ΔΙΑΧΩΡΙΣΤΙΚΗ ΑΝΑΛΥΣΗ

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

Εισαγωγή

Η διαχωριστική ανάλυση είναι μια χρήσιμη στατιστική τεχνική που σκοπό έχει την διάκριση των διαφορών μεταξύ δύο ή περισσότερων αντικειμένων σε σχέση με πολλές ανεξάρτητες μεταβλητές (μεταβλητές πρόβλεψης) ταυτοχρόνως

Η αρχή στην οποία στηρίζεται είναι ο γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών έτσι ώστε να επιτευχθεί η άριστη ένταξη μιας μεταβλητής σε κάποια από τις διακριθείς ομάδες.

Εισαγωγή

Η διαχωριστική ανάλυση είναι μια στατιστική τεχνική η οποία έχει δυο στόχους:

1. Εύρεση διαχωριστικών εξισώσεων κατάλληλων ώστε να είναι δυνατή η ταξινόμηση μιας νέας παρατήρησης σε ομάδες που είναι ήδη γνωστές.
2. Η ερμηνεία των εξισώσεων πρόβλεψης με σκοπό να διακρίνεται η σχέση μεταξύ των ομάδων.

Εισαγωγή

Αρχικά, τα δεδομένα που διατίθενται θα πρέπει να είναι μέλη δύο ή περισσότερων αμοιβαία ασύμβατων ομάδων

Ο διαχωρισμός των ομάδων θα πρέπει να είναι αυστηρά διαμορφωμένος ώστε κάθε παρατήρηση να ανήκει σε μια και μόνο ομάδα όπως αναφέραμε και παραπάνω.

Γι’ αυτό κάθε ομάδα θα πρέπει να έχει συγκεκριμένα χαρακτηριστικά τα οποία θα είναι διαφορετικά μεταξύ τους. Στην περίπτωση όπου δυο ομάδες δεν έχουν ευδιάκριτα χαρακτηριστικά υπάρχει μεγάλος κίνδυνος να γίνει λάθος ταξινόμηση (misclassification)

ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

- Αρχικά το μέγεθος του δείγματος θα πρέπει να είναι όσο το δυνατόν πιο μεγάλο έτσι ώστε να εξασφαλιστεί η αποτελεσματικότητα της διαδικασίας και η ορθότητα των συμπερασμάτων.
- Οι μεταβλητές θα πρέπει να ακολουθούν κανονική κατανομή και να είναι ανεξάρτητες και ασυσχέτιστες μεταξύ τους.
- Ενδέχεται κάποιες μεταβλητές να περιέχουν ακραίες παρατηρήσεις οι οποίες είναι πιθανό να διαστρεβλώσουν και να επηρεάσουν δραματικά τα συμπεράσματα.

ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

Τρόποι υπολογισμού του μεγέθους του δείγματος

Έστω ένα διάνυσμα παρατηρήσεων $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$ όπου με n αναφερόμαστε στο συνολικό μέγεθος του δείγματος, σ είναι η τυπική απόκλιση και μ η μέση τιμή. Για πληθυσμούς που η κατανομή τους εμφανίζει αριστερή ασυμμετρία δηλαδή έχει μεγάλη δεξιά ουρά, ένας εμπειρικός κανόνας για την επιλογή του κατάλληλου μεγέθους δείγματος είναι ο εξής

$$n > 25G_1^2$$

Όπου η τιμή G_1^2 είναι γνωστή ως μέτρο ασυμμετρίας του Fisher και ορίζεται ως

$$G_1 = \frac{1}{n\sigma^3}\left(\sum_{i=1}^{n} {y_i}^3 - 3\mu\sum_{i=1}^{n} {y_i}^2 + 2\mu^3\right)$$

ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

Κανόνας του Lehr

Ο βασικός κανόνας για δύο ισομεγέθη πληθυσμούς που ακολουθούν κανονική κατανομή με ίσες διασπορές και μέσες τιμές μ_0 και μ_1ορίζεται ως

$$n = \frac{16}{\Delta^2}$$

Όπου $\Delta = \frac{\mu_0 - \mu_1}{\sigma}$ είναι το εκτιμώμενο σφάλμα δειγματοληψίας.

Ο τύπος που απαιτείται για την σύγκριση των μέσων δύο πληθυσμών, μ_1 και μ_2 , με κοινή διασπορά δίνεται από τον τύπο

$$n = \frac{2\left(Z_{1-a/2} + Z_{1-\beta}\right)^2}{\Delta^2}$$

ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

Όπου n είναι το απαιτούμενο μέγεθος δείγματος, Z_a είναι μια σταθερά που η τιμή της προκύπτει ανάλογα με την μορφή του ελέγχου (μονόπλευρος ή δίπλευρος) και το a-επίπεδο σημαντικότητας που ορίζεται ως η πιθανότητα η τιμή του ελέγχου να πάρει μια τιμή τόσο ακραία ή περισσότερο ακραία από αυτήν που πήρε στο συγκεκριμένο δείγμα κάτω από την συγκεκριμένη υπόθεση, ενδεικτικά:

a ε.σ.	5%	1%	0,1%
2-πλευρος	1,96	2,5758	3,2905
1-πλευρος	4,65	2,33	

ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

Κανόνας για δείγματα από Poisson κατανομή

Έστω Y_i ακολουθεί Poisson κατανομή με μέσο θ_i, i=0,1. Υποθέτουμε πως συγκρίνονται οι μέσοι από δυο πληθυσμούς που ακολουθούν την Poisson κατανομή. Έστω θ_0 και θ_1 οι μέσοι των πληθυσμών. Ο ενδεδειγμένος αριθμός παρατηρήσεων ανά δείγμα είναι

$$n = \frac{4}{\left(\sqrt{\theta_0} - \sqrt{\theta_1}\right)^2}$$

Είναι γνωστό πως η κατανομή του $\sqrt{Y_i}$ είναι προσεγγιστικά κανονική $N(\mu_i = \sqrt{\theta_i}, \sigma^2 = 0.25)$. Έτσι ο αρχικός του Lehr παίρνει την παραπάνω μορφή.

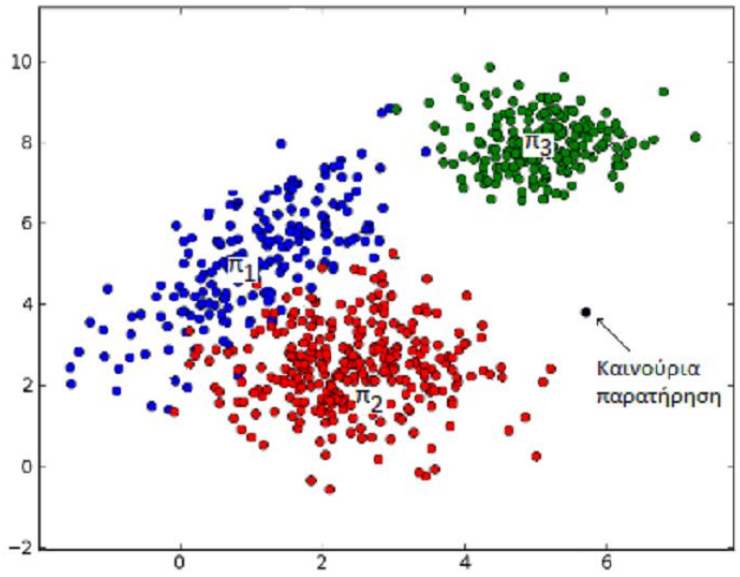
ΠΡΟΥΠΟΘΕΣΕΙΣ ΧΡΗΣΗΣ

Οι υποθέσεις μπορούν να ορισθούν ως:

1. Χρειαζόμαστε δύο ή περισσότερες ομάδες.
2. Τουλάχιστον δυο περιπτώσεις σε κάθε ομάδα ($n_i \geq 2$)
3. Ο αριθμός των διαχωριστικών μεταβλητών είναι μικρότερος από το συνολικό αριθμό περιπτώσεων όλων των ομάδων μειωμένος κατά δυο μονάδες
   ($0<p< (n-2)$)
4. Οι διαχωριστικές μεταβλητές είναι μετρήσιμες
5. Οι μη διαχωριστικές μεταβλητές πρέπει να είναι γραμμικός συνδυασμός των άλλων διαχωριστικών μεταβλητών
6. Οι πίνακες διασποράς-συνδιασποράς πρέπει να είναι ίσοι μεταξύ τους εξαιρουμένων κάποιων περιπτώσεων
7. Κάθε ομάδα αποτελείται από πληθυσμό που ακολουθεί κανονική κατανομή για τις ανεξάρτητες μεταβλητές.

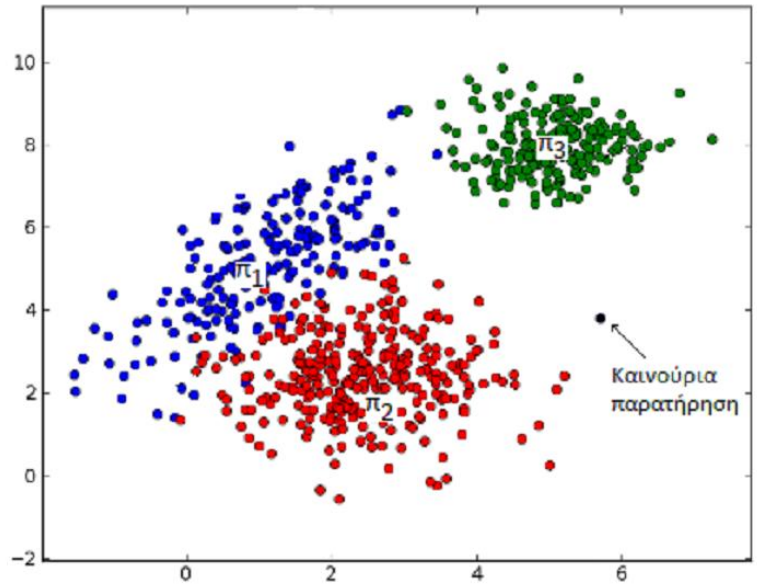
ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

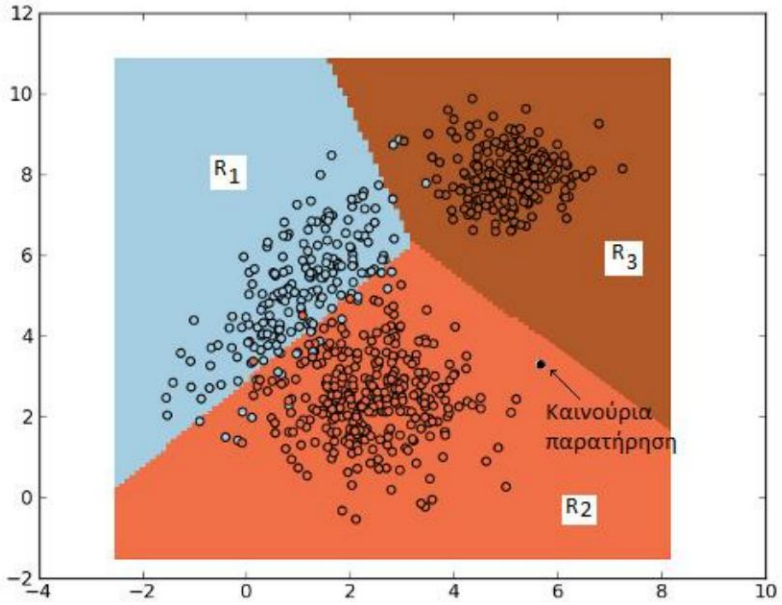
Οι κανόνες κατάταξης δημιουργούνται από τα δεδομένα που έχουμε στην διάθεσή μας. Η κάθε πολυδιάστατη παρατήρηση μπορεί να πάρει την μορφή διανύσματος $\boldsymbol{x}' = \left[x_1, x_2, \ldots, x_p \right]$ διάστασης $p \times 1$. Οι τιμές των παρατηρήσεων των μεταβλητών εξετάζονται ξεχωριστά για την διερεύνηση των διαφορών μεταξύ των ομάδων με σκοπό την δημιουργία του κατάλληλου κανόνα κατάταξης. Οι ομάδες οι οποίες χαρακτηρίζονται ως πληθυσμοί συμβολίζονται συνήθως με $\pi_i, i = 1,2, \ldots, g$ και περιγράφονται από τις αντίστοιχες συναρτήσεις πυκνότητας πιθανότητας $f_i(\boldsymbol{x})$. Το σύνολο όλων των παρατηρήσεων χαρακτηρίζεται ως δειγματικός χώρος και συμβολίζεται με Ω.



ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Όπως παρατηρούμε κάθε παρατήρηση ανήκει σε έναν πληθυσμό π_i, που διακρίνεται με ξεχωριστό χρώμα.





Μια καινούρια παρατήρηση για να ταξινομηθεί σε έναν από τους τρεις πληθυσμούς εξετάζεται με κάποιον από τους κανόνες που προαναφέραμε και που θα αναλύσουμε αργότερα. Σκοπός των κανόνων ταξινόμησης είναι να ταξινομήσουν τις καινούριες παρατηρήσεις στους πληθυσμούς με το δυνατότερο ελάχιστο κόστος και την μεγαλύτερη δυνατή ακρίβεια.

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Στην περίπτωση που ο πληθυσμός χωρίζεται από τον κανόνα κατάταξης σε δυο περιοχές που συμβολίζονται ως R_1 και R_2, μια παρατήρηση που βρίσκεται στην περιοχή R_1 αντιστοιχεί στην ομάδα π_1. Δεδομένου πως κάθε παρατήρηση θα πρέπει να προορίζεται σε έναν από τους δύο πληθυσμούς οι δύο περιοχές θα πρέπει να είναι αμοιβαία αποκλειόμενες και έτσι

$$R_1 = \Omega - R_2$$

Γενικεύοντας στην περίπτωση που μια περιοχή χωρίζεται σε g υποπληθυσμούς, R_i i=1,2,...,g μια παρατήρηση που βρίσκεται στην περιοχή R_i, αντιστοιχεί στον πληθυσμό π_i. Για τις αμοιβαία αποκλειόμενες περιοχές θα ισχύει

$$R_j = \Omega - \sum_{i=1}^{g} R_i,\ \text{j=1,...,g, i} \neq \text{j}$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Απόσταση Mahalanobis (Mahalanobis distance)

Σκοπός της απόστασης είναι να μετρήσει πόσο απέχουν δύο παρατηρήσεις, να ποσοτικοποιήσει δηλαδή αν μοιάζουν ή όχι οι παρατηρήσεις

Η απόσταση Mahalanobis είναι ένα μέτρο που μετράει την απόσταση μεταξύ μιας παρατήρησης και ενός πληθυσμού π

Για καθεμία από τις N παρατηρήσεις ενός συνόλου από p-διάστατες μεταβλητές $\boldsymbol{x}' = \left[x_1, x_2, ..., x_p\right]$ υπολογίζουμε την τετραγωνική απόστασης Mahalanobis, D_i. Αν θεωρήσουμε πως $\boldsymbol{\mu}$ είναι η μέση τιμή και πως ο $\boldsymbol{\Sigma}$ είναι ο πίνακας διασποράς-συνδιασποράς τότε το τετράγωνο της απόστασης Mahalanobis δίνεται από τον τύπο

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

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

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

Επιθυμούμε να μειώσουμε την απόσταση αυτή, δηλαδή να μειώσουμε την απόσταση της παρατήρησης από τον μέσο του πληθυσμού. Ο κανόνας για δύο πληθυσμούς γίνεται:

- Εάν $(\boldsymbol{x} - \mu_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \mu_1) < (\boldsymbol{x} - \mu_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \mu_2)$ ταξινομούμε την παρατήρηση $\boldsymbol{x}$ στον πληθυσμό π_1
- Εάν $(\boldsymbol{x} - \mu_1)' \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \mu_1) > (\boldsymbol{x} - \mu_2)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \mu_2)$ ταξινομούμε την παρατήρηση $\boldsymbol{x}$ στον πληθυσμό π_2

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Κανόνας μέγιστης πιθανοφάνειας

Ο κανόνας μέγιστης πιθανοφάνειας είναι ο πιο απλός κανόνας κατάταξης και βασίζεται στην ιδέα της πιθανοφάνειας γιατί κατατάσσει κάθε πολυμεταβλητή παρατήρηση στον πληθυσμό από τον οποίο είναι πιο πιθανό να προέρχεται. Βασίζεται σε πληθυσμούς που έχουν γνωστούς πυκνότητες.

Έστω R_i περιοχές όπου $i=1,\ldots,g$ το πλήθος των ομάδων.

$$R_i = \{x: f_i(\boldsymbol{x}) > f_j(\boldsymbol{x}) \quad j = 1, \ldots, g \ \mu\varepsilon \ j \neq i \ \}$$

Δηλαδή κατατάσσω την παρατήρηση $\boldsymbol{x}$ στον πληθυσμό i αν $\frac{f(\boldsymbol{x}|i)}{f(\boldsymbol{x}|j)} > 1$. Η λογική του κανόνα είναι να βρίσκει τη τιμή της πιθανοφάνειας της κάθε παρατήρησης στην κάθε ομάδα και όπου έχουμε την μεγαλύτερη πιθανοφάνεια θα είναι και η πιο πιθανή περιοχή για να κατατάξουμε την παρατήρηση.

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Εάν οι παρατηρήσεις ακολουθούν κανονική κατανομή, ο κανόνας μέγιστης πιθανοφάνειας μοιάζει με τον κανόνα της απόστασης Mahalanobis. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής, είναι της μορφής:

$$f(\boldsymbol{x}|i) = 2\pi^{-\frac{g}{2}}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\mu_i)\right\}$$

Επιθυμούμε να μεγιστοποιήσουμε την πιθανότητα το οποίο επιτυγχάνεται ελαχιστοποιώντας τον λογάριθμο της πιθανοφάνειας, μιας και ο λογάριθμος είναι αύξουσα συνάρτηση

$$\log f(\boldsymbol{x}|i) = -\frac{g}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}_i|) - \frac{1}{2}(\boldsymbol{x}-\mu_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\mu_i)$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Απομονώνοντας τον σταθερό όρο $-\frac{g}{2}\log(2\pi)$ και ελαχιστοποιώντας την αρνητική λογαριθμική πιθανότητα ο κανόνας ταξινόμησης μιας καινούριας παρατήρησης για δύο κανονικούς πληθυσμούς γίνεται:

- Εάν $\log(|\boldsymbol{\Sigma}_1|) + (\boldsymbol{x} - \mu_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \mu_1) > \log(|\boldsymbol{\Sigma}_2|) + (\boldsymbol{x} - \mu_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \mu_2)$ ταξινομούμε την παρατήρηση **x** στον πληθυσμό π_1
- Εάν $\log(|\boldsymbol{\Sigma}_1|) + (\boldsymbol{x} - \mu_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \mu_1) < \log(|\boldsymbol{\Sigma}_2|) + (\boldsymbol{x} - \mu_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \mu_2)$ ταξινομούμε την παρατήρηση **x** στον πληθυσμού π_2

Αυτός ο κανόνας έχει το μειονέκτημα πως δεν λαμβάνει υπόψιν του τα διαφορετικά μεγέθη της κάθε ομάδας, δηλαδή τις πιθανότητες να πάρουμε παρατήρηση από την κάθε ομάδα. Επίσης ένα άλλο μειονέκτημα είναι πως αν οι τιμές των πιθανοφανειών για μια παρατήρηση είναι ίσες τότε η συγκεκριμένη παρατήρηση δεν μας δίνει πληροφορία για την ομάδα στην οποία πρέπει να την κατατάξουμε.

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Κανόνας bayes

Ο κανόνας του bayes προϋποθέτει την ύπαρξη μιας πιθανότητας για κάθε πληθυσμό, $P(\pi_i) = p_i$. Αν συμβολίσουμε με π_i την πιθανότητα να πάρουμε μια παρατήρηση από τον i πληθυσμό, όπου $i=1,2$, τότε ο κανόνας του bayes χρησιμοποιεί για τον υπολογισμό των εκ των υστέρων πιθανοτήτων, η παρατήρηση $\mathbf{x}$ να προήλθε από τον πληθυσμό π_i την δεσμευμένη πιθανότητα:

$$P(\pi_i|\boldsymbol{x}) = \frac{P(\pi_i, \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|\pi_i)P(\pi_i)}{\sum_{i=1}^{2} P(\boldsymbol{x}|\pi_i)P(\pi_i)}$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Δεδομένων των εκ των προτέρων πιθανοτήτων και των παρατηρήσεων **x**, οι εκ των υστέρων πιθανότητες ώστε το **x** να προέρχεται από τον πληθυσμό π_1 δίνεται από τον τύπο:

$$P(\pi_1|\boldsymbol{x}) = \frac{P(\eta\ \pi\alpha\rho\alpha\tau\acute{\eta}\rho\eta\sigma\eta\ \boldsymbol{x}\ \nu\alpha\ \alpha\nu\acute{\eta}\kappa\epsilon\iota\ \sigma\tau o\ \pi_1)}{P(\nu\alpha\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\omega\ \tau\eta\nu\ \boldsymbol{x})}$$

$$= \frac{P(\nu\alpha\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\omega\ \tau\eta\nu\ \boldsymbol{x}|\pi_1)P(\pi_1)}{P(\nu\alpha\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\omega\ \tau\eta\nu\ \boldsymbol{x}|\pi_1)P(\pi_1) + P(\nu\alpha\ \pi\alpha\rho\alpha\tau\eta\rho\acute{\eta}\sigma\omega\ \tau\eta\nu\ \boldsymbol{x}|\pi_2)P(\pi_2)}$$

$$= \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})}$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Για κάθε τιμή του **x** οι τιμές των εκ των προτέρων πιθανοτήτων έχουν άθροισμα την μονάδα, $P(\pi_1) + P(\pi_2) = p_1 + p_2 = 1$. Αντίστοιχα λοιπόν, για την εκ των υστέρων πιθανότητα η παρατήρηση **x** να ανήκει στην ομάδα π_2:

$$P(\pi_2|\boldsymbol{x}) = 1 - P(\pi_1|\boldsymbol{x}) = 1 - \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} = \frac{p_2 f_2(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})}$$

Ο κανόνας για δύο πληθυσμούς διαμορφώνεται ως εξής:

- Εάν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{p_2}{p_1}$ ταξινομούμε την παρατήρηση **x** στον πληθυσμό π_1
- Εάν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1}$ ταξινομούμε την παρατήρηση **x** στον πληθυσμό π_2

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Γενίκευση για g πληθυσμούς

Γενικεύοντας τον κανόνα για i ομάδες όπου τώρα $i=1,\ldots,g$ και θεωρώντας R_i περιοχές, ο κανόνας γίνεται

$$R_i = \{\boldsymbol{x}: p_i f_i(\boldsymbol{x}) > p_j f_j(\boldsymbol{x}) \quad j = 1, \ldots, g \ \mu\varepsilon \ j \neq i \ \}$$

Αν $p_i = \frac{1}{g}$ για κάθε υποπληθυσμό ο κανόνας ταυτίζεται με τον κανόνα μέγιστης πιθανοφάνειας. Η ουσιαστική διαφορά είναι ότι σταθμίζουμε τις πιθανοφάνειες με βάση πόσο πιθανό είναι να έχουν προέλθει από κάθε υποπληθυσμό.

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Κανόνας ελαχιστοποίησης κόστους ταξινόμησης

Για τον κανόνα ελαχιστοποίησης κόστους ταξινόμησης για δύο πληθυσμούς υποθέτουμε και πάλι δυο ομάδες π_1 και π_2 και οι συναρτήσεις πυκνότητας πιθανότητας αντίστοιχα θα είναι $f_1(\boldsymbol{x})$ και $f_2(\boldsymbol{x})$. Ο διαχωρισμός των δύο ομάδων γίνεται με την βοήθεια των τιμών που θα παίρνουν κάποιες μεταβλητές **X,** έτσι έχουμε το διάνυσμα $\boldsymbol{x}_{p\times 1}{}' = [x_1, x_2, .., x_p]$.

Το βασικό πρόβλημα στα προβλήματα ταξινόμησης είναι ότι όλοι οι κανόνες δίνουν και κάποια λάθη, μέλημά μας είναι να βρούμε κανόνες που να ελαχιστοποιούν την πιθανότητα να κάνουμε λάθος κατάταξη. Αυτό μπορεί να συμβεί βρίσκοντας το αναμενόμενο ελάχιστο κόστος για την μέθοδο λανθασμένης κατάταξης.

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Για να δημιουργηθεί αυτός ο κανόνας θα πρέπει να ορίσουμε έναν τύπο ο οποίος θα μας δίνει το αναμενόμενο κόστος λανθασμένης κατάταξης έτσι ώστε να κατατάξουμε την παρατήρηση x στην ομάδα που θα έχει το μικρότερο αναμενόμενο κόστος.

$$ECM = p_j \sum_{i=1}^{g} c(i|j)P(i|j)$$

όπου

- $c(i|j)$: το κόστος να κατατάξουμε την παρατήρηση x στην i ομάδα ενώ ανήκει στην j ομάδα
- $P(i|j)$: η πιθανότητα να κατατάξουμε την παρατήρηση x στην i ομάδα ενώ ανήκει στην j ομάδα
- p_j : η εκ των προτέρων πιθανότητα να ανήκει μια παρατήρηση x στην j ομάδα

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Στην περίπτωση των δύο πληθυσμών, το αναμενόμενο κόστος λανθασμένης κατάταξης μιας παρατήρησης x στον πληθυσμό π_1 ενώ ανήκει στον πληθυσμό π_2 υπολογίζεται ως εξής:

$$ECM_1 = p_1[c(1|1)P(1|1) + c(2|1)P(2|1)] = p_1[0 \cdot P(1|1) + c(2|1)P(2|1)]$$
$$= p_1 c(2|1)P(2|1)$$

Με όμοιο τρόπο βρίσκουμε και το αντίστοιχο αναμενόμενο κόστος λανθασμένης κατάταξης μιας παρατήρησης **x** στον πληθυσμό π_2 ενώ ανήκει στον πληθυσμό π_1

$$ECM_2 = p_2[c(1|2)P(1|2) + c(2|2)P(2|2)] = p_1[c(1|2)P(1|1) + 0 \cdot P(2|2)]$$
$$= p_2 c(1|2)P(1|2)$$

Το συνολικό αναμενόμενο κόστους λανθασμένης κατάταξης είναι:

$$ECM = ECM_1 + ECM_2 = p_1 c(2|1)P(2|1) + p_2 c(1|2)P(1|2)$$

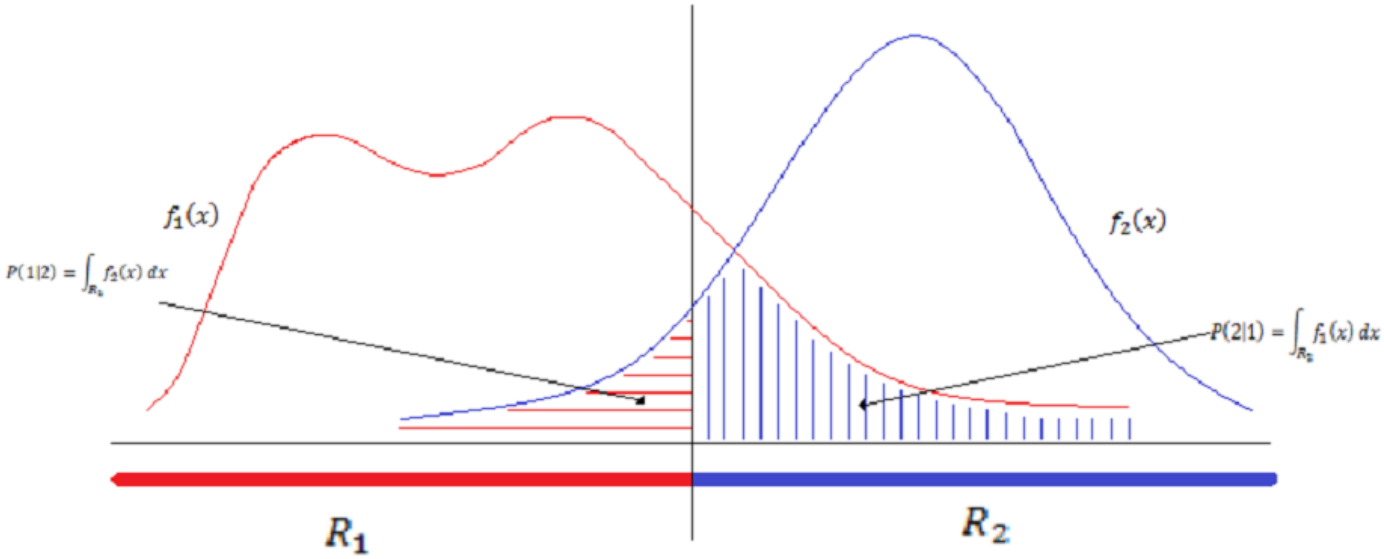
ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Παρατηρούμε πως οι $P(2|1)$ και $P(1|2)$ είναι δεσμευμένες πιθανότητες που ισούνται με

$$P(2|1) = P(\boldsymbol{x}\epsilon R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})\, d\boldsymbol{x} \quad (1)$$

$$P(1|2) = P(\boldsymbol{x}\epsilon R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})\, d\boldsymbol{x} \quad (2)$$

Το ολοκλήρωμα στην σχέση (1) παριστάνει τον όγκο που σχηματίζεται από την συνάρτηση πυκνότητας $f_1(\boldsymbol{x})$, πάνω στην περιοχή R_2



ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Πρωτίστως είδαμε ότι για την περίπτωση που έχουμε δύο ομάδες το αναμενόμενο κόστος λανθασμένης κατάταξης είναι $ECM = p_1 c(2|1) P(2|1) + p_2 c(1|2) P(1|2)$ το οποίο με την βοήθεια των σχέσεων (1) και (2) θα γίνει

$$ECM = p_1 c(2|1) \int_{R_2} f_1(x)\, dx + p_2 c(1|2) \int_{R_1} f_2(x)\, dx \quad (3)$$

Επίσης γνωρίζουμε ότι $\Omega = R_1 \cup R_2$ οπότε έχουμε

$$\int_{\Omega} f_1(\boldsymbol{x}) d\boldsymbol{x} = 1 \Rightarrow \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} = 1 \Rightarrow \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} = 1 - \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x} \quad (4)$$

Από τις σχέσεις (3) και (4) έχουμε

$$ECM = p_1 c(2|1) \left[1 - \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x}\right] + p_2 c(1|2) \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

$$= p_1 c(2|1) - p_1 c(2|1) \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 c(1|2) \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} =$$

$$= p_1 c(2|1) - \int_{R_1} p_1 c(2|1) f_1(\boldsymbol{x}) d\boldsymbol{x} + \int_{R_1} p_2 c(1|2)\, f_2(\boldsymbol{x}) d\boldsymbol{x}$$

$$= p_1 c(2|1) + \int_{R_1} [p_2 c(1|2)\, f_2(\boldsymbol{x}) - p_1 c(2|1) f_1(\boldsymbol{x})] d\boldsymbol{x}$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

$$p_1 c(2|1) + \int_{R_1} [p_2 c(1|2)\, f_2(\boldsymbol{x}) - p_1 c(2|1) f_1(\boldsymbol{x})] d\boldsymbol{x}$$

Στην παραπάνω σχέση βλέπουμε ότι p_1, p_2, $c(1|2)$, $c(2|1)$ είναι μη αρνητικές ποσότητες και οι συναρτήσεις πυκνότητας πιθανότητας $f_1(\boldsymbol{x})$ και $f_2(\boldsymbol{x})$ είναι οι μόνες που εξαρτώνται από το **x**. Επίσης είναι και μη αρνητικές για κάθε τιμή του διανύσματος **x**. Συνεπώς το αναμενόμενο κόστος λανθασμένης κατάταξης θα ελαχιστοποιηθεί ότι η περιοχή R_1 πάρει εκείνες τις τιμές του διανύσματος **x** για τις οποίες το παραπάνω ολοκλήρωμα γίνει μικρότερο ή ίσο του μηδενός. Δηλαδή έχουμε ότι:

$$\int_{R_1} [p_2 c(1|2)\, f_2(\boldsymbol{x}) - p_1 c(2|1) f_1(\boldsymbol{x})] d\boldsymbol{x} \leq 0 \Rightarrow$$

$$p_2 c(1|2)\, f_2(\boldsymbol{x}) - p_1 c(2|1) f_1(\boldsymbol{x}) \leq 0 \Rightarrow$$

$$p_2 c(1|2)\, f_2(\boldsymbol{x}) \leq p_1 c(2|1) f_1(\boldsymbol{x}) \Rightarrow$$

$$\frac{p_2 c(1|2)\, f_2(\boldsymbol{x})}{p_1 c(2|1) f_1(\boldsymbol{x})} \leq \frac{p_1 c(2|1) f_1(\boldsymbol{x})}{p_1 c(2|1) f_1(\boldsymbol{x})} \Rightarrow$$

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2 c(1|2)}{p_1 c(2|1)}$$

ΔΙΑΧΩΡΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ

Ο κανόνας ταξινόμησης μιας καινούριας παρατήρησης για δύο πληθυσμούς λοιπόν διαμορφώνεται ως εξής:

- Εάν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2 c(1|2)}{p_1 c(2|1)}$ ταξινομούμε την παρατήρηση **x** στον πληθυσμό π_1
- Εάν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \leq \frac{p_2 c(1|2)}{p_1 c(2|1)}$ ταξινομούμε την παρατήρηση **x** στον πληθυσμό π_2

Κανόνες Διαχωρισμού	Κατάταξη στην ομάδα π_1
Απόσταση Mahalanobis	$\frac{D_1^2}{D_2^2} < 1$
Κανόνας μέγιστης πιθανοφάνειας	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > 1$
Κανόνας του bayes	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{\pi_2}{\pi_1}$
Κανόνας ελαχιστοποίησης κόστους λανθασμένης κατάταξης	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{\pi_2}{\pi_1}\frac{c(1\|2)}{c(2\|1)}$