Ινστιτούτο Σχεδιασμού και Ανάλυσης Πειραμάτων του Γ.Π.Α.

3ος Κύκλος Διαλέξεων-Απρίλιος 2024

Πολυμεταβλητή Ανάλυση στην Πράξη

Μ. Κούτρας
Καθηγητής Τμήματος Στατιστικής & Ασφαλιστικής Επιστήμης
Δντης ΠΜΣ στην Εφαρμοσμένη Στατιστική
Πανεπιστήμιο Πειραιώς

Διαχωριστική Ανάλυση

Εισαγωγικά

Διαχωριστική Ανάλυση: η βασική ιδέα

- Να κατατάξει άτομα σε έναν από πολλούς διαθέσιμους γνωστούς πληθυσμούς με βάση τις τιμές συγκεκριμένων χαρακτηριστικών.

- Οι (συνήθως πολυδιάστατες) παρατηρήσεις-χαρακτηριστικά υποτίθεται ότι ακολουθούν γνωστές κατανομές για κάθε διαθέσιμο πληθυσμό.

Discriminant analysis (διαχωριστική ή διακριτική ή ταξινομική ανάλυση)

Το γενικό μοντέλο

- Υποθέτουμε ότι έχουμε k γνωστούς πληθυσμούς $\Pi_1, \Pi_2, ..., \Pi_k$ με $k \geq 2$.
- Τα χαρακτηριστικά που μελετάμε περιγράφονται από μια p-διάστατη τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2, ..., X_p)$.
- Για κάθε πληθυσμό Π_i, $i = 1,2,..., k$, η p-διάστατη τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2, ..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_i(\mathbf{x}) = f_i(x_1, x_2, ..., x_p)$.

Εφαρμογές της ΔΑ:
Credit scoring (χρηματοοικονομικά)

- Οι τράπεζες ενδιαφέρονται να εντοπίσουν

  'καλούς' (Π_1) ή 'κακούς' (Π_2) πελάτες

  πριν τη χορήγηση δανείου ή κάποιας πιστωτικής κάρτας.

- Με χρήση ιστορικών στοιχείων (σχετικά με άτομα που έλαβαν δάνειο από την τράπεζα), η τράπεζα θα ήθελε να σχηματίσει κανόνες ώστε να κατατάξει έναν καινούριο πελάτη σε μια από τις δύο κατηγορίες

- Έτσι μπορεί πλέον να αποφασίσει αν θα χορηγήσει το δάνειο ή θα αρνηθεί τη χορήγηση του δανείου.

Εφαρμογές της ΔΑ:
έλεγχος καταλληλόλητας τροφίμων

- Κάθε τρόφιμο κατατάσσεται σε 2 κατηγορίες (πληθυσμοί Π_1 , Π_2):

  κατάλληλο προς χρήση ή ακατάλληλο (αλλοιωμένο).

  Η κάθε κατηγορία σχετίζεται με τις τιμές κάποιων χαρακτηριστικών ($X_1, X_2, \ldots, X_p$).

- Με χρήση δεδομένων από τρόφιμα τα οποία έχουν ήδη χαρακτηρισθεί (γνωρίζουμε σε ποιον ακριβώς πληθυσμό ανήκουν), θα μας ενδιέφερε να σχηματίσουμε κανόνες ώστε να κατατάσσουμε ένα τρόφιμο σε μια από τις *k*=2 κατηγορίες

- Όταν μας φέρνουν ένα νέο τρόφιμο, μπορούμε πλέον με βάση τις τιμές των χαρακτηριστικών $X_1, X_2, \ldots, X_p$ να αποφασίζουμε αν είναι κατάλληλο προς χρήση ή ακατάλληλο .

Εφαρμογές της ΔΑ:
έλεγχος ποιότητας τροφίμων

- Κάθε τρόφιμο κατατάσσεται σε διάφορες ποιότητες (πληθυσμοί Π_1 , Π_2, … , Π_k). Κάθε ποιότητα σχετίζεται με τις τιμές κάποιων χαρακτηριστικών (X_1, X_2,…, X_p).
- Με χρήση δεδομένων από τρόφιμα τα οποία έχουν ήδη καταταχθεί ως προς την ποιότητα (γνωρίζουμε σε ποιον ακριβώς πληθυσμό ανήκουν), θα μας ενδιέφερε να σχηματίσουμε κανόνες ώστε να κατατάσσουμε έναν τρόφιμο σε μια από τις *k* κατηγορίες
- Όταν μας φέρνουν ένα νέο προϊόν, μπορούμε πλέον με βάση τις τιμές των μετρήσιμων χαρακτηριστικών X_1, X_2 ,…, X_p να αποφασίζουμε ποια είναι η κατηγορία ποιότητας του τροφίμου.

Εφαρμογές της ΔΑ:
ταξινόμηση κρασιών σε ποικιλίες

- Τα κρασιά κατατάσσονται σε διάφορες ποικιλίες ή ποιότητες (πληθυσμοί Π_1 , Π_2, … , Π_k). Για κάθε ποικιλία έχουν καταγραφεί κάποια χαρακτηριστικά τα (X_1, X_2,…, X_p) οποία διαφέρουν από ποικιλία σε ποικιλία.
- Με χρήση δεδομένων από κρασιά τα οποία γνωρίζουμε σε ποια ποικιλία ανήκουν θα μας ενδιέφερε να σχηματίσουμε «λογικούς» κανόνες ώστε να κατατάσσουμε έναν κρασί σε μια από τις k κατηγορίες
- Όταν μας φέρνουν ένα νέο προϊόν, μπορούμε πλέον με βάση τις τιμές των μετρήσιμων χαρακτηριστικών X_1, X_2 ,…, X_p να αποφασίζουμε ποια είναι η ποικιλία του κρασιού.

Εφαρμογές της ΔΑ:
ταξινόμηση φυτών σε ποικιλίες

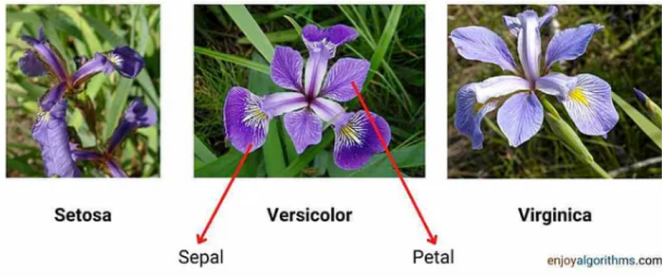
Fisher Iris Data: Δείγμα 30 παρατηρήσεων

Μεταβλητές

SL : Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL : Μήκος Πετάλου
PW:Πλάτος Πετάλου

Ποικιλίες

a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica



α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2
3	a	4.7	3.2	1.3	.2
4	a	4.6	3.1	1.5	.2
5	a	5.0	3.6	1.4	.2
6	a	5.4	3.9	1.7	.4
7	a	4.6	3.4	1.4	.3
8	a	5.0	3.4	1.5	.2
9	a	4.4	2.9	1.4	.2
10	a	4.9	3.1	1.5	.1
11	b	7.0	3.2	4.7	1.4
12	b	6.4	3.2	4.5	1.5
13	b	6.9	3.1	4.9	1.5
14	b	5.5	2.3	4.0	1.3
15	b	6.5	2.8	4.6	1.5
16	b	6.3	3.3	4.7	1.6
17	b	4.9	2.4	3.3	1.0
18	b	6.6	2.9	4.6	1.3
19	b	5.2	2.7	3.9	1.4
20	b	5.7	2.8	4.1	1.3
21	c	6.3	3.3	6.0	2.5
22	c	5.8	2.7	5.1	1.9
23	c	7.1	3.0	5.9	2.1
24	c	6.3	2.9	5.6	1.8
25	c	6.5	3.0	5.8	2.2
26	c	7.6	3.0	6.6	2.1
27	c	4.9	2.5	4.5	1.7
28	c	7.3	2.9	6.3	1.8
29	c	6.7	2.5	5.8	1.8
30	c	7.2	3.6	6.1	1.5

Άλλες εφαρμογές της ΔΑ

- Έλεγχος αυθεντικότητας τροφίμων
- Καθορισμός αν ένα καύσιμο είναι νοθευμένο ή όχι
- Ταξινόμηση αγροτικών εκτάσεων με βάση κάποια χαρακτηριστικά (εδάφους, κλίματος κτλ)
- Ταξινόμηση ζώων ή φυτών σε κατηγορίες με βάση κάποια μετρήσιμα χαρακτηριστικά

Σκοπός της διαχωριστικής ανάλυσης (ΔΑ)

- Να διαμορφώσει κάποιο κανόνα (διαχωριστικός κανόνα) με τον οποίο θα μπορούμε να κατανείμουμε κάθε παρατήρηση σε έναν από του τους k διαθέσιμους πληθυσμούς.
- Ο κανόνας θα βασίζεται σε κάποιες συναρτήσεις του $\mathbf{X}' = (X_1, X_2, ..., X_p)$.
- Ψάχνουμε για ένα διαχωριστικό κανόνα που μπορεί να κατατάξει σωστά όσο το δυνατόν περισσότερες παρατηρήσεις.

ΔΑ για $k=2$ πληθυσμούς

Χαρακτηριστικά που μελετάμε: $\mathbf{X}'=(X_1, X_2, ..., X_p)$ (p-διάστατη τυχαία μεταβλητή)

Πληθυσμός Π_1: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}'=(X_1, X_2, ..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_1(\mathbf{x})=f_1(x_1, x_2, ..., x_p)$ πχ $\mathbf{X} \sim N(\boldsymbol{\mu}_1, \Sigma_1)$.

Πληθυσμός Π_2: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}'=(X_1, X_2, ..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_2(\mathbf{x})=f_2(x_1, x_2, ..., x_p)$ πχ $\mathbf{X} \sim N(\boldsymbol{\mu}_2, \Sigma_2)$.

Η συνάρτηση απόφασης

Λήψη απόφασης: Θα χρησιμοποιήσουμε μια συνάρτηση $g(\mathbf{x})$ (ή και περισσότερες, αν χρειάζεται) και με βάση την τιμή της θα αποφασίσουμε.

Για Παράδειγμα:

$$R_1 = \{\mathbf{x} : g(\mathbf{x}) > c\}$$

- αν $g(\mathbf{x}) > c$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_1
- αν $g(\mathbf{x}) \le c$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_2

$$R_2 = \{\mathbf{x} : g(\mathbf{x}) \le c\}$$

Συνολική πιθανότητα λανθασμένης ταξινόμησης (total probability of misclassification)

Δική μας απόφαση

Πραγματικότητα	Π_1	Π_2
Π_1	$P(1\|1)$	$P(2\|1)$
Π_2	$P(1\|2)$	$P(2\|2)$

$$TPM = P(2|1) + P(1|2) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$

Η συνολική πιθανότητα λανθασμένης ταξινόμησης

$$TPM = 1 - \int_{R_1} [f_1(\mathbf{x}) - f_2(\mathbf{x})] d\mathbf{x}$$

Επομένως, ο κανόνας που οδηγεί στην **ελαχιστοποίηση** της *TPM* καθορίζεται επιλέγοντας το R_1 ως

$$R_1 : \mathbf{x} \text{ τέτοια ώστε να ισχύει } f_1(\mathbf{x}) - f_2(\mathbf{x}) > 0 .$$

Ένας απλός κανόνας απόφασης

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1$$

- αν $f_1(\mathbf{x}) > f_2(\mathbf{x})$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_1
- αν $f_1(\mathbf{x}) \leq f_2(\mathbf{x})$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_2

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \leq 1$$

Παράδειγμα

X: ο αριθμός των σημαδιών που παρατηρούνται σε ένα φυτό λόγω προσβολής από κάποια αρρώστια

Π_1: υγιή φυτά. Ο αριθμός σημαδιών σε αυτά ακολουθεί κατανομή Poisson με μέση τιμή λ_1

Π_2: άρρωστα φυτά. Ο αριθμός σημαδιών σε αυτά ακολουθεί κατανομή Poisson με μέση τιμή λ_2 $(\lambda_2 >> \lambda_1)$

Στόχος: Να αποφασίσουμε αν ένα φυτό είναι άρρωστο ή όχι

Πως θα αποφασίσουμε: Μετρώντας τον αριθμό X των σημαδιών που έχει

Παράδειγμα

Π_1: Υγιή φυτά $\quad f_1(x) = e^{-\lambda_1}\frac{\lambda_1^x}{x!}$

Π_2: Άρρωστα φυτά $\quad f_2(x) = e^{-\lambda_2}\frac{\lambda_2^x}{x!}$

Έχουμε

$$\frac{f_1(x)}{f_2(x)} = e^{-(\lambda_1-\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^x$$

οπότε

$$\frac{f_1(x)}{f_2(x)} > 1 \Leftrightarrow e^{-(\lambda_1-\lambda_2)}\left(\frac{\lambda_1}{\lambda_2}\right)^x > 1 \Leftrightarrow x < (\lambda_1-\lambda_2)/\ln(\frac{\lambda_1}{\lambda_2})$$

Παράδειγμα

Κανόνας απόφασης

Αν $X=x$ το άτομο ταξινομείται στον πληθυσμό

- Π_1 αν ισχύει $x < (\lambda_1 - \lambda_2)/\ln(\frac{\lambda_1}{\lambda_2})$
- Π_2 αν ισχύει $x \geq (\lambda_1 - \lambda_2)/\ln(\frac{\lambda_1}{\lambda_2})$

Παράδειγμα

- Π_1 αν ισχύει $x < (\lambda_1 - \lambda_2)/\ln(\frac{\lambda_1}{\lambda_2})$
- Π_2 αν ισχύει $x \geq (\lambda_1 - \lambda_2)/\ln(\frac{\lambda_1}{\lambda_2})$

	$\lambda_1 = 2$	$\lambda_2 = 7$
x	$f(x) = P(X = x)$	$f(x) = P(X = x)$
0	0.135	0.001
1	0.271	0.006
2	0.271	0.022
3	0.180	0.052
4	0.090	0.091
5	0.036	0.128
6	0.012	0.149
7	0.003	0.149
8	0.001	0.130
9	0.000	0.101
10	0.000	0.071
11	0.000	0.045
12	0.000	0.026
13	0.000	0.014
14	0.000	0.007
15	0.000	0.003
	1.00	1.00

cut point= 3.991178

Ένα πιο ρεαλιστικό μοντέλο

Έστω ότι θέλουμε να ταξινομήσουμε άτομα ενός πληθυσμού (φυτά, ζώα κτλ) ως προς την ύπαρξη ή μη μιας ασθένειας σε αυτό.

Π_1: Υγιές άτομο (ως προς την ασθένεια)

Π_2: Ασθενές άτομο (ως προς την ασθένεια)

Ας υποθέσουμε ακόμη ότι

- Το ποσοστό των ατόμων του (συνολικού) πληθυσμού που πάσχουν από την ασθένεια είναι μόλις 10%.
- Το κόστος που προκύπτει αν δεν εντοπίσουμε έναν ασθενή είναι 20 φορές μεγαλύτερο από το κόστος να μην εντοπίσουμε σωστά έναν υγιή.

Ένα πιο ρεαλιστικό μοντέλο: Πρόσθετοι συμβολισμοί

Δική μας απόφαση

Πραγματικότητα	Π_1	Π_2
$\Pi_1(p_1)$	$P(1\|1)$	$P(2\|1)$
$\Pi_2(p_2)$	$P(1\|2)$	$P(2\|2)$

p_i :
εκ των προτέρων πιθανότητα μια παρατήρηση **x** να προέρχεται από τον πληθυσμό Π_i

Πίνακας κόστους

Δική μας απόφαση

Πραγματικότητα	Π_1	Π_2
$\Pi_1(p_1)$	0	$C(2\|1)$
$\Pi_2(p_2)$	$C(1\|2)$	0

$C(i \mid j)$:
το κόστος που προκύπτει αν κατατάξουμε την παρατήρηση στον πληθυσμό Π_i ενώ η παρατήρηση ανήκει στον πληθυσμό Π_j

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

Ελαχιστοποίηση του *ECM* : ο κανόνας απόφασης

Επομένως

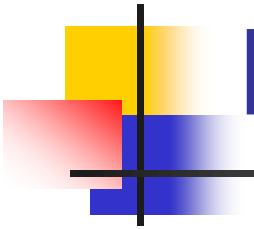
$$R_1 : \mathbf{x} \text{ τέτοια ώστε να ισχύει } C(2|1)p_1 f_1(\mathbf{x}) - C(1|2)p_2 f_2(\mathbf{x}) > 0$$

ή ισοδύναμα

$$R_1 : \mathbf{x} \text{ τέτοια ώστε να ισχύει } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{C(1|2)p_2}{C(2|1)p_1}$$

δηλαδή

- αν $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{C(1|2)p_2}{C(2|1)p_1}$ η παρατήρηση ταξινομείται στον πληθυσμό Π_1
- αν $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \leq \frac{C(1|2)p_2}{C(2|1)p_1}$ η παρατήρηση ταξινομείται στον πληθυσμό Π_2



Παράδειγμα

Έστω ότι θέλουμε να ταξινομήσουμε άτομα ενός πληθυσμού ως προς την ύπαρξη ή μη μιας ασθένειας.

Π_1: Υγιές άτομο (ως προς την ασθένεια)

Π_2: Ασθενές άτομο (ως προς την ασθένεια)

Η ταξινόμηση θα γίνει με βάση το αποτέλεσμα X κάποιας εξέτασης που γίνεται στο προς ταξινόμηση άτομο.

Παράδειγμα

Ας υποθέσουμε ακόμη ότι

- Το ποσοστό των ατόμων του (συνολικού) πληθυσμού που πάσχουν από την ασθένεια είναι 10% ($p_1 = 0.90, p_2 = 0.10$).
- Το κόστος που προκύπτει αν δεν εντοπίσουμε έναν ασθενή είναι 20 φορές μεγαλύτερο από το κόστος να μην εντοπίσουμε σωστά έναν υγιή ($C(1|2) = 20C(2|1)$).
- $X \sim N(\mu_1, \sigma_1^2)$ όταν το άτομο είναι υγιές
- $X \sim N(\mu_2, \sigma_2^2)$ όταν το άτομο ασθενεί.

Παράδειγμα 3

Π_1: Κακός πελάτης

$$f_1(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2}$$

Π_2: Καλός πελάτης

$$f_2(x) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2}$$

Έχουμε

$$\frac{f_1(x)}{f_2(x)} = \frac{\sigma_2}{\sigma_1} \exp\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 + \frac{1}{2\sigma_2^2}(x-\mu_2)^2\}$$

$$\frac{f_1(x)}{f_2(x)} = \frac{\sigma_2}{\sigma_1}\exp\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2 + \frac{1}{2\sigma_2^2}(x-\mu_2)^2\}$$

Κανόνας απόφασης

$g(x)$

$$\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)x^2 - 2x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right)+\left(\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right) < -2\ln\left[\frac{\sigma_1}{\sigma_2}\frac{C(1|2)p_2}{C(2|1)p_1}\right]$$

c

Κανόνας απόφασης

Αφού $p_1 = 0.10, p_2 = 0.9$ και $C(2|1) = 20C(1|2)$ προκύπτει ότι

$$c = -2\ln\frac{9\sigma_1}{20\sigma_2}.$$

$\mu_1, \mu_2, \sigma_1, \sigma_2$?

$$g(x) = \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x^2 - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right)$$

$$c = -2\ln\left[\frac{\sigma_1}{\sigma_2}\frac{C(1|2)p_2}{C(2|1)p_1}\right]$$

- αν $g(x) < c$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_1 (Υγιές άτομο)
- αν $g(x) \geq c$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_2 (Ασθενές άτομο)

Ειδική περίπτωση Ι

Έστω ότι ισχύει

$$p_1 = p_2 = 0.5 \text{ και } C(1|2) = C(2|1).$$

Αν πάρουμε $\mu_1 = 0, \mu_2 = 1, \sigma_1 = 1, \sigma_2 = 1/2$ τότε θα έχουμε

$$g(x) = -3x^2 + 8x - 4 \text{ (τετραγωνική συνάρτηση)}$$
$$\mathrm{c} = -2\ln 2 = -1.4$$

και προκύπτει ο κανόνας

- αν $-3x^2 + 8x - 4 < -1.4$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_1
- αν $-3x^2 + 8x - 4 \geq -1.4$ τότε η παρατήρηση ταξινομείται στον πληθυσμό Π_2

Ειδική περίπτωση ΙΙ

Αν ισχύει $\sigma_1 = \sigma_2$ τότε θα έχουμε

$$g(x) = -2x\left(\frac{\mu_1 - \mu_2}{\sigma}\right) + \left(\frac{\mu_1^2 - \mu_2^2}{\sigma}\right) = -2a(x - b)$$

(γραμμική συνάρτηση)

$$c = -2\ln\left[\frac{C(1|2)p_2}{C(2|1)p_1}\right]$$

όπου

$$a = \frac{\mu_1 - \mu_2}{\sigma}\,,\; b = \frac{\mu_1 + \mu_2}{2}$$

Η πολυδιάστατη περίπτωση

Πληθυσμός Π_1: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2, ..., X_p)$ ακολουθεί την $N(\boldsymbol{\mu}_1, \Sigma_1)$.

Πληθυσμός Π_2: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}' = (X_1, X_2, ..., X_p)$ ακολουθεί την $N(\boldsymbol{\mu}_2, \Sigma_2)$.

Για ένα άτομο παρατηρούμε την τυχαία μεταβλητή $\mathbf{X}$ και διαπιστώνουμε ότι $\mathbf{X} = \mathbf{x}$.

Ερώτημα: Το άτομο προέρχεται από τον Πληθυσμό Π_1 ή από τον Πληθυσμό Π_2;

Ταξινόμηση σε έναν από δύο κανονικούς πληθυσμούς: Ομοσκεδαστικοί πληθυσμοί

Αφού $\Sigma_1 = \Sigma_2 = \Sigma$ θα έχουμε

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

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

Κανόνας απόφασης

Ταξινόμησε μια παρατήρηση στον πληθυσμό Π_1 αν και μόνο αν

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > \ln c$$

όπου

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

Ταξινομικός κανόνας της ελάχιστης απόστασης

Όταν c =0 η τελευταία συνθήκη ισχύει αν και μόνο αν

$$(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) < (\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)$$

Ισοδύναμα, αν συμβολίσουμε με

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\Sigma^{-1}(\mathbf{x}-\mathbf{y})}$$

τη λεγόμενη **απόσταση Mahalanobis** μεταξύ των **x** και **y**, καταλήγουμε στον κανόνα:

Ταξινομικός κανόνας της ελάχιστης απόστασης

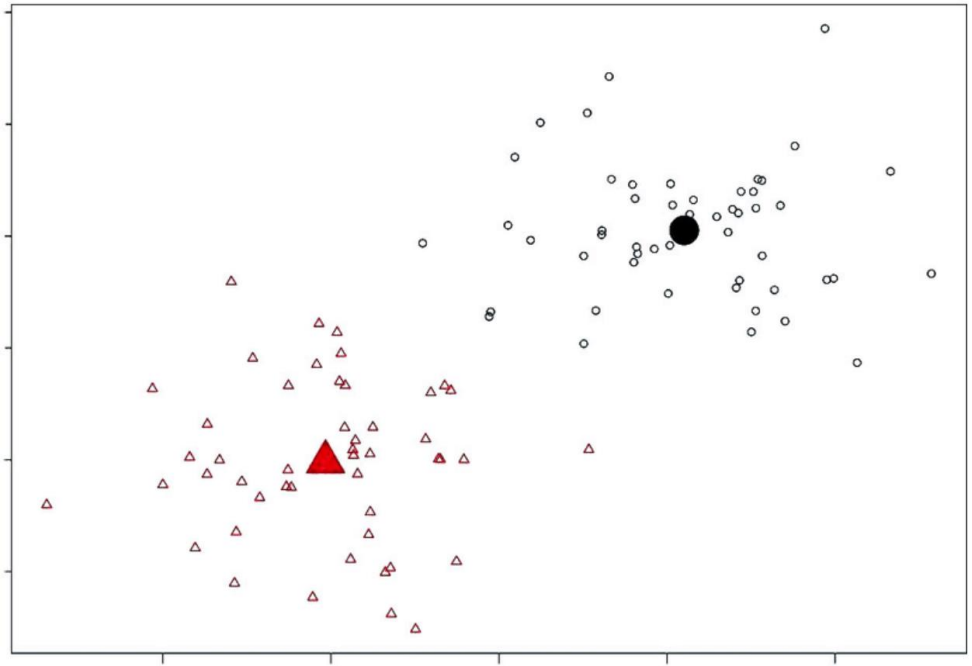
Ταξινόμησε μια παρατήρηση στον πληθυσμό Π_1 αν και μόνο αν

$$d(\mathbf{x},\boldsymbol{\mu}_1) < d(\mathbf{x},\boldsymbol{\mu}_2)$$

Οι πιθανότητες λανθασμένης ταξινόμησης

$$P(2|1) = \Phi\left(\frac{\ln c - (\delta^2/2)}{\sqrt{\delta^2}}\right) \qquad P(1|2) = 1 - \Phi\left(\frac{\ln c + (\delta^2/2)}{\sqrt{\delta^2}}\right)$$

$$\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$



Συνολική πιθανότητα λανθασμένης ταξινόμησης

$$TPM = P(2|1) + P(1|2) = 1 + \Phi\left(\frac{\ln c - (\delta^2/2)}{\sqrt{\delta^2}}\right) - \Phi\left(\frac{\ln c + (\delta^2/2)}{\sqrt{\delta^2}}\right)$$

Ειδική περίπτωση: $p_1 = p_2 = 0.5$ και $C(1|2) = C(2|1)$.

Τότε $c = 1$ οπότε

$$P(2|1) = \Phi\left(-\frac{\delta}{2}\right) = 1 - \Phi\left(\frac{\delta}{2}\right), \qquad P(1|2) = 1 - \Phi\left(\frac{\delta}{2}\right)$$

$$TPM = P(2|1) + P(1|2) = 2\left[1 - \Phi\left(\frac{\delta}{2}\right)\right]$$

Ταξινόμηση σε έναν από δύο κανονικούς πληθυσμούς: Μη ομοσκεδαστικοί πληθυσμοί

$\Sigma_1 \neq \Sigma_2$

$$f_1(\mathbf{x}) = \frac{1}{|2\pi\Sigma_1|^{1/2}} \exp\left\{-\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1))\right\}$$

$$f_2(\mathbf{x}) = \frac{1}{|2\pi\Sigma_2|^{1/2}} \exp\left\{-\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2))\right\}$$

Κανόνας απόφασης

Ταξινόμησε μια παρατήρηση στον πληθυσμό Π_1 αν και μόνο αν

$$\mathbf{x}' A\mathbf{x} + \mathbf{b}'\mathbf{x} + \gamma > \ln c$$

όπου

$$A = -\frac{1}{2}(\Sigma_1^{-1} - \Sigma_2^{-1}), \quad \mathbf{b}' = \boldsymbol{\mu}_1'\Sigma_1^{-1} - \boldsymbol{\mu}_2'\Sigma_2^{-1}$$

$$\gamma = -\frac{1}{2}(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2) + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|}, \qquad c = \frac{C(1|2)p_2}{C(2|1)p_1}.$$

Έλεγχος της αποτελεσματικότητας του μοντέλου

Ένας απλό τρόπος είναι να χρησιμοποιήσουμε το μοντέλο που διαμορφώσαμε για να κατατάξουμε τις παρατηρήσεις του δείγματος και να βρούμε το ποσοστό των σωστών κατατάξεων. Στην περίπτωση της τέλειας διαμέρισης περιμένουμε το ποσοστό των σωστών προβλέψεων να είναι 1. Πρακτικά αυτό δεν θα συμβεί ποτέ.

Το πρόβλημα που έχει αυτή η προσέγγιση είναι ότι χρησιμοποιούμε τις ίδιες παρατηρήσεις για να φτιάξουμε το μοντέλο και για να δούμε την αποτελεσματικότητά του. Με αυτό τον τρόπο συνήθως **υπερεκτιμάμε την αποτελεσματικότητα του μοντέλου**.

Έλεγχος της αποτελεσματικότητας του μοντέλου: cross-validation

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε την εξής τεχνική (cross-validation) για να εκτιμήσουμε το ποσοστό επιτυχίας της ανάλυσης.

- Χωρίζουμε το δείγμα σε δύο κομμάτια, το ένα είναι το δείγμα εκμάθησης (**training set**) και το άλλο το δείγμα επικύρωσης (**test set**).
- Βρίσκουμε το μοντέλο ταξινόμησης (κανόνα απόφασης) χρησιμοποιώντας τις παρατηρήσεις της πρώτης ομάδας
- Ταξινομούμε τις παρατηρήσεις του δείγματος επικύρωσης
- Υπολογίζουμε το ποσοστό των παρατηρήσεων του δείγματος επικύρωσης οι οποίες ταξινομούνται σωστά από τον κανόνα

Έλεγχος της αποτελεσματικότητας του μοντέλου: cross-validation

Σημείωση: Μπορούμε να επαναλάβουμε τη διαδικασία πολλές φορές και να υπολογίζουμε τον μέσο όρο των πιθανοτήτων ορθής ταξινόμησης.

Εναλλακτική μέθοδος: jackknife (παραλείπουμε κάθε φορά μία παρατήρηση και στη συνέχεια την ταξινομούμε)

Ταξινόμηση σε έναν από k≥2 πληθυσμούς

Πληθυσμός Π_1: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}'=(X_1, X_2,..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_1(\mathbf{x}) = f_1(x_1, x_2,..., x_p)$ πχ $\mathbf{X} \sim N(\boldsymbol{\mu}_1, \Sigma_1)$.

Πληθυσμός Π_2: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}'=(X_1, X_2,..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_2(\mathbf{x}) = f_2(x_1, x_2,..., x_p)$ πχ $\mathbf{X} \sim N(\boldsymbol{\mu}_2, \Sigma_2)$.

Πληθυσμός Π_3: Περιέχει άτομα για τα οποία η p-διάστατη τυχαία μεταβλητή $\mathbf{X}'=(X_1, X_2,..., X_p)$ ακολουθεί μία κατανομή με από κοινού συνάρτηση πυκνότητας $f_3(\mathbf{x}) = f_2(x_1, x_2,..., x_p)$ πχ $\mathbf{X} \sim N(\boldsymbol{\mu}_3, \Sigma_3)$.

Για ένα άτομο παρατηρούμε την τυχαία μεταβλητή $\mathbf{X}$ και διαπιστώνουμε ότι $\mathbf{X} = \mathbf{x}$.

Ερώτημα: Το άτομο προέρχεται από τον Πληθυσμό Π_1, από τον Πληθυσμό Π_2 ή από τον Πληθυσμό Π_3;

Ταξινόμηση σε έναν από k≥2 πληθυσμούς

Αναμενόμενο κόστος λανθασμένης ταξινόμησης μίας παρατήρησης που προέρχεται

- από τον πληθυσμό Π_1: $ECM_1 = C(2|1)P(2|1) + C(3|1)P(3|1)$
- από τον πληθυσμό Π_2: $ECM_2 = C(1|2)P(1|2) + C(3|2)P(3|2)$
- από τον πληθυσμό Π_3: $ECM_3 = C(1|3)P(1|3) + C(2|3)P(2|3)$

Συνολικό αναμενόμενο κόστος λανθασμένης ταξινόμησης μίας παρατήρησης:

$$ECM = p_1 ECM_1 + p_2 ECM_2 + p_3 ECM_3$$

Εύρεση των περιοχών κατάταξης στους 3 πληθυσμούς:

Ελαχιστοποίηση του *ECM*

Ταξινόμηση σε έναν από k≥2 πληθυσμούς: ειδική περίπτωση

Αν έχουμε ίσα κόστη λανθασμένων ταξινομήσεων και $p_1 = p_2 = p_3$, η επιλογή του πληθυσμού στο οποίο θα καταταχθεί μια παρατήρηση **x** γίνεται κοιτάζοντας για ποιόν πληθυσμό i_0 ισχύει

$$f_{i_0}(\mathbf{x}) = \max\left\{f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})\right\}.$$

Αντίστοιχα αποτελέσματα ισχύουν για k πληθυσμούς.

Διαχωριστική Ανάλυση

Μία μελέτη περίπτωσης (case study)

Fisher Iris Data

N=150 παρατηρήσεις

Ποικιλίες

a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica

Μεταβλητές

SL: Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL: Μήκος Πετάλου
PW: Πλάτος Πετάλου

Fisher Iris Data: Δείγμα 30 παρατηρήσεων

Μεταβλητές

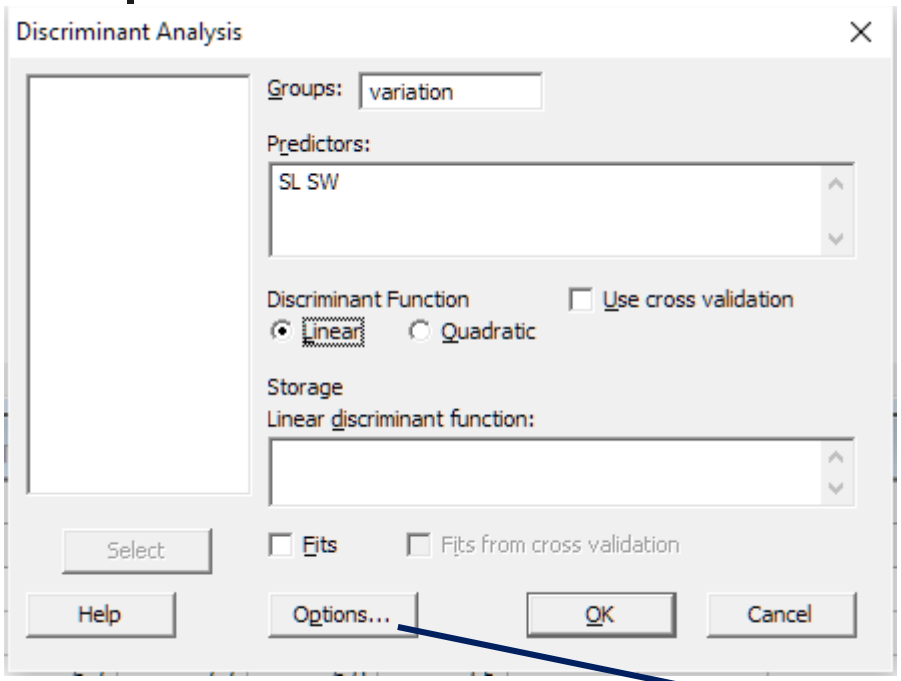
SL : Μήκος Σεπάλου
SW: Πλάτος Σεπάλου
PL : Μήκος Πετάλου
PW:Πλάτος Πετάλου

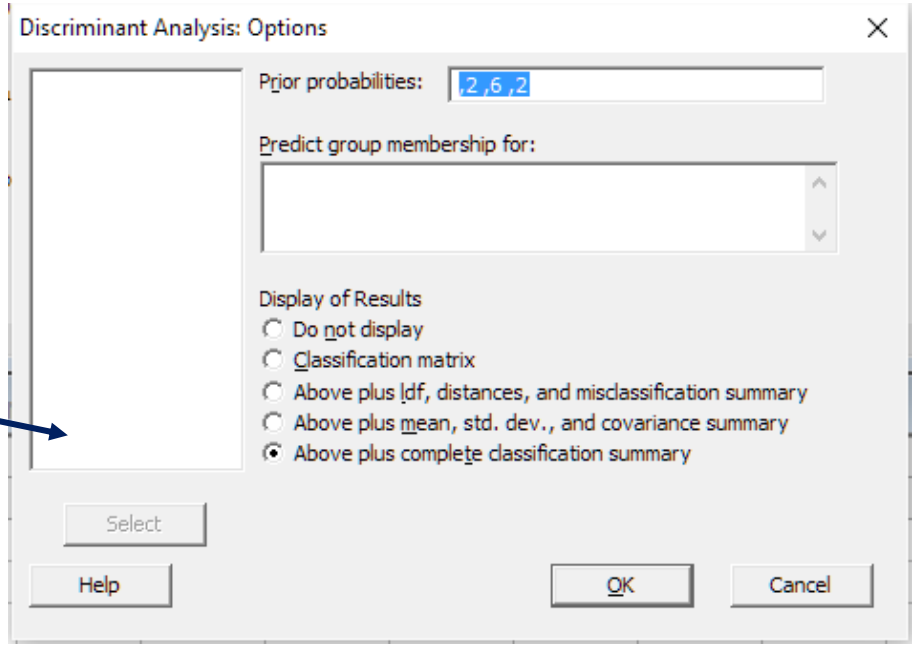
Ποικιλίες

a: Iris Setosa
b: Iris Versicolor
c: Iris Virginica

α/α	Ποικιλία	SL	SW	PL	PW
1	a	5.1	3.5	1.4	.2
2	a	4.9	3.0	1.4	.2
3	a	4.7	3.2	1.3	.2
4	a	4.6	3.1	1.5	.2
5	a	5.0	3.6	1.4	.2
6	a	5.4	3.9	1.7	.4
7	a	4.6	3.4	1.4	.3
8	a	5.0	3.4	1.5	.2
9	a	4.4	2.9	1.4	.2
10	a	4.9	3.1	1.5	.1
11	b	7.0	3.2	4.7	1.4
12	b	6.4	3.2	4.5	1.5
13	b	6.9	3.1	4.9	1.5
14	b	5.5	2.3	4.0	1.3
15	b	6.5	2.8	4.6	1.5
16	b	6.3	3.3	4.7	1.6
17	b	4.9	2.4	3.3	1.0
18	b	6.6	2.9	4.6	1.3
19	b	5.2	2.7	3.9	1.4
20	b	5.7	2.8	4.1	1.3
21	c	6.3	3.3	6.0	2.5
22	c	5.8	2.7	5.1	1.9
23	c	7.1	3.0	5.9	2.1
24	c	6.3	2.9	5.6	1.8
25	c	6.5	3.0	5.8	2.2
26	c	7.6	3.0	6.6	2.1
27	c	4.9	2.5	4.5	1.7
28	c	7.3	2.9	6.3	1.8
29	c	6.7	2.5	5.8	1.8
30	c	7.2	3.6	6.1	1.5

MINITAB


Discriminant Analysis
Groups: variation
Predictors:
SL SW
Discriminant Function
Linear
Quadratic
Use cross validation
Storage
Linear discriminant function:
Select
Fits
Fits from cross validation
Help
Options...
OK
Cancel


Discriminant Analysis: Options
Prior probabilities: ,2 ,6 ,2
Predict group membership for:
Display of Results
Do not display
Classification matrix
Above plus ldf, distances, and misclassification summary
Above plus mean, std. dev., and covariance summary
Above plus complete classification summary
Select
Help
OK
Cancel

MINITAB output (εκτιμήσεις)

Variable	Pooled Mean	Means for Group a	b	c
SL	5,8433	4,8600	6,1000	6,5700
SW	3,0400	3,3100	2,8700	2,9400

Variable	Pooled StDev	StDev for Group a	b	c
SL	0,6482	0,2914	0,7272	0,8042
SW	0,3285	0,3071	0,3401	0,3373

Pooled Covariance Matrix

	SL	SW
SL	0,4202	
SW	0,1343	0,1079

MINITAB output (Linear)

Discriminant Analysis: variation vs SL; SW

Linear Method for Response: variation
Predictors: SL; SW

Group	a	b	c
Count	10	10	10

Summary of classification

	True Group		
Put into Group	a	b	c
a	10	0	0
b	0	6	4
c	0	4	6
Total N	10	10	10
N correct	10	6	6
Proportion	1,000	0,600	0,600

N = 30 N Correct = 22
Proportion Correct = 0,733

MINITAB output (LDF)

Linear Discriminant Function for Groups

	a	b	c
Constant	-51,843	-50,794	-56,793
SL	2,929	9,992	11,505
SW	27,025	14,159	12,925

Η ταξινόμηση γίνεται με βάση τον **κανόνα της ελάχιστης απόστασης** (του Mahalanobis). Η LDF αποτελεί ένα μέρος της απόστασης αυτής το οποίο μεγιστοποιείται όταν ελαχιστοποιείται το τετράγωνο της απόστασης. Επομένως η παρατήρηση κατατάσσεται στην ομάδα με τη **μεγαλύτερη τιμή της LDF**

MINITAB output (Linear)

Discriminant Analysis: variation versus SL; SW; PL; PW

Linear Method for Response: variation

Predictors: SL; SW; PL; PW

Group	a	b	c
Count	10	10	10

Summary of classification

	True Group		
Put into Group	a	b	c
a	10	0	0
b	0	10	0
c	0	0	10
Total N	10	10	10
N correct	10	10	10
Proportion	1,000	1,000	1,000

N = 30 N Correct = 30

Proportion Correct = 1,000

MINITAB output (Quadratic)

Discriminant Analysis: variation versus SL;SW; PL; PW

Quadratic Method for Response:variation

Predictors: SL; SW; PL; PW

Group	a	b	c
Count	10	10	10

Summary of classification

	True Group		
Put into Group	a	b	c
a	10	0	0
b	0	10	0
c	0	0	10
Total N	10	10	10
N correct	10	10	10
Proportion	1,000	1,000	1,000

N = 30 N Correct = 30

Proportion Correct = 1,000