

Διαχωρισμός
 παρατηρήσεων
 σε 2 classes και παρατήρηση
 όπως 2 classes based p-variables
 fitable.

Linear

Discriminant Analysis (LDA)

$$X = (X_1, X_2, \dots, X_p)', \quad p \geq 2$$

$$\Pi_1 \equiv N_p(\mu_1, \Sigma)$$

Αιχμηροί
 και
 ίδια διασπορά

$$\Pi_2 \equiv N_p(\mu_2, \Sigma)$$

$$\mu_1, \mu_2, \Sigma = \text{ΑΓΝΟΣΤΑ}$$



$p=2$

$$X \sim N \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \text{Var} X_1 & \text{Cov} X_1 X_2 \\ \text{Cov} X_1 X_2 & \text{Var} X_2 \end{pmatrix}$$

\downarrow 2×1 \downarrow 2×1 \downarrow 2×2

Πρόβλημα ταξινόμησης

H_0 : Αθώος
 H_1 : Ένοχος

^{Κρίση}
 $H_0: X \in \Pi_1$

vs.

^{Υπόθεση}
 $H_1: X \in \Pi_2$

Έστω $f(x | \mu_i, \Sigma)$ η σ.η.π. του $\Pi_i \implies$

X ταξινομείται στον Π_1 αν

$$\eta = \frac{f(x | \mu_1, \Sigma)}{f(x | \mu_2, \Sigma)} > C$$

Credit Scoring

Λίγος πιθανοφάνεια
Likelihood Ratio Test

$f(x | \mu_1, \Sigma)$ για τον H_0
 $f(x | \mu_2, \Sigma)$ για τον H_1

$X \in \Pi_1$ αν $\frac{f(x | \mu_1, \Sigma)}{f(x | \mu_2, \Sigma)} > 1$

1.1
0.9

$$A_v \quad f \sim N_p(\mu_i, \Sigma) \Rightarrow \frac{\exp\{\dots\}}{\exp\{\dots\}} > C$$

op Titou I: $\lambda_{\text{θωος}} \rightarrow \text{logno (w)kath}$
 op Titou II: $\text{evoxh} \rightarrow \lambda_{\text{θωovno}}$
 Elyxapocovnoy = $\alpha = \begin{matrix} 0.01 \\ 0.08 \\ 0.10 \end{matrix}$

$$(X - \mu_1)' \Sigma^{-1} (X - \mu_1) < (X - \mu_2)' \Sigma^{-1} (X - \mu_2) + C \Rightarrow \exp\{\dots\} < C \cdot \exp\{\dots\}$$

$$-X' \Sigma^{-1} \mu_1 - \mu_1' \Sigma^{-1} X + \mu_1' \Sigma^{-1} \mu_1 < -X' \Sigma^{-1} \mu_2 - \mu_2' \Sigma^{-1} X + \mu_2' \Sigma^{-1} \mu_2 + C$$

$$-2\mu_1' \Sigma^{-1} X + \mu_1' \Sigma^{-1} \mu_1 < -2\mu_2' \Sigma^{-1} X + \mu_2' \Sigma^{-1} \mu_2 + C \Rightarrow$$

$$(\mu_1 - \mu_2)' \Sigma^{-1} X > \mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2 + \underbrace{\mu_1' \Sigma^{-1} \mu_2 - \mu_2' \Sigma^{-1} \mu_1}_{=0} + C$$

$$(\mu_1 - \mu_2)' \Sigma^{-1} X > (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) + C$$

$$\boxed{(\mu_1 - \mu_2)' \Sigma^{-1} X > C}$$

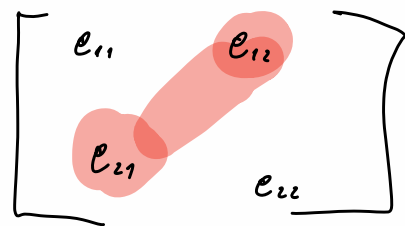
Το C υποδειγνται μέσω των σφαημάτων:

$$e_{12} = P[\underline{X \in \pi_1} / X \in \pi_2]$$

$$e_{21} = P[X \in \pi_2 / X \in \pi_1]$$

$$e_{12} = P[(\mu_1 - \mu_2)' \Sigma^{-1} X > c / X \in \pi_2]$$

$$e_{21} = P[(\mu_1 - \mu_2)' \Sigma^{-1} X \leq c / X \in \pi_1]$$



$$X \sim N(\mu, \sigma^2)$$

$$aX \sim N(a\mu, a^2\sigma^2)$$

$$X \in \Pi_1 \Rightarrow (\mu_1 - \mu_2)' \Sigma^{-1} X \sim N_p \left((\mu_1 - \mu_2)' \Sigma^{-1} \mu_2, \underbrace{(\mu_1 - \mu_2)' \Sigma^{-1} \Sigma \Sigma^{-1}}_{\Delta^2} (\mu_1 - \mu_2) \right)$$

$$X \in \Pi_2 \Rightarrow (\mu_1 - \mu_2)' \Sigma^{-1} X \sim N_p \left((\mu_1 - \mu_2)' \Sigma^{-1} \mu_2, \Delta^2 \right)$$

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\|X\|^2 = X'X$$

$$(zW)' = W'z$$

Apa

$$e_{12} = P\left[Z > \frac{C - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_2}{\Delta} \right] = \underline{1 - \Phi\left(\frac{C - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_2}{\Delta}\right)}$$

$$e_{21} = P\left[Z < \frac{C - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_1}{\Delta} \right] = \underline{\Phi\left(\frac{C - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_1}{\Delta}\right)}$$

Απαιτούμε:

→ Δόξη συμμετρίας

$$e_{12} = e_{21} \Rightarrow$$

$$1 - \Phi\left(\frac{c - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_2}{\Delta}\right) = \Phi\left(\frac{c - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_1}{\Delta}\right) \Rightarrow$$

$$\Phi\left(\frac{-c + (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_2}{\Delta}\right) \equiv \Phi\left(\frac{c - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_1}{\Delta}\right) \Rightarrow$$

$$-c + (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_2 \equiv c - (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} \mu_1 \Rightarrow$$

$$c = \frac{1}{2} (\mu_1 - \mu_2)' \bar{\Sigma}^{-1} (\mu_1 + \mu_2)$$

Αρα $X \in \Pi_1$ όταν

$$(\mu_1 - \mu_2)' \Sigma^{-1} X > \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)$$



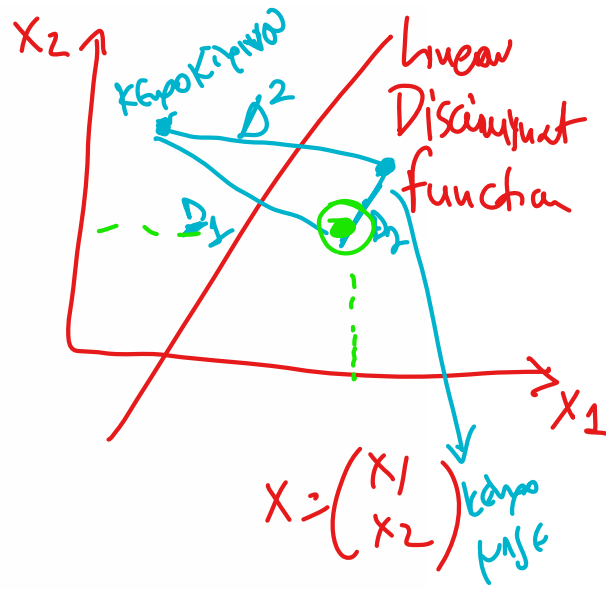
$$(X - \mu_1)' \Sigma^{-1} (X - \mu_1) < (X - \mu_2)' \Sigma^{-1} (X - \mu_2)$$

$$\| \Sigma^{-1/2} (X - \mu_1) \|^2 < \| \Sigma^{-1/2} (X - \mu_2) \|^2$$

$$\| \Delta_1 \|^2 < \| \Delta_2 \|^2$$

Απόσταση X από Π_1

Απόσταση X από Π_2



$$\|X\|^2 = X \cdot X$$

$$\| \Sigma^{-1/2} (X - \mu_1) \|^2 = (X - \mu_1)' \Sigma^{-1/2} \cdot \Sigma^{-1/2} (X - \mu_1)$$

Άρα η μικρότερη απόσταση οδηγεί στην παράσταση χ :

$$\chi \in \Pi_i \text{ αν } \Delta_L^2 = \min \{ \Delta_1^2, \Delta_2^2 \}, i=1,2.$$

Γενικώς σε k παραστάσεις:

$$\chi \in \Pi_i \text{ αν } \Delta_L^2 = \min \{ \Delta_1^2, \dots, \Delta_k^2 \}$$

Exams

ΕΧΘΛΙΟ

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) = \left\| \Sigma^{-1/2} (\mu_1 - \mu_2) \right\|^2$$

Απόσταση μεταξύ π_1 & π_2

= Mahalanobis Distance

λαμβάνει υπόψη της και τη 2η ροπή όχι μόνο την 1η (μέση τιμή)

Αν $\mu_1, \mu_2, \Sigma = \text{γνωστά} \Rightarrow$

$$Cov(X_1, X_2) = E\left(\begin{matrix} X_1 - \mu_1 \\ 1 \end{matrix}\right) (X_2 - \mu_2)$$

$$= \frac{1}{n} \sum_{i=1}^n (X_{i,1} - \bar{X}_1) (X_{i,2} - \bar{X}_2)$$

$$\left\{ (\hat{\mu}_1 - \hat{\mu}_2)' \bar{S}^{-1} X > \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' \bar{S}^{-1} (\hat{\mu}_1 + \hat{\mu}_2) \right\}$$

όπου

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_{i,1}, \quad \hat{\mu}_2 = \bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_{i,2}$$

$$\hat{E}X_1 = \frac{1}{n} \sum X_{i,1} = \bar{X}_1$$

$$\hat{E}X_2 = \frac{1}{n} \sum X_{i,2} = \bar{X}_2$$

$$S_1 = \frac{1}{n-1} \sum (X_{i,1} - \bar{X}) (X_{i,1} - \bar{X})', \quad S_2 = \frac{1}{m-1} \sum (X_{i,2} - \bar{X}) (X_{i,2} - \bar{X})'$$

Επίσης έχει
και τις ανισότητες
χρησιμότητας για απόδοσή
επιπέδου (MNH)
εξίσωσή του.

$$S = \frac{(n-1)S_1 + (m-1)S_2}{n+m-2} = \begin{pmatrix} \hat{Var}X_1 & \hat{Cov}X_1X_2 \\ \hat{Cov}X_1X_2 & \hat{Var}X_2 \end{pmatrix}$$

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \equiv d' \Sigma^{-1} d = \sum_i \sum_j s^{ij} d_j \cdot d_i = \sum_i l_i \cdot d_i$$

$$= d' \cdot l = l' \cdot d \Rightarrow (\mu_1 - \mu_2)' \Sigma^{-1} = \underline{l}' = \sum_j s^{ij} d_j$$

or for evaluation:

$$\hat{l}' = (\bar{x}_1 - \bar{x}_2) \cdot \hat{\Sigma}^{-1} = (\hat{\mu}_1 - \hat{\mu}_2)' S^{-1}$$

$$\hat{l}' \cdot \hat{x} > \frac{1}{2} (\hat{\mu}_1 - \hat{\mu}_2)' S^{-1} (\hat{\mu}_1 + \hat{\mu}_2) \Rightarrow$$

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2)' S^{-1} \hat{\mu}_1}{2} + \frac{(\hat{\mu}_1 - \hat{\mu}_2)' S^{-1} \hat{\mu}_2}{2} \Rightarrow$$

$$\hat{l}' \cdot \hat{x} > \frac{\hat{l}' \hat{\mu}_1 + \hat{l}' \hat{\mu}_2}{2} \approx$$

$$\hat{l}' \cdot \hat{x} > \frac{1}{2} (\hat{l}' \bar{x}_1 + \hat{l}' \bar{x}_2)$$

$$l = (\hat{\mu}_1 - \hat{\mu}_2) \cdot S^{-1}$$

$\tilde{X}_{11}, \dots, \tilde{X}_{1m_1}$ δείγμα από $N_p(\underline{\mu}_1, \Sigma)$.

$\tilde{X}_{21}, \dots, \tilde{X}_{2n_2}$ δείγμα από $N_p(\underline{\mu}_2, \Sigma)$.

$H_0: \tilde{X}, \tilde{X}_{11}, \dots, \tilde{X}_{1m_1} \in \Pi_1$ & $\tilde{X}_{21}, \dots, \tilde{X}_{2n_2} \in \Pi_2$

vs.

$H_0: \tilde{X}_{11}, \dots, \tilde{X}_{1m_1} \in \Pi_1$ & $\tilde{X}, \tilde{X}_{21}, \dots, \tilde{X}_{2n_2} \in \Pi_2$

2023

Παράδειγμα (Johnson and Wichern ~~2007~~, Example 11.3): $p = 2$, $n_1 = 30$, $n_2 = 22$, $X_1 =$

Αντιαμορροφιλικός Παράγοντας

Αντιγόνο τύπου AHF (μοιάζει με Αντιαμορροφιλικός Παράγοντας)

\log_{10} (AHF activity), $X_2 = \log_{10}$ (AHF - like antigen). Μας ενδιαφέρει η ανίχνευση
(υπεύθυνος πήξη αίματος)

φορέων hemophilia A στον γυναικείο πληθυσμό. Από τα δεδομένα προέκυψε ότι τα διανύσματα των δειγματικών μέσων τιμών για τους υπό-πληθυσμούς υγιών και μη-υγιών γυναικών είναι, αντίστοιχα,

$$\bar{\mathbf{x}}_1 = (-0.0065, -0.0390)', \bar{\mathbf{x}}_2 = (-0.2483, -0.0262)'$$

Άσκηση 11.32?

Επίσης, ο S_{pooled}^{-1} είναι ίσος με

$$S_{\text{pooled}}^{-1} \equiv S^{-1} = \begin{pmatrix} 131 & -90 \\ -90 & 108 \end{pmatrix},$$

και άρα

$$l. = (\bar{x}_1 - \bar{x}_2)' S^{-1} = S^{-1}(\bar{x}_1 - \bar{x}_2) = (37.5438, -28.8036)',$$

δηλ. η LDF είναι

$$l'x = 37.61x_1 - 28.92x_2,$$

με

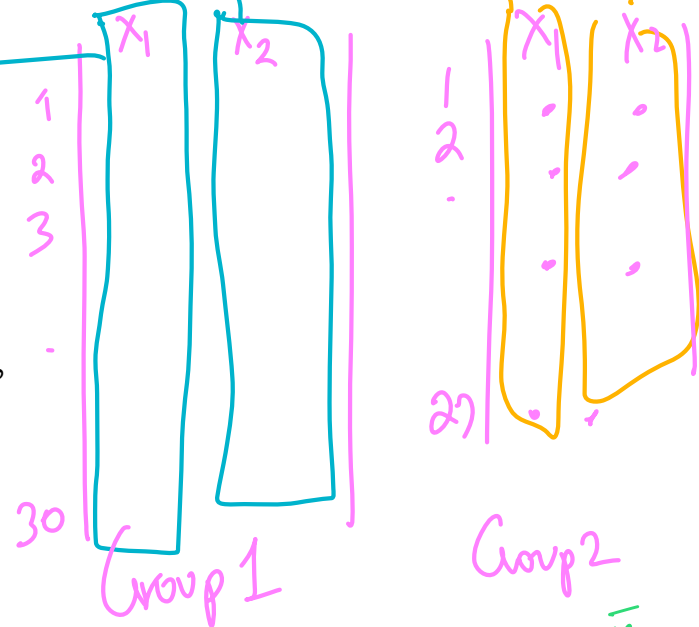
$$\bar{y}_1 = l'\bar{x}_1 = 0.8793, \bar{y}_2 = l'\bar{x}_2 = -10.0768 \text{ και } \bar{y} = -4.5987.$$

- Αν έρθει μια νέα παρατήρηση με διάνυσμα $x_0 = (-0.210, -0.044)'$, αυτή θα τοποθετηθεί στον Π2 (δηλ. η γυναίκα ταξινομείται ως φορέας της hemophilia A) αφού

$$l'x_0 = 37.54 \cdot (-0.210) - 28.92 \cdot (-0.044) = -6.6168 < -4.5987.$$

- Αξίζει να σημειωθεί ότι η παραπάνω εφαρμογή της LDF του Fisher, υποθέτει ίσα κόστη εσφαλμένης ταξινόμησης στους δύο πληθυσμούς αλλά και ίσες εκ των προτέρων πιθανότητες ταξινόμησης σε αυτούς.

$$l'x_0 = (37.54, -28.92) \cdot \begin{pmatrix} -0.210 \\ -0.044 \end{pmatrix} = -6.6168$$



$$\frac{l'\bar{x}_1 + l'\bar{x}_2}{2} = \frac{\bar{y}_1 + \bar{y}_2}{2}$$

- Έστω τώρα ότι οι πιθανότητες ταξινόμησης είναι γνωστές και ίσες με $p_1 = 0.75$ και $p_2 = 0.25$, αντίστοιχα. Έστω επίσης ότι τα κόστη εσφαλμένης ταξινόμησης είναι επίσης ίσα. Τότε, η παρατήρηση με διάνυσμα \mathbf{x}_0 ταξινομείται στον Π_1 αν

$$\mathbf{x}_0' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \frac{1}{2} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > \log(p_2/p_1).$$

Άρα, με άμεση αντικατάσταση, διαπιστώνουμε ότι

$$\mathbf{x}'_0 \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = -6.6168,$$

$$\frac{1}{2} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = -4.5987, \quad \log(p_2/p_1) = -1.0986$$

και τελικά, αφού

$$-6.6168 - (-4.5987) = -2.0181 \not\approx -1.0986,$$

η παρατήρηση αυτή θα ταξινομηθεί στον Π2.