \min_{t=1,...,K} d_{im_t} .$$

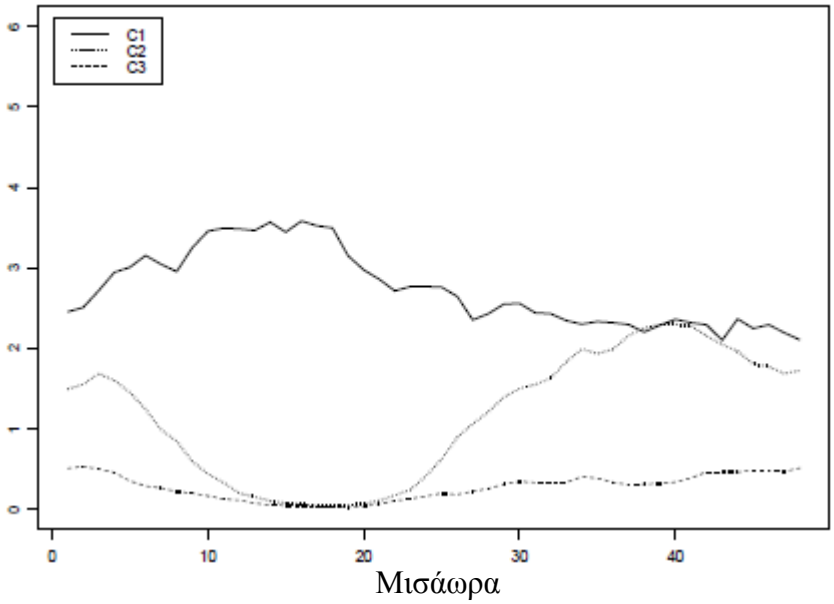
Έπειτα, κάθε αντικείμενο ταξινομείται στην συστάδα με το κοντινότερο medoid. Δηλαδή, το αντικείμενο *i* ταξινομείται στην συστάδα C_j, με το medoid, m_j να είναι το κοντινότερο στο αντικείμενο *i*, δηλαδή $d_{im_j} \leq d_{im_t}$, για όλα τα $t \in \{1,2,...K\}$. Στην συγκεκριμένη μελέτη, οι χρήστες είναι τα αντικείμενα και οι μεταβλητές είναι τα μισάωρα διαστήματα του ρυθμού μεταφοράς δεδομένων. Τα δεδομένα που περιγράφηκαν παραπάνω χρησιμοποιουνταί για να σχηματιστούν οι συστάδες σε δύο διαφορετικές καταστάσεις: δεδομένου όλου του αριθμού των χρηστών και δεδομένου του μισού αριθμού των χρηστών, οι οποίοι επιλέγονται τυχαία. Τα αποτελέσματα της πρώτης ανάλυσης χρησιμοποιούνται για να αξιολογήσουν τις μεθόδους ταξινόμησης. Η δεύτερη περίπτωση εξομοιώνει ένα σύστημα όπου μόνο τα μισά δεδομένα, τα οποία ονομάζονται σύνολο δεδομένων εκπαίδευσης (training set), είναι αρχικά διαθέσιμο για να αποκτηθούν οι συστάδες. Έπειτα, το σύστημα ταξινομεί και τους υπόλοιπους μισούς χρήστες (το υπόλοιπο μισό σύνολο δεδομένων από το αρχικό σύνολο) σε μια από τις προηγούμενες συστάδες.
Το σύνολο δεδομένων μετασχηματίζεται σύμφωνα με:

$$Y_j = \ln(1 + X_j)$$

για *j*=1,…,48. Αυτός ο μετασχηματισμός βοηθάει στη εξάλειψη των ανωμαλιών στην μεταβλητότητα της χρήσης του Internet στα μισάωρα διαστήματα, οι οποίες φαίνεται σε αυξάνονται με καθημερινή μέση χρήση του διαδικτύου.
Οι δύο διαχωρισμοί που βασίζονται σε ολόκληρο το σύνολο δεδομένων και στο σύνολο δεδομένων εκπαίδευσης έχουν παρόμοια ερμηνεία. Η Εικόνα 6.6.2 και η Εικόνα 6.6.3 αναπαριστούν τον μέσο όρο του μισάωρου ρυθμού μεταφοράς για κάθε συστάδα, για τον διαχωρισμό που αποκτάται από όλο το σύνολο δεδομένων και απο το σύνολο δεδομένων εκπαίδευσης, αντίστοιχα. Έτσι, η πρώτη συστάδα, C_1, περιέχει χρήστες με υψηλό ρυθμό μεταφοράς σε όλες τις χρονικές περιόδους της ημέρας, οι χρήστες στη C_2 έχουν χαμηλό ρυθμό μεταφοράς κατά τις πρωινές ώρες και υψηλό ρυθμό μεταφοράς κατά τις απογευματινές ώρες και η C_3 περιέχει χρήστες με χαμηλό ρυθμό μεταφορές όλη τη διάρκεια της ημέρας.



Εικόνα 6.6.2: Δειγματικός μέσος για κάθε συστάδα, για όλο το σύνολο δεδομένων



Εικόνα 6.6.3: Δειγματικός μέσος για κάθε συστάδα, για το σύνολο δεδομένων εκπαίδευσης

Σύμφωνα με τον Πίνακα 6.6.1, η συστάδα C_3 έχει το υψηλότερο ποσοστό χρηστών ενώ η C_1 έχει το χαμηλότερο. Οι βασικές διαφορές στους δύο διαχωρισμούς είναι στις κλάσεις C_2 και C_3. Η ερμηνεία των συστάδων και των δύο διαχωρισμών συνοψίζονται στον Πίνακα 6.6.2. Ο Πίνακας 6.6.3 είναι ένας συνεκτικός πίνακας για τους δύο διαχωρισμούς. Όλοι οι 1216 χρήστες της C_3 στον διαχωρισμό που βασίζεται στο σύνολο δεδομένων εκπαίδευσης παραμένουν στην ίδια συστάδα με τον διαχωρισμό που βασίζεται σε όλο το σύνολο δεδομένων. Ωστόσο, περίπου το 18% (≈ 272/1504 x 100%) των χρηστών της C_3 στον διαχωρισμό που βασίζεται σε όλο το σύνολο δεδομένων και που ανήκουν στο σύνολο δεδόμενων εκπαίδευσης ταξινομήθηκαν στην C_2 στον διαχωρισμό που βασίζεται στο σύνολο δεδομένων εκπαίδευσης.

	Σύνολο δεδομένων		Σύνολο δεδομένων εκπαίδευσης	
	Μέγεθος	**Ποσοστό**	**Μέγεθος**	**Ποσοστό**
C1	145	4.23%	87	5.07%
C2	266	7.75%	413	24.07%
C3	3021	88.02%	1216	70.86%

Πίνακας 6.6.1: Μεγέθη συστάδων

Συστάδα	**Ερμηνεία**
C1	Αριθμός χρηστών με υψηλό ρυθμό μεταφοράς σε όλες τις χρονικές περιόδους
C2	Αριθμός χρηστών με χαμηλό/υψηλό ρυθμό μεταφοράς κατά τις πρωινές/απογευματινές ώρες
C3	Αριθμός χρηστών με χαμηλό ρυθμό μεταφοράς σε όλες τις χρονικές περιόδους

Πίνακας 6.6.2: Ερμηνεία των συστάδων

Συνολικό σύνολο δεδομένων	Σύνολο δεδομένων εκπαίδευσης			
	C1	C2	C3	
C1	68	4	0	73
C2	3	137	0	140
C3	16	272	1216	1504
	87	413	1216	1716

Πίνακας 6.6.3: Συνεκτικός πίνακας των δύο διαχωρισμών

Για να αναπαραστίσουμε τις διαφορές μεαξύ των διαχωρισμών, υπολογίζουμε τα πρώτα δύο συστατικά διαχωρισμού (principal components), που βασίζονται στον πίνακα συσχέτισης του ολόκληρου συνόλου δεδομένων, και επεκτείνοντας τις παρατηρήσεις (που σχετίζονται με τους 1716 χρήστες από το σύνολο δεδομένων εκπαίδευσης) σε δύο ορθογώνιες κατευθύνσεις. Το *k*-οστό συστατικό διαχωρισμού (PC*k*) ορίζεται ως ο γραμμικός συνδυασμός

$$Z_k = a_{k1}X_1 + a_{k2}X_2 + ... + a_{kp}X_p$$

Δεδομένου ενός συνόλου *n* παρατηρήσεων στις τυχαίες μεταβλητές $X_1, X_2, ..., X_p$. Οι συντελεστές του Z_k, $a_k = (a_{k1}, a_{k2}, ..., a_{kp})^t$, έχουν ενιαία Ευκλείδια νόρμα, μέγιστη διασπορά και PC*k*, *k*≥2, είναι ασυσχέτιστα με τα προηγούμενα PC, το οποίο σημαίνει ότι $a_i^t a_j = 0$ εάν $i \neq j$ και $a_i^t a_i = 1$. Έτσι, το πρώτο συστατικό διαχωρισμού είναι ο γραμμικός συνδυασμός των παρατηρούμενων μεταβλητών με μέγιστη διασπορά. Το δεύτερο συστατικό διαχωρισμού επαληθεύει παρόμοιο βέλτιστο κριτήριο και είναι ασυσχέτιστο με το PC*1*, κοκ. Ως αποτέλεσμα, τα συστατικά διαχωρισμού χαρακτηρίζονται από φθίνουσα διασπορά, δηλαδή $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, όπου λ_r είναι η

διασπορά του PC*r και p* είναι ο μέγιστος αριθμός των PC (*n*>*p*).
Το διάνυσμα των συντελεστών του *k*-οστού συστατικού διαχωρισμού, α_k, είναι το ιδιοδιάνυσμα που σχετίζεται με την *k*-οστή μεγαλύτερη ιδιοτιμή, λ_k, του πίνακα συνδυακύμανσης των παρατηρούμενων μεταβλητών. Επομένως, η *k*-οστή μεγαλύτερη ιδιοτιμή του πίνακα συνδυακύμανσης είναι η διασπορά του PC*k*, δηλαδή $\lambda_k=\mathrm{Var}(Z_k)$.
Το κλάσμα την συνολικής διακύμανσης των πρώτων *r* συστατικών διαχωρισμού δίνεται από την σχέση

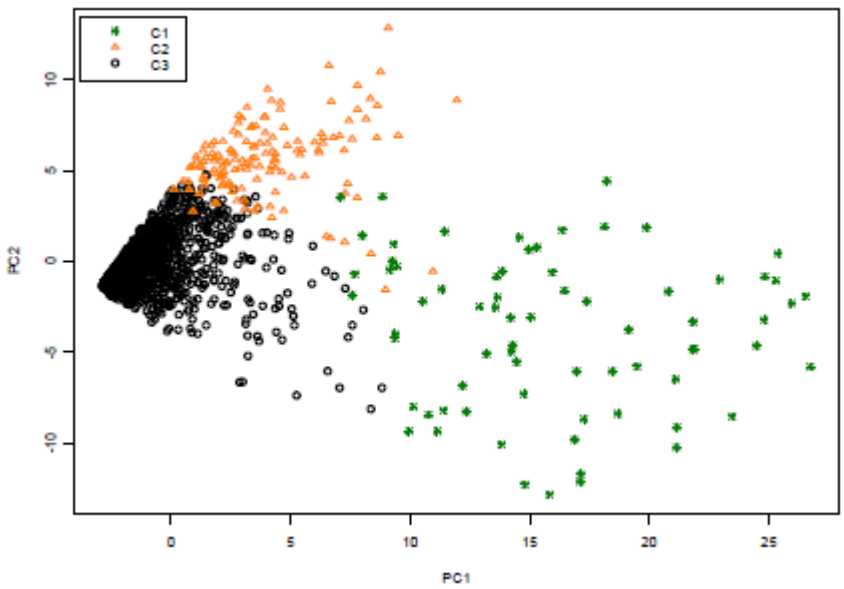
$$\frac{\lambda_1 + ... + \lambda_r}{\lambda_1 + ... + \lambda_p}.$$

Εάν το αποτέλεσμα του κλάσματος είναι κοντά στο 1, τότε υπάρχει τόση πληροφορία στα πρώτα *r* συστατικά διαχωρισμού όση και στις αρχικές *p* μεταβλητές.
Μόλις ορισθούν οι συντελεστές των συστατικών διαχωρισμού, ο βαθμός της μονάδας *i* στο PC*j* δίνεται από τον τύπο

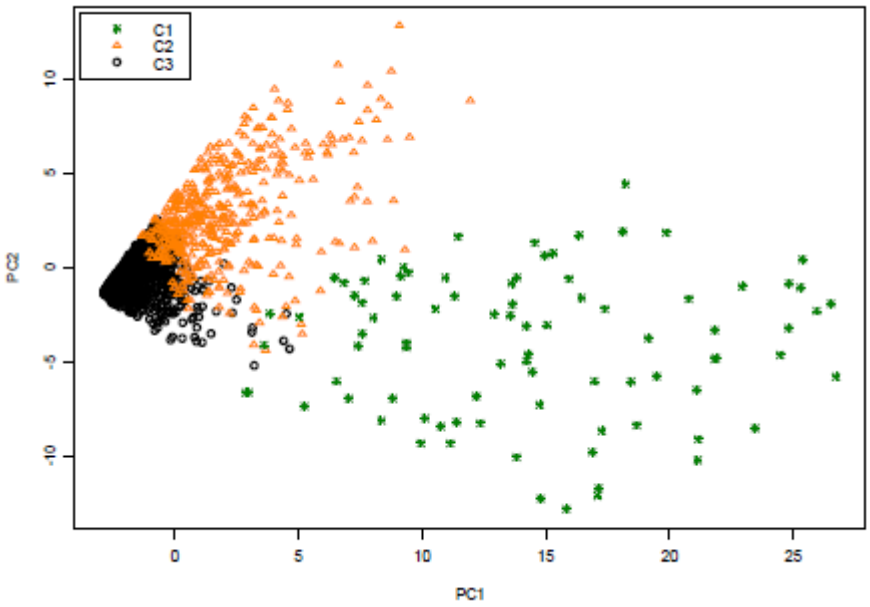
$$z_{ij} = a_{j1}x_{i1} + a_{j2}x_{i2} + ... + a_{jp}x_{ip}$$

όπου $x_i = (x_{i1},...,x_{ip})^t$ είναι τα δεδομένα για τη μονάδα *i*.

Στην περίπτωση μας, το πρώτο συστατικό διαχωρισμού (PC*1*) μπορεί να ερμηνευθεί ως ο μέσος όρος της χρήσης του Internet καθ'όλη τη διάρκεια της ημέρας και το δεύτερο συστατικό διαχωρισμού (PC*2*) ως το μέτρο αντίθεσης μεταξύ της πρωινής και της απογευματινής χρήσης. Στην Εικόνα 6.6.4 αναπαριστάται ο διαχωρισμός που βασίζεται σε όλο το σύνολο δεδομένων ενώ στην Εικόνα 6.6.5 ο διαχωρισμός που βασίζεται στο σύνολο δεδομένων εκπαίδευσης. Αυτά τα γραφήματα αναπαριστούν το γεγονός ότι ένας μεγάλος αριθμός χρηστών που ταξινομούνται στην συστάδα C_3, με τον διαχωρισμό που βασίζεται σε όλο το σύνολο δεδομένων, ταξινομούνται στην C_2 με τον διαχωρισμό που βασίζεται στο σύνολο δεδομένων εκπαίδευσης.



Εικόνα 6.6.4: Βαθμοί των χρηστών (σύνολο δεδομένων εκπαίδευσης) στα δύο πρώτα συστατικά διαχωρισμού, που αποκτήθηκαν από το συνολικό σύνολο δεδομένων. Οι συστάδες σχηματίσθηκαν λαμβάνοντας υπ'όψιν όλους τους χρήστες

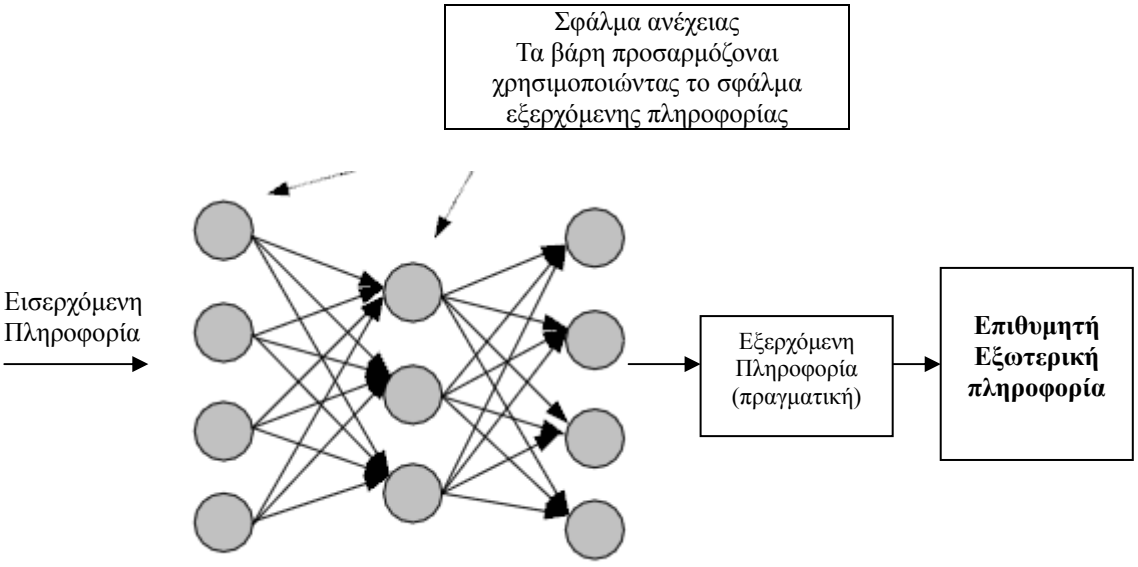


Εικόνα 6.6.5: Βαθμοί των χρηστών (σύνολο δεδομένων εκπαίδευσης) στα δύο πρώτα συστατικά διαχωρισμού, που αποκτήθηκαν από το συνολικό σύνολο δεδομένων. Οι συστάδες σχηματίσθηκαν λαμβάνοντας υπ'όψιν τους χρήστες από το σύνολο δεδομένων εκπαίδευσης

Ταξινόμηση χρησιμοποιώντας Νευρωνικά Δίκτυα

Γενικά, τα νευρωνικά δίκτυα (Neural Networks) περιλαμβάνουν πολλά επίπεδα (layers) με νευρώνες (neurons): το επίπεδο εισερχόμενης πληροφορίας λαμβάνει δεδομένα, ένα ή περισσότερα επίπεδα τα οποία λαμβάνουν δεδομένα μόνο από την προηγούμενη μονάδα, και ένα επίπεδο εξερχόμενης πληροφορίας το οποίο λαμβάνει τα αποτελέσματα από το προηγούμενο επίπεδο επεξεργασμένων μονάδων. Κάθε εισερχόμενη τιμή μια επεξεργασμένης μονάδας είναι πολλαπλασιασμένο με ένα βάρος και το άθροισμα όλων εκείνων των τιμών μαζί με μια βαθμωτή μεροληψία (scalar bias) για κάθε νευρώνα εφαρμόζονται σε μια συνάρτηση (ήδη ορισμένη για κάθε στρώμα), παράγοντας την εξερχόμενη τιμή κάθε νευρώνα. Υπάρχουν τρεις κύριες τοπολογίες σύνδεσης που ορίζουν την κίνηση των δεδομένων μεταξύ του επιπέδου εισερχόμενης πληροφορίας, των κρυφών και του επιπέδου εξερχόμενης πληροφορίας: δίκτυα τροφοδοσίας προς τα εμπρός διάδοσης (feed-forward network), δίκτυα περιορισμένης επανάληψης (limited recurrent network) και δίκτυα πλήρους επανάληψης (fully recurrent network). Τα δίκτυα τροφοδότησης προς τα εμπρός διάδοσης είναι κατάλληλα για να λύσουν προβλήματα όπου όλη η πληροφορία μπορεί να παρουσιαστεί στο νευρωνικό δίκτυο αμέσως.
Η εφαρμογή ενός νευρωνικού δικτύου ως προς την επίλυση ενός προβλήματος περιλαμβάνει δύο φάσεις: την φάση εκπαίδευσης (training phase) και την φάση εξέτασης (test phase). Στην φάση εκπαίδευσης, το σετ δεδομένων εισέρχεται σε ένα νευρωνικό δίκτυο το οποίο προσαρμόζει επαναληπτικά τα βάρη και την μεροληψία του δικτύου με σκοπό να παράγει ένα αποτέλεσμα το οποίο αντιστοιχεί, σύμφωνα με έναν βαθμό ακρίβειας, σε ένα ήδη γνωστό αποτέλεσμα, το οποίο είναι το σύνολο δεδομένων εξέτασης. Στη φάση εξέτασης, νέα δεδομένα εισάγονται στο δίκτυο και ένα νέο αποτέλεσμα αποκτάται, βασισμένο στις παραμέτρους του δικτύου, οι οποίες υπολογίζονται κατά την φάση εκπαίδευσης. Η φάση εξέτασης σε ένα νευρωνικό δίκτυο αποσκοπεί στο να τεθεί ο σωστός συνδυασμός των διαχωριστικών συναρτήσεων που θα χρησιμοποιηθούν για την σωστή ταξινόμηση των δεδομένων. Υπάρχουν δύο μαθησιακά παραδείγματα (learning paradigms), το ελεγχόμενο και το μη-ελεγχόμενο (supervised and non-supervised learning) και πολλοί μαθησιακοί αλγόριθμοι (learning algorithms) οι οποίοι μπορούν να εφαρμοστούν και αυτό εξαρτάται απαραίτητα από τον τύπο του προβλήματος που πρόκειται να επιλυθεί.

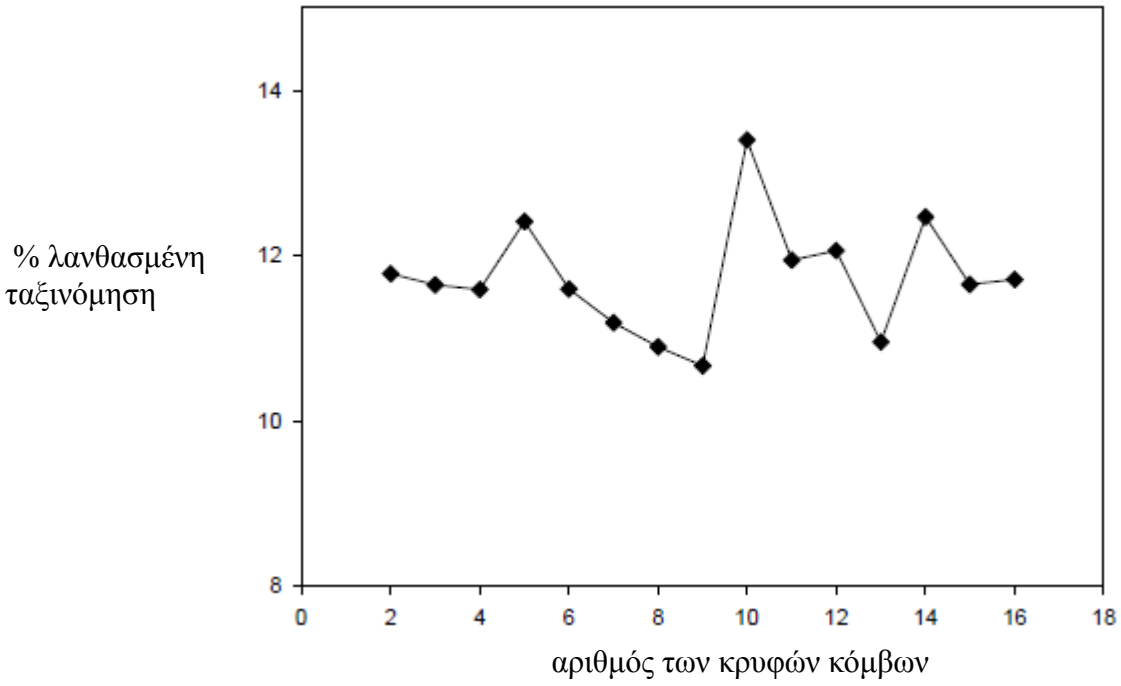
Ο συνδυασμός της τοπολογίας, του μαθησιακού παραδείγματος και του μαθησιακού αλγορίθμου ορίζουν ένα μοντέλο νευρωνικού δικτύου (neural network model). Η προς τα πίσω διάδοση (back propagation) είναι ένας κατάλληλος μαθησιακός αλγόριθμος για την κατάρτιση ενός δικτύου τροφοδότησης προς τα εμπρός διάδοσης για την ταξινόμηση διανυσμάτων, την μοντελοποίηση και πρόβλεψη χρονοσειρών. Είναι γενικά ένας προτεινόμενος μαθησιακός αλγόριθμος, ο οποίος είναι 'δυνατός' αλλά και αρκετά πολύπλοκος ως προς τα πλαίσια των υπολογισμών που απαιτούνται για την κατάρτιση. Ένα νευρωνικό δίκτυο τροφοδότησης προς τα πίσω διάδοσης χρησιμοποιεί τοπολογία προς τα εμπρός διάδοσης, ελεγχόμενης κατάρτισης και τον μαθησιακό αλγόριθμο προς τα πίσω διάδοσης. Ένα δίκτυο τροφοδότησης προς τα πίσω διάδοσης με ένα κρυφό επίπεδο επεξεργασμένων στοιχείων μπορεί να μοντελοποιήσει κάθε συνεχής συνάρτηση μέχρι ένα βαθμό ακριβείας (δίνοντας αρκετά επεξεργασμένα στοιχεία στο κρυφό επίπεδο).
Ο αλγόριθμος προς τα πίσω διάδοσης αποτελείται από τρία βήματα (Εικόνα 6.6.6). Το εισερχόμενο διάνυσμα παρουσιάζεται στο επίπεδο εισερχόμενης πληροφορίας του δικτύου. Αυτά τα δεδομένα διαδίδονται μέσω του δικτύου μέχρι να φτάσουν στο επίπεδο εξερχόμενης πληροφορίας. Μέσω αυτής της διαδικασίας παράγεται το πραγματικό ή το προβλεπόμενο εξερχόμενο διάνυσμα. Καθώς η διάδοση προς τα πίσω είναι ένας ελεγχόμενος μαθησιακός αλγόριθμος, τα επιθυμητά αποτελέσματα δίνονται ως μέρος του συνόλου δεδομένων εκπαίδευσης. Η εξερχόμενη πληροφορία του πραγματικού δικτύου αφαιρείται από την επιθυμητή εξερχόμενη πληροφορία και έτσι παράγεται ένα σφάλμα παρατήρησης. Αυτό το σφάλμα παρατήρησης αποτελεί τη βάση για το βήμα της διάδοσης προς τα πίσω, σύμφωνα με την οποία τα σφάλματα περνάνε προς τα πίσω μέσω του νευρωνικού δικτύου υπολογίζοντας τη συνεισφορά κάθε κρυφής μονάδας επεξεργασίας και παράγοντας την απαραίτητη προσαρμογή που απαιτείται για να παραχθεί η σωστή εξερχόμενη πληροφορία. Έπειτα, προσαρμόζονται τα βάρη την σύνδεσης. Δύο μεγάλοι μαθησιακοί παράγοντες χρησιμοποιούνται για να ελέγξουν την διαδικασία εκπαίδευσης ενός δικτύου τροφοδότησης προς τα πίσω διάδοσης: ο μαθησιακός ρυθμός χρησιμοποιείται για να καθοριστεί εάν το νευρωνικό δίκτυο πρόκειται να κάνει μεγάλες προσαρμογές μετά από κάθε προσπάθεια μάθησης ή εάν πρόκειται να κάνει μόνο μικρές προσαρμογές και η ορμή χρησιμοποιείται για ελέχξει πιθανές ταλαντώσεις στα βάρη, οι οποίες μπορεί να προκλήθηκαν από εναλλασσόμενα προσημασμένα σφάλματα. Αυτές οι δύο παράμετροι συνήθως παράγουν την μεγαλύτερη επίδραση στον χρόνο και στην επίδοση του δικτύου εκπαίδευσης.



Εικόνα 6.6.6: Δίκτυα προς τα πίσω διάδοσης

Στην περίπτωση μας, κάθε διάνυσμα εισερχόμενης πληροφορίας πρέπει να περιέχει τον ρυθμό

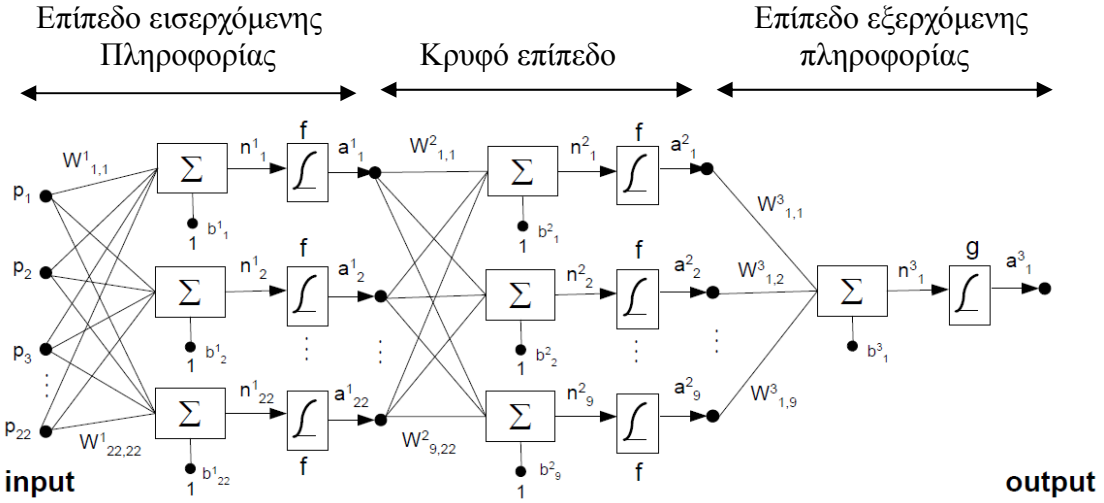
μεταφοράς σε κάθε ένα από τα *k*-οστά μισάωρα διανύσματα, $X_k, k = 1,2,...,48$ δηλαδή πρέπει να έχει μια διάσταση των 48 στοιχείων και πρέπει να ταξινομηθεί σε μία από τις 3 συστάδες. Για να μειώσουμε τις διαστάσεις των αρχικών δεδομένων χρησιμοποιήθηκε Ανάλυση κατά Κύρια Συστατικά (Principal Component Analysis, PCA). Έτσι, η διάσταση κάθε διανύσματος εισερχόμενης πληροφορίας μειώθηκε σε 22 εξαλείφοντας τα αρχικά συστατικά που συνεισφέρουν λιγότερο από 0.5% στην συνολική μεταβλητότητα του συνόλου δεδομένων.
Για ένα πρόβλημα τέτοιας διαστάσεως, ένα δίκτυο τροφοδότησης προς τα εμπρός πίσω-διάδοσης με τρία επίπεδα φαίνεται να είναι κατάλληλο. Το επίπεδο εισερχόμενης πληροφορίας θα έχει 22 νευρώνες, σύμφωνα με την διαστατοποίηση των εισερχόμενων διανυσμάτων, και το επίπεδο της εξερχόμενης πληροφορίας θα έχει 1 νευρώνα. Ο αριθμός των κόμβων σε ένα κρυφό επίπεδο επιλέγεται εμπειρικά έτσι ώστε η συνάρτηση, η οποία έχει μέσο τετραγωνικό σφάλμα για δίκτυα τροφοδότησης προς τα εμπρός διάδοσης, να ελαχιστοποιείται. Αρχικά, θεωρήθηκαν νευρωνικά δίκτυα με μεταβλητό αριθμό νευρώνων στο κρυφό επίπεδο. Για να καταρτιστεί κάθε νευρωνικό δίκτυο χρησιμοποιήθηκε το σύνολο δεδομένων εκπαίδευσης και ελέγχθηκε χρησιμοποιώντας το σύνολο δεδομένων εξέτασης. Η Εικόνα 6.6.7 δείχνει το ποσοστό των λανθασμένων ταξινομήσεων (συγκρινόμενο με την ταξινόμηση κατά συστάδες βάσει του συνόλου δεδομένων εκπαίδευσης) όταν τα δίκτυα εκπαίδευσης (με διαφορετικό αριθμό νευρώνων στο κρυφό επίπεδο) χρησιμοποιούνται για να ταξινομήσουν το σύνολο δεδομένων εξέτασης. Σύμφωνα μ'αυτά τα αποτελέσματα, είναι φανερό ότι αυξάνοντας τον αριθμό των κρυφών κόμβων πέρα από 9 δεν παρατηρείται καμία βελτίωση στην εκτέλεση, γι'αυτό το λόγο ο αριθμός των κρυφών κόμβων που θα χρησιμοποιηθούν για την ταξινόμηση επιλέχθηκε να είναι 9.



Εικόνα 6.6.7: Ποσοστό λανθασμένης ταξινόμησης προς τον αριθμό των κόμβων στο κρυφό επίπεδο

Η δομή του νευρωνικού δικτύου φαίνεται στο γράφημα 6.6.8. Το μέγεθος $w_{j,k}^i$ αναπαριστά την τιμή του βάρους που αναφέρεται στο επίπεδο *i*, όπου *i* =1,2,3, το οποίο πολλαπλασιάζεται με το εισερχόμενο *k* του νευρώνα *j*, όπου *j* και *k* είναι διαφορετικής τάξεως σύμφωνα με το επίπεδο του δικτύου. Η βαθμίδα b_j^i αναπαριστά την μεροληψία που συσχετίζεται με τον νευρώνα *j* του επιπέδου *i*.

Για τα επίπεδα εισερχόμενης πληροφορίας και κρυφά επίπεδα, χρησιμοποιείται μια λογαριθμική-σιγμοειδής (log-sigmoid) συνάρτηση μεταφοράς, παράγοντας αποτελέσματα μεταξύ 0 και 1 καθώς η εισερχόμενη πληροφορία του νευρώνα παίρνει τιμές από το $-\infty$ μέχρι το $+\infty$. Για το επίπεδο εξωτερικής πληροφορίας, χρησιμοποιείται μια γραμμική συνάρτηση μεταφοράς. Πολλαπλά στρώματα νευρώνων με μη γραμμικές συναρτήσεις μεταφοράς επιτρέπουν στο δίκτυο να μάθει τις μη-γραμμικές και γραμμικές σχέσεις μεταξύ των εισερχόμενων και εξερχόμενων διανυσμάτων. Ένα νευρωνικό δίκτυο που περιλαμβάνει μεροληψίες, ένα σιγμοειδές επίπεδο και ένα γραμμικό εξερχόμενο επίπεδο μπορεί να προσεγγίσει κάθε συνάρτηση με πεπερασμένο αριθμό ασυνεχειών. Για να βελτιωθεί η γενίκευση του νευρωνικού δικτύου χρησιμοποιείται η αυτοματοποιημένη Μπευζιανή κανονικοποίηση (automated Bayesian regularization).



Εικόνα 6.6.8: Αρχιτεκτονική του Νευρωνικού Δικτύου που χρησιμοποιήθηκε για την ταξινόμηση των χρηστών του Διαδικτύου

Χρησιμοποιώντας τους παραπάνω συμβολισμούς, το αποτέλεσμα κάθε νευρώνα δίνεται από το τύπο: $a_j^1 = f(w_{j,k}^1 p_k + b_j^1),\ j,k = 1,...,22$ όπου η ποσότητα p_k δίνει την k-οστή εισερχόμενη πληροφορία, για το στρώμα εισερχόμενης πληροφορίας, $a_j^2 = f(w_{j,k}^2 a_k^1 + b_j^2),\ j = 1,...,9, k = 1,...,22$ για το μεσσαίο στρώμα και $a_1^3 = g(w_{1,k}^3 a_k^2),\ k = 1,...,9$ για το στρώμα εξερχόμενης πληροφορίας.

Αποτελέσματα ταξινόμησης

Το αρχικό σύνολο δεδομένων, που περιέχει 3432 χρήστες, χωρίζεται σε δύο υποσύνολα ίδιου μεγέθους. Πρώτα, η ταξινόμηση των χρήστων στο πρώτο σύνολο δεδομένων, που ονομάζεται σύνολο δεδομένων εκπαίδευσης, χρησιμοποιείται για την εκτίμηση των διαχωριστικών συναρτήσεων στην περίπτωση της διαχωριστικής ανάλυσης, και για να εκπαιδεύσει το νευρωνικό δίκτυο. Έπειτα, οι χρήστες του δεύτερου συνόλου, που ονομάζεται σύνολο δεδομένων εξέτασης, ταξινομούνται σε μία από τις 3 ομάδες οι οποίες έχουν καθοριστεί από τον διαχωρισμό σύμφωνα με το σύνολο δεδομένων εκπαίδευσης.
Για κάθε από τις τρεις συστάδες, έχουμε υπολογίσει το αντίστοιχο διάνυσμα των δειγματικών μέσων, ο οποίος ονομάζεται και κεντροειδής της συστάδας. Ακολούθως, ο απλούστερος κανόνας ταξινόμησης που μπορεί να σχεδιαστεί για να ταξινομηθούν οι χρήστες του συνόλου δεδομένων εξέτασης είναι να ταξινομηθεί κάθε χρήστης στην συστάδα που σχετίζεται με τον κοντινότερο κεντροειδή.

Ξεκινώντας από την ταξινόμηση των χρηστών από το σύνολο δεδομένων εκπαίδευσης, η γραμμική διαχωριστική ανάλυση χρησιμοποιείται για να ταξινομηθούν οι χρήστες του συνόλου δεδομένων εξέτασης. Δηλαδή, η διαχωριστική συνάρτηση του Fisher βασισμένη στο σύνολο δεδομένων εκπαίδευσης χρησιμοποιείται για να ταξινομήσει του χρήστες του συνόλου δεδομένων εξέτασης. Καθώς οι 3 συστάδες του συνόλου δεδομένων εκπαίδευσης έχουν πολύ διαφορετικά μεγέθη, οι ρυθμοί σφάλματος της ταξινόμησης η οποία βασίζεται στην διαχωριστική ανάλυση υπολογίζονται σύμφωνα με τις εκ των προτέρων πιθανότητες αναλόγως με τα μεγέθη των ομάδων. Ο προφανής ρυθμός σφάλματος είναι 6.82%. Καθώς είναι γνωστό ότι αυτός ο ρυθμός υποτιμά τον πραγματικό ρυθμό σφάλματος, επίσης υπολογίστηκε ένας leave-one-out ρυθμός σφάλματος, οδηγώντας στο ρυθμό σφάλματος του 7.98%. Για να υπολογιστεί αυτός ο ρυθμός σφάλματος ένας χρήστης του συνόλου δεδομένων εκπαίδευσης αφαιρείται και οι υπόλοιποι 1715 χρησιμοποιούνται για την εκτίμηση των διαχωριστικών συναρτήσεων. Έπειτα, η παρατήρηση που αφαιρέθηκε, ταξινομείται. Καθώς και οι δύο ρυθμοί σφάλματος είναι χαμηλοί, περιμένουμε καλά αποτελέσματα στην ταξινόμηση του συνόλου δεδομένων εξέτασης.
Πριν την χρήση του νευρωνικού δικτύου διεξάγεται μια PCA ανάλυση στα δεδομένα που συσχέτιζονται με το σύνολο δεδομένων εκπαίδευσης και αποκτήθηκαν 22 μεταβλητές, οι οποίες εξηγούν περισσότερο από το 0.5% της συνολικής μεταβλητότητας. Έτσι η ανάλυση του νευρωνικού δικτύου εφαρμόζεται σ'αυτές και όχι στο αρχικό σύνολο δεδομένων των 58 μεταβλητών.
Τα αποτελέσματα της ταξινόμησης των χρηστών του συνόλου δεδομένων εκπαίδευσης, το οποία βασίζονται στις διαφορετικές μεθόδους ταξινόμησης (διαχωριστική ανάλυση, νευρωνικά δίκτυα και της απόσταση από τον κοντινότερο κεντροειδή) συνοψίζονται στον Πίνακα 6.6.4. Τα αποτελέσματα αυτών των μεθόδων ταξινόμησης συγκρίνοται με τον διαχωρισμό που αποκτήθηκε μέσω της ανάλυσης κατά συστάδες, η οποία διεξάχθηκε σε όλα τα δεδομένα, δηλαδή λαμβάνοντας υπ'όψιν και τους 3432 χρήστες. Η σύγκριση (Πίνακας 6.6.4) για να βρεθεί ποια μέθοδος δίνει την καλύτερη ταξινόμηση των χρηστών γίνεται με βάση ένα μέγεθος που δίνει το ποσοστό των χρηστών που είναι ταξινομημένοι στην ίδια ομάδα (λαμβάνοντας ως μέτρο αναφοράς τον διαχωρισμό που αποκτάται λαμβάνοντας υπ'όψιν ολόκληρο το σύνολο δεδομένων).

Μέθοδος ταξινόμησης	C1	C2	C3	Ποσοστό ομοιότητας
Νευρωνικά δίκτυα	92	275	1349	88.00%
Διαχωριστική ανάλυση	68	256	1392	92.13%
Απόσταση από κεντροειδείς	75	315	1326	88.69%
Συστάδες που αποκτήθηκαν από το συνολικό σύνολο δεδομένων	73	126	1517	

Πίνακας 6.6.4: Ταξινόμηση των χρηστών του συνόλου δεδομένων εξέτασης

Συνολικό σύνολο δεδομένων	NN			
	C1	C2	C3	
C1	57	11	5	73
C2	9	113	4	126
C3	26	151	1340	1517
	92	275	1349	1716

Πίνακας 6.6.5: Διασταύρωση της ταξινόμησης της μεθόδου του Νευρωνικού δικτύου και του αρχικού διαχωρισμού με βάση το συνολικό σύνολο δεδομένων

Συνολικό σύνολο δεδομένων	Διαχωριστική ανάλυση			
	C1	C2	C3	
C1	63	10	0	73
C2	0	126	0	126
C3	5	120	1392	1517
	68	256	1392	1716

Πίνακας 6.6.6: Διασταύρωση της ταξινόμησης της Διαχωριστικής ανάλυσης και του αρχικού διαχωρισμού με βάση το συνολικό σύνολο δεδομένων

Συνολικό σύνολο δεδομένων	Κεντροειδής			
	C1	C2	C3	
C1	72	1	0	73
C2	2	124	0	126
C3	1	190	1326	1517
	75	315	1326	1716

Πίνακας 6.6.7: Διασταύρωση της ταξινόμησης σύμφωνα με τον κοντινότερο κεντροειδή και του αρχικού διαχωρισμού με βάση το συνολικό σύνολο δεδομένων

Από τον Πίνακα 6.6.4 συμπεραίνουμε ότι η μέθοδος της διαχωριστικής ανάλυσης είναι η μέθοδος ταξινόμησης που δίνει τα καλύτερα αποτελέσματα. Από την ανάλυση των Πινάκων 6.6.5, 6.6.6 και 6.6.7, διαπιστώνουμε ότι η πλειοψηφία του σφάλματος στην ταξινόμηση που παράγεται από τις τρεις μεθόδους ταξινόμησης προέρχεται από χρήστες που ανήκουν στην τρίτη συστάδα, C3, ενώ λανθασμένα ταξινομούνται στην συστάδα C2. Αυτό σημαίνει ότι οι διαδικασίες ταξινόμησης έχουν προβλήματα στο να διαχωρίσουν χρήστες με μικρά ποσοστό μεταφοράς κατά την διάρκεια όλου του 24ωρου με τους χρήστες με χαμηλά ποσοστά μεταφοράς στις πρωινές ώρες και του χρήστες με υψηλά ποσοστά μεταφοράς στις απογευματινές ώρες. Αυτό το αποτέλεσμα είναι αναμενόμενο, καθώς ο διαχωρισμός που έγινε σύμφωνα με το σύνολο δεδομένων εκπαίδευσης αποκαλύπτει το ίδιο πρόβλημα. Οι χρήστες από την ομάδα C2 ταξινομήθηκαν από την διαχωριστική ανάλυση στην σωστή συστάδα.

ΚΕΦΑΛΑΙΟ 7^{Ο}

Διαχωριστική ανάλυση στο SPSS Clementine

7.1 Συμβολισμοί

Οι συμβολισμοί που χρησιμοποιούνται παρακάτω είναι οι εξής:

g αριθμός των ομάδων

p αριθμός των μεταβλητών

q αριθμός των επιλεγμένων μεταβλητών

X_{ijk} τιμή της μεταβλητής i για την περίπτωση k στην ομάδα j

f_{jk} βάρος της περίπτωσης k στην ομάδα j

m_j αριθμός των περιπτώσεων στην ομάδα j

n_j άθροισμα των βαρών στην ομάδα j

n συνολικό άθροισμα των βαρών

7.2 Βασικά στατιστικά μεγέθη

Μέσος (mean)

$$\bar{x}_{ij} = (\sum_{k=1}^{m_j} f_{jk} x_{ijk}) / n_j \quad \text{(μεταβλητή } i \text{ στην ομάδα } j\text{)}$$

Διασπορά (variance)

$$s_{ij}^2 = \frac{(\sum_{k=1}^{m_j} f_{jk} x_{ijk}^2 - n_j \bar{x}_{ij}^2)}{(n_j - 1)} \quad \text{(μεταβλητή } i \text{ στην ομάδα } j\text{)}$$

Άθροισμα τετραγώνων μεταξύ των ομάδων (within-groups sums of squares W)

$$w_{il} = \sum_{j=1}^{g} \sum_{k=1}^{m_j} f_{jk} x_{ijk} x_{ijk} - \sum_{j=1}^{g} (\sum_{k=1}^{m_j} f_{jk} x_{ijk})(\sum_{k=1}^{m_j} f_{jk} x_{ljk}) / n_j \qquad i, l = 1, ..., p$$

Όλικο άθροισμα τετραγώνων (total sums of squares T)

$$t_{il} = \sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk} x_{ijk} x_{ljk} - (\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk} x_{ijk})(\sum_{j=1}^{g}\sum_{k=1}^{m_j} f_{jk} x_{ljk}) / n$$

7.3 Κανόνες επιλογής μεταβλητών

Η επιλογή των μεταβλητών μπορεί να γίνει άμεσα ή με την μπρος-πίσω μέθοδο (stepwise).

7.3.1 Άμεση μέθοδος

Στην άμεση μέθοδο επιλογής μεταβλητών, μια μεταβλητή περιλαμβάνεται στην ανάλυση όταν και μετά την εισαγωγή της στην ανάλυση, η τιμή ανέχειας κάθε μεταβλητής παραμένει μεγαλύτερη ή ίση από το επιτρεπτό όριο (=0.001).

7.3.2 Μπρος-πίσω επιλογή μεταβλητών (stepwise variable selection)

Οι κάτωθι κανόνες ακολουθούνται για την επιλογή μεταβλητών με την μπρος-πίσω μέθοδο

- Πρώτα εισάγονται κατάλληλες μεταβλητές που είναι στατιστικά σημαντικές για το μοντέλο.
- Η σειρά εισαγωγής των κατάλληλων μεταβλητών, που είναι το ίδιο στατιστικά σημαντικές, καθορίζεται από την τιμή τους σε κάθε κριτήριο εισαγωγής που εφαρμόζεται. Η μεταβλητή με την "καλύτερη" τιμή στο κριτήριο που χρησιμοποιείται εισάγεται πρώτα.
- Μια μεταβλητή είναι κατάλληλη να αφαιρεθεί εάν η F-τιμή είναι μικρότερη από την θεωρητική τιμή F για την αφαίρεση μεταβλητών. Εάν περισσότερες από μια μεταβλητές πρέπει να αφαιρεθούν, η μεταβλητή η οποία αφαιρείται είναι αυτή που είναι λιγότερο στατιστικά σημαντική και αφηνεί τις μεταβλητές που είναι απαραίτητες για το μοντέλο. Η αφαίρεση των μεταβλητών συνεχίζεται μέχρι να μην χρειάζεται αφαιρεθεί καμία άλλη μεταβλητή.
- Μια μεταβλητή που έχει μηδενικό επίπεδο σημαντικότητας δεν εισάγεται ποτέ στο μοντέλο.

Κριτήρια ακαταλληλότητας

Μια μεταβλητή θεωρείται ακατάλληλη για να εισαχθεί στην ανάλυση εάν:

- Η τιμή της ανέχειας για κάθε μεταβλητή στην ανάλυση είναι μικρότερη από το όριο της ανέχειας, εάν εισαχθεί.
- Η τιμή της F για την εισαγωγή είναι μικρότερη από την F-τιμή για μια μεταβλητή.
- Εάν χρησιμοποιείται το κριτήριο πιθανότητας, το επίπεδο σημαντικότητας που σχετίζεται με την F υπερβαίνει την πιθανότητα εισαγωγής.

7.4 Υπολογισμοί κατά την επιλογή μεταβλητών

Κατά την επιλογή μεταβλητών, ο πίνακας W αντικαθίσταται σε κάθε βήμα από έναν νέο συμμετρικό πίνακα W^*. Εάν οι πρώτες q μεταβλητές έχουν περιληφθεί στην ανάλυση, ο W μπορεί να γραφτεί ως:

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

Όπου W_{11} είναι ένας qxq πίνακας. Ο πίνακας W^* ορίζεται ως:

$$W^* = \begin{bmatrix} -W_{11}^{-1} & W_{11}^{-1}W_{12} \\ W_{21}W_{11}^{-1} & W_{22} - W_{21}W_{11}^{-1}W_{12} \end{bmatrix} = \begin{bmatrix} W_{11}^* & W_{12}^* \\ W_{21}^* & W_{22}^* \end{bmatrix}$$

Επιπλέον, όταν γίνεται επιλογή μεταβλητών με την μπρος-πίσω επιλογή, ο πίνακας T αντικαθίσταται από τον πίνακα T^*, ο οποίος ορίζεται με όμοιο τρόπο.

Ανέχεια (Tolerance)

$$TOL_i = \begin{cases} \mathbf{0} & \text{εάν } w_{ii} = \mathbf{0} \\ w_{ii}^* / w_{ii} & \text{εάν η μεταβλητή } i \text{ δεν είναι στην ανάλυση και } w_{ii} \neq \mathbf{0} \\ -\mathbf{1}/(w_{ii}^* w_{ii}) & \text{εάν η μεταβλητή } i \text{ είναι στην ανάλυση και } w_{ii} \neq \mathbf{0} \end{cases}$$

Εάν η ανέχεια μιας μεταβλητής είναι μικρότερη ή ίση σε σύγκριση με ένα συγκεκριμένο όριο ανέχειας, ή η εισαγωγή της στην ανάλυση θα είχε ως αποτέλεσμα την μείωση της ανέχειας μιας άλλης μεταβλητής στην εξίσωση στο όριο ή κάτω από το όριο, τότε τα ακόλουθα στατιστικά μεγέθη δεν υπολογίζονται για την συγκεκριμένη μεταβλητή ή για κάθε σύνολο δεδομένων που την περιλαμβάνει.

F-to-Remove

$F_i = \dfrac{(w_{ii}^* - t_{ii}^*)(n-q-g+1)}{t_{ii}^*(g-1)}$ με *g-1* και *n-q-g+1* βαθμούς ελευθερίας

F-to-Enter

$F_i = \dfrac{(t_{ii}^* - w_{ii}^*)(n-q-g)}{w_{ii}^*(g-1)}$ με *g-1* και *n-q-g* βαθμούς ελευθερίας

Wilks' Lambda για τον έλεγχο των μέσων των ομάδων

$\Lambda = |W_{11}|/|T_{11}|$ με *q, g-1* και *n-g* βαθμούς ελευθερίας

Rao's V (Lawley-Hotelling Trace)

$$V = -(n-g)\sum_{i=1}^{q}\sum_{l=1}^{q} w_{il}^{*}(t_{il} - w_{il})$$

Όταν το *n-g* είναι αρκετά μεγάλο, το *V* υπό την μηδενική υπόθεση συμπεριφέρεται όπως η χ^2 με *q(g-1)* βαθμούς ελευθερίας. Όταν εισάγεται μία μεταβλητή, η αλλαγή στο *V*, εάν είναι θετική, ακολουθεί την χ^2 με *g-1* βαθμούς ελευθερίας.

Συναρτήσεις ταξινόμησης (classification functions)

Όταν επιλεγχθούν *q* μεταβλητές, οι συναρτήσεις ταξινόμησης (γνωστές και ως γραμμικές διαχωριστικές συναρτήσεις του Fisher) μπορούν να υπολογιστούν χρησιμοποιώντας

$$b_{ij} = (n-g)\sum_{l=1}^{q} w_{il}^{*}\bar{x}_{lj} \quad i = 1,2,...,q, j = 1,2,...,g$$

για τους συντελεστές, και

$$a_j = \log p_j - \frac{1}{2}\sum_{i=1}^{q} b_{ij}\bar{x}_{ij} \quad j = 1,2,...,q$$

για την σταθερά, όπου p_j είναι η προηγούμενη πιθανότητα για την ομάδα *j*.

7.5 Κανονικές διαχωριστικές συναρτήσεις (canonical discriminant functions)

Οι συντελεστές της κανονικής διαχωριστικής συνάρτησης καθορίζονται λύνοντας το γενικό πρόβλημα ιδιοτιμών

(T-W)V=λWV

όπου **V** είναι ο μη-τυποποιημένος πίνακας των συντελεστών της διαχωριστικής συνάρτησης και λ είναι ο διαγώνιος πίνακας των ιδιοτιμών. Το σύστημα λύνεται ως ακολούθως:

Σχηματίζεται η αποσύνθεση Cholesky,

W=LU

όπου **L** είναι ο μικρότερος τριγωνικός πίνακας, και **U=L΄.**

Σχηματίζεται ο συμμετρικός πίνακας $\mathbf{L^{-1}BU^{-1}}$ και λύνεται το σύστημα

$(\mathbf{L^{-1}(T-W)U^{-1}-\lambda I)(UV)=0}$

χρησιμοποιώντας τριδιαγωνοποίηση και την QL μέθοδο. Το αποτέλεσμα του παραπάνω συστήματος είναι οι *m* ιδιοτιμές, όπου *m=min(q,g-1)* και τα ορθογώνια ιδιοδιανύσματα **UV**. Τα ιδιοδιανύσματα του αρχικού συστήματος επαληθεύουν την

V=U^{-1}(UV)

Για κάθε μία από τις ιδιοτιμές, οι οποίες διατάσσονται σε φθίνουσα σειρά, υπολογίζονται τα ακόλουθα στατιστικά μεγέθη.

Ποσοστό της διασποράς μεταξύ των ομάδων (Percentage of Between-Groups Variance)

$$\frac{100\lambda_k}{\sum_{k=1}^{m}\lambda_k}$$

Κανονική συσχέτιση (Canonical Correlation)

$$\sqrt{\lambda_k /(1+\lambda_k)}$$

Wilks' Lambda

Ελέγχεται η σημαντικότητα των διαχωριστικών συναρτήσεις (μετά τις *k* πρώτες)

$$\Lambda_k = \prod_{i=k+1}^{m} 1/(1+\lambda_i) \quad k = 0,1,...,m-1$$

Το επίπεδο σημαντικότητας βασίζεται στην τιμή

$$\chi^2 = -(n-(q+g)/2-1)\ln\Lambda_k$$

με (*q-k*)(*g-k-1*) βαθμούς ελευθερίας

Πίνακας των τυποποιημένων κανονικών διαχωριστικών συντελεστών D (The standardized canonical discriminant coefficient matrix D)

Ο πίνακας των τυποποιημένων κανονικών διαχωριστικών συντελεστών D δίνεται από την σχέση

$$D = S_{11}^{-1}V$$

όπου

$$S = diag(\sqrt{w_{11}},\sqrt{w_{22}},...,\sqrt{w_{pp}})$$

S_{11} είναι η διαμέριση που περιλαμβάνει τις πρώτες *q* γραμμές και στήλες του *S*

V είναι ο πίνακας των ιδιοδιανυσμάτων τέτοιος ώστε $V'W_{11}V = I$

Σχέσεις μεταξύ των κανονικών διαχωριστικών συναρτήσεων και των διαχωριστικών μεταβλητών

Οι σχέσεις μεταξύ των κανονικών διαχωριστικών συναρτήσεων και των διαχωριστικών μεταβλητών δίνονται από τον παρακάτω τύπο:

$$R = S_{11}^{-11} W_{11} V$$

Εάν κάποιες μεταβλητές δεν εισαχθούν στην ανάλυση ($q<p$), τα ιδιοδιανύσματα επεκτείνονται με μηδενικά έτσι ώστε να συμπεριληφθούν και οι προαναφερθέν μεταβλητές στον πίνακα συσχέτισης. Οι μεταβλητές για τις οποίες $W_{ii}=0$ εξαιρούνται από το S και το W γι'αυτόν τον υπολογισμό, και το p αναπαριστά τον αριθμό των μεταβλητών με μη μηδενική μεταβλητότητα μέσα στην ομάδα.

Μη-τυποποιημένοι συντελεστές (Unstandardized Coefficients)

Οι μη-τυποποιημένοι συντελεστές υπολογίζονται από τους τυποποιημένους χρησιμοποιώντας τη σχέση

$$B = \sqrt{(n-g)} S_{11}^{-1} D .$$

Οι σχετικές σταθερές δίνονται από τον εξής τύπο:

$$a_k = -\sum_{i=1}^{q} b_{il} \bar{x}_{i\bullet}$$

Οι κεντροειδείς των ομάδων εκτιμούνται στους μέσους κάθε ομάδας ως εξής:

$$\bar{f}_{kj} = a_k + \sum_{i=1}^{q} b_{ik} \bar{x}_{ij}$$

Έλεγχος ισότητας διασποράς

Το μέγεθος Box's M χρησιμοποιείται για να ελεγχθεί η ισότητα των πινάκων για την διασπορά των ομάδων και ορίζεται ως εξής:

$$M = (n-g)\log\left|C'\right| - \sum_{j=1}^{g} (n_j - 1)\log\left|C^{(j)}\right|$$

όπου

C'=συγκεντρωτικός πίνακας διασποράς αφαιρώντας τις ομάδες με μοναδιαίο πίνακα διασποράς

$C^{(j)}$ =πίνακας διασποράς για την ομάδα j

$$\log\left|C^{(j)}\right| = 2\sum_{i=1}^{p} \log l_{ii} - p\log(n_j - 1)$$

όπου l_{ii} είναι η i-διαγώνια εισαγωγή του L έτσι ώστε $(n_j - 1)C^{(j)} = L'L$. Ομοίως,

$$\log\left|C'\right| = 2\sum_{i=1}^{p} \log l_{ii} - p\log(n' - g)$$

όπου

$(n' - g)C' = L'L$ και

n΄ = άθροισμα των βαρών των περιπτώσεων σε όλες τις ομάδες μη αμελητέους πίνακες διασποράς.

Τα $C΄$ και $C^{(j)}$ προκύπτουν από την αποσύνθεση του Cholesky. Εάν κάποιο από τα διαγώνια στοιχεία της αποσύνθεσης είναι μικρότερο από 10^{-11}, ο πίνακας θεωρείται πολύ μικρός και δεν συμπεριλαμβάνεται στην ανάλυση.

Το επίπεδο σημαντικότητας προκύπτει από την F κατανομή με t_1 και t_2 βαθμούς ελευθερίας χρησιμοποιώντας το

$$F = \begin{cases} M/b & \text{εάν } e_2 > e_1^2 \\ \dfrac{t_2 M}{t_1(b-M)} & \text{εάν } e_2 < e_1^2 \end{cases}$$

όπου

$$e_1 = \left(\sum_{j=1}^{g} \frac{1}{n_j - 1} - \frac{1}{n-g}\right)\left(\frac{2p^2+3p-1}{6(g-1)(p+1)}\right)$$

$$e_2 = \left(\sum_{j=1}^{g} \frac{1}{(n_j-1)^2} - \frac{1}{(n-g)^2}\right)\left(\frac{(p-1)(p+2)}{6(g-1)}\right)$$

$$t_1 = (g-1)p(p+1)/2$$

$$t_2 = (t_1+2)/\left|e_2 - e_1^2\right|$$

$$b = \begin{cases} \dfrac{t_1}{1-e_1-t_1/t_2} & \text{εάν } e_2 > e_1^2 \\ \dfrac{t_2}{1-e_1-2/t_2} & \text{εάν } e_2 < e_1^2 \end{cases}$$

Εάν το $e_1^2 - e_2$ είναι μηδενικό, ή πολύ μικρότερο από το e_2, το t_2 δεν μπορεί να υπολογιστεί ή δεν μπορεί να υπολογιστεί με ακρίβεια. Εάν

$$e_2 = e_2 + 0.0001(e_2 - e_1^2)$$

χρησιμοποιείται το στατιστικό χ^2 του Bartlett και όχι το στατιστικό F με

$$\chi^2 = \mathrm{M}(1-e_1)$$

με t_1 βαθμούς ελευθερίας.

Για τον έλεγχο του ομαδικού πίνακα διασποράς των κανονικών διαχωριστικών συναρτήσεων, η διαδικασία είναι όμοια. Οι πίνακες $C΄$ και $C^{(j)}$ αντικαθίστανται από τους D_j και $D΄$, όπου

$$D_j = B' C^{(j)} B$$

είναι ο ομαδικός πίνακας διασποράς των διαχωριστικών συναρτήσεων.

7.6 Το παραγόμενο μοντέλο (Generated model)

Η βασική διαδικασία για την ταξινόμηση μιας περίπτωσης είναι ως εξής:

- Εάν **X** είναι ένα *1xq* διάνυσμα των διαχωριστικών μεταβλητών για μια περίπτωση, το *1xm* διάνυσμα των τιμών της κανονικής διαχωριστικής ανάλυσης είναι
  f=XB+a

- Η απόσταση chi-square από κάθε κεντροειδή υπολογίζεται ως

$$\chi_j^2 = (f - \bar{f}_j) D_j^{-1} (f - \bar{f}_j)'$$

όπου *Dj* είναι ο πίνακας διασποράς των κανονικών διαχωριστικών συναρτήσεων για μία ομάδα *j* και $\bar{f}_j$ είναι το διάνυσμα του κεντροειδή της ομάδας. Εάν η περίπτωση είναι μέλος της ομάδας *j*, το χ_j^2 ακολουθεί χ^2 κατανομή με *m* βαθμούς ελευθερίας και το $P(X|G)$ είναι το επίπεδο σημαντικότητας μιας χ_j^2.

- Ο κάνονας ταξινόμησης δίνεται από την πιθανότητα:

$$P\langle G_j \,|\, X\rangle = \frac{p_j |D_j|^{-1/2} e^{-\chi_i^2/2}}{\sum_{j=1}^{g} P_j |D_j|^{-1/2} e^{-\chi_\xi^2/2}}$$

όπου p_j είναι η εκ των προτέρων πιθανότητα για την ομάδα *j*. Κάθε περίπτωση ταξινομείται στην ομάδα για την οποία το μέγεθος $P\langle G_j \,|\, X\rangle$ είναι μέγιστο.

Ο πραγματικός υπολογισμός του $P\langle G_j \,|\, X\rangle$ είναι ο εξής:

$g_j = \log P_j - \frac{1}{2}(\log|D_j| + \chi_j^2)$ και

$$P\langle G_j \,|\, X\rangle = \begin{cases} \dfrac{\exp(g_j - \max_j g_j)}{\sum_{j=1}^{g} \exp(g_j - \max_j g_j)} & \text{εάν } g_j - \max_j g_j > -46 \\ 0 & \text{διαφορετικά} \end{cases}$$

Εάν οι ατομικοί πίνακες διασποράς δεν χρησιμοποιούνται στην ταξινόμηση, ο συγκεντρωτικός πίνακας διασποράς των διαχωριστικών συναρτήσεων αντικαθίσταται για το D_j στον παραπάνω υπολογισμό. Αυτό έχει ως αποτέλεσμα την απλοποίηση των υπολογισμών.

Εάν κάποιος από τους D_j είναι ιδιόμορφος (singular), ένας ψευτό-αντίστροφος πίνακας της μορφής

$$\begin{bmatrix} D_{j11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

αντικαθιστά τον D_j και το $|D_{j11}|$ αντικαθιστά το $|D_j|$. Ο D_{j11} είναι υποπίνακας του D_j του

οποίου οι σειρές και οι στήλες αντιστοιχίζονται στις συναρτήσεις που δεν εξαρτώνται από τις συναρτήσεις που προηγήθηκαν. Δηλαδή, η συνάρτηση 1 θα εξαχθεί μόνο εάν η τάξη του D_j=0, η συνάρτηση 2 θα διεξαχθεί μόνο εάν εξαρτάται από την συνάρτηση 1 και κοκ. Η επιλογή του ψευδο-αντίστροφου δεν είναι βέλτιστη για την αριθμητική σταθερότητα του D_{j11}^{-1}, αλλά μεγιστοποιεί την διαχωριστική δύναμη των εναπομείναντων συναρτήσεων.

7.7 Επικύρωση (cross-validation)

Για την ανάλυση που ακολυθεί χρησιμοποιούνται οι κάτωθι συμβολισμοί:

$\vec{X}_{jk}$ $(X_{1jk},.....,X_{qjk})^T$

$\vec{M}_j$ Δειγματικός μέσος για την *j*-οστή ομάδα

$$\vec{M}_j = \frac{1}{n_j}\sum_{k=1}^{m_j} f_{jk}\vec{X}_{jk}$$

$\vec{M}_{jk}$ Δειγματικός μέσος για την *j*-οστή ομάδα εξαιρώντας το σημείο $\vec{X}_{jk}$

Σ Συγκεντρωτικός δειγματικός πίνακας διασποράς

Σ_j Δειγματικός πίνακας διασποράς της *j*-οστής ομάδας

Σ_{jk} Συγκεντρωτικός δειγματικός πίνακας διασποράς χωρίς το σημείο $\vec{X}_{jk}$

$d_0^2(\vec{a},\vec{b})$ $(\vec{a}-\vec{b})^T\Sigma_{jk}^{-1}(\vec{a}-\vec{b})^T$

Η επικύρωση εφαρμόζεται μόνο στη γραμμική διαχωριστική ανάλυση. Κάθε περίπτωση, έστω η $\vec{X}_{jk}$, αφαιρείται μια φορά και λειτουργεί ως δεδομένο εξέτασης. Οι περιπτώσεις που απομένουν λειτουργούν ως ένα νέο σύνολο δεδομένων.
Έπειτα, υπολογίζονται οι ποσότητες $d_0^2(\vec{X}_{jk},\vec{M}_{jk})$ και $d_0^2(\vec{X}_{jk},\vec{M}_i)$ $(i=1,....,g.i\neq j)$. Εάν υπάρχει κάποιο $i(i\neq j)$ τέτοιο ώστε $(\log(P_i)-d_0^2(\vec{X}_{jk},M_i)/2>\log(P_j)-d_0^2(\vec{X}_{jk},\vec{M}_{jk})/2)$, η περίπτωση που αφαιρέθηκε, η $\vec{X}_{jk}$, είναι λανθασμένα ταξινομημένη. Η εκτίμηση του προβλεπόμενου ρυθμού σφάλματος είναι το κλάσμα του αθροίσματος των βαρών της λανθασμένα ταξινομημένης περίπτωσης προς το άθροισμα των βαρών όλων των περιπτώσεων.
Για να μειωθεί ο χρόνος υπολογισμού, χρησιμοποιείται η γραμμική διαχωριστική μέθοδος στη θέση της κανονικής διαχωριστικής μεθόδου. Η θεωρητική λύση παραμένει ίδια και για τις δύο μεθόδους.

7.8 Εφαρμογή στο SPSS Clementine

Ταξινόμηση πελατών σύμφωνα με την χρήση των τηλεπικοινωνιών

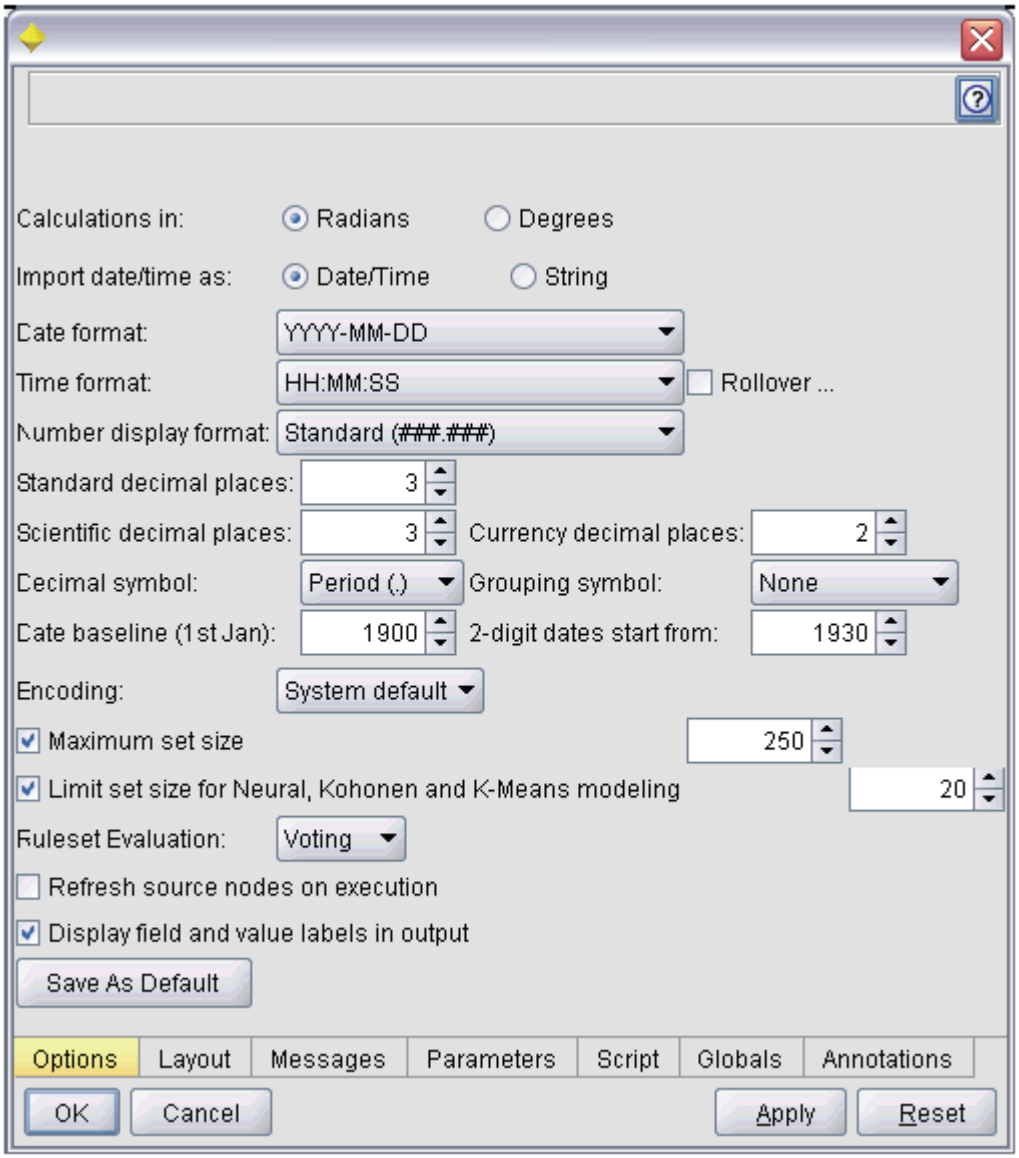
Στην εφαρμογή που ακολουθεί υποθέτουμε ότι έχουμε έναν παροχέα τηλεπικοινωνιών ο οποίος χωρίζει τους πελάτες του σύμφωνα με τα πρότυπα χρήσης των υπηρεσιών και κατηγοροποιεί τους πελάτες σε τέσσερις ομάδες. Εάν χρησιμοποιηθούν δημογραφικά δεδομένα για να προβλεφθεί η ιδιότητα μέλων των ομάδων, μπορεί να εξατομικευθούν προσφορές για μελλοντικούς πελάτες.
Το stream που χρησιμοποιείται ονομάζεται telco_cutstcat_discriminant.str. Η μεταβλητή cutstat μπορεί να πάρει τέσσερις πιθανές τιμές, οι οποίες ανταποκρίνονται στις τέσσερις ομάδες των πελατών.

Τιμή	**Χαρακτηρισμός**
1	Βασική υπηρεσία (Basic service)
2	E-service
3	Πρόσθετη υπηρεσία (Plus service)
4	Συνολική υπηρεσία (Total service)

Δημιουργώντας το stream

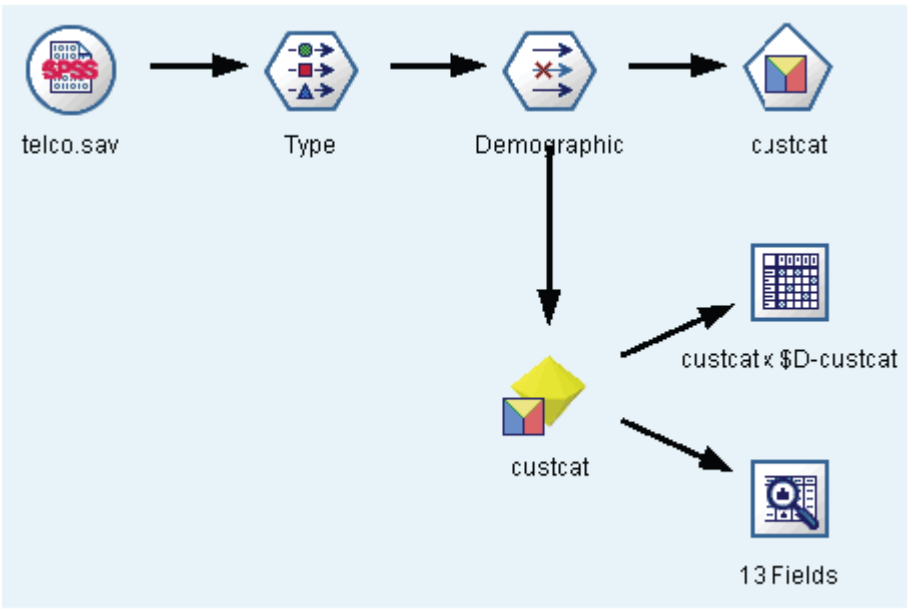
- Αρχικά, ορίζουμε τις ιδιότητες του stream για να φανούν οι μεταβλητές και οι τιμές των χαρακτηριστικών στην έξοδο. Από το μενού επιλέγουμε File >> Stream Properties.

- Επιλέγουμε το πεδίο Display και στην έξοδο τις τιμές των χαρακτηριστικών και πατάμε OK.



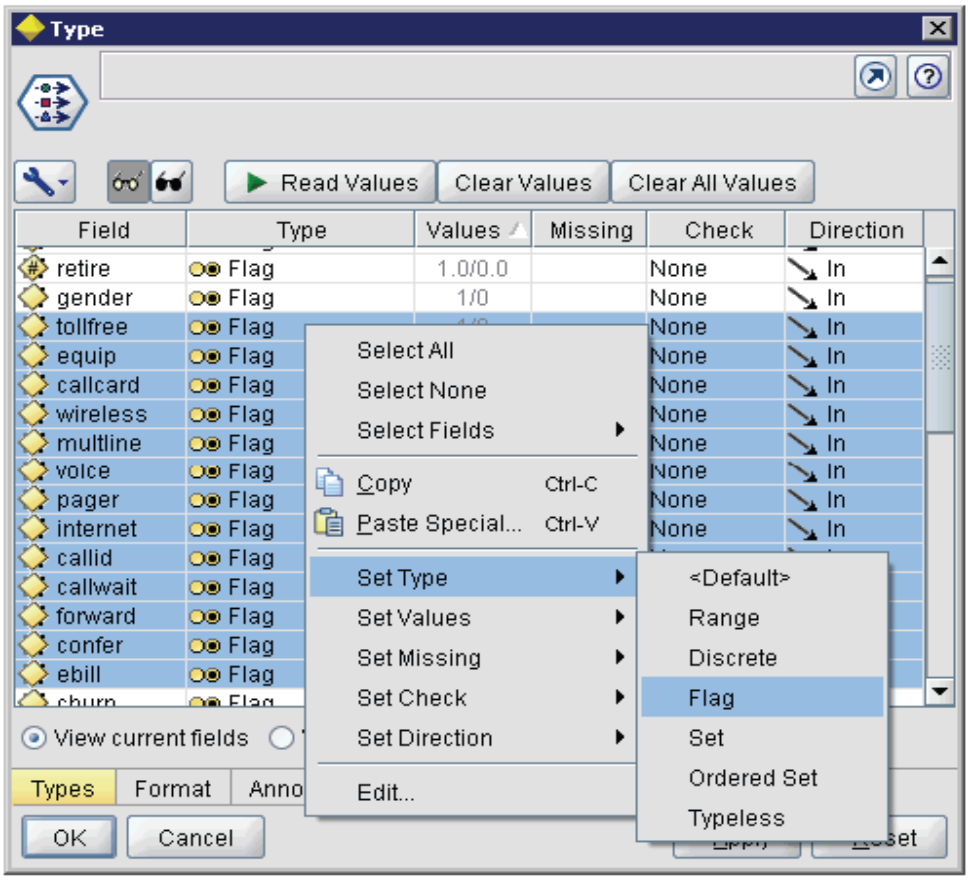
Εικόνα 7.8.1: Ιδιότητες Stream

- Προσθέτουμε ένα SPSS αρχείο με τον κόμβο της πηγής στο telco.sav.



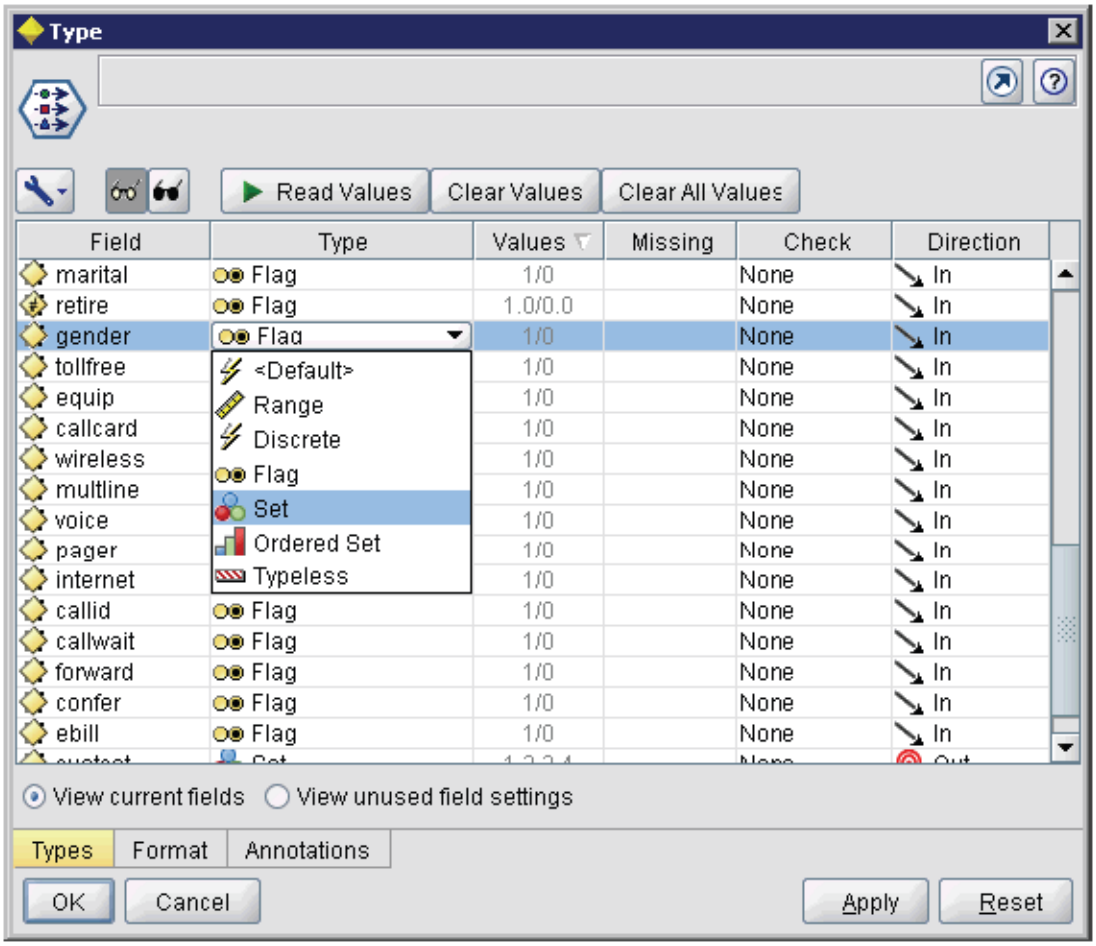
Εικόνα 7.8.2: Δειγματικό stream για να ταξινομηθούν οι πελάτες χρησιμοποιώντας την διαχωριστική ανάλυση

- Προσθέτουμε έναν κόμβο Type για να ορίσουμε τα πεδία, βεβαιώνοντας ότι όλοι οι τύποι είναι σωστά τοποθετημένοι. Για παράδειγμα, τα περισσότερα πεδία με τιμές 0 και 1 μπορούν να θεωρηθούν ως flags, αλλά κάποια πεδία, όπως για παράδειγμα το φύλο, λαμβάνουν δύο τιμές.



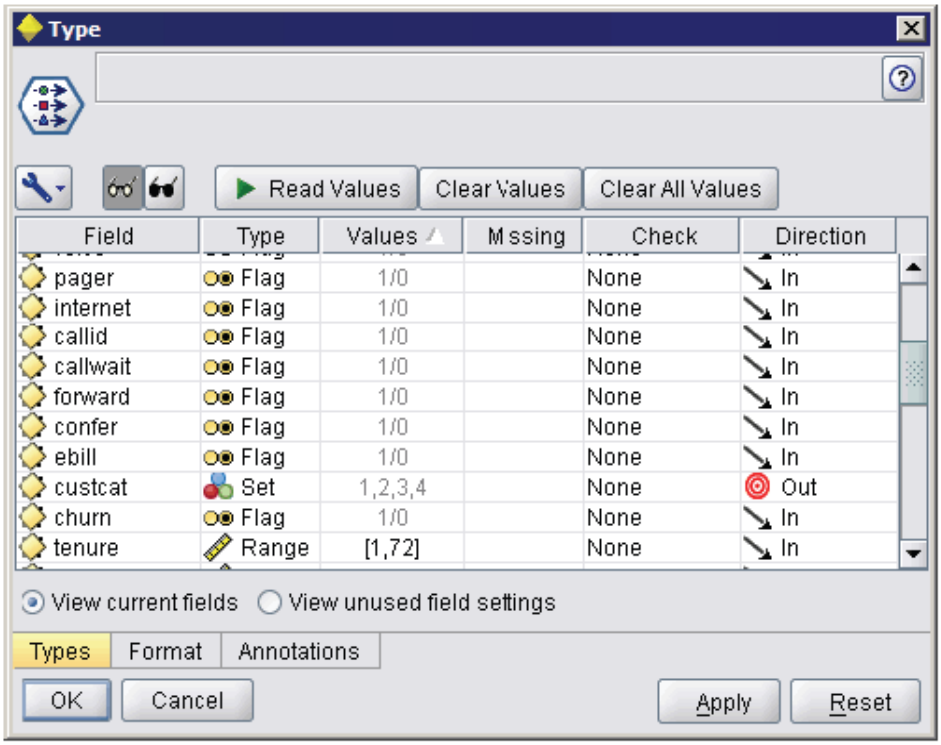
Εικόνα 7.8.3: Ορισμός του τύπου για πολλαπλά πεδία

Σημείωση: Για να αλλάξουμε τις ιδιοτήτητες για πολλαπλά πεδίο με ίδιες τιμές (όπως για παράδειγμα 0/1), κάνουμε κλικ στη στήλη Values και πατώντας το shift επιλέγουμε όλα τα πεδία που θέλουμε να αλλάξουμε. Με δεξί κλικ αλλάζουμε τον τύπο στα επιλεγμένα πεδία.
Καθώς το φύλο (gender) δεν θεωρείται flag αλλά set, επιλέγοντας την τιμή Type την αλλάζουμε σε Set.



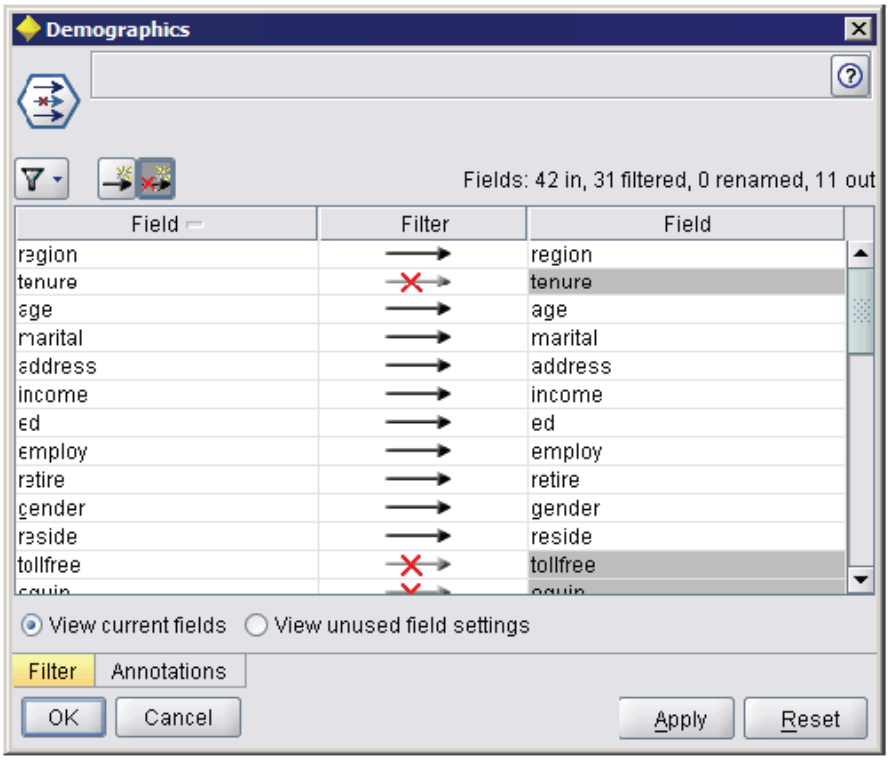
Εικόνα 7.8.4: Αλλαγή του τύπου του φύλου από flag σε set

- Ορίζουμε την κατεύθυνση για το πεδίο custcat ως Out. Όλα τα υπόλοιπα πεδία θα πρέπει να έχουν κατεύθυνση In.



Εικόνα 7.8.5: Ορισμός της κατεύθυνσης των πεδίων

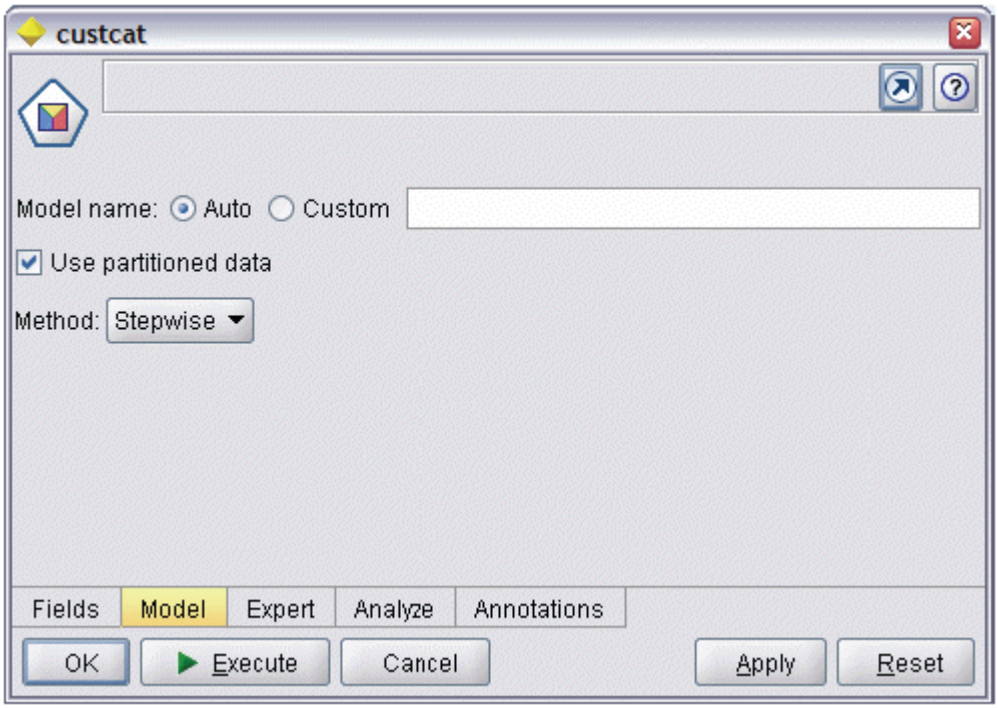
Καθώς αυτό το παράδειγμα επικεντρώνεται σε δημογραφικά δεδομένα, χρησιμοποιούμε έναν κόμβο Filter για να συμπεριλάβουμε μόνο τα σχετικά πεδία (region, age, marital, address, income, ed, employ, retire, gender, reside και custcat). Τα υπόλοιπα πεδία μπορούν να εξαιρεθούν από τον σκοπό της ανάλυσης.



Εικόνα 7.8.6: Φιλτράρισμα στα δημογραφικά πεδία

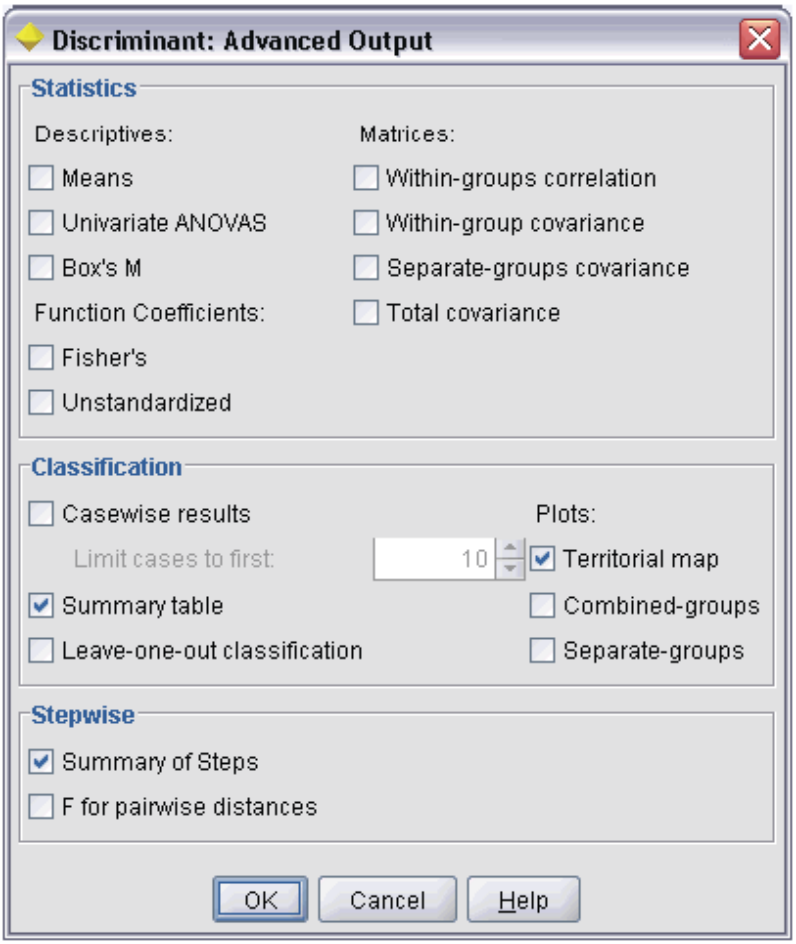
(Εναλλακτικά, θα μπορούσαμε αντί να τα εξαιρέσουμε αυτά τα πεδία να ορίσουμε ως κατεύθυνση None)

- Στον διαχωριστικό κόμβο, κάνουμε κλικ στο πεδίο Mode και επιλέγουμε Stepwise method.



Εικόνα 7.8.7: Επιλογές του μοντέλου

- Στο πεδίο Expert, θέτουμε τον κόμβο στο Expert και κάνουμε κλικ στο Output
- Στις επιλογές του Advanced Output επιλέγουμε τα Summary table, Territorial map και Summary of Steps.

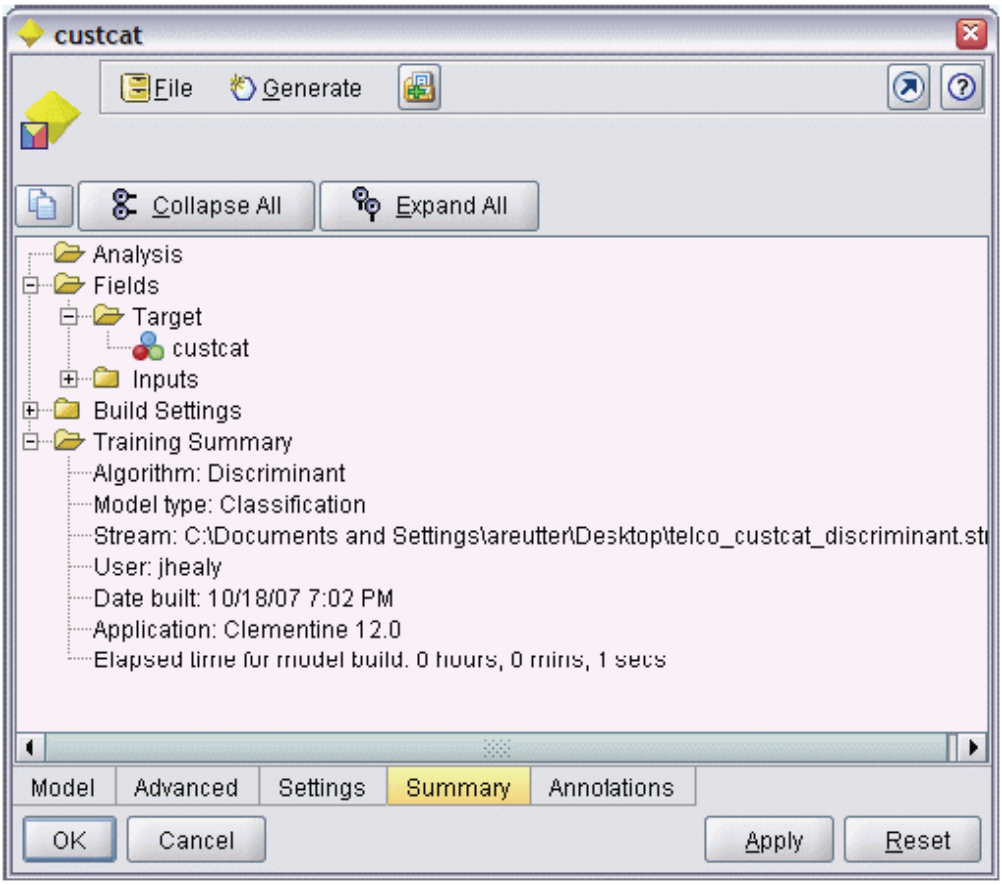


Εικόνα 7.8.8: Επιλογές για τα αποτελέσματα στην έξοδο

Έλεγχος του παραγόμενου μοντέλου

- Εκτελούμε τον κόμβο για να παράγουμε το μοντέλο, το οποίο προστίθεται στο Models. Για να δούμε λεπτομερώς το μοντέλο, κάνουμε δεξί κλικ στο παραγόμενο κόμβο του μοντέλου και επιλέγουμε Browse.

Το πεδίο Summary δείχνει τον στόχο και την συμπληρωμένη λίστα με τα δεδομένα (πεδία πρόβλεψης) που υποβλήθηκαν.



Εικόνα 7.8.9: Σύνοψη του μοντέλου με τον στόχο και τα πεδία εισαγωγής

Για λεπτομέρειες των αποτελεσμάτων της διαχωριστικής ανάλυσης, επιλέγουμε το πεδίο Advanced.

Ανάλυση με μπρος-πίσω βήματα (Stepwise discriminant analysis)

Βήμα 0	Ανέχεια	Ελάχιστη Ανέχεια	F to enter	Wilks' Lambda
Ηλικία (σε χρόνια)	1.000	1.000	7.521	.978
Οικογενειακή κατάσταση	1.000	1.000	3.500	.990
Χρόνια στην τωρινή διεύθυνση	1.000	1.000	8.433	.975
Εισόδημα νοικοκυριού (σε χιλιάδες)	1.000	1.000	6.689	.980
Επίπεδο εκπαίδευσης	1.000	1.000	61.454	.844
Συνταξιοδοτημένος	1.000	1.000	3.005	.991
Χρόνια με τον τωρινό εργοδότη	1.000	1.000	16.976	.951
Φύλο	1.000	1.000	.373	.999
Αριθμός ατόμων στο νοικοκυριό	1.000	1.000	3.976	.988

Εικόνα 7.8.10: Μεταβλητές που δεν συμπεριλαμβάνονται στην ανάλυση, βήμα 0

Όταν έχουμε πολλές μεταβλητές πρόβλεψης, η μέθοδος με μπρος-πίσω βήματα (stepwise method) μπορεί να είναι αποτελεσματική καθώς επιλέγει αυτόματα τις 'καλύτερες' μεταβλητές για το μοντέλο. Η μέθοδος με μπρος-πίσω βήματα ξεκινάει με ένα μοντέλο που δεν περιλαμβάνει καμία από τις μεταβλητές πρόβλεψης. Σε κάθε βήμα, η μεταβλητή πρόβλεψης με το μεγαλύτερο F-to-

enter που υπερβαίνει το κριτήριο εισαγωγής (εξ'ορισμού F=3.84) προστίθεται στο μοντέλο.

Βήμα 3	Ανέχεια	Ελάχιστη Ανέχεια	F to enter	Wilks' Lambda
Ηλικία (σε χρόνια)	.535	.535	.252	.795
Οικογενειακή κατάσταση	.605	.593	1.507	.792
Χρόνια στην τωρινή διεύθυνση	.776	.771	3.514	.787
Εισόδημα νοικοκυριού (σε χιλιάδες)	.688	.657	.687	.794
Συνταξιοδοτημένος	.917	.880	.353	.795
Φύλο	.997	.931	.395	.795

Εικόνα 7.8.11: Μεταβλητές που δεν συμπεριλαμβάνονται στην ανάλυση, βήμα 3

Οι μεταβλητές που στο τελευταίο βήμα έχουν F-to-enter τιμή μικρότερη από 3.84 δεν συμπεριλαμβάνονται στην ανάλυση.

Βήμα		Ανέχεια	F to remove	Wilks' Lambda
1	Επίπεδο εκπαίδευσης	1.000	61.454	
2	Επίπεδο εκπαίδευσης	.953	59.108	.951
	Χρόνια με τον τωρινό εργοδότη	.953	14.933	.844
3	Επίπεδο εκπαίδευσης	.951	60.046	.940
	Χρόνια με τον τωρινό εργοδότη	.934	15.824	.834
	Αριθμός ατόμων στο νοικοκυριό	.979	4.841	.807

Εικόνα 7.8.12: Μεταβλητές στην ανάλυση

Ο πίνακας 7.8.12 δείχνει τα στατιστικά μεγέθη για τις μεταβλητές οι οποίες προστέθηκαν στην ανάλυση σε κάθε βήμα. Η ανέχεια είναι η αναλογία της διασποράς μιας μεταβλητής που δεν επεξηγείται από τις άλλες ανεξάρτητες μεταβλητές στη διαχωριστική εξίσωση. Μια μεταβλητή με πολύ μικρή ανέχεια μπορεί να δώσει λίγη πληροφορία στο μοντέλο και μπορεί να προκαλέσει υπολογιστικά προβλήματα.
Οι τιμές F-to-remove είναι χρήσιμες για να περιγράψουν τι θα συμβεί εάν μια μεταβλητή αφαιρεθεί από το παρόν μοντέλο (δεδομένου ότι οι υπόλοιπες μεταβλητές θα παραμείνουν στο μοντέλο).

Σημείωση για τις μεθόδους με μπρος-πίσω βήματα

Οι μέθοδοι με μπρος-πίσω βήματα είναι πολύ άνετες στην χρήση αλλά έχουν και κάποιους περιορισμούς. Καθώς οι μέθοδοι με μπρος-πίσω βήματα επιλέγουν μοντέλα αποκλειστικά και μόνο σύμφωνα με την στατιστική αξία των μεταβλητών, μπορεί να επιλεχθούν μεταβλητές πρόβλεψης οι οποίες δεν είναι πρακτικά σημαντικές. Εάν έχουμε κάποια εμπειρία με τα δεδομένα και περιμένουμε ποιες μεταβλητές θα είναι σημαντικές, θα πρέπει να χρησιμοποιήσουμε αυτή την πληροφορία και να αποφύγουμε να χρησιμοποιήσουμε τις μεθόδους με μπρος-πίσω βήματα. Εάν, ωστόσο, έχουμε πολλές μεταβλητές πρόβλεψης και δεν γνωρίζουμε από που να ξεκινήσουμε, μπορούμε να εφαρμόσουμε την μπρος-πίσω μέθοδο και να προσαρμόσουμε το επιλεγμένο μοντέλο.

Έλεγχος του μοντέλου

Συνάρτηση	Ιδιοτιμή	% Διακύμανση	Συσσώρευση %	Κανονική συσχέτιση
1	.198	.802	80.2	.407
2	.048	19.4	99.6	.214
3	.001	.4	100.0	.031

Εικόνα 7.8.13: Ιδιοτιμές

Σχεδόν όλη η διασπορά που εξηγείται από το μοντέλο οφείλεται στις δύο πρώτες διαχωριστικές συναρτήσεις. Συνολικά προσαρμόζονται τρεις συναρτήσεις, αλλά καθώς η τρίτη εξίσωση έχει μικρή ιδιοτιμή, μπορεί να παραληφθεί.

Έλεγχος συνάρτησης	Wilks' lambda	Chi-square	β.ε.	p-τιμή
1 μέσω 3	.796	227.345	9	.000
2 μέσω 3	.953	47.486	4	.000
3	.999	.929	1	.335

Εικόνα 7.814: Wilks' lambda

Το Wilks' lambda επιβεβαιώνει ότι μόνο οι δύο πρώτες συναρτήσεις είναι χρήσιμες. Ο έλεγχος εξετάζει την υπόθεση ότι οι μέσοι των συναρτήσεων είναι ίσοι μεταξύ των ομάδων. Ο έλεγχος για την τρίτη συνάρτηση έχει p-τιμή μεγαλύτερη από 0.10. Γι'αυτό το λόγο αυτή η συνάρτηση 3 δεν συνεισφέρει στο μοντέλο.

Πίνακας δομής

	Συνάρτηση		
	1	2	3
Επίπεδο εκπαίδευσης	.966*	-.090	-.244
Χρόνια με τον τωρινό εργοδότη	-.182	.964*	-.193
Ηλικία (σε χρόνια)[a]	-.162	.598*	-.285
Εισόδημα νοικοκυριού (σε χιλιάδες)[a]	.109	.514*	-.190
Χρόνια στην τωρινή διεύθυνση[a]	-.151	.394*	-.214
Συνταξιοδοτημένος[a]	-.108	.230*	-.137
Φύλο[a]	.008	.054*	.009
Αριθμός ατόμων στο νοικοκυριό	.232	.097	.968*
Οικογενειακή κατάσταση[a]	.132	.134	.600*

*. Η μεγαλύτερη κατά απόλυτη τιμή συσχέτιση κάθε μεταβλητής με κάθε διαχωριστική συνάρτηση
a. Αυτή η μεταβλητή δεν συμπεριλαμβάνεται στην ανάλυση

Εικόνα 7.8.15: Πίνακας δομής

Όταν υπάρχουν περισσότερες από μία διαχωριστικές συναρτήσεις, το αστεράκι (*) δείχνει την μεγαλύτερη συσχέτιση σε απόλυτη τιμή κάθε μεταβλητής με μία από τις κανονικές συναρτήσεις. Σε

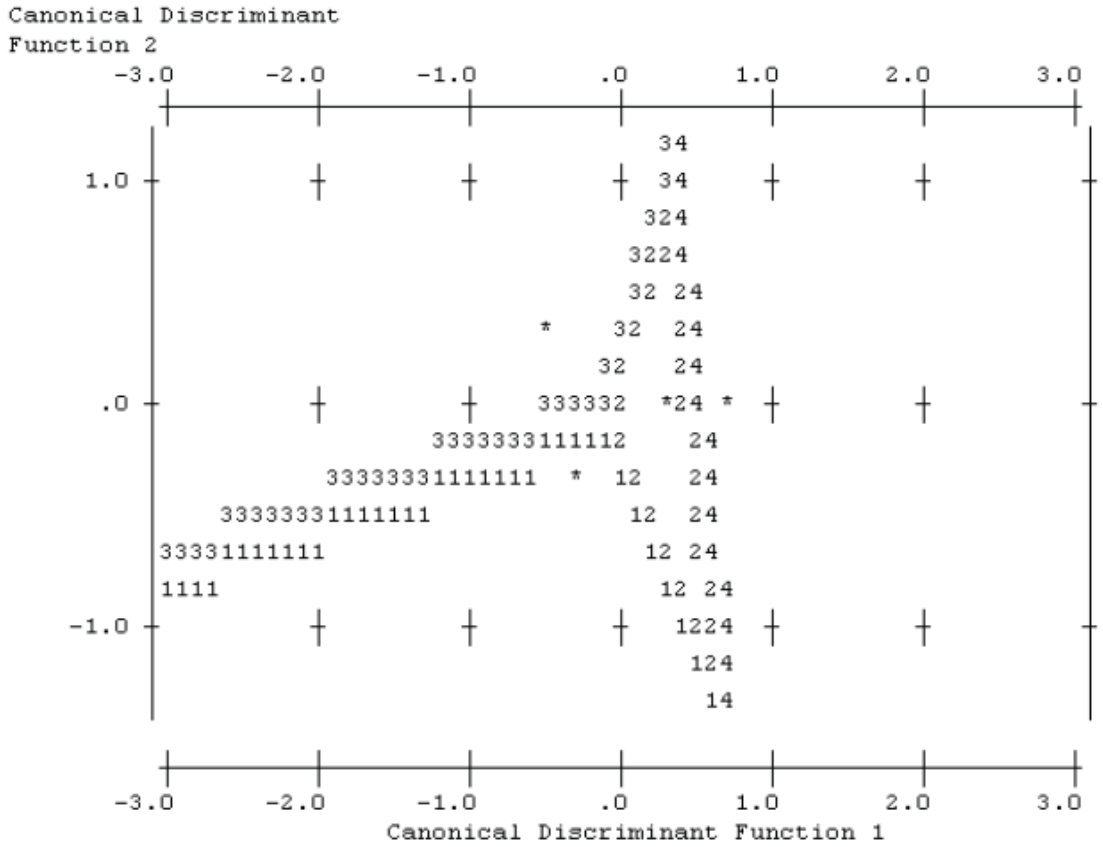
κάθε συνάρτηση, οι μαρκαρισμένες μεταβλητές κατατάσσονται σύμφωνα με το μέγεθος της συσχέτισης.

- Το ‘επίπεδο μόρφωσης’ (level of education) είναι η μόνη μεταβλητή που συσχετίζεται με την πρώτη συνάρτηση

- Τα ‘χρόνια με τον παρόν εργοδότη’ (years with current employer), η ‘ηλικία’ (age), το ‘εισόδημα σε χιλιάδες στο νοικοκυριό’ (household income in thousands), τα ‘χρόνια στην παρούσα διεύθυνση’ (years in current address), ‘το αν είναι συνταξιοδοτημένος ή όχι’ (retired), και το ‘φύλο’ (gender) συσχετίζονται με την δεύτερη συνάρτηση. Το φύλο και το αν είναι συνταξιοδοτημένος ή όχι είναι πιο αδύναμες μεταβλητές από τις υπόλοιπες.

- Ο ‘αριθμός των ανθρώπων στο νοικοκυριό’ (number of people in household) και ‘η οικογενειακή κατάσταση’ (marital status) συσχετίζονται με την τρίτη συνάρτηση, αλλά καθώς δεν χρησιμοποιείται αυτή η συνάρτηση, δεν είναι χρήσιμες ούτε αυτές οι μεταβλητές για την πρόβλεψη.

Χάρτης territorial



Εικόνα 7.8.16: Χάρτης territorial

Ο χάρτης territorial είναι χρήσιμος στη μελέτη των σχέσεων μεταξύ των ομάδων και των διαχωριστικών συναρτήσεων. Συνδυάζοντας τα αποτελέσματα του πίνακα δομής, ο χάρτης territorial δίνει μια γραφική ερμηνεία της σχέσης μεταξύ των μεταβλητών πρόβλεψης και των ομάδων. Η πρώτη συνάρτηση, χωρίζει την ομάδα 4 (Πελάτες συνολικής υπηρεσίας) από τις υπόλοιπες ομάδες. Καθώς ‘το επίπεδο μόρφωσης’ συσχετίζεται θετικά με την πρώτη συνάρτηση, οι ‘Πελάτες συνολικής υπηρεσίας’ είναι υψηλά μορφωμένοι. Η δεύτερη συνάρτηση χωρίζει τις ομάδες 1 και 3 (Πελάτες βασικής υπηρεσίας και Πελάτες πρόσθετης υπηρεσίας) από τις υπόλοιπες. Οι

‘Πελάτες πρόσθετης υπηρεσίας’ τείνουν να δουλεύουν περισσότερο και είναι μεγαλύτεροι σε ηλικία από τους ‘Πελάτες βασικής υπηρεσίας’. Οι ‘Πελάτες του e-service’ δεν διαχωρίζονται καλά από τους υπόλοιπους, αν και ο χάρτης δείχνει ότι τείνουν να είναι καλά μορφωμένοι με μία μέση προυπηρεσία.
Εφ’όσον οι κεντροειδείς των ομάδων, οι οποίοι συμβολίζονται με αστεράκι (*), είναι κοντά ο ένας με τον άλλον, δεν υπάρχει σαφής διαχωρισμός των ομάδων.
Ο χάρτης territorial δίνει μια συνοπτική εικόνα του διαχωριστικού μοντέλου, όπου σχεδιάστηκαν μόνο οι δύο πρώτες διαχωριστικές συναρτήσεις, , καθώς η τρίτη συνάρτηση δεν είναι σημαντική.

Αποτελέσματα ταξινόμησης

		Ομάδα πρόβλεψης				
	Κατηγορία πελατών	**Βασική υπηρεσία**	**e-service**	**Πρόσθετη υπηρεσία**	**Συνολική υπηρεσία**	**Σύνολο**
	Βασική υπηρεσία	125	11	61	69	266
	e-service	49	15	58	95	217
	Πρόσθετη υπηρεσία	102	14	112	53	281
	Συνολική υπηρεσία	40	16	37	143	236
%	**Βασική υπηρεσία**	47.0	4.1	22.9	25.9	100.0
	e-service	22.6	6.9	26.7	43.8	100.0
	Πρόσθετη υπηρεσία	36.3	5.0	39.9	18.9	100.0
	Συνολική υπηρεσία	16.9	6.8	15.7	60.6	100.0

a. Το 39.5% των αρχικών περιπτώσεων είναι σωστά ταξινομημένες

Εικόνα 7.8.17: Αποτελέσματα ταξινόμησης

Σύμφωνα με τα παρατηρούμενα δεδομένα, το ‘μηδενικό’ μοντέλο (δηλαδή, χωρίς μεταβλητές πρόβλεψης) θα ταξινομούσε όλους τους πελάτες στην ομάδα, ‘Πρόσθετη υπηρεσία’. Δηλαδή, το μηδενικό μοντέλο θα ταξινομούσε σωστά το 281/1000=28.1% των παρατηρήσεων. Ενώ, το μοντέλο που χρησιμοποιήσαμε το 39.9% των παρατηρήσεων είναι σωστά ταξινομημένες στην ομάδα ‘Πρόσθετη υπηρεσία’. Το μοντέλο υπερτερεί στην αναγνώριση των πελατών για την ‘Συνολική υπηρεσία’. Ωστόσο, δεν είναι καλό στην ταξινόμηση των πελατών για την ‘E-service’ και ίσως θα πρέπει να βρεθεί κάποια άλλη μεταβλητή πρόβλεψης για να διαχωριστούν αυτοί οι πελάτες.

Σύνοψη

Έχει δημιουργηθεί ένα διαχωριστικό μοντέλο το οποίο ταξινομεί τους πελάτες σε μία από τις τέσσερις ομάδες ‘χρήση υπηρεσιών’, σύμφωνα με τις δημογραφικές πληροφορίες για κάθε πελάτη. Χρησιμοποιώντας τον πίνακα δομής και τον χάρτη territorial, αναγνωρίζουμε ποιες μεταβλητές είναι χρήσιμες για τον διαχωρισμό των πελατών. Τα αποτελέσματα ταξινόμησης δείχνουν ότι το μοντέλο υστερεί στην ταξινόμηση των πελατών της υπηρεσίας ‘e-service’. Γι’αυτό το λόγο πρέπει να βρεθεί μια άλλη μεταβλητή πρόβεψης που να ταξινομεί καλύτερα αυτούς τους πελάτες. Ωστόσο, εάν δεν ενδιαφερόμαστε να αναγνωρίσουμε τους πελάτες του ‘e-service’, αυτό το μοντέλο μπορεί να θεωρηθεί αρκετά ακριβές. Εάν η μέγιστη επιστροφή της επένδυσης προέρχεται από τους πελάτες ‘Πρόσθετης υπηρεσίας’ ή ‘Συνολικής υπηρεσίας’, το μοντέλο δίνει την πληροφορία που χρειαζόμαστε.
Επίσης, πρέπει να τονιστεί ότι αυτά τα αποτελέσματα βασίζονται μόνο σε δοκιμαστικό σύνολο δεδομένων. Για να αξιολογήσουμε πόσο καλά το μοντέλο γενικεύεται και σε άλλα δεδομένα, μπορούμε να χρησιμοποιήσουμε ένα κόμβο Partition όπου αποθηκεύεται ένα υποσύνολο αποτελεσμάτων το οποίο θα χρησιμοποιηθεί για έλεγχο και επικύρωση.

ΚΕΦΑΛΑΙΟ 8^{Ο}

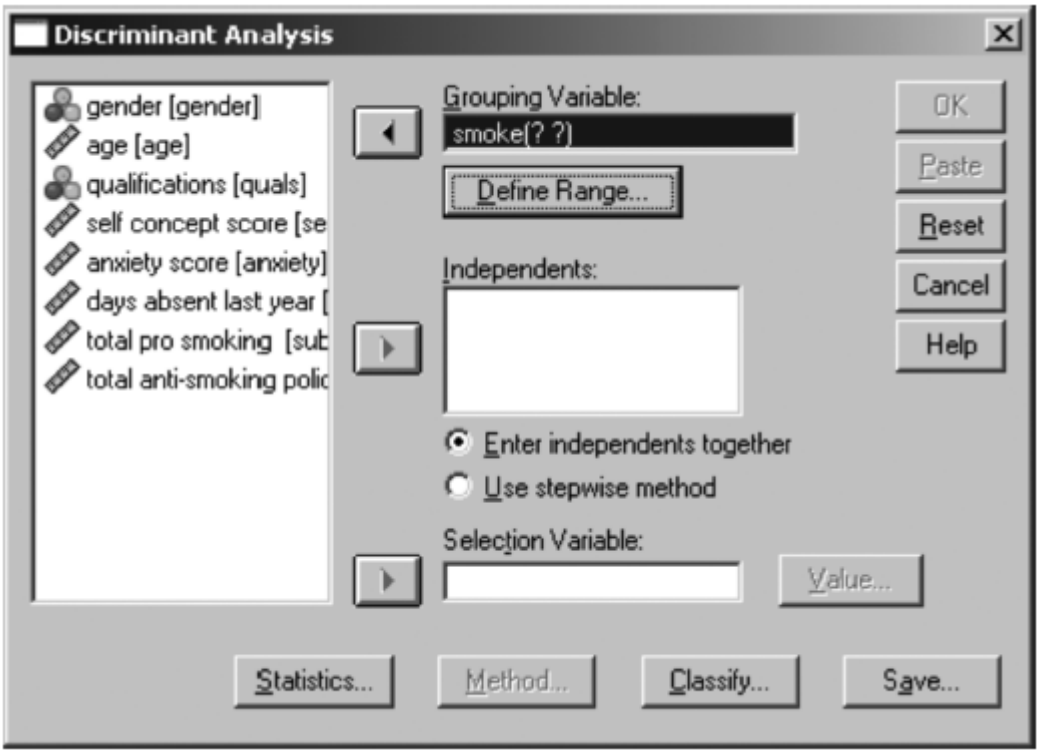
Διαχωριστική ανάλυση στο SPSS

Η ανάλυση που ακολουθεί αφορά μία εφαρμογή της Διαχωριστικής ανάλυσης στο SPSS.

Τα αντικείμενα που θέλουμε να διαχωρίσουμε είναι οι υπάλληλοι μιας εταιρείας. Η κατηγορική μεταβλητή που χρησιμοπιείται είναι το 'κάπνισμα' (smoke) που δείχνει εάν ένας υπάλληλος είναι καπνιστής ή όχι. Οι υπόλοιπες μεταβλητές είναι η ηλικία (age), οι μέρες απουσίας από την δουλειά λόγω ασθένειας τον τελευταίο χρόνο (days absent sick from work last year), ο βαθμός αυτοεκτίμησης κάθε υπαλλήλου (self-concept score), ο βαθμός άγχους (anxiety score) και ο βαθμός της αντι-καπνιστικής στάσης (attitudes to anti-smoking at work score) στον χώρο εργασίας.
Ο στόχος της ανάλυσης είναι να καθοριστεί εάν αυτές οι μεταβλητές μπορούν να διαχωρίσουν τους καπνιστές από τους μη-καπνιστές. Συνεπώς, ακολουθεί μια απλή διαχωριστική ανάλυση με δύο ομάδες.

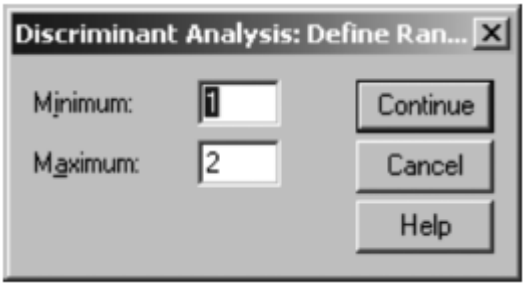
Για να ξεκινήσουμε την ανάλυση στο SPSS επιλέγουμε

1. **Analyse >> Classify >> Discriminant.**
2. Ως κατηγορική μεταβλητή επιλέγουμε την μεταβλητή 'smoke' και την τοποθετούμε στο **Grouping Variable Box** (Εικόνα 8.1).



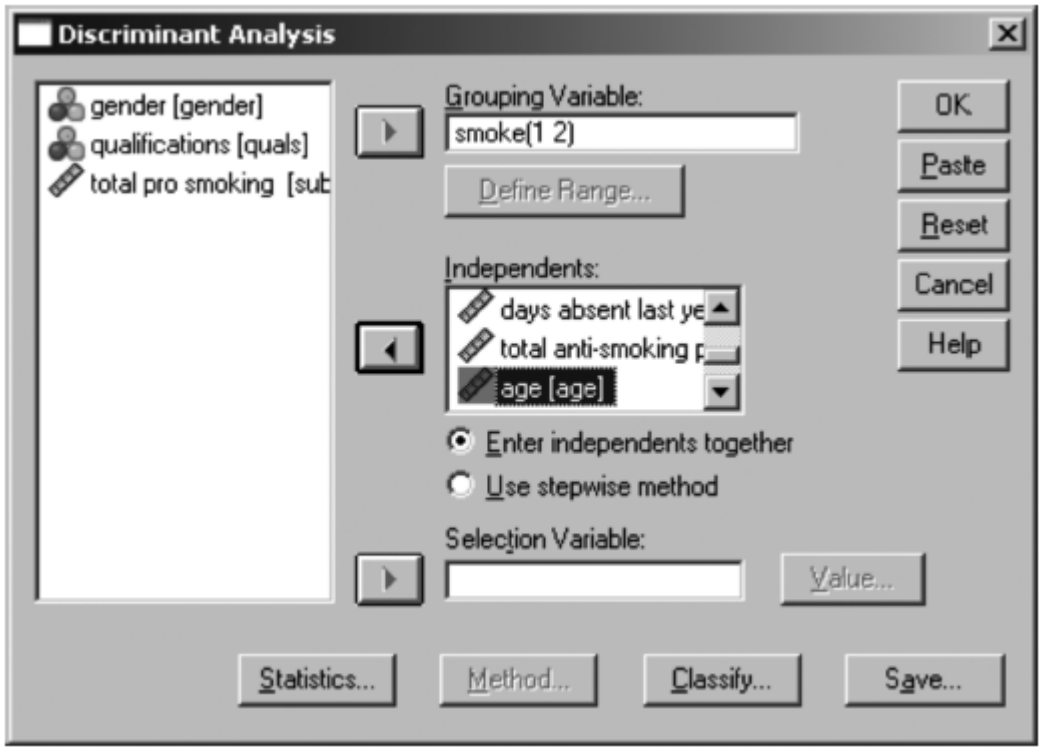
Εικόνα 8.1: Παράθυρο διαχωριστικής ανάλυσης

3. Επιλέγουμε **Define Range** και εισάγουμε το ελάχιστο και το μέγιστο αριθμό ομάδων (στην περίπτωση μας είναι 1 και 2) (Εικόνα 8.2).
4. Επιλέγουμε **Continue.**



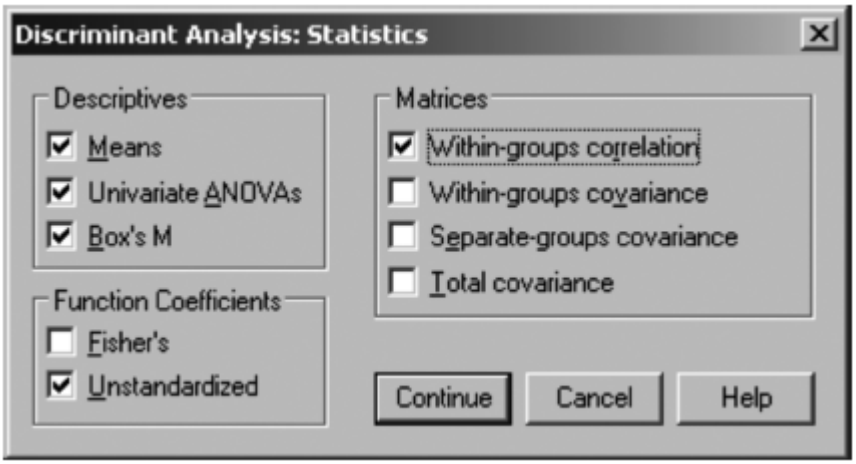
Εικόνα 8.2: Παράθυρο Define range

5. Επιλέγουμε τις μεταβλητές πρόβλεψης και τις εισάγουμε στην παράθυρο **Independents** (Εικόνα 8.3) και επιλέγουμε **Enter Independents Together**. Εάν επιλέξουμε να κάνουμε ανάλυση με μπροπίσω βήματα (sterpwise analysis) μπορούμε αντί για Enter Independents Together να επιλέξουμε **Use Stepwise Method**.



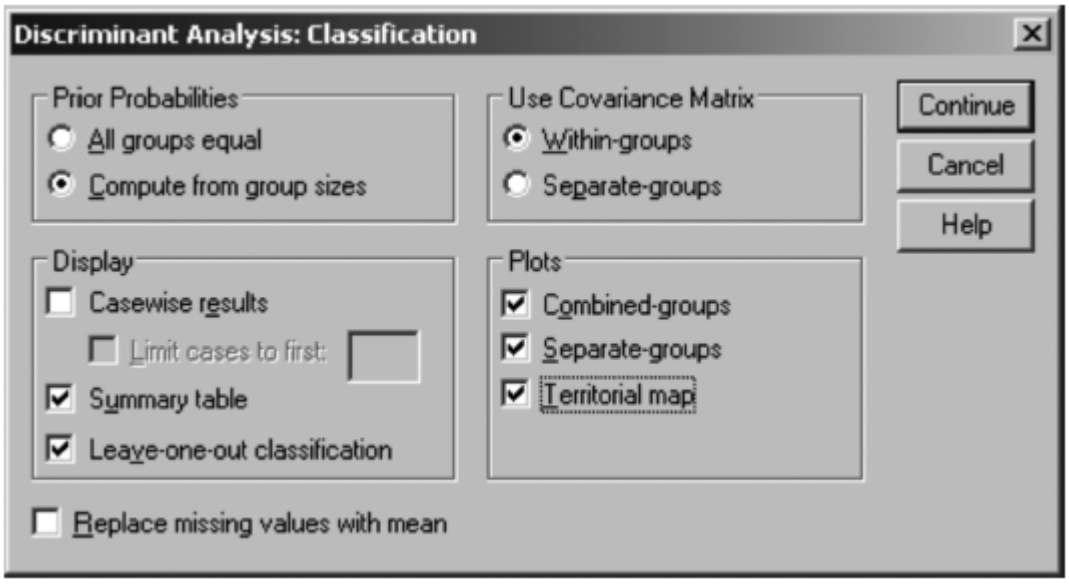
Εικόνα 8.3: Παράθυρο διαχωριστικής ανάλυσης

6. Στο **Statistics** επιλέγουμε τις ποσότητες **Means, Univariate Anovas, Box's M, Unstandardized** και **Within-Groups Correlation** (Εικόνα 8.4).



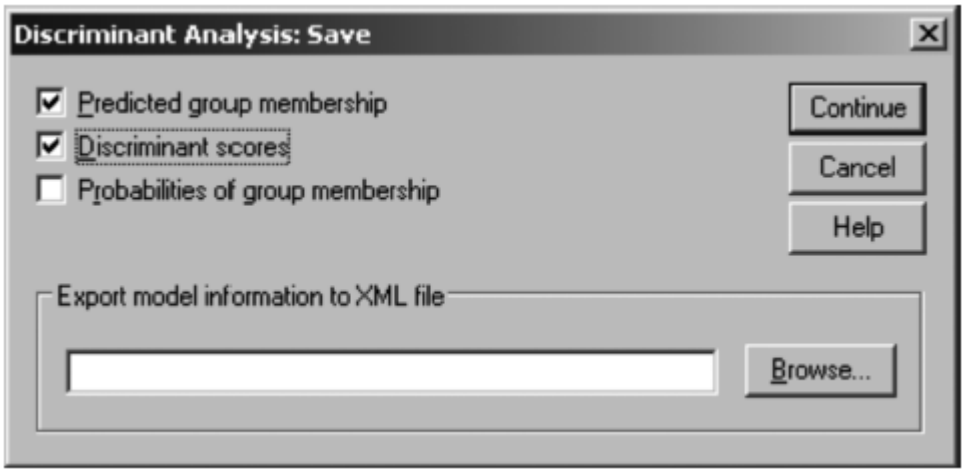
Εικόνα 8.4: Παράθυρο στατιστικών διαχωριστικής ανάλυσης

7. **Continue** >> **Classify**, επιλέγουμε **Compute From Group Sizes, Summary Table, Leave One Out Classification, Within Groups**, και όλα τα **Plots** (Εικόνα 8.5).



Εικόνα 8.5: Παράθυρο ταξινόμησης διαχωριστικής ανάλυσης

8. **Continue** >> **Save** και επιλέγουμε **Predict Group Membership** και **Discriminant Scores** (Εικόνα 8.6).
9. **Ok.**



Εικόνα 8.6 Παράθυρο αποθήκευσης διαχωριστικής ανάλυσης

Ερμηνεία αποτελεσμάτων

Πίνακες στατιστικών των ομάδων (Group statistics tables)

Στην διαχωριστική ανάλυση προσπαθούμε να προβλέψουμε την ιδιότητα μέλους μιας ομάδας, και γι'αυτό πρώτα εξετάζουμε εάν υπάρχουν σημαντικές διαφορές μεταξύ των ομάδων για κάθε μια από τις ανεξάρτητες μεταβλητές χρησιμοποιώντας τον μέσο κάθε ομάδας και τα αποτελέσματα του πίνακα ANOVA. Εάν δεν υπάρχουν στατιστικά σημαντικές διαφορές μεταξύ των ομάδων δεν είναι σημαντικό να προχωρήσουμε περαιτέρω την ανάλυση. Για να δούμε το πως συμπεριφέρονται οι μεταβλητές μπορούμε να μελετήσουμε τον μέσο κάθε ομάδας και τις τυπικές τους αποκλίσεις. Για παράδειγμα, οι διαφορές μέσων μεταξύ των βαθμών που αφορούν την αυτοεκτίμηση και των βαθμών που αφορούν το άγχος αναπαρίσταται στον Πίνακα 8.1 όπου φαίνεται ότι είναι καλοί διαχωριστές.

Στατιστικά ομάδων					
Καπνιστής ή όχι		**Μέσος**	**Τυπική απόκλιση**	**Χωρίς βάρος**	**Με βάρος**
Μη-καπνιστές	Ηλικία	38.7665	9.23647	257	257.00
	Βαθμός αυτοεκτίμησης	46.6148	11.16826	257	257.00
	Βαθμός άγχους	19.6848	5.23565	257	257.00
	Ημέρες απουσίας τελευταίου έτους	4.8482	5.39643	257	257.00
	Συνολική αντι-καπνιστική πολιτική	22.6770	2.56036	257	257.00
Καπνιστές	Ηλικία	36.1934	8.52325	181	181.00
	Βαθμός αυτοεκτίμησης	28.2818	6.54159	181	181.00
	Βαθμός άγχους	28.5028	7.25153	181	181.00
	Ημέρες απουσίας τελευταίου έτους	8.3481	7.53107	181	181.00
	Συνολική αντι-καπνιστική πολιτική	20.6409	3.15670	181	181.00
Σύνολο	Ηλικία	37.7032	9.02823	438	438.00
	Βαθμός αυτοεκτίμησης	39.0388	13.12921	438	438.00
	Βαθμός άγχους	23.3288	7.52428	438	438.00
	Ημέρες απουσίας τελευταίου έτους	6.2945	6.58773	438	438.00
	Συνολική αντι-καπνιστική πολιτική	21.8356	2.99204	438	438.00

Πίνακας 8.1: Στατιστικά ομάδων

Ο Πίνακας 8.2 αποδεικνύει ότι είναι στατιστικά σημαντικές οι διαφορές μεταξύ των μέσων για τις ομάδες των καπνιστών και μη-καπνιστών για όλες τις ανεξάρτητες μεταβλητές, καθώς και ότι οι βαθμοί της αυτοεκτίμησης και του άγχους δίνουν μεγάλες F-τιμές.

Έλεγχος ισότητας των μέσων των ομάδων				
	Wilks' lambda	F-τιμή	Βαθμοί ελευθερίας	p-τιμή
Ηλικία	.980	8.781	1	.003
Βαθμός αυτοεκτίμησης	.526	392.672	1	.000
Βαθμός άγχους	.666	218.439	1	.000
Ημέρες απουσίας τελευταίου έτους	.931	32.109	1	.000
Συνολική αντι-καπνιστική πολιτική	.887	55.295	1	.000

Πίνακας 8.2: Έλεγχος ισότητας των μέσων των ομάδων

Από τους συγκεντρωτικούς πίνακες ανάμεσα στις ομάδες (Pooled Within-Group Matrices) (Πίνακας 8.3) συνίσταται η χρήση αυτών των ανεξάρτητων μεταβλητών καθώς η αυτοσυσχέτιση μεταξύ των μεταβλητών είναι χαμηλή.

Πίνακας λογαρίθμων και του ελέγχου Box's M

Στην ανάλυση ANOVA, η κύρια υπόθεση είναι ότι οι διακυμάνσεις είναι ισοδύναμες για κάθε ομάδα, ενώ στην διαχωριστική ανάλυση η βασική υπόθεση είναι ότι οι πίνακες διακύμανσης και συνδιακύμανσης είναι ισοδύναμοι. Ο έλεγχος του Box's M εξετάζει την μηδενική υπόθεση εάν οι πίνακες συνδιακύμανσης δεν διαφέρουν μεταξύ των ομάδων (between groups) που δημιουργούνται από την εξαρτημένη μεταβλητή. Αυτό που επιθυμούμε είναι να διατηρηθεί η μηδενική υπόθεση, δηλαδή να μη διαφέρουν οι ομάδες.
Για να διατηρηθεί αυτή η υπόθεση, οι λογάριθμοι των κατηγοριών της εξαρτημένης μεταβλητής πρέπει να είναι ίσοι. Εξετάζοντας το Box's M, αναζητούμε ένα μη σημαντικό M τέτοιο ώστε να αποδεικνύει ομοιότητα και έλειψη των σημαντικών διαφορών. Στην περίπτωσή μας οι λογάριθμοι είναι σχεδόν ίσοι και η τιμή του Box's M είναι 176.474 με F=11.615 το οποίο είναι σημαντικό με p< .000 (Πίνακες 8.4 και 8.5). Ωστόσο, σε μεγάλα δείγματα, ένα σταστικά σημαντικό αποτέλεσμα δεν μπορεί να θεωρηθεί καταλυτικό. Όταν υπάρουν τρεις ή περισσότερες ομάδες, και το M είναι σημαντικό, οι ομάδες με πολύ μικρό λογάριθμο θα πρέπει να παραληφθούν από την ανάλυση.

Συγκεντρωτικοί πίνακες ανάμεσα στις ομάδες						
		Ηλικία	Βαθμός αυτοεκτίμησης	Βαθμός άγχους	Ημέρες απουσίας τελεύταιου έτους	Συνολική αντικαπνιστική πολιτική
Συσχέτιση	Ηλικία	1.000	-.118	.060	.042	.061
	Βαθμός αυτοεκτίμησης	-.118	1.000	.042	-.143	-.044
	Βαθμός άγχους	.060	.042	1.000	.118	.137
	Ημέρες απουσίας τελευαίου έτους	.042	.143	.118	1.000	.116
	Συνολική αντι-καπνιστική πολιτική	.061	-.044	.137	.116	1.000

Πίνακας 8.3: Συγκεντρωτικοί πίνακες ανάμεσα στις ομάδες

Λογάριθμοι κατηγοριών		
Καπνιστής ή όχι	Τάξη	Λογάριθμος
Μη-καπνιστής	5	17.631
Καπνιστής	5	18.058
Συγκεντρωτικό ανάμεσα στις ομάδες	5	18.212

Πίνακας 8.4: Λογάριθμοι των κατηγοριών των εξαρτημένων μεταβλητών

Αποτέλεσματα εξέτασης		
Box's M		176.474
F		11.615
	Βαθμοί ελευθερίας	15
	p-τιμή	.000

Πίνακας 8.5: Αποτελέσματα του Box's M ελέγχου

Πίνακας ιδιοτιμών

Ο πίνακας ιδιοτιμών δίνει πληροφορίες για κάθε μία από τις διαχωριστικές συναρτήσεις που παράγονται. Ο μέγιστος αριθμός των διαχωριστικών συναρτήσεων που παράγονται είναι ο αριθμός των ομάδων μειωμένος κατά 1. Εδώ, χρησιμοποιούμε δύο ομάδες, "καπνιστές" και "μη καπνιστές", και γι'αυτό έχουμε μόνο μία συνάρτηση. Η κανονική συσχέτιση είναι η πολλαπλή συσχέτιση μεταξύ των μεταβλητών που χρησιμοποιούνται για την πρόβλεψη (predictors) και της διαχωριστικής συνάρτησης. Στο παράδειγμα μας (Πίνακας 8.6), η κανονική συσχέτιση είναι 0.802, το οποίο σημαίνει ότι το μοντέλο εξηγεί το 80,2% της μεταβλητότητας της εξαρτημένης μεταβλητής, δηλαδή εάν ο ερωτώμενος είναι καπνιστής ή όχι.

Ιδιοτιμές				
Συνάρτηση	Ιδιοτιμή	Ποσοστό διακύμανσης	Ποσοστό συσσώρευσης	Κανονική συσχέτιση
1	1,806	100.0	100.0	.802

Πίνακας 8.6: Πίνακας ιδιοτιμών

Wilks' lamba

Η ποσότητα Wilks' lamba δίνει την σημαντικότητα της διαχωριστικής συνάρτησης. Ο Πίνακας 8.7 αποδεικνύει την στατιστικά σημαντική συνάρτηση ($p<.000$) και δείχνει την αναλογία της συνολικής μεταβλητότητας που δεν έχει επεξηγηθεί, δηλαδή είναι το αντίστροφο μέγεθος του τετραγώνου της κανονικής συσχέτισης. Συνεπώς, εδώ η ποσότητα Wilks' lamba παίρνει την τιμή 0.356.

Wilks' Lambda				
Έλεγχος συνάρτησης	Wilks' lambda	Chi-square	Βαθμοί ελευθερίας	p-τιμή
1	.356	447.227	5	.000

Πίνακας 8.7: Πίνακας Wilks' lambda

Πίνακας τυποποιημένων συντελεστών κανονικής διαχωριστικής συνάρτησης

Ο Πίνακας 8.8 δείχνει την σπουδαιότητα κάθε μεταβλητής που χρησιμοποιείται για την πρόβλεψη. Το πρόσημο δείχνει την κατεύθυνση της σχέσης. Ο βαθμός της αυτοεκτίμησης είναι η ισχυρότερη μεταβλητή πρόβλεψης ενώ το χαμηλό άγχος είναι η δεύτερη σημαντική μεταβλητή πρόβλεψης. Αυτές οι δύο μεταβλητές με μεγάλους συντελεστές ξεχωρίζουν ως οι ισχυρότερες για την ταξινόμηση καπνιστών και μη καπνιστών στις ομάδες. Η ηλικία, η απουσία από την δουλειά και ο

βαθμός της στάσης προς την απαγόρευση του καπνίσματος είναι λιγότερο κατάλληλες για την ταξινόμηση.

Συντελεστές τυποποιημένης κανονικής διαχωριστικής συνάρτησης	
	Συνάρτηση 1
Ηλικία	.212
Βαθμός αυτοεκτίμησης	.763
Βαθμός άγχους	-.614
Ημέρες απουσίας τελευταίου έτους	-.073
Συνολική αντι-καπνιστική πολιτική	.378

Πίνακας 8.8: Πίνακας τυποποιημένων συντελεστών κανονικής διαχωριστικής συνάρτησης

Πίνακας δομής

Ο Πίνακας 8.9 παρέχει έναν άλλο τρόπο που δίνει την σημαντικότητα των μεταβλητών πρόβλεψης. Πολλοί ερευνητές χρησιμοποιούν τον πίνακα δομής γιατί θεωρούν ότι είναι πιο ακριβής σε σύγκριση με τον πίνακα τυποποιημένων συντελεστών της διαχωριστικής συνάρτησης. Ο πίνακας δομής δείχνει την συσχέτιση κάθε μεταβλητής με την διαχωριστική συνάρτηση. Οι συντελεστές Pearson ονομάζονται συντελεστές δομής ή διαχωριστικά φορτία (loadings) και συμπεριφέρονται όπως τα φορτία στην παραγοντική ανάλυση (factor analysis).

Στην συγκεκριμένη περίπτωση έχουμε ότι οι δείκτες της αυτοεκτίμησης και του άγχους διαχωρίζουν καλύτερα τους μη-καπνιστές από τους καπνιστές. Αντιθέτως, η απουσία από την δουλειά δεν συνεισφέρει στην διαχωριστική συνάρτηση, δηλαδή είναι ο πιο αδύναμος συντελεστής και δεν σχετίζεται με την συμπεριφορά των υπαλλήλων προς το κάπνισμα.

Πίνακας Δομής	
	Συνάρτηση 1
Ηλικία	.706
Βαθμός αυτοεκτίμησης	-.527
Βαθμός άγχους	.265
Ημέρες απουσίας τελευταίου έτους	-.202
Συνολική αντι-καπνιστική πολιτική	.103

Πίνακας 8.9: Πίνακας δομής

Πίνακας συντελεστών κανονικής διαχωριστικής συνάρτησης

Οι μη τυποποιημένοι συντελεστές (*b*) χρησιμοποιούνται για να δημιουργήσουν την διαχωριστική συνάρτηση. Σ'αυτή την περιπτώση έχουμε ότι (Πίνακας 8.10):

D= (.024 x ηλικία) + (.080 x αυτοεκτίμηση) + (-.100 x άγχος) + (-.312 x μέρες απουσίας) + (.134 x βαθμός αντι καπνιστών) - 4.543

Συντελεστές κανονικής διαχωριστικής συνάρτησης	
	Συνάρτηση 1
Ηλικία	.204
Βαθμός αυτοεκτίμησης	.080
Βαθμός άγχους	-.100
Ημέρες απουσίας τελευταίου έτους	-.012
Συνολική αντι-καπνιστική πολιτική	.134
Σταθερά	-4,543

Πίνακας 8.10: Συντελεστές κανονικής διαχωριστικής συνάρτησης

Οι συντελεστές της διαχωριστικής συνάρτησης *b* ή τυποποιημένη μορφή βήτα δείχνουν την μερική συνεισφορά κάθε μεταβλητής στην διαχωριστική συνάρτηση ελέγχοντας όλες τις μεταβλητές στην εξίσωση. Μπορούν να χρησιμοποιηθούν για να αξιολογήσουν την συνεισφορά κάθε ανεξάρτητης μεταβλητής στην διαχωριστική συνάρτηση και επομένως δίνουν πληροφορίες για την σημαντικότητα κάθε μεταβλητής. Εάν υπάρχουν πολλές ψευδομεταβλητές (dummy), όπως στην Ανάλυση παλινδρόμησης, δεν μπορούν να χρησιμοποιηθούν οι συντελεστές *b* και οι ψευδομεταβλητές πρέπει να αξιολογηθούν ως μια ομάδα μέσω της ιεραρχικής διαχωριστικής ανάλυσης, πρώτα χωρίς τις ψευδομεταβλητές και μετά με αυτές. Η διαφορά στο τετράγωνο της κανονικής συσχέτισης δείχνει την επίδραση των ψευδομεταβλητών.

Πίνακας κεντροειδών κατά ομάδα (Group centroids table)

Ένας άλλο μέτρο που μπορεί να χρησιμοποιηθεί για την ερμηνεία των αποτελεσμάτων της διαχωριστικής ανάλυσης είναι ο μέσος των μεταβλητών πρόβλεψης για κάθε ομάδα. Αυτοί οι μέσοι ονομάζονται κεντροειδείς (centroids) (Πίνακας 8.11). Στην περίπτωση μας, οι μη-καπνιστές έχουν μέσο 1.125 ενώ οι καπνιστές έχουν μέσο -1.598. Μια καινούρια περίπτωση της οποίας ο βαθμός είναι πλησιέστερος σε έναν κεντροειδή θεωρείται ότι ανήκει σ'αυτή την ομάδα.

Κεντροειδείς ομάδων	
Καπνιστής ή όχι	Συνάρτηση 1
Μη καπνιστής	1.125
Καπνιστής	-1.598

Πίνακας 8.11: Κεντροειδείς ομάδων

Πίνακας ταξινόμησης (Classification table)

Ο πίνακας ταξινόμησης, που επίσης ονομάζεται και πίνακας σύγχυσης-συνάφειας (confusion table), είναι ένας πίνακας στον οποίο οι γραμμές είναι οι παρατηρούμενες κατηγορίες της εξαρτημένης μεταβλητής και οι στήλες είναι οι προβλεπόμενες κατηγορίες. Όταν υπάρχει σωστή πρόβλεψη όλες

οι περιπτώσεις βρίσκονται στην διαγώνιο. Το ποσοστό των περιπτώσεων στην διαγώνιο είναι το ποσοστό των σωστά ταξινομημένων περιπτώσεων. Η δύναμη της διαχωριστικής συνάρτησης φαίνεται καλύτερα όταν διασταυρωθεί με τα υπόλοιπα δεδομένα και όχι μόνο με τα δεδομένα που χρησιμοποιήθηκαν για να παραχθεί ο κανόνας ταξινόμησης. Η διασταύρωση ονομάζεται 'jack-knife' ταξινόμηση, στην οποία ταξινομούνται όλες οι περιπτώσεις εκτός από μια η οποία χρησιμοποιείται για να παραχθεί μια διαχωριστική συνάρτηση και ταξινομείται τελευταία. Η διαδικασία επαναλαμβάνεται έτσι ώστε να 'μένει έξω' μια περίπτωση κάθε φορά. Ο τρόπος της διασταύρωσης δίνει μια πιο αξιόπιστη συνάρτηση και αυτό συμβαίνει γιατί η περίπτωση που χρησιμοποιούμε να προβλέψουμε σε ποια κατηγορία ανήκει, δεν χρησιμοποιείται ως μέρος της διαδικασίας ταξινόμησης.
Τα αποτελέσματα ταξινόμησης (Πίνακας 8.12) δείχνουν ότι το 91.6% των περιπτώσεων ταξινομήθηκαν σωστά στις ομάδες 'καπνιστές' και 'μη-καπνιστές'. Οι μη-καπνιστές ταξινομήθηκαν με καλύτερη ακρίβεια (92.6%) σε σύγκριση με τους καπνιστές (90.6%).

Αποτελέσματα ταξινόμησης					
			Ομάδα πρόβλεψης		
		Καπνιστής ή όχι	Μη-καπνιστής	Καπνιστής	Σύνολο
Αρχικό	Ποσοστό %	Μη-καπνιστής	238	19	257
		Καπνιστής	17	164	100.0
		Μη-καπνιστής	92.6	7.4	100.0
		Καπνιστής	9.4	90.6	100.0
Επικυρωμένο	Ποσοστό %	Μη-καπνιστής	238	19	257
		Καπνιστής	17	164	181
		Μη-καπνιστής	92.6	7.4	100.0
		Καπνιστής	9.4	90.6	100.0

Πίνακας 8.12: Αποτελέσματα ταξινόμησης

Επίσης μπορούν να χρησιμοποιηθούν γραφήματα για να εξεταστεί η αποδοτικότητα της διαχωριστικής συνάρτησης. Για παράδειγμα, το ιστόγραμμα ή το θηκόγραμμα είναι εναλλακτικοί τρόποι για να αναπαρασταθεί η κατανομή των βαθμών της διαχωριστικής συνάρτησης για κάθε ομάδα. Από τα ιστογράμματα (25.10 και 25.11) γίνεται φανερός ο διαχωρισμός καθώς υπάρχει μικρή κάλυψη των δύο γραφημάτων. Δηλαδή, τα γραφήματα δείχνουν ότι η συνάρτηση διαχωρίζει σωστά τις περιπτώσεις, όπως συμπεράναμε και από τους προηγούμενους πίνακες.

Καινούργιες περιπτώσεις

Η απόσταση Mahalanobis είναι η απόσταση μεταξύ μιας περίπτωσης και του κεντροειδή η οποία χρησιμοποιείται για να αναλύσει περιπτώσεις για κάθε ομάδα. Μια νέα περίπτωση θα ταξινομηθεί στην ομάδα στην οποία έχει την μικρότερη απόσταση από τον κεντροειδή της.

Παρουσίαση αποτελεσμάτων

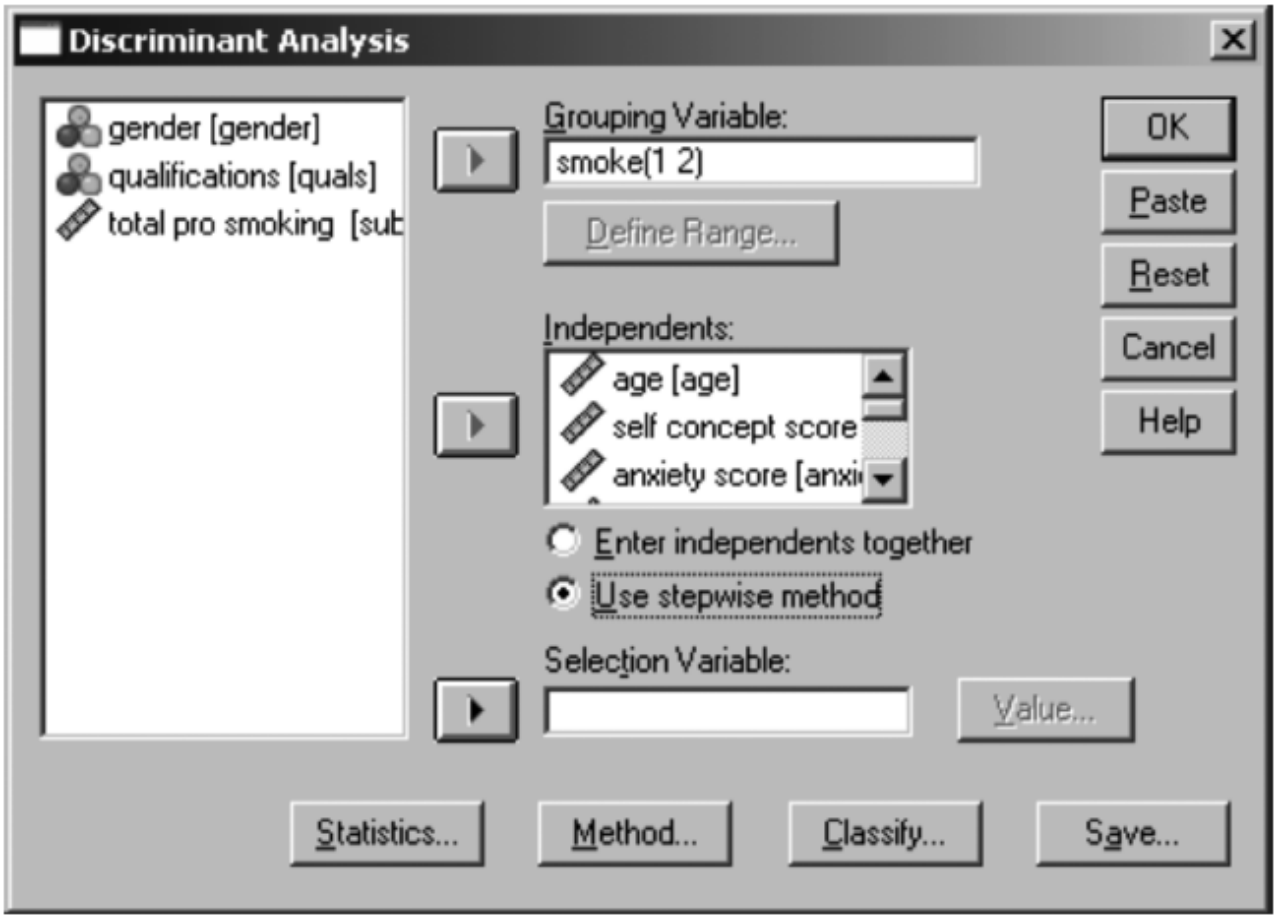
Η μέθοδος της διαχωριστικής ανάλυσης διεξάχθηκε για να προβλεφθεί εάν ένας υπάλληλος είναι καπνιστής ή όχι. Οι μεταβλητές που χρησιμοποιούνται για την πρόβλεψη είναι η ηλικία, οι μέρες απουσίας από την δουλειά τον προηγούμενο χρόνο, ο βαθμός της αυτοεκτίμησης, ο βαθμός του άγχους και η αντι-καπνιστική στάση στον χώρο εργασίας. Σημειώνονται σημαντικές διαφορές για τους μέσους των μεταβλητών στην διαχωριστική συνάρτηση. Ενώ οι λογάριθμοι των κατηγοριών της εξαρτημένης μεταβλητής ήταν όμοιοι, από το γράφημα των Box's M διαπιστώνουμε ότι η

υπόθεση της ισότητας των πινάκων συνδιακύμανσης δεν ισχύει. Ωστόσο, δεδομένου ότι έχουμε μεγάλο δείγμα, η μέθοδος μπορεί να συνεχιστεί. Η διαχωριστική συνάρτηση αποκαλύπτει την συσχέτιση μεταξύ των ομάδων και των μεταβλητών που χρησιμοποιούνται για την πρόβλεψη, ωστόσο με βαθύτερη ανάλυση μόνο δύο από τις μεταβλητές είναι οι πιο σημαντικές, ο βαθμός της αυτοεκτίμησης (.706) και ο βαθμός του άγχους (-.527) ενώ αντιθέτως η ηλικία και οι μέρες απουσίας από την δουλειά είναι οι πιο αδύναμες. Σύμφωνα με την διασταύρωση των παρατηρήσεων για την ταξινόμηση το 91.8% των παρατηρήσεων είναι σωστά ταξινομημένες.

Διαχωριστική ανάλυση με μπρος-πίσω βήματα (stepwise discriminant analysis)

Με την μπρος-πίσω διαχωριστική ανάλυση μπορούμε με βρούμε το καλύτερο σύνολο μεταβλητών που θα χρησιμοποιηθούν για την πρόβλεψη. Στην μπρος-πίσω διαχωριστική ανάλυση, η μεταβλητή που εισάγεται πρώτη είναι αυτή που συνεισφέρει περισσότερο στην συνάρτηση, έπειτα η δεύτερη μέχρις ότου καμία μεταβλητή να μην συνεισφέρει επιπλέον στο R^2.
Στην περίπτωση μας, θα εισάγουμε τις ίδιες μεταβλητές σε κάθε βήμα την φορά έτσι ώστε να δούμε ποιος συνδιασμός αποτελεί το καλύτερο σύνολο μεταβλητών που θα χρησιμοποιηθούν για την πρόβλεψη.

1. **Analyze >> Classify >> Discriminant**.
2. Στο παράθυρο **Grouping Variable** εισάγουμε την εξαρτημένη μεταβλητή. Κάνουμε click στο **Define Range** και εισάγουμε το ελάχιστο και το μέγιστο αριθμό ομάδων.
3. Κάνουμε κλικ στο **Continue** και εισάγουμε στο παράθυρο **Independents box** τις μεταβλητές για την πρόβλεψη και επιλέγουμε **Use Stepwise Methods**. Αυτή είναι η σημαντικότερη διαφορά από το προηγούμενο παράδειγμα (Εικόνα 7).
4. **Statistics** >> **Means, Univariate Anovas, Box'M, Unstandarized and Within Groups Correlation**.
5. **Classify** και επιλέγουμε **Compute From Group Sizes, Summary Table, Leave One Out Classification, Within Groups**, και όλα τα **Plots**.
6. **Continue** >> **Save** και επιλέγουμε **Predicted Group Membership** και **Discriminant Scores**.
7. **OK**



Εικόνα 8.7: Διαχωριστική ανάλυση με μέθοδο μπρος-πίσω βήματα

Πίνακες από την μπρος-πίσω στατιστική ανάλυση

Ο Πίνακας 8.11 δείχνει ότι έγιναν τέσσερα βήματα, όπου στο καθένα προστίθεται και μια μεταβλητή. Επομένως οι τέσσερις μεταβλητές, βαθμός αυτοεκτίμησης, βαθμός άγχους, συνολική αντι-καπνιστική πολιτική και ηλικία συμπεριλαμβάνονται στην ανάλυση καθώς είναι οι μεταβλητές που συνεισφέρουν περισσότερο στη διαχωριστική συνάρτηση.

Μεταβλητές στην ανάλυση				
Βήμα		Ανέχεια	F to remove	Rao's V
1	Δείκτης αυτοεκτίμησης	1.000	392.672	
2	Δείκτης αυτοεκτίμησης	.998	277.966	218.439
	Δείκτης άγχους	.988	128.061	392.672
3	Δείκτης αυτοεκτίμησης	.996	255.631	309.665
	Δείκτης άγχους	.979	138.725	461.872
	Αντι-καπνιστική στάση	.979	45.415	636.626
4	Δείκτης αυτοεκτίμησης	.982	264.525	320.877
	Δείκτης άγχους	.976	139.844	485.614
	Αντι-καπνιστική στάση	.977	41.295	677.108
	Ηλικία	.980	12.569	748.870

Πίνακας 8.11: Μεταβλητές που συμπεριλαμβάνονται στην ανάλυση

Πίνακας Wilks' lambda

Ο Πίνακας 8.12 επιβεβαιώνει ότι όλες οι μεταβλητές πρόβλεψης είναι στατιστικά σημαντικές με $p<.000$ και πρέπει να εισέλθουν στην διαχωριστική συνάρτηση.
Οι υπόλοιποι πίνακες οι οποίοι δίνουν τους συντελεστές της διαχωριστικής συνάρτησης, τους κεντροειδείς για κάθε ομάδα και την ταξινόμηση, καθώς και ο πίνακα δομής, είναι ίδιοι με πριν.

Wilks' Lambda									
Βήματα	Αριθμός μεταβλητών	Lambda	Β.ε 1	Β.ε. 2	Β.ε.3	Στατιστικό	Β.ε.1	Β.ε.2	Σιγμοειδής
1	1	.526	1	1	436	392.672	1	436.00	.000
2	2	.406	2	1	436	317.583	2	435.00	.000
3	3	.368	3	1	436	248.478	3	434.00	.000
4	4	.358	4	1	436	194.468	4	433.00	.000

Πίνακας 8.12: Wilks' lambda

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Christensen, R., (2001), *Advanced Linear Modeling*, Second edition, Springer-Verlang, New York.

[2] Fernandez, G.C.J., *Discriminant Analysism A powerful Classification Techique in Data Mining,* SUGI 27, Statistics and Data analysis, paper 247.

[3] Friel, C.M. (2004). *Discriminant analysis*.
http://www.shsu.edu/~icc_cmf/cj_742/stats7.doc

[4] Johnson, A. R., Wichern W. D., (1998). *Applied Multivariate Statistical Analysis*. Fifth edition, Prentice Hall, New Jersey.

[5] Nogueiro, A., Rosario de Oliveira, M., Salvador, P., Valadas, R., Pacheso, A., (2005). *Classification of Internet Users using Discriminant Analysis and Neural Networks.* Appeared in Next Generation Internet (NGI) Networks Conference, pp. 341-348

[6] Tinsley, H.E.A., Brown, S.D. Editors. *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press, San Diego (2000).

ΠΡΟΣΘΕΤΕΣ ΒΙΒΛΙΟΓΡΦΙΚΕΣ ΠΗΓΕΣ

[1] Clementine 12.0 Algorithms Guide, Chapter 18

[2] Clementine 12.0 Applications Guide, Chapter 24

[3] Chapter 25. Discriminant Analysis pdf document from
http://www.uk.sagepub.com/burns/website material/