Πολυμεταβλητή Ανάλυση
Διακρίνουσα ανάλυση (Discriminant analysis)

Λουκία Μελιγκοτσίδου

Τμήμα Μαθηματικών, ΕΚΠΑ

Περιεχόμενα

Ταξινόμηση (Classification)

- Η **ταξινόμηση (Classification)** είναι το πρόβλημα της κατάταξης καινούριων παρατηρήσεων σε ήδη υπάρχουσες ομάδες (κατηγορίες) βασισμένοι σε υπάρχοντα δεδομένα για τα οποία η κατάταξη είναι γνωστή
- Συνήθως είναι διαθέσιμα δεδομένα από κάθε μια από τις ομάδες.
- Αυτά περιλαμβάνουν μια σειρά από μεταβλητές και πρέπει να φτιάξουμε έναν κανόνα κατάταξης ώστε όταν έχουμε καινούριες παρατηρήσεις να μπορούμε να τις κατατάξουμε σε ομάδες

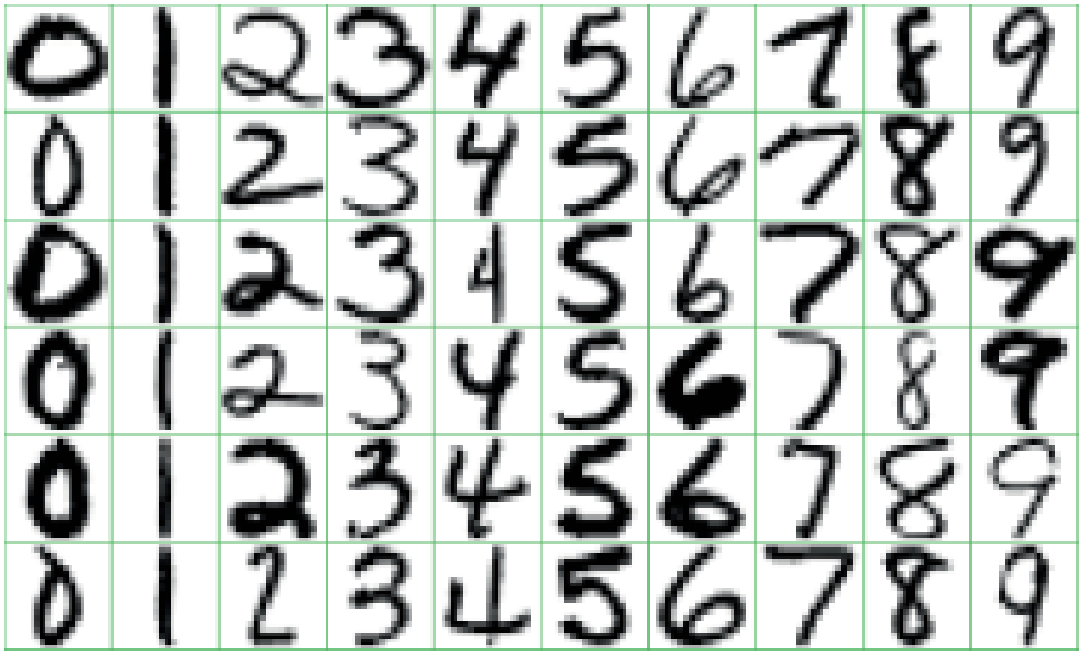
Παραδείγματα εφαρμογών

- Ιατρική διάγνωση με βάση κάποια συμπτώματα του ασθενούς
- Credit Risk: Στα χρηματοοικονομικά οι τράπεζες ενδιαφέρονται να εντοπίσουν **καλούς** και **κακούς** πελάτες πριν τη χορήγηση δανείου ή πιστωτικής κάρτας
- Marketing: όπου ζητείται ο διαχωρισμός επιτυχημένων και αποτυχημένων διαφημιστικών εκστρατειών.
- Αναποφάσιστοι: Θέλουμε να δημιουργήσουμε κανόνες ώστε ο αναποφάσιστος να κατατάσσεται σε κάποια ομάδα ψήφου.

Παραδείγματα εφαρμογών (συνέχεια)

- Δορυφορικές φωτογραφίες: Με βάση κάποια προηγούμενα δεδομένα μπορεί κανείς να κατατάξει το είδος της βλάστησης με βάση την εικόνα από το δορυφόρο
- Κοινωνικές επιστήμες: υπάρχει έντονο ενδιαφέρον να κατατάξουμε ομάδες πληθυσμού σε συγκεκριμένες κοινωνικές ομάδες με βάση μία σειρά από χαρακτηριστικά που έχουν

Handwritten digits



- Έχοντας δεδομένα για τα οποία ξέρουμε ποιο ψηφίο απεικονίζεται θέλουμε να φτιάξουμε έναν κανόνα ώστε να μπορούμε να αναγνωρίζουμε/κατατάξουμε καινούρια δεδομένα (ψηφία).

Προβλήματα που θα δούμε

- Τι μέθοδοι είναι διαθέσιμες;
- Χρειαζόμαστε πολλές μεταβλητές για έναν καλό κανόνα κατάταξης;
- Πως θα μετρήσουμε αν ο κανόνας που φτιάξαμε είναι καλός (goodness of fit);

Μέθοδοι

Τεράστια ποικιλία

- Διακρίνουσα Ανάλυση
- Μη παραμετρική διακριτική ανάλυση
- k-nn
- Παλινδρόμηση
- Λογιστική και πολυωνυμική παλινδρόμηση
- Δέντρα αποφάσεων
- SVM
- Νευρωνικά δίκτυα

Θα ασχοληθούμε κυρίως με την Διακρίνουσα ανάλυση

Διακρίνουσα Ανάλυση

- Η Διακρίνουσα ή διαχωριστική ανάλυση έχει σκοπό τη δημιουργία κανόνων κατάταξης από τα υπάρχοντα δεδομένα έτσι ώστε να είμαστε σε θέση να κατατάξουμε καινούριες παρατηρήσεις στις υπό εξέταση ομάδες.
- Στην πραγματικότητα πρόκειται για μια μεγάλη οικογένεια διαφορετικών μεθόδων.
- Αποτελεί μια παραδοσιακή μέθοδο με αρκετά διαφορετικές ερμηνείες

Ταξινόμηση μιας παρατήρησης σε έναν από δύο πληθυσμούς

- Έστω $G = 1, 2 \rightarrow$ δηλώνει την ομάδα (πληθυσμό) 1 ή 2.
- Θέλουμε να κατατάξουμε μια νέα παρατήρηση $\boldsymbol{x} \in \mathbb{R}^p$ στην ομάδα 1 $(G = 1)$ ή 2 $(G = 2)$
- Αυτό σημαίνει ότι
  - Εάν $G = 1, \boldsymbol{X} \sim f_1(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_1)
  - Εάν $G = 2, \boldsymbol{X} \sim f_2(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_2)

Ταξινόμηση μιας παρατήρησης σε έναν από δύο πληθυσμούς

- Έστω $G = 1, 2 \rightarrow$ δηλώνει την ομάδα (πληθυσμό) 1 ή 2.
- Θέλουμε να κατατάξουμε μια νέα παρατήρηση $\boldsymbol{x} \in \mathbb{R}^p$ στην ομάδα 1 $(G = 1)$ ή 2 $(G = 2)$
- Αυτό σημαίνει ότι
  - Εάν $G = 1, \boldsymbol{X} \sim f_1(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_1)
  - Εάν $G = 2, \boldsymbol{X} \sim f_2(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_2)

Κανόνας κατάταξης

Θέλουμε να βρούμε R_1 και R_2 διαμέριση του $\mathbb{R}^p$ τέτοια ώστε

- Εάν $\boldsymbol{x} \in R_1 \rightarrow^{\text{κατατάσσουμε}}$ ομάδα 1
- Εάν $\boldsymbol{x} \in R_2 \rightarrow^{\text{κατατάσσουμε}}$ ομάδα 2

Κανόνας πιθανοφάνειας

Κατατάσσουμε την νέα παρατήρηση $\boldsymbol{x}$ στην ομάδα 1 εάν

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1$$

διαφορετικά την κατατάσσουμε στην ομάδα 2. Επομένως,

$$R_1 = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1 \right\}$$

.

Με απλά λόγια

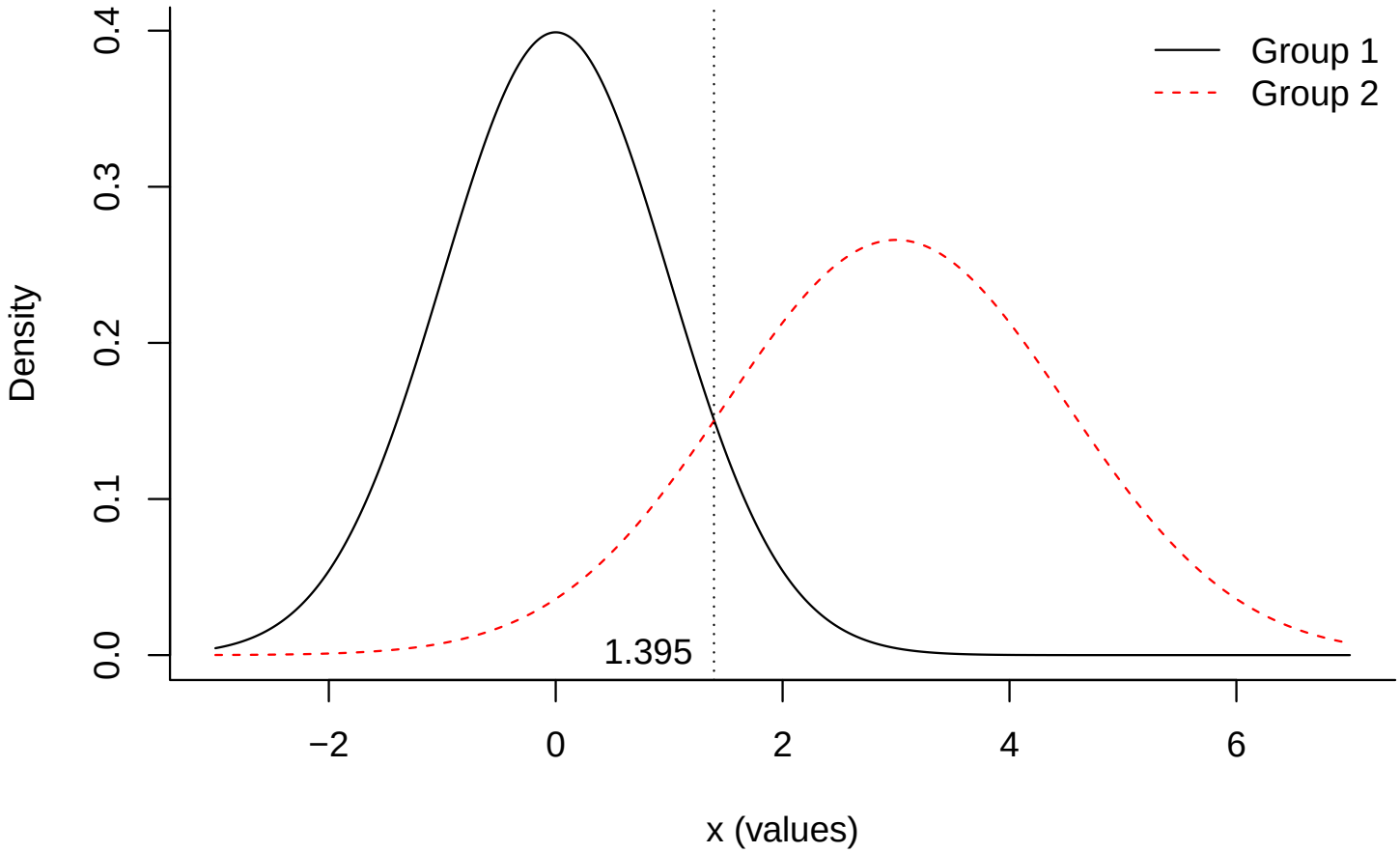
Ας υποθέσουμε ότι έχουμε δυο ομάδες, (προφανώς γενικεύεται και σε παραπάνω)

- Εκτίμησε από τα δεδομένα που έχεις μια πιθανοφάνεια (πυκνότητα πιθανότητας) για την πρώτη ομάδα. Αυτό ισοδυναμεί με το να εκτιμήσω πχ τις παραμέτρους μιας κατανομής που πιστεύω ότι περιγράφει τα δεδομένα μου.
- Κάνε το ίδιο και για τη δεύτερη ομάδα.
- Όταν έρθει καινούρια παρατήρηση κατάταξε εκεί που έχει μεγαλύτερη πιθανότητα να ανήκει.

Ένα απλό παράδειγμα

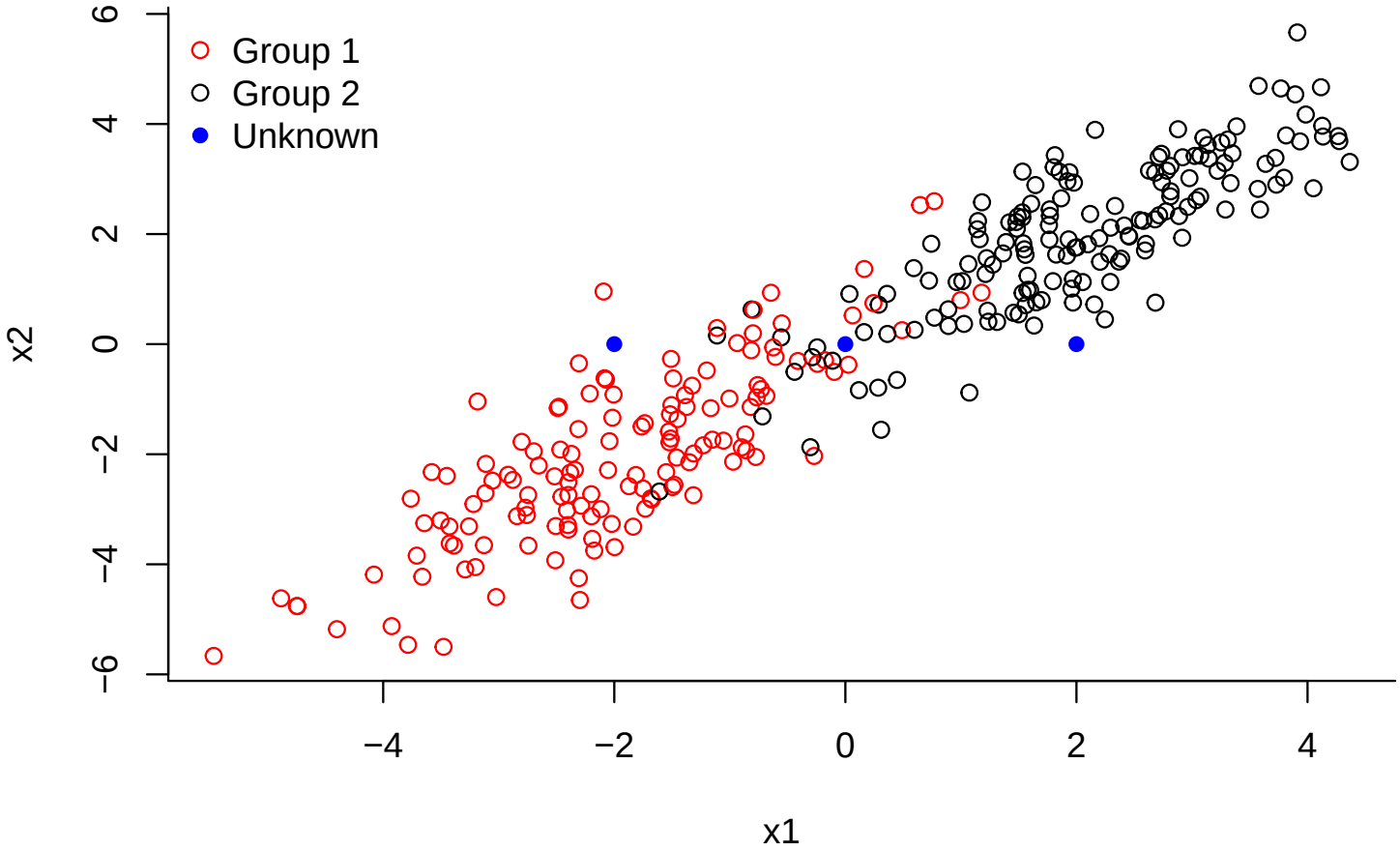
- Ας υποθέσουμε πως έχουμε δύο πληθυσμούς.
- Ο πρώτος πληθυσμός ακολουθεί $N(0, 1)$ ενώ ο δεύτερος $N(3, 1.5)$.
- Και έστω ότι έχουμε μια καινούρια παρατήρηση με τιμή 4 και θέλουμε να την κατατάξουμε σε έναν από τους δύο πληθυσμούς.
- Επειδή $f(4|\mu = 0, \sigma = 1) = 0.000134$ ενώ $f(4|\mu = 3, \sigma = 1.5) = 0.212965$, θα κατατάξουμε την παρατήρηση στο δεύτερο πληθυσμό

Ένα απλό παράδειγμα



- Για τιμές $x > 1.395$ η πυκνότητα του 2ου πληθυσμού είναι μεγαλύτερη συνεπώς κατατάσσουμε εκεί την παρατήρηση, ενώ αντίστοιχα για τιμές $x < 1.395$ κατατάσσουμε στον πρώτο πληθυσμό

Ένα απλό παράδειγμα



- Χρησιμοποιώντας τα δεδομένα θέλουμε να κατατάξουμε τις μπλε νέες παρατηρήσεις

Κανόνας Μεγίστης Πιθανοφάνειας - 2 πληθυσμοί

Παρατήρηση

Δεν είναι απαραίτητο να έχουμε την ίδια οικογένεια κατανομών σε κάθε πληθυσμό, δηλαδή θα μπορούσαν οι f_1 και f_2 να ανήκουν σε άλλη οικογένεια (πχ η μία να είναι κανονική κατανομή και η άλλη t-κατανομή)

Παράδειγμα Poisson

Ας υποθέσουμε πως η τμ X παίρνει τιμές $0, 1, \ldots$. Οι δυο πληθυσμοί ακολουθούν κατανομή Poisson με παραμέτρους θ και μ, αντίστοιχα. Τότε ο κανόνας μεγίστης πιθανοφάνειας δίνεται ως εξής: Εάν

$$\frac{f_1(x|\theta)}{f_2(x|\mu)} = \frac{\exp(-\theta)\theta^x}{\exp(-\mu)\mu^x} \geq 1$$

κατατάσσουμε στην ομάδα 1, αλλιώς στην ομάδα 2. Εναλλακτικά ο κανόνας μπορεί να γραφτεί ως

$$(\mu - \theta) + x \log\left(\frac{\theta}{\mu}\right) \geq 0.$$

Παράδειγμα: Κανονικοί πληθυσμοί

Έστω ότι η κατανομή μιας τυχαίας μεταβλητής X στα δύο επίπεδα μιας κατηγορικής μεταβλητής είναι κανονική

$$\begin{aligned} \text{Group 1}: &\quad N(\mu_1, \sigma_1^2) \\ \text{Group 2}: &\quad N(\mu_2, \sigma_2^2) \end{aligned}$$

$$\begin{aligned} f_1(x|\mu_1, \sigma_1^2) &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right\} \\ f_2(x|\mu_2, \sigma_2^2) &= \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right\} \end{aligned}$$

Ο κανόνας μεγίστης πιθανοφάνειας θα κατατάξει στην πρώτη ομάδα όταν

$$\frac{f_1(x|\mu_1, \sigma_1^2)}{f_2(x|\mu_2, \sigma_2^2)} \geq 1$$

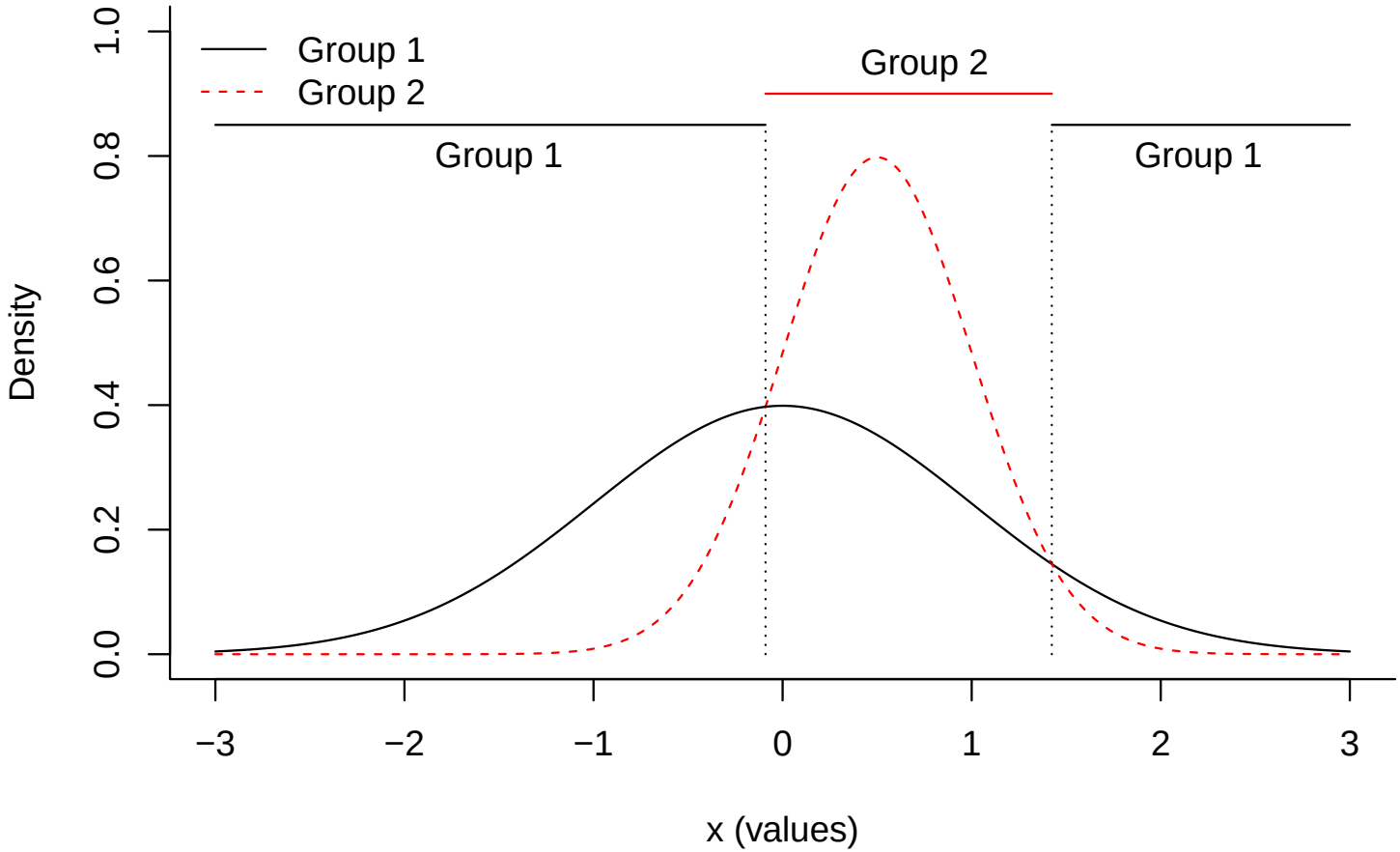
Παράδειγμα: Κανονικοί πληθυσμοί (συνέχεια)

Ο κανόνας μεγίστης πιθανοφάνειας θα κατατάξει στην πρώτη ομάδα όταν $\frac{f_1(x|\mu_1,\sigma_1^2)}{f_2(x|\mu_2,\sigma_2^2)} \geq 1$

$$\begin{aligned}
&\Leftrightarrow \quad \frac{\sigma_2}{\sigma_1}\exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2+\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right\}\geq 1\\
&\Leftrightarrow \quad -\frac{1}{\sigma_1^2}(x-\mu_1)^2+\frac{1}{\sigma_2^2}(x-\mu_2)^2\geq -2\log\left(\frac{\sigma_2}{\sigma_1}\right)\\
&\Leftrightarrow \quad x^2\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)-2x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)+\left(\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right)\leq 2\log\left(\frac{\sigma_2}{\sigma_1}\right)
\end{aligned}$$

- Ανίσωση δευτέρου βαθμού ως προς x
- Μπορεί να έχουμε ένα κριτήριο που χωρίζει την ευθεία των πραγματικών σε 3 μέρη, και να κατατάσσουμε στην πρώτη ομάδα στα δύο ακραία και στη δεύτερη το μεσαίο.

Διαχωρισμος δυο πληθυσμών $N(0, 1)$ και $N(0.5, 0.5^2)$



- Οι περιοχές κατάταξης δεν είναι συνεχόμενες, κατατάσσουμε στον πληθυσμό 1 όταν $x < -0.0904$ και όταν $x > 1.4238$ στον πληθυσμό 2.
- $R_1 = (\infty, -0.0904) \cup (1.4238, \infty)$ ενώ $R_2 = [-0.0904, 1.4238]$

Δυσταξινόμηση

- Έστω $G = 1, 2 \rightarrow$ δηλώνει την ομάδα (πληθυσμό) 1 ή 2.
- Νέα παρατήρηση $\boldsymbol{x} \in$ στην ομάδα 1 $(G = 1)$ ή 2 $(G = 2)$;
  - Εάν $G = 1, \boldsymbol{X} \sim f_1(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_1)
  - Εάν $G = 2, \boldsymbol{X} \sim f_2(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_2)

Δυσταξινόμηση

- Έστω $G = 1, 2 \rightarrow$ δηλώνει την ομάδα (πληθυσμό) 1 ή 2.
- Νέα παρατήρηση $\boldsymbol{x} \in$ στην ομάδα 1 $(G = 1)$ ή 2 $(G = 2)$;
  - Εάν $G = 1, \boldsymbol{X} \sim f_1(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_1)
  - Εάν $G = 2, \boldsymbol{X} \sim f_2(\boldsymbol{x})$ (συνάρτηση πυκνότητας f_2)

Θέλουμε να βρούμε R_1 και R_2 διαμέριση του $\mathbb{R}^p$ τέτοια ώστε

- Εάν $\boldsymbol{x} \in R_1 \rightarrow^{\text{κατατάσσουμε}}$ ομάδα 1
- Εάν $\boldsymbol{x} \in R_2 \rightarrow^{\text{κατατάσσουμε}}$ ομάδα 2

Δυσταξινόμηση

$$\begin{aligned}
\pi_{21} &= P(\text{η } \boldsymbol{x} \text{ ταξινομείται στην ομάδα 2 ενώ προέρχεται από την ομάδα 1}) \\
&= P(\boldsymbol{X} \in R_2 | G = 1) = \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} \\
\pi_{12} &= P(\text{η } \boldsymbol{x} \text{ ταξινομείται στην ομάδα 1 ενώ προέρχεται από την ομάδα 2}) \\
&= P(\boldsymbol{X} \in R_1 | G = 2) = \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}$$

Εκ των προτέρων πιθανότητες (prior distributions)

- Ο κανόνας της μέγιστης πιθανοφάνειας που μόλις είδαμε δεν λαμβάνει υπόψη του τα διαφορετικά μεγέθη των υπο-πληθυσμών, δηλαδή της εκ των προτέρων πιθανότητας να πάρουμε παρατήρηση από κάθε υπο-πληθυσμό

Εκ των προτέρων πιθανότητες:

$$
\begin{aligned}
p_1 &= \text{εκ των προτέρων πιθανότητα η } \boldsymbol{x} \text{ να προέρχεται από την ομάδα 1} \\
p_2 &= 1 - p_1 \\
p_1 &= P(G = 1) = 1 - p_2 = 1 - P(G = 2)
\end{aligned}
$$

Οι εκ των προτέρων πιθανότητες μπορεί να βασίζονται σε

- προηγούμενη γνώση (πριν δούμε τα δεδομένα), π.χ. εάν είναι γνωστό ότι ο πληθυσμός 1 είναι διπλάσιος του 2, $p_1 = 2p_2 \Rightarrow p_1 = 2/3$ και $p_2 = 1/3$
- προσωπική εμπειρία

Εαν δεν υπάρχει εκ των προτέρων πληροφορία, $p_1 = p_2 = 0.5$.

Δυσταξινόμηση

- Τα ενδεχόμενα $\{\boldsymbol{X} \in R_1, G = 1\}$ και $\{\boldsymbol{X} \in R_2, G = 2\}$ δηλώνουν ορθή ταξινόμηση.
- Δυσταξινόμηση: $\{\boldsymbol{X} \in R_2, G = 1\}$ και $\{\boldsymbol{X} \in R_1, G = 2\}$

Ολική πιθανότητα δυσταξινόμησης (total probability of misclassification, TPM)

$$\begin{aligned} TPM &= P(\boldsymbol{X} \in R_2, G = 1) + P(\boldsymbol{X} \in R_1, G = 2) \\ &= P(\boldsymbol{X} \in R_2 | G = 1)P(G = 1) + P(\boldsymbol{X} \in R_1 | G = 2)P(G = 2) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \\ &= p_1 \left(1 - \int_{R_1} f_1(\boldsymbol{x}) d\boldsymbol{x}\right) + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \\ &= p_1 + \int_{R_1} \{p_2 f_2(\boldsymbol{x}) - p_1 f_1(\boldsymbol{x})\} d\boldsymbol{x} \end{aligned}$$

- Θέλουμε το ολοκλήρωμα να γίνει ελάχιστο άρα η R_1 να περιλαμβάνει μόνο τα στοιχεία για τα οποία $p_2 f_2(\boldsymbol{x}) - p_1 f_1(\boldsymbol{x}) \leq 0$

Ελαχιστοποίηση *TPM*

Ολική πιθανότητα δυσταξινόμησης (total probability of misclassification, TPM)

$$TPM = p_1 + \int_{R_1} \{p_2 f_2(\boldsymbol{x}) - p_1 f_1(\boldsymbol{x})\}\, d\boldsymbol{x}$$

η οποία ελαχιστοποιείται διαλέγοντας την R_1 (σύνολο των $\boldsymbol{x}$ τα οποία θα κατατάσσαμε στην 1η ομάδα)

$$\begin{aligned} R_1 &= \{\boldsymbol{x} \in \mathbb{R}^p : p_2 f_2(\boldsymbol{x}) - p_1 f_1(\boldsymbol{x}) \leq 0\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^p : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1}\right\} \end{aligned}$$

Κανόνας ταξινόμησης

Εάν $\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1}$, ταξινόμηση την νέα παρατήρηση στην ομάδα 1, διαφορετικά στην ομάδα 2

- Likelihood ratio > prior ratio
- $p_1 = p_2 \rightarrow$ κανόνας μέγιστης πιθανοφάνειας.

Κανόνας Bayes

Ο κανόνας που ελαχιστοποιεί την ολική πιθανότητα λανθασμένης ταξινόμησης επιλέγει τον πληθυσμό με την μεγαλύτερη εκ τών υστέρων πιθανότητα

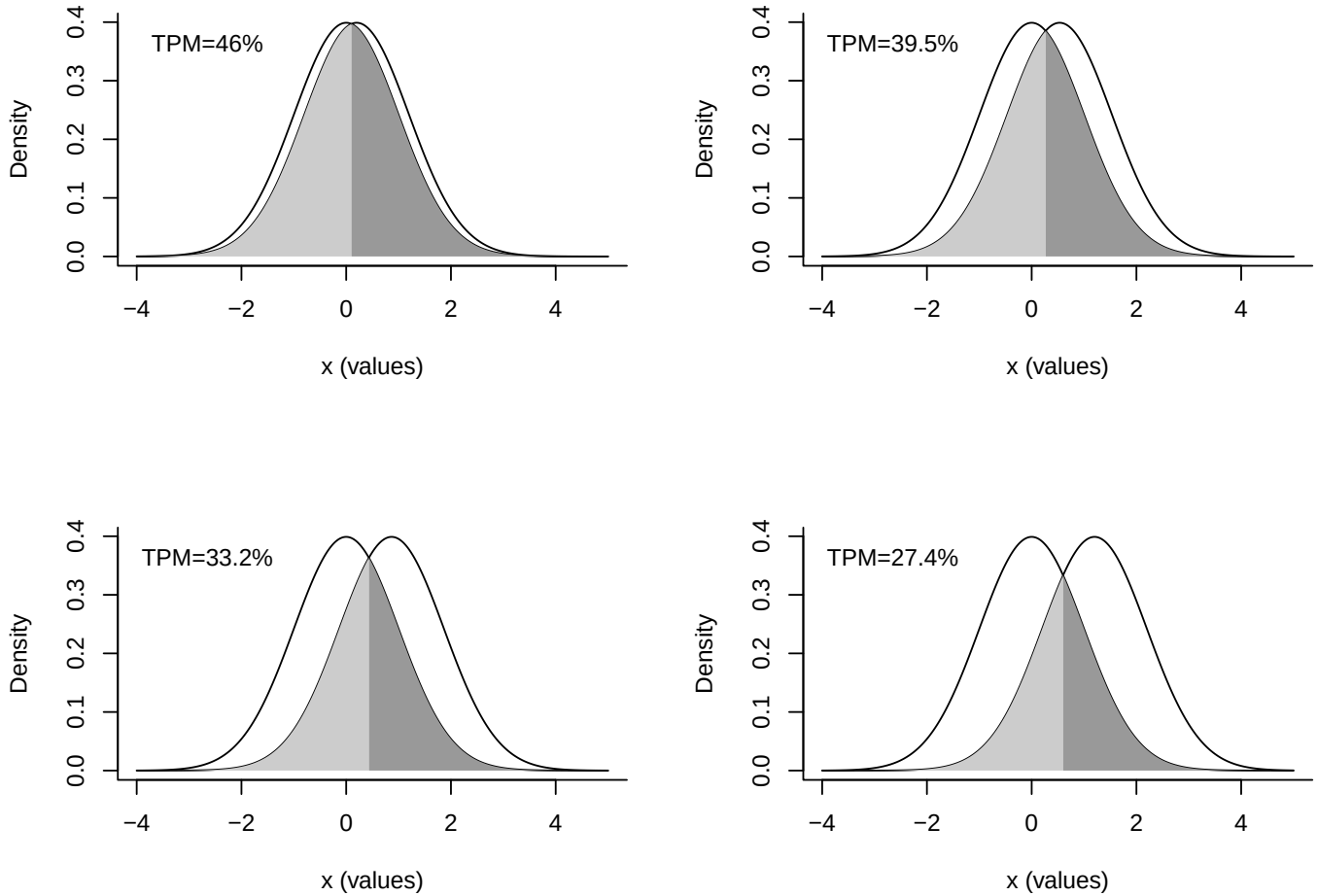
- Η $\boldsymbol{X}$ έχει μεικτή κατανομή: $f(x) = p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})$
- Η εκ των υστέρων πιθανότητα η νέα παρατήρηση $\boldsymbol{x}$ να προέρχεται από την ομάδα 1 είναι

$$\begin{aligned} P(G = 1|\boldsymbol{X} = \boldsymbol{x}) &= \frac{P(G = 1, \boldsymbol{X} = \boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} = \frac{p_1 f_1(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} \\ &= 1 - \frac{p_2 f_2(\boldsymbol{x})}{p_1 f_1(\boldsymbol{x}) + p_2 f_2(\boldsymbol{x})} \\ &= 1 - P(G = 2|\boldsymbol{X} = \boldsymbol{x}) \end{aligned}$$

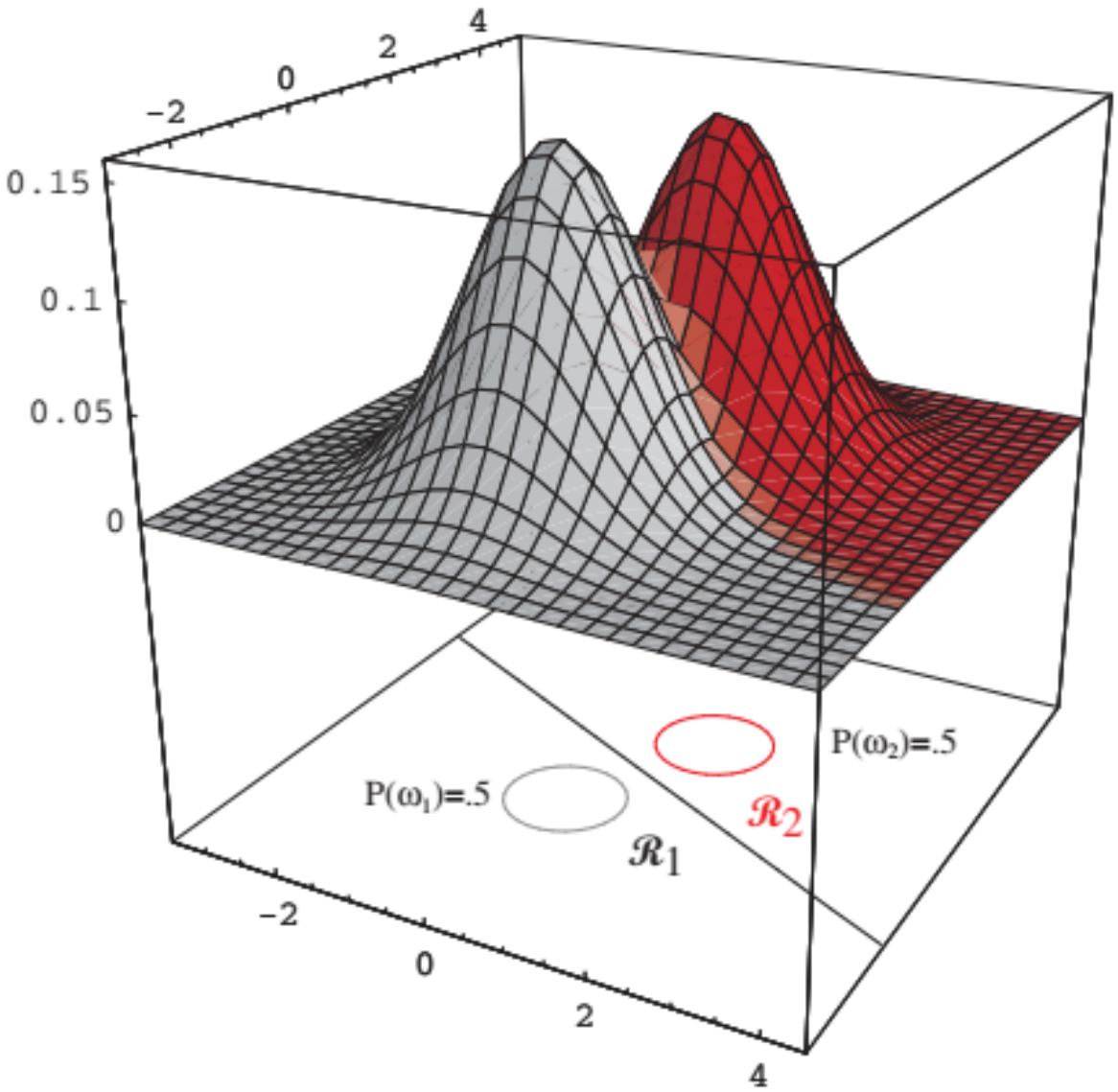
- $P(G = 1|\boldsymbol{X} = \boldsymbol{x}) \geq P(G = 2|\boldsymbol{X} = \boldsymbol{x}) \Leftrightarrow \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1}$.

Δυσταξινόμηση: Παραδείγματα

Όσο μεγαλύτερη είναι η επικάλυψη τόσο μεγαλύτερο είναι το σφάλμα. Αν δηλαδή οι υποπληθυσμοί μοιάζουν αρκετά είναι δύσκολο να φτιάξουμε έναν πολύ καλό κανόνα κατάταξης



Περισσότερες διαστάσεις



Ελαχιστοποίηση κόστους λανθασμένης κατάταξης

- Σε πολλά παραδείγματα αν κάνουμε λανθασμένη κατάταξη αυτό σημαίνει ένα κόστος.
- Για παράδειγμα στην περίπτωση του credit risk αν κάποιος πελάτης λανθασμένα θεωρηθεί καλός αλλά στην πραγματικότητα δεν είναι, τότε το κόστος της τράπεζας είναι το ποσό του δανείου που δεν θα αποπληρωθεί.
- Μπορούμε να δημιουργήσουμε κανόνες κατάταξης που να λαμβάνουν υπόψη αυτό το κόστος.
- Έστω c_{12} και c_{21} το κόστος αν κατατάξουμε στον πληθυσμό 1 ενώ η παρατήρηση πραγματικά ανήκει στον πληθυσμό 2 και αντίστοιχα.
- Το συνολικό κόστος λανθασμένης κατάταξης (expected cost of misclassification, ECM) είναι

$$ECM = p_1\pi_{21}c_{21} + p_2\pi_{12}c_{12}$$

Ελαχιστοποίηση κόστους λανθασμένης κατάταξης

$$ECM = p_1 \pi_{21} c_{21} + p_2 \pi_{12} c_{12}$$

Με ανάλογο τρόπο, αποδεικνύεται ότι το αναμενόμενο κόστος λανθασμένης ταξινόμησης ελαχιστοποιείται για

$$R_1 = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c_{12} p_2}{c_{21} p_1} \right\}$$

Παράδειγμα

- Έστω μια ασθένεια με ποσοστό στον πληθυσμό 2%.
- Το κόστος να μην εντοπίσουμε σωστά τον ασθενή είναι 10 φορές μεγαλύτερο από το κόστος να μην εντοπίσουμε σωστά έναν υγιή.
- Στην ουσία αυτό σημαίνει ότι προτιμούμε να στείλουμε άδικα κάποιον για περαιτέρω θεραπεία παρά να μην εντοπίσουμε κάποιον ασθενή και να τον αφήσουμε χωρίς θεραπεία.
- Τότε 1ος πληθυσμός είναι οι υγιείς και 2ος οι ασθενείς,

$$c_{12} = c(\text{Υγιής}|\text{Ασθενής}) = 10c(\text{Ασθενής}|\text{Υγιής}) = 10c_{21}$$

- Αν

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{10 \times 0.02}{0.98}$$

τότε θεωρούμε το άτομο ως Υγιή (1η ομάδα) διαφορετικά θεωρούμε το άτομο ως ασθενή (2η ομάδα) και του χορηγούμε περαιτέρω θεραπεία

Σύγκριση

Όταν δουλεύουμε με 2 πληθυσμούς ο κανόνες για να κατατάξουμε στον πρώτο πληθυσμό γίνεται

Κανόνας μεγ. Πιθανοφάνειας	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1$
Κανόνας Bayes	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1}$
Ελάχιστο κόστος	$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c_{12}p_2}{c_{21}p_1}$

Ταξινόμηση μιας παρατήρησης σε έναν από δύο κανονικούς πληθυσμούς

- Έστω $\boldsymbol{X}|G=1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ και $\boldsymbol{X}|G=2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
- Υποθέτουμε $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

$$f_1(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right\}$$

$$f_2(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \Rightarrow$$

$$\begin{aligned}
\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \\
&= \exp\left\{ -\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \right. \\
&+ \left. \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \frac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} \right\}
\end{aligned}$$

Ταξινόμηση μιας παρατήρησης σε έναν από δύο κανονικούς πληθυσμούς (συνέχεια)

Άρα,

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \exp\left\{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right\}$$

Επιπλέον, επειδή $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$, καταλήγουμε

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c_{12}p_2}{c_{21}p_1}$$

$$\Leftrightarrow \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \log\left(\frac{c_{12}p_2}{c_{21}p_1}\right)$$

Ταξινόμηση μιας παρατήρησης σε έναν από δύο κανονικούς πληθυσμούς (συνέχεια)

Κανόνας κατάταξης

Μια νέα παρατήρηση, κατατάσσεται στον πληθυσμό 1 εάν

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \log\left(\frac{c_{12} p_2}{c_{21} p_1}\right)$$

διαφορετικά κατατάσσεται στον πληθυσμό 2.

- Θέτοντας $\boldsymbol{L} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, ο κανόνας παίρνει απλούστερη μορφή

$$\boldsymbol{L}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{L}^\top(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq k_0 \text{ (κάποια σταθερά)}$$

- Η παραπάνω συνάρτηση είναι **γραμμική** ως προς τις μεταβλητές.
- Στην πράξη αυτό σημαίνει πως η **συνάρτηση** είναι μια ευθεία που χωρίζει το χώρο σε δύο μέρη
- $\rightarrow$ Γραμμική Διακρίνουσα Ανάλυση (**Linear Discriminant Analysis**)

Παρατηρήσεις

- Λόγος λογαριθμικών πιθανοφανειών που χρησιμοποιούμε $\rightarrow$ την απόσταση Mahalanobis της παρατήρησης από το μέσο κάθε ομάδας.
- Δεν είναι εύκολο να πάρουμε τόσο εύχρηστους τύπους για κάθε κατανομή. Στην πράξη απλά υπολογίζουμε τους λόγους πιθανοφανειών.
- Στην περίπτωση που δεν χρησιμοποιήσουμε την υπόθεση των ίσων διακυμάνσεων τότε ο κανόνας που προκύπτει δεν είναι γραμμικός αλλά τετραγωνικός (quadratic) και άρα στο επίπεδο διαχωρίζουμε τις δυο ομάδες όχι με μια γραμμή αλλά με μια καμπύλη

Στην πράξη ταξινόμηση μέσω ενός δείγματος

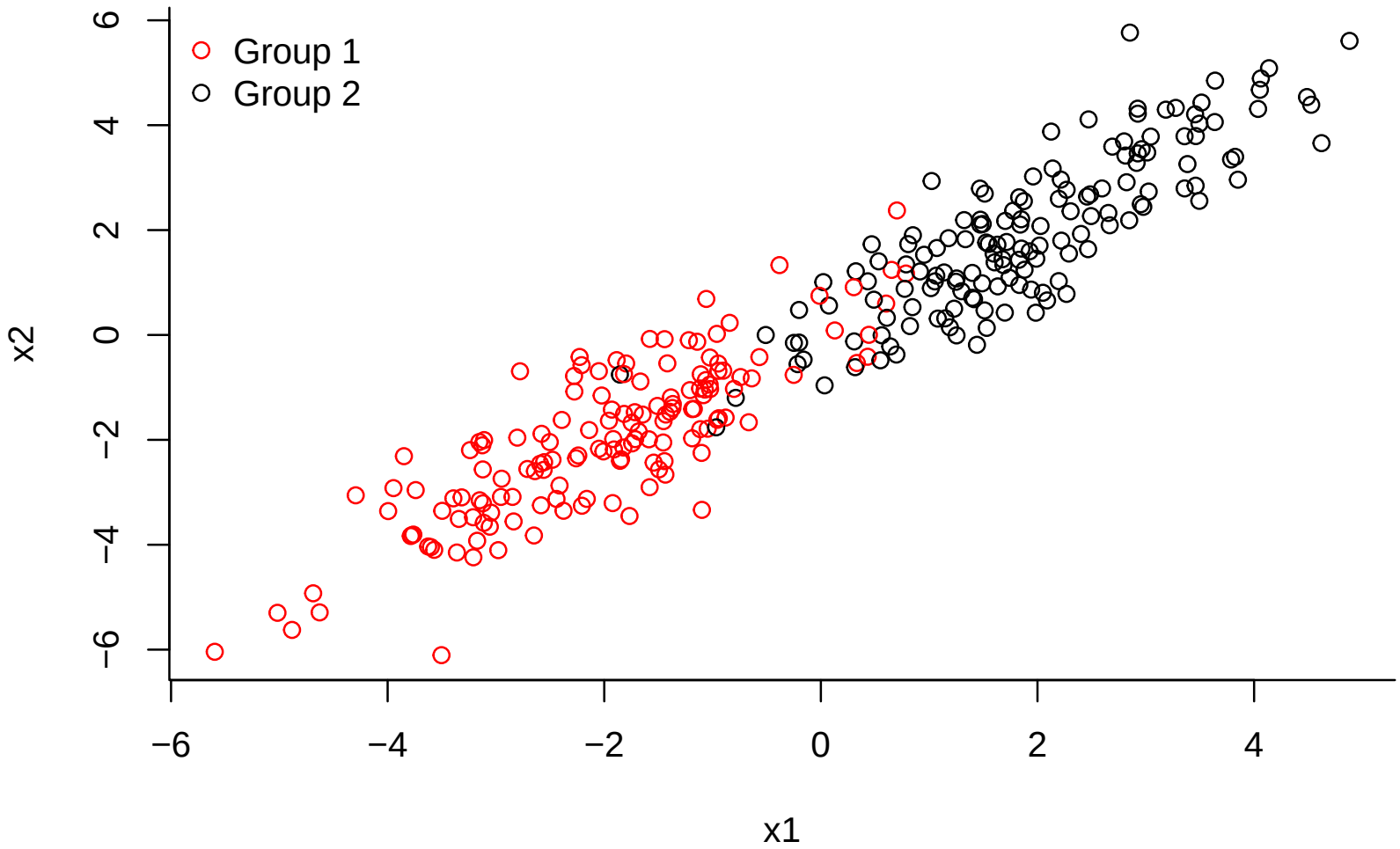
Στην πράξη δεν γνωρίζουμε τις παραμέτρους τών πληθυσμών, άρα πρέπει να τις εκτιμήσουμε. Έστω

$$\boldsymbol{x}_1^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)} \text{ τυχαίο δείγμα από την } N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$
$$\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)} \text{ τυχαίο δείγμα από την } N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

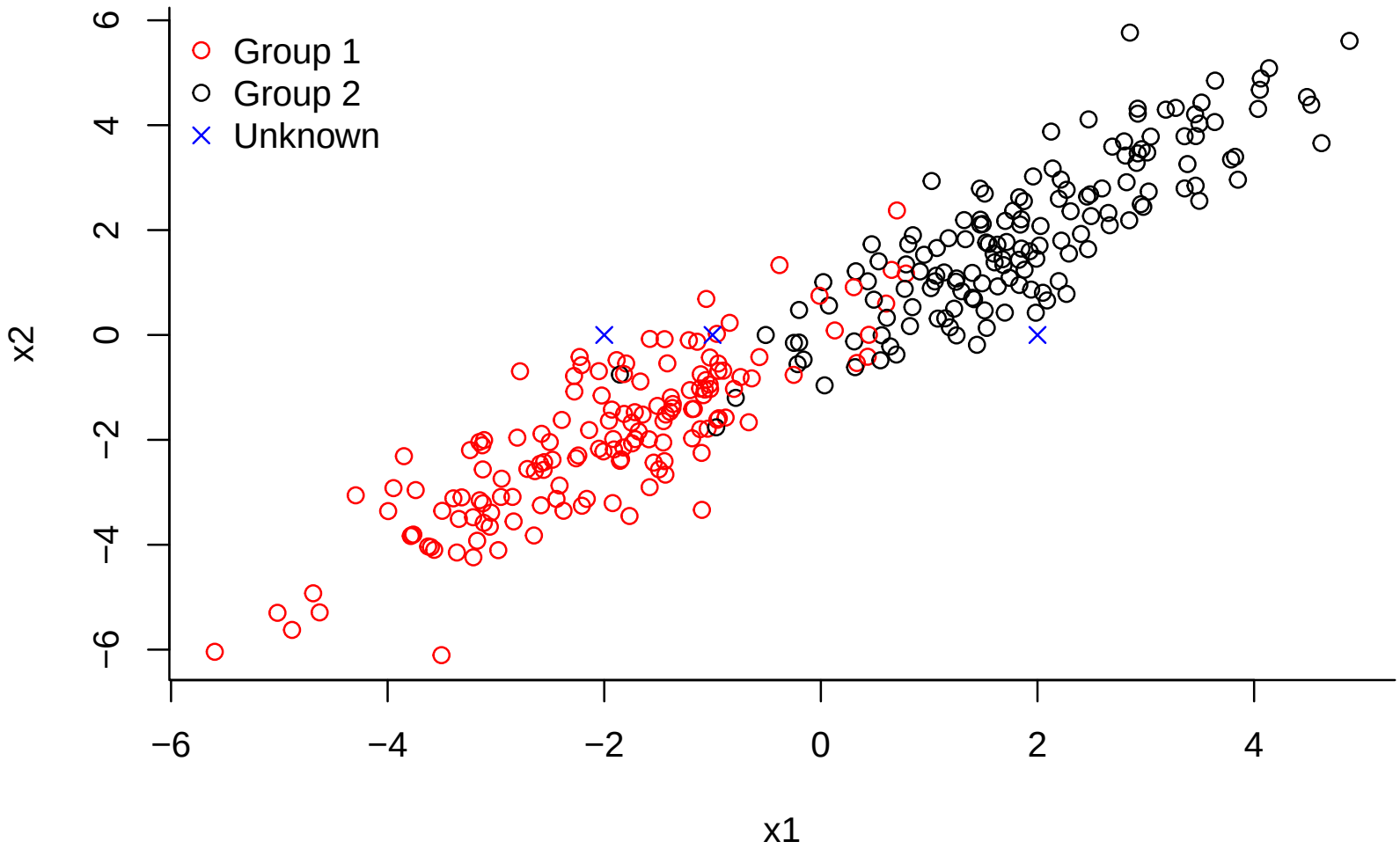
Παράμετρος	Εκτίμηση
$\boldsymbol{\mu}_1$	$\overline{\boldsymbol{x}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{x}_i^{(1)}$
$\boldsymbol{\mu}_2$	$\overline{\boldsymbol{x}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} \boldsymbol{x}_i^{(2)}$
$\boldsymbol{\Sigma}$	$\boldsymbol{S}_p = \frac{1}{n_1+n_2-2}\Big\{ \sum_{i=1}^{n_1}(\boldsymbol{x}_i^{(1)} - \overline{\boldsymbol{x}}^{(1)})(\boldsymbol{x}_i^{(1)} - \overline{\boldsymbol{x}}^{(1)})^\top$
	$+ \sum_{i=1}^{n_2}(\boldsymbol{x}_i^{(2)} - \overline{\boldsymbol{x}}^{(2)})(\boldsymbol{x}_i^{(2)} - \overline{\boldsymbol{x}}^{(2)})^\top \Big\}$

- Χρησιμοποιούμε τις εκτιμήσεις στην θέση των παραμέτρων
- $\boldsymbol{S}_p \rightarrow$ pooled covariance matrix (αντίστοιχο της κοινής διακύμανσης του t-test).

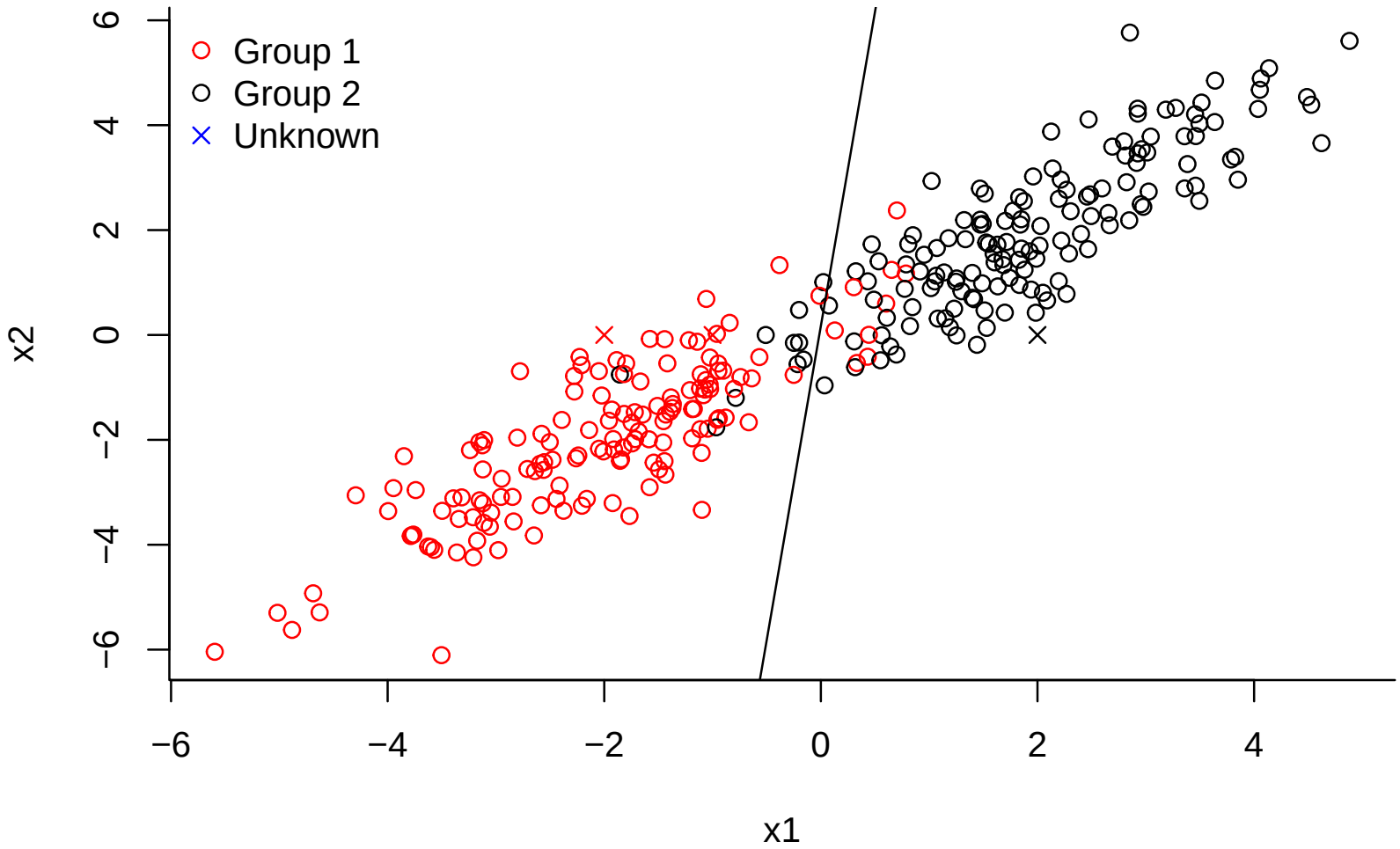
Δυο ομάδες



Δεδομένα



Η διακρίνουσα γραμμή



Αξιολόγηση συναρτήσεων ταξινόμησης

Ολική πιθανότητα λανθασμένης ταξινόμησης ενός κανόνα με R_1 και R_2

$$TPM = p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}$$

- Στην πράξη δεν μπορεί όμως να εκτιμηθεί επειδή οι πυκνότητες $f_1(\boldsymbol{x})$ και $f_2(\boldsymbol{x})$ είναι άγνωστες
- Εάν υποθέσουμε ότι $\boldsymbol{X}|G = 1$ και $\boldsymbol{X}|G = 2$ ακολουθούν την πολυδιάστατη κανονική κατανομή και εκτιμήσουμε τις παραμέτρους τους από το δείγμα, οι υπολογισμοί μπορούν να γίνουν, αλλά πρέπει να επιβεβαιωθεί ότι η προσέγγιση της κανονικότητας είναι αξιόπιστη - δύσκολο σε μεγάλες διαστάσεις.

Αξιολόγηση συναρτήσεων ταξινόμησης

Για την αξιολόγηση τού κανόνα ταξινόμησης χρησιμοποιούμε την **φαινομενική πιθανότητα λανθασμένης ταξινόμησης** (apparent error rate, APER) που ορίζεται ως το ποσοστό τών παρατηρήσεων στο εκπαιδευτικό δείγμα τα οποία ταξινομούνται λανθασμένα από τον κανόνα

Αξιολόγηση συναρτήσεων ταξινόμησης

Για την αξιολόγηση τού κανόνα ταξινόμησης χρησιμοποιούμε την **φαινομενική πιθανότητα λανθασμένης ταξινόμησης** (apparent error rate, APER) που ορίζεται ως το ποσοστό τών παρατηρήσεων στο εκπαιδευτικό δείγμα τα οποία ταξινομούνται λανθασμένα από τον κανόνα

		ταξινόμηση από τον κανόνα		
		Ομάδα 1	Ομάδα 2	
Πραγματικός	Ομάδα 1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
πληθυσμός	Ομάδα 2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Αξιολόγηση συναρτήσεων ταξινόμησης

Για την αξιολόγηση τού κανόνα ταξινόμησης χρησιμοποιούμε την **φαινομενική πιθανότητα λανθασμένης ταξινόμησης** (apparent error rate, APER) που ορίζεται ως το ποσοστό τών παρατηρήσεων στο εκπαιδευτικό δείγμα τα οποία ταξινομούνται λανθασμένα από τον κανόνα

		ταξινόμηση από τον κανόνα		
		Ομάδα 1	Ομάδα 2	
Πραγματικός	Ομάδα 1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
πληθυσμός	Ομάδα 2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Η φαινομενική πιθανότητα λανθασμένης ταξινόμησης είναι

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

Αξιολόγηση συναρτήσεων ταξινόμησης

Για την αξιολόγηση τού κανόνα ταξινόμησης χρησιμοποιούμε την **φαινομενική πιθανότητα λανθασμένης ταξινόμησης** (apparent error rate, APER) που ορίζεται ως το ποσοστό τών παρατηρήσεων στο εκπαιδευτικό δείγμα τα οποία ταξινομούνται λανθασμένα από τον κανόνα

		ταξινόμηση από τον κανόνα		
		Ομάδα 1	Ομάδα 2	
Πραγματικός	Ομάδα 1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
πληθυσμός	Ομάδα 2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

Η φαινομενική πιθανότητα λανθασμένης ταξινόμησης είναι

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

- Παρ' όλο που το *APER* είναι ένας λογικός εκτιμητής της πραγματικής δυσταξινόμησης, την υποεκτιμά. Για να διορθωθεί αυτό το πρόβλημα απαιτούνται `μεγάλα' n_1, n_2.
- Η `αισιοδοξία' τού εκτιμητή δικαιολογείται από το γεγονός ότι το ίδιο δείγμα αξιολογεί τον κανόνα που εκείνο υπέδειξε

Αξιολόγηση συναρτήσεων ταξινόμησης (Cross-validation)

Εάν έχουμε «αρκετές» παρατηρήσεις, διαιρούμε το δείγμα σε δύο μέρη:

- Πρώτο μέρος (συνήθως $\simeq 50\%$): εκπαιδευτικό δείγμα (training sample)
- Δεύτερο μέρος (συνήθως $\simeq 50\%$): επικυρωτικό δείγμα (validation sample).

Ο κανόνας ταξινόμησης κατασκευάζεται από το εκπαιδευτικό δείγμα και αξιολογείται από το επικυρωτικό δείγμα όπως προηγουμένως.
Αυτή η προσέγγιση έχει δύο μειονεκτήματα:

1. Απαιτεί «μεγάλα» δείγματα
2. Δεν χρησιμοποιούνται όλες οι παρατηρήσεις για την αξιολόγηση και επομένως μπορεί να χάνεται σημαντική πληροφορία

Δεύτερη εναλλακτική προσέγγιση: jackknifing

Η μέθοδος jackknife (Leave-One-Out Cross-Validation) είναι μία γενική μέθοδος που βασίζεται στην εξής απλή ιδέα: Η διαδικασία εφαρμόζεται n φορές (όσο είναι το συνολικό μέγεθος δείγματος) παραλείποντας κάθε φορά μία παρατήρηση με την σειρά

1. Παραλείποντας μια παρατήρηση, κατασκευάζουμε τον κανόνα ταξινόμησης από τις υπόλοιπες παρατηρήσεις.
2. Βάσει αυτού τού κανόνα, ταξινομούμε την παρατήρηση που παραλείψαμε σε κάποιον από τους δύο πληθυσμούς
3. Επαναλαμβάνουμε για όλες τις παρατηρήσεις

Αποδεικνύεται ότι το συνολικό ποσοστό λανθασμένα ταξινομημένων παρατηρήσεων είναι αμερόληπτος εκτιμητής της πραγματικής πιθανότητας λάθους ταξινόμησης

Η προσέγγιση του Fisher

Αν έχουμε πολλές μεταβλητές/διαστάσεις τότε θα μπροούσαμε να επιλύσουμε το πρόβλημα πιο απλά κάνοντας το εξής:

- Να βρούμε εναν γραμμικό συνδυασμό των μεταβλητών, τέτοιο ωστε υπολογίζοντας αυτα τα σκορ οι δυο ομάδες μας να ειναι όσο γινεται πιο μακριά η μία από την άλλη
- Σκοπός, δηλαδή είναι η εύρεση ενος γραμμικού μετασχηματισμού των δεδομένων

$$U(\boldsymbol{x}) = \boldsymbol{L}^{\top}\boldsymbol{x}$$

- ο οποίος μεγιστοποιεί την πιθανοφάνεια και το t-test που συγκρίνει τις 2 ομάδες

Η διαχωριστική ανάλυση του Fisher

- Μετατροπή των χαρακτηριστικών $\boldsymbol{x}$ σε μονοδιάστατα σκορ

$$U(\boldsymbol{x}) = L^\top \boldsymbol{x}$$

- $U(\boldsymbol{x})$ ονομάζεται διαχωριστική/διακρίνουσα συνάρτηση.
- Τα σκορ των δύο ομάδων θα πρέπει να είναι όσο το δυνατόν πιο απομακρυσμένα έτσι ώστε να μπορούμε εύκολα με βάση αυτά τα σκορ να κάνουμε διαχωρισμό και ταξινόμηση των δύο ομάδων.
- Ο Fisher πρότεινε τη χρήση γραμμικών συνδυασμών για τη δημιουργία αυτών των σκορ χωρίς να γίνει κάποια υπόθεση για την κατανομή των ομάδων.
- Η γραμμικότητα υιοθετήθηκε για λόγους ευκολίας.
- Υπέθεσε **ισότητα των πινάκων συνδιακύμανσης** (χωρίς να το λέει ρητά) αφού χρησιμοποίησε τη συνδυασμένη κοινή (pooled) εκτίμηση $\boldsymbol{S}_p$.

Η διαχωριστική ανάλυση του Fisher (2)

- Έστω τα σκορ $U^{(1)}$ και $U^{(2)}$ για τις 2 ομάδες.
- Μέτρο απόκλισης των σκορ των δύο ομάδων = απόσταση των μέσων τιμών $(\overline{U}^{(1)} - \overline{U}^{(2)})$.
- Ο Fisher χρησιμοποίησε την ποσότητα:

$$D = \frac{|\overline{U}^{(1)} - \overline{U}^{(2)}|}{s_U}$$

όπου s_U^2 δηλώνει την κοινή (pooled) διακύμανση

$$s_U^2 = \frac{\sum_{i=1}^{n_1}(U_i^{(1)} - \overline{U}^{(1)})^2 + \sum_{i=1}^{n_2}(U_i^{(2)} - \overline{U}^{(2)})^2}{n_1 + n_2 - 2}$$

- **Σκοπός**: η μεγιστοποίηση της ποσότητας D ή αντίστοιχα της απόστασης D^2.

Η διαχωριστική ανάλυση του Fisher (3)

- Θέλουμε να βρούμε τον γραμμικό συνδυασμό $\boldsymbol{L}$ ώστε

$$D^2 = (\overline{U}^{(1)} - \overline{U}^{(2)})^2 / s_U^2$$

να γίνει μέγιστο.

- Ανισότητα Cauchy-Schwarz: $(\boldsymbol{a}^\top \boldsymbol{b})^2 \leq (\boldsymbol{a}^\top \boldsymbol{a})(\boldsymbol{b}^\top \boldsymbol{b}), \forall \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$.
- Επειδή $\boldsymbol{S}_p$ είναι θετικά ορισμένος, θέτοντας $\boldsymbol{a} = \boldsymbol{S}_p^{1/2}\boldsymbol{L}$ και $\boldsymbol{b} = \boldsymbol{S}_p^{-1/2}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})$

$$\begin{aligned}
\left\{\boldsymbol{L}^\top \boldsymbol{S}_p^{1/2}\boldsymbol{S}_p^{-1/2}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})\right\}^2 &\leq (\boldsymbol{L}^\top \boldsymbol{S}_p \boldsymbol{L})(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)}) \\
\Rightarrow \frac{\left\{\boldsymbol{L}^\top(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})\right\}^2}{\boldsymbol{L}^\top \boldsymbol{S}_p \boldsymbol{L}} &\leq (\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})
\end{aligned}$$

Η διαχωριστική ανάλυση του Fisher (4)

Παρατηρείστε ότι $\overline{U}^{(1)} = \boldsymbol{L}^\top \overline{\boldsymbol{x}}^{(1)}$ και $\overline{U}^{(2)} = \boldsymbol{L}^\top \overline{\boldsymbol{x}}^{(2)}$. Επίσης,

$$
\begin{aligned}
s_U^2 &= \frac{\sum_{i=1}^{n_1}(U_i^{(1)} - \overline{U}^{(1)})^2 + \sum_{i=1}^{n_2}(U_i^{(2)} - \overline{U}^{(2)})^2}{n_1 + n_2 - 2} \\
&= \frac{\sum_{i=1}^{n_1}\left\{\boldsymbol{L}^\top(\boldsymbol{x}_i^{(1)} - \overline{\boldsymbol{x}}^{(1)})\right\}^2 + \sum_{i=1}^{n_2}\left\{\boldsymbol{L}^\top(\boldsymbol{x}_i^{(2)} - \overline{\boldsymbol{x}}^{(2)})\right\}^2}{n_1 + n_2 - 2} \\
&= \frac{\boldsymbol{L}^\top\left\{\sum_{i=1}^{n_1}(\boldsymbol{x}_i^{(1)} - \overline{\boldsymbol{x}}^{(1)})(\boldsymbol{x}_i^{(1)} - \overline{\boldsymbol{x}}^{(1)})^\top\right\}\boldsymbol{L}}{n_1 + n_2 - 2} \\
&+ \frac{\boldsymbol{L}^\top\left\{\sum_{i=1}^{n_2}(\boldsymbol{x}_i^{(2)} - \overline{\boldsymbol{x}}^{(2)})(\boldsymbol{x}_i^{(2)} - \overline{\boldsymbol{x}}^{(2)})^\top\right\}\boldsymbol{L}}{n_1 + n_2 - 2} \\
&= \boldsymbol{L}\boldsymbol{S}_p\boldsymbol{L}
\end{aligned}
$$

Η διαχωριστική ανάλυση του Fisher (5)

Επομένως, έχουμε αποδείξει ότι για κάθε γραμμικό μετασχηματισμό $\boldsymbol{L}$

$$\begin{aligned} D^2 &= \frac{(\overline{U}^{(1)} - \overline{U}^{(2)})^2}{s_U^2} = \frac{\left\{\boldsymbol{L}^\top(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})\right\}}{\boldsymbol{L}\boldsymbol{S}_p\boldsymbol{L}} \\ &\leq (\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)}) \end{aligned}$$

Παίρνοντας $\boldsymbol{L} = \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})$, επιβεβαιώνεται εύκολα ότι $D^2 = (\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})$. Άρα, η $\boldsymbol{L} = \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})$ είναι ο βέλτιστος γραμμικός μετασχηματισμός

- Κρίσιμη τιμή: Μέση τιμή των 2 σκορ

$$\boldsymbol{m} = \frac{1}{2}(\overline{U}^{(1)} + \overline{U}^{(2)})$$

- **Διαχωριστική (διακρίνουσα) συνάρτηση**: $\boldsymbol{L}^\top \boldsymbol{x} - \boldsymbol{m}$

Η διαχωριστική ανάλυση του Fisher (6)

- **Διαχωριστικός κανόνας**: $\boldsymbol{L}^\top \boldsymbol{x} - \boldsymbol{m}$
- **Διαφορετική προσέγγιση αλλά ίδιο αποτέλεσμα** με μέθοδο μέγιστης πιθανοφάνειας για 2 πληθυσμούς ($k_0 = 0$)

$$\boldsymbol{L}^\top \boldsymbol{x} - \boldsymbol{m} = (\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}^{(1)} - \overline{\boldsymbol{x}}^{(2)})^\top \boldsymbol{S}_p^{-1}(\overline{\boldsymbol{x}}^{(1)} + \overline{\boldsymbol{x}}^{(2)})$$

- Η παραπάνω προσέγγιση δεν προϋποθέτει κανονικότητα. Για το λόγο αυτό μπορούμε να χρησιμοποιήσουμε και μη κανονικές μεταβλητές ακόμα και κατηγορικές. Αυτός είναι ο λόγος που αυτή η προσέγγιση αναφέρεται πιο πολύ στη βιβλιογραφία
- Υπάρχουν όμως οι προϋποθέσεις της γραμμικότητας και της ομοσκεδαστικότητας (οι οποίες πολλές φορές δεν ισχύουν στην πράξη).

Ταξινόμηση μίας παρατήρησης σε έναν από $k \geq 2$ πληθυσμούς

- Έστω $G \in \{1, 2, \ldots, k\}$ (k διαφορετικοί πληθυσμοί)
- Θέλουμε να ταξινομήσουμε μια παρατήρηση $\boldsymbol{x} \in \mathbb{R}^p$ σε έναν εκ των k πληθυσμών
- $\boldsymbol{X}|G = i \sim f_i, i = 1, 2, \ldots, k.$
- Εκ των προτέρων πιθανότητες, $p_i = P(G = i), i = 1, 2, \ldots, k$, και $\sum_{i=1}^{k} p_i = 1$
- $c_{ij} \rightarrow$ κόστος ταξινόμησης της παρατήρησης στον i πληθυσμό ενώ προέρχεται από τον j ($c_{ii} = 0$)
- Η $\boldsymbol{x}$ κατατάσσεται στον i πληθυσμό εάν $\boldsymbol{x} \in \mathbb{R}^p$, όπου $\{R_1, \ldots, R_k\}$ διαμέριση του $\mathbb{R}^p$
- $\pi_{ij} = P(\boldsymbol{X} \in R_i|G = j) = \int_{R_i} f_j(\boldsymbol{x})d\boldsymbol{x}$
- $\pi_{jj} = 1 - \sum_{i \neq j} \pi_{ij}$

Αναμενόμενο κόστος δυσταξινόμησης

Το (δεσμευμένο) αναμενόμενο κόστος δυσταξινόμησης μίας παρατήρησης $\boldsymbol{x}$ από τον πληθυσμό j είναι

$$ECM(j) = \sum_{i=1}^{k} c_{ij}\pi_{ij} = \sum_{i=1}^{k} c_{ij} \int_{R_i} f_j(\boldsymbol{x}) d\boldsymbol{x}$$

ενώ το αναμενόμενο κόστος δυσταξινόμησης μίας παρατήρησης ξ είναι

$$ECM = \sum_{j=1}^{k} p_j ECM(j) = \sum_{i=1}^{k} \int_{R_i} \left\{ \sum_{j=1}^{k} p_j c_{ij} f_j(\boldsymbol{x}) d\boldsymbol{x} \right\}$$

Το αναμενόμενο κόστος δυσταξινόμησης ελαχιστοποιείται αν επιλέξουμε

$$R_i = \left\{ \boldsymbol{x} \in \mathbb{R}^p : \sum_{j=1}^{k} p_j c_{ij} f_j(\boldsymbol{x}) = \min_{1 \le \nu \le k} \sum_{j=1}^{k} p_j c_{\nu j} f_j(\boldsymbol{x}) \right\}$$

Αναμενόμενο κόστος δυσταξινόμησης

Ισοδύναμα, υπολογίζουμε σκορ καταχώρισης στην i ομάδα

$$W_i = -\sum_{j=1}^{k} p_j c_{ij} f_j(\boldsymbol{x}),\ i = 1, 2, \ldots, k$$

και καταχωρούμε την παρατήρηση μας στην ομάδα με το μεγαλύτερο σκορ.

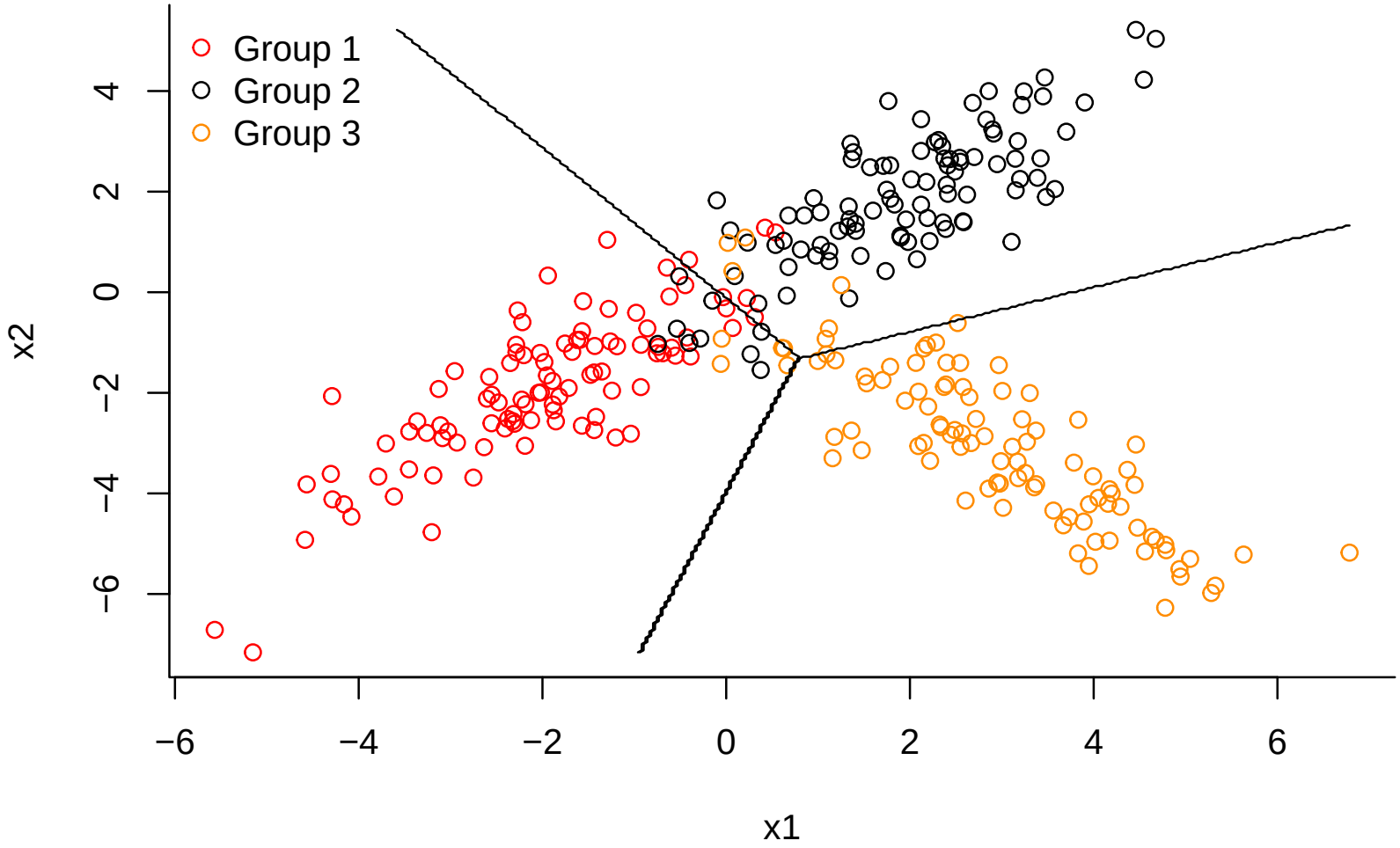
Παρατήρηση

Εάν τα κόστη είναι σταθερά, ο προηγούμενος κανόνας είναι ισοδύναμος με την μεγιστοποίηση τής εκ τών υστέρων πιθανότητας

$$P(G = i|\boldsymbol{X} = \boldsymbol{x}) = \frac{p_i f_i(\boldsymbol{x})}{\sum_{\nu=1}^{k} p_\nu f_\nu(\boldsymbol{x})}$$

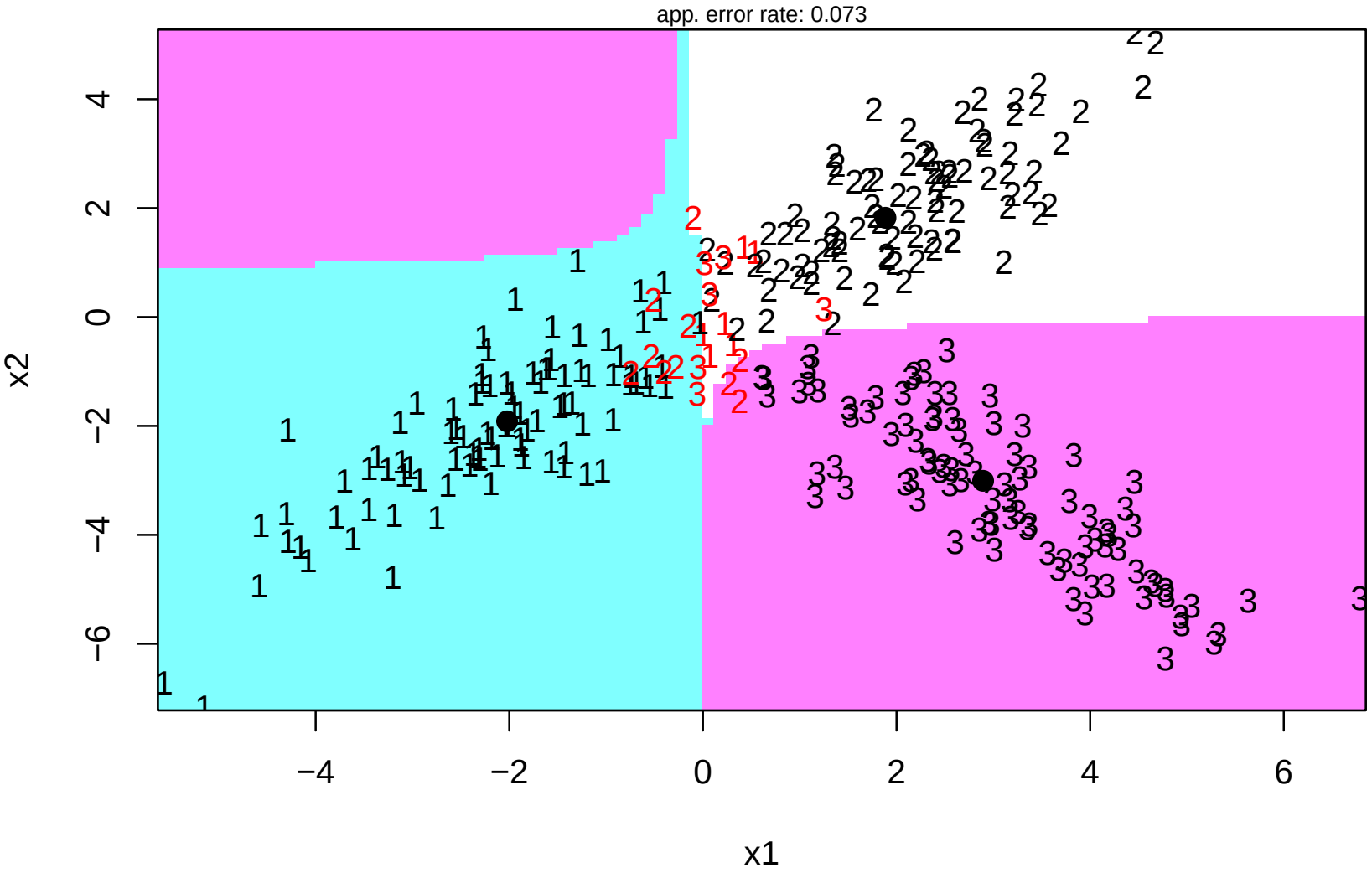
Γραμμική διακρίνουσα ανάλυση με 3 ομάδες

Πίνακες συνδιακύμανσης ίσοι



Τετραγωνική διακρίνουσα ανάλυση με 3 ομάδες

Πίνακες συνδιακύμανσης διαφορετικοί



Η μέθοδος του Fisher για $k \geq 2$ πληθυσμούς

- Έστω $k \geq 2$ πληθυσμοί με μέσες τιμές $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ και **κοινό** πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}$ (θετικά ορισμένο).
- Δείγμα

$$\begin{aligned}
&\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \ldots, \boldsymbol{x}_{n_1}^{(1)} \sim G = 1 \\
&\boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \ldots, \boldsymbol{x}_{n_2}^{(2)} \sim G = 2 \\
&\vdots \\
&\boldsymbol{x}_1^{(k)}, \boldsymbol{x}_2^{(k)}, \ldots, \boldsymbol{x}_{n_k}^{(k)} \sim G = k
\end{aligned}$$

- Για ευκολία, δουλεύουμε με μετασχηματισμένες (μονοδιάστατες) παρατηρήσεις από τον κάθε πληθυσμό, $y_i^{(j)} = \boldsymbol{a}^\top \boldsymbol{x}_i^{(j)}$, όπου $\boldsymbol{a} \in \mathbb{R}^p$.
- Το ερώτημα είναι πως θα επιλεγεί το $\boldsymbol{a}$...

Η μέθοδος του Fisher για $k \geq 2$ πληθυσμούς

- Δείγμα μετασχηματισμένων παρατηρήσεων

$$
\begin{aligned}
y_1^{(1)}, y_2^{(1)}, \ldots, y_{n_1}^{(1)} &\sim G = 1 \\
y_1^{(2)}, y_2^{(2)}, \ldots, y_{n_2}^{(2)} &\sim G = 2 \\
&\vdots \\
y_1^{(k)}, y_2^{(k)}, \ldots, y_{n_k}^{(k)} &\sim G = k
\end{aligned}
$$

- Θα θέλαμε οι μέσες τιμές $\overline{y}^{(1)}, \ldots, \overline{y}^{(k)}$ να είναι όσο το δυνατόν πιο διαφορετικες μεταξύ τους

Η μέθοδος του Fisher: sum of squares

Μία «λογική» προσέγγιση για να εξετάσουμε την ισότητα τών μέσων τιμών τών μετασχηματισμένων πληθυσμών είναι να συγκρίνουμε το άθροισμα τών τετραγώνων τών αποκλίσεων μεταξύ τών δειγμάτων (between sum of squares)

$$\begin{aligned} SSB &= \sum_{j=1}^{k} n_j(\overline{y}^{(j)} - \overline{y})^2 = \sum_{j=1}^{k} n_j(\boldsymbol{a}^\top \overline{\boldsymbol{x}}^{(j)} - \boldsymbol{a}^\top \overline{\boldsymbol{x}})^2 = \sum_{j=1}^{k} n_j \left\{\boldsymbol{a}^\top(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}})\right\}^2 \\ &= \boldsymbol{a}^\top \left\{\sum_{j=1}^{k} n_j(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}})^\top\right\} \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a} \end{aligned}$$

με το άθροισμα τών τετραγώνων τών αποκλίσεων εντός τών δειγμάτων (within sum of squares)

$$\begin{aligned} SSW &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_i^{(j)} - \overline{y}^{(j)})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\boldsymbol{a}^\top \boldsymbol{x}_i^{(j)} - \boldsymbol{a}^\top \overline{\boldsymbol{x}}^{(j)})^2 \\ &= \boldsymbol{a}^\top \left\{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}^{(j)})^\top\right\} \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a} \end{aligned}$$

Η μέθοδος του Fisher: one-way ANOVA

- Όπως στην one{way ANOVA

$$SSB = \boldsymbol{a}^\top \left\{ \sum_{j=1}^{k} n_j (\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}^{(j)} - \overline{\boldsymbol{x}})^\top \right\} \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}$$

$$SSW = \boldsymbol{a}^\top \left\{ \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}^{(j)})(\boldsymbol{x}_i^{(j)} - \overline{\boldsymbol{x}}^{(j)})^\top \right\} \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}$$

- Είναι φανερό ότι

$$\boldsymbol{W} = \sum_{j=1}^{k} (n_j - 1) S_j = (n_1 + n_2 + \ldots n_k - k) \boldsymbol{S}_p$$

δηλαδή είναι ένα πολλαπλάσιο τού κοινού εκτιμητή τού $\boldsymbol{\Sigma}$

Η ιδέα του Fisher

Η υπόθεση τής ισότητας τών μέσων τιμών αμφισβητείται όσο ο λόγος

$$Q(\boldsymbol{a}) = \frac{SSB}{SSW} = \frac{\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}}$$

παίρνει μεγαλύτερες τιμές.

- Η ιδέα του Fisher ήταν να χρησιμοποιήσουμε το $\boldsymbol{a}$ που μεγιστοποιεί το $Q(\boldsymbol{a})$ γιατί έτσι διαχωρίζονται περισσότερο οι πληθυσμοί
- Εάν πολλαπλασιάσουμε το $\boldsymbol{a}$ με ένα σταθερό αριθμό $c \neq 0$, έχουμε τον ίδιο λόγο $\rightarrow Q(\boldsymbol{a}) = Q(c\boldsymbol{a})$.
- Επομένως, χωρίς βλάβη της γενικότητας, μπορούμε να υποθέσουμε ότι $\boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a} = 1$

Διακρίνουσα του Fisher

Αποτέλεσματα

Αποδεικνύεται ότι ο πίνακας $\boldsymbol{W}^{-1}\boldsymbol{B}$ έχει $s \leq \min(k-1, p)$ μη μηδενικές ιδιοτιμές $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0$ και έστω $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_s$ τα αντίστοιχα ιδιοδιανύσματα, κανονικοποιημένα ώστε $\boldsymbol{e}_i^\top \boldsymbol{W}\boldsymbol{e}_i = 1$. Τότε ισχύει ότι

$$\max_{\boldsymbol{a}\in\mathbb{R}^p:\ \boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}=1} \frac{\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}} = \frac{\boldsymbol{e}_1^\top \boldsymbol{B}\boldsymbol{e}_1}{\boldsymbol{e}_1^\top \boldsymbol{W}\boldsymbol{e}_1} = \lambda_1$$

$$\max_{\boldsymbol{a}\in\mathbb{R}^p:\ \boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}=1, \boldsymbol{a}\perp\boldsymbol{e}_1\ldots,\boldsymbol{e}_r} \frac{\boldsymbol{a}^\top \boldsymbol{B}\boldsymbol{a}}{\boldsymbol{a}^\top \boldsymbol{W}\boldsymbol{a}} = \frac{\boldsymbol{e}_{r+1}^\top \boldsymbol{B}\boldsymbol{e}_{r+1}}{\boldsymbol{e}_{r+1}^\top \boldsymbol{W}\boldsymbol{e}_{r+1}} = \lambda_{r+1}$$

για $r = 1, 2, \ldots, s-1$.

Διακρίνουσα του Fisher: Στην πράξη

Έστω $y_1 = \boldsymbol{e}_1^\top \boldsymbol{x}, \ldots, y_s = \boldsymbol{e}_s^\top \boldsymbol{x}$ οι s δειγματικές διακρίνουσες κατά Fisher και έστω ότι αποφασίζουμε να χρησιμοποιήσουμε $r \leq s$ από αυτές για την ταξινόμηση μίας νέας παρατήρησης σε κάποιον από τους k πληθυσμούς

Κανόνας

Μία νέα παρατήρηση $\boldsymbol{x}$ ταξινομείται στον πληθυσμό i για τον οποίον

$$\sum_{j=1}^{r}(\boldsymbol{e}_j^\top \boldsymbol{x} - \boldsymbol{e}_j^\top \overline{\boldsymbol{x}}^{(i)})^2 = \sum_{j=1}^{r}\left\{\boldsymbol{e}_j^\top(\boldsymbol{x} - \overline{\boldsymbol{x}}^{(i)})\right\}^2 = \min_{1\leq\nu\leq k}\sum_{j=1}^{r}\left\{\boldsymbol{e}_j^\top(\boldsymbol{x} - \overline{\boldsymbol{x}}^{(\nu)})\right\}^2$$

- Κάθε νέα παρατήρηση κατατάσσεται στον πληθυσμό απο του οποίου τη μέση τιμή `απέχει' λιγότερο βάσει των διακρίνουσων συναρτήσεων.
- **Προσοχή**, ο κανόνας διαχωρισμού εδώ είναι διαφορετικός και συμπίπτει με τη μέθοδο πιθανοφανειών (κανονικών κατανομών) μόνο για $k = 2$.

Διακρίνουσα του Fisher: Ερμηνεία

Ερμηνεία των παραπάνω διαχωριστικών συναρτήσεων:

- 1η διαχωριστική συνάρτηση μεγιστοποιεί τις διαφορές των μέσων σε μια διάσταση.
- 2η διαχωριστική συνάρτηση μεγιστοποιεί την απόσταση των μέσων σε μια κατεύθυνση ορθογώνια στην 1η
- 3η μας δείχνει την απόσταση σε μια 3η διάσταση ανεξάρτητη των άλλων 2 κ.ο.κ.