ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ

ΚΡΗΤΙΚΟΥ ΚΑΤΕΡΙΝΑ
ΝΙΚΑΚΗ ΚΑΤΕΡΙΝΑ
ΝΙΚΟΛΑΪΔΟΥ ΧΡΥΣΑ

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ

Είναι τεχνικές που έχουν σκοπό:

- ✓ τον εντοπισμό χαρακτηριστικών των οποίων οι αριθμητικές τιμές επιτυγχάνουν όσο το δυνατόν καλύτερο διαχωρισμό των παρατηρήσεων σε 2 ή περισσότερες κλάσεις
- ✓ Την εύρεση ενός κανόνα για την ένταξη νέων παρατηρήσεων σε μια από τις κλάσεις.

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

- Καλούμε τις κλάσεις π_1 και π_2.
- Ταξινομούμε τις παρατηρήσεις με βάση τις μετρήσεις p χαρακτηριστικών που εκφράζονται με το διάνυσμα:$X' = [X_1, X_2, \ldots, X_p]$ όπου $X_1, X_2, \ldots, X_p$ τυχαίες μεταβλητές.
- Οι 2 πληθυσμοί έχουν συναρτήσεις πυκνότητας πιθανότητας(σ.π.π.) $f_1(x)$ και $f_2(x)$ αντίστοιχα.
- Το σύνολο όλων των δυνατών μετρήσεων διαχωρίζεται σε 2 περιοχές R_1 και R_2.Αν η νέα παρατήρηση κείται στην περιοχή R_1 τότε εντάσσεται στον πληθυσμό π_1 διαφορετικά στον πληθυσμό π_2.

ΠΑΡΑΔΕΙΓΜΑ

Θεωρούμε 2 ομάδες σε μία πόλη:

Π_1:άτομα που κατέχουν μηχανή θερισμού,

Π_2:άτομα που δεν την κατέχουν.

Ένας κατασκευαστής θέλει να ταξινομήσει τις οικογένειες με βάση τα:

X_1:εισόδημα και X_2:διαθέσιμο κομμάτι γης. Λαμβάνουμε τυχαίο δείγμα $n_1=12$ και $n_2=12$ κάτοχοι και μη κάτοχοι αντίστοιχα.

TABLE 11.1

π_1: Riding-mower owners		π_2: Nonowners	
x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)	x_1 (Income in \$1000s)	x_2 (Lot size in 1000 ft^2)
60.0	18.4	75.0	19.6
85.5	16.8	52.8	20.8
64.8	21.6	64.8	17.2
61.5	20.8	43.2	20.4
87.0	23.6	84.0	17.6
110.1	19.2	49.2	17.6
108.0	17.6	59.4	16.0
82.8	22.4	66.0	18.4
69.0	20.0	47.4	16.4
93.0	20.8	33.0	18.8
51.0	22.0	51.0	14.0
81.0	20.0	63.0	14.8

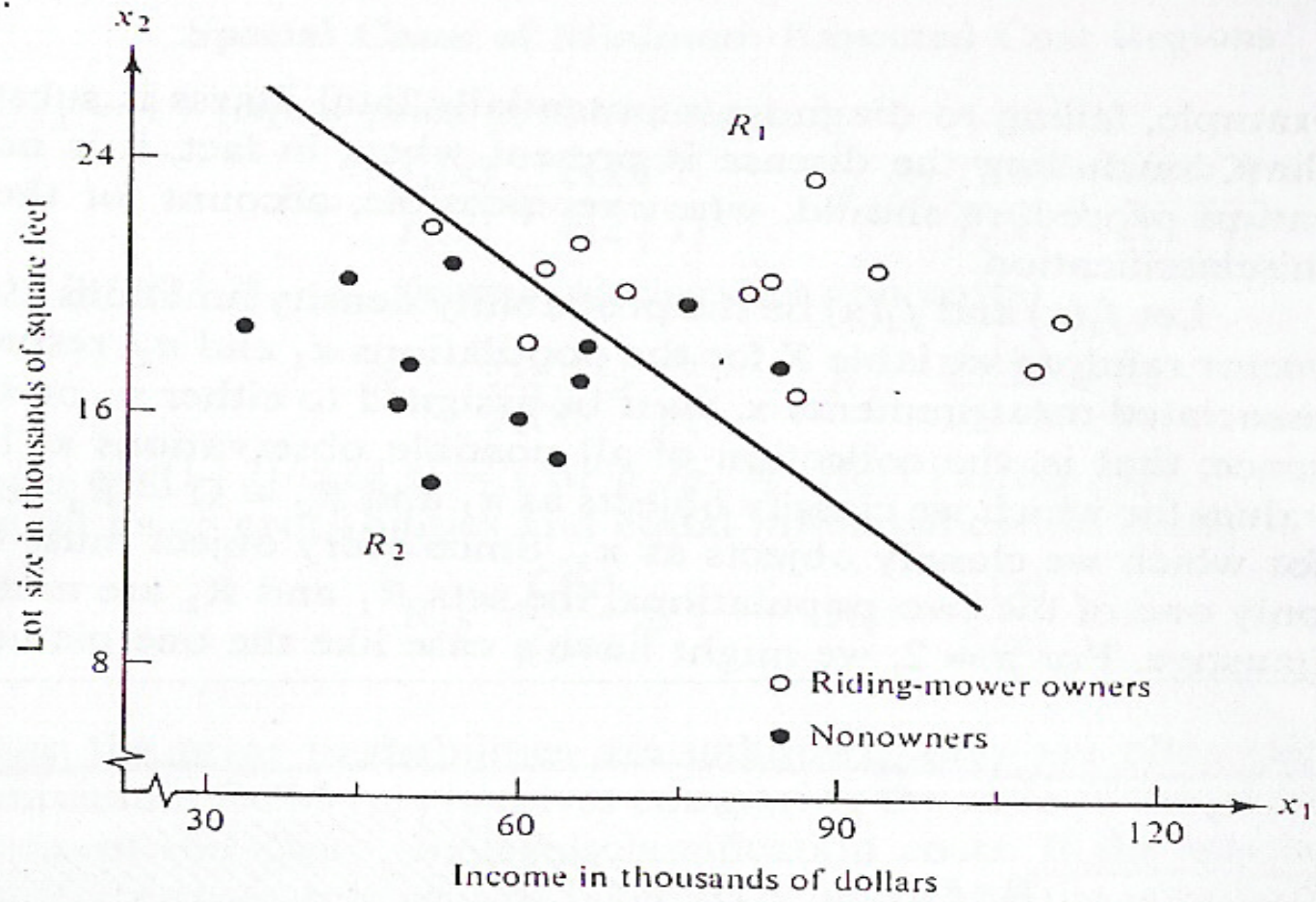

Figure 11.1 Income and lot size for riding-mower owners and nonowners.

ΠΑΡΑΤΗΡΗΣΕΙΣ

1) Παρατηρούμε ότι οι κάτοχοι μηχανής τείνουν να έχουν μεγαλύτερο εισόδημα και έκταση γης από τους μη, αν και το εισόδημα φαίνεται να διαχωρίζει καλύτερα τις 2 κλάσεις.

2) Υπάρχει μερική επικάλυψη μεταξύ των 2 κλάσεων.

- ✓ Στόχος μας είναι να δημιουργήσουμε έναν καλό κανόνα κατάταξης.

ΠΡΟΥΠΟΘΕΣΕΙΣ

1) Ένας καλός κανόνας κατάταξης πρέπει να οδηγεί σε λίγες λανθασμένες ταξινομήσεις.
2) Πρέπει να λαμβάνουμε υπόψιν τις «εκ των προτέρων» πιθανότητες μια παρατήρηση να ανήκει σε έναν από τους 2 πληθυσμούς.
3) Τέλος οφείλουμε να συνυπολογίζουμε το κόστος των λανθασμένων κατατάξεων.

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

Έστω Ω: δειγματικός χώρος

R_1={x$\in$ Ω/x ανήκει στην π_1 κλάση}

R_2=Ω-R_1

p_1, p_2 οι «εκ των προτέρων» πιθανότητες των π_1, π_2 αντίστοιχα.

Η δεσμευμένη πιθανότητα P(2|1) δηλαδή μια παρατήρηση να ανήκει στον πληθυσμό π_2, ενώ προέρχεται από τον π_1 είναι:

$$P(2|1)=(x\in R_2|\pi_1)=\int_{R_2=\Omega-R_1} f_1(x)\,dx$$

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

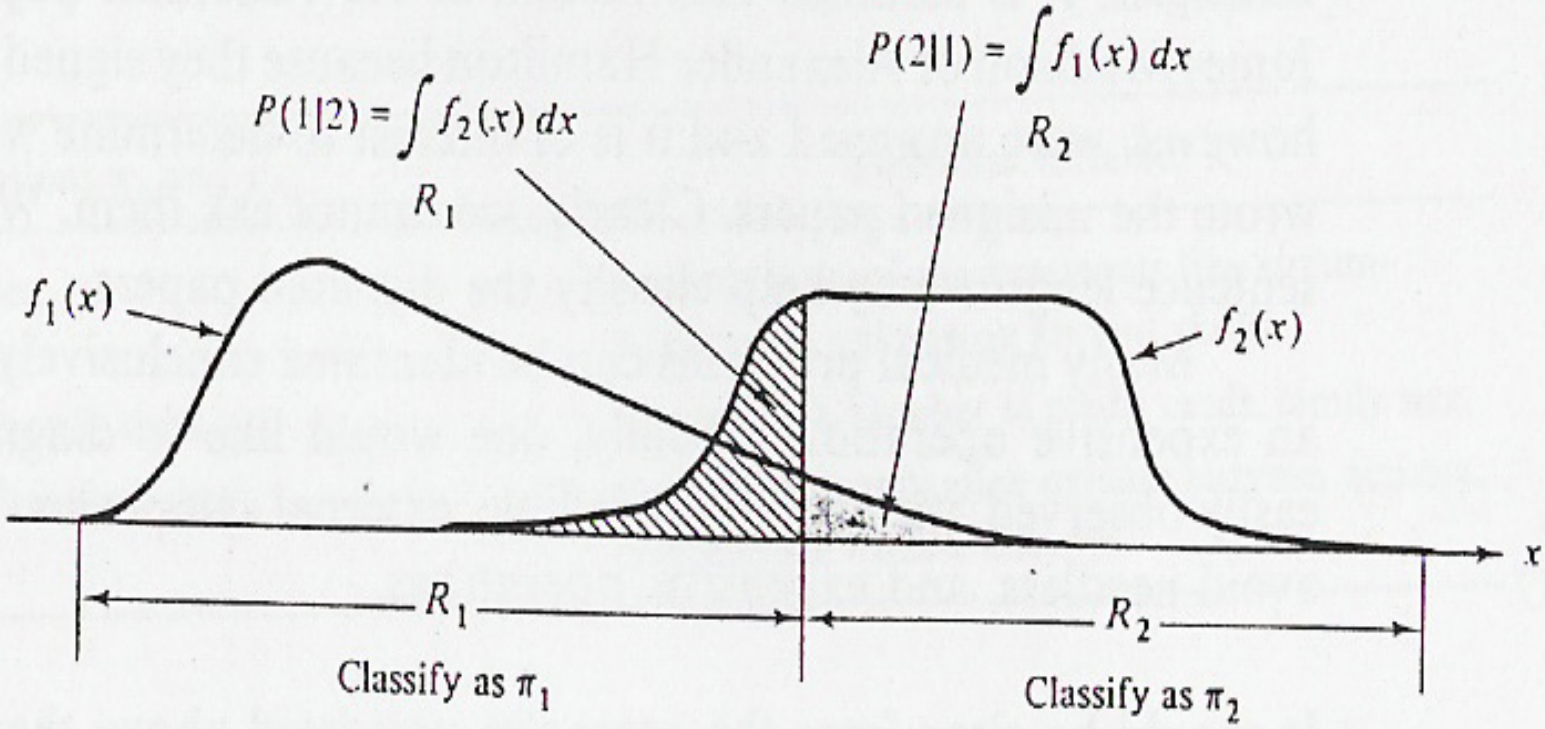

Figure 11.3 Misclassification probabilities for hypothetical classification regions when $p = 1$.

- P(σωστά ταξινομημένη στην π_1)=

  P(η παρατήρηση να προέρχεται από την π_1 και να έχει σωστά ταξινομηθεί στη π_1) = $P(x \in R_1|\pi_1)P(\pi_1)=P(1|1)\ p_1$
- P(λανθασμένα ταξινομημένη στην π_1)=

  P(η παρατήρηση να προέρχεται από την π_2 και να έχει ταξινομηθεί λάθος στην π1)=

  $P(x \in R_1|\pi_2)\ P(\pi_2) = P(1|2)\ p_2$
- P(σωστά ταξινομημένη στην π_2)=

  P(η παρατήρηση να προέρχεται από την π_2 και να έχει σωστά ταξινομηθεί στη π_2)=

  $P(x \in R_2|\pi_2)P(\pi_2) = P(2|2)\ p_2$
- P(λανθασμένα ταξινομημένη στην π_2)=P(η παρατήρηση να προέρχεται από την π_1 και να έχει ταξινομηθεί λάθος στην π_2)=

  $P(x \in R_2|\pi_1)\ P(\pi_1)=P(2|1)\ p_1$

ΤΑ ΚΟΣΤΗ

Αν δεν λάβουμε υπόψιν το κόστος προκαλούνται προβλήματα. Το κόστος της λανθασμένης ταξινόμησης απεικονίζεται στον ακόλουθο πίνακα:

	Classify as:		
True population:		π_1	π_2
	π_1	0	C(2/1)
	π_2	C(1/2)	0

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

Το αναμενόμενο κόστος λανθασμένων κατατάξεων (ECM) παρέχεται από τον τύπο:

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

- ✓ Ένας καλός κανόνας οφείλει να ελαχιστοποιεί το ECM.

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

ΚΑΝΟΝΑΣ ΚΑΤΑΤΑΞΗΣ:

R1: $\frac{f_1(x)}{f_2(x)} \geq \frac{C(1/2)}{C(2/1)} \frac{P_2}{P_1}$

R2: $\frac{f_1(x)}{f_2(x)} < \frac{C(1/2)}{C(2/1)} \frac{P_2}{P_1}$

ΠΑΡΑΔΕΙΓΜΑ

Ένας ερευνητής διαθέτει αρκετά δεδομένα ώστε να εκτιμήσει τις συναρτήσεις πυκνότητας πιθανότητας $f_1(x)$ και $f_2(x)$. Υποθέτουμε ότι $c(2|1)=5$ μονάδες και $c(1|2)=10$ μονάδες. Επιπλέον το 20% των αντικειμένων ανήκουν στον π_2. Έτσι, οι εκ των προτέρων πιθανότητες είναι: $p_1=0,8$ και $p_2=0,2$.

Προκύπτει ο κανόνας:

R_1: $f_1(x)/f_2(x) \geq (10/5)(0,2/0,8)=0,5$

R_2: $f_1(x)/f_2(x) < (10/5)(0,2/0,8)=0,5$

Για μια νέα παρατήρηση x_0 προκύπτουν:

$f_1(x_0)=0.3$ και $f_2(x_0)=0.4$

$$\frac{f_1(X_0)}{f_2(X_0)} = 0.75 > \frac{C(1|2)}{C(2|1)} \frac{P_2}{P_1} = 0.5$$

ΚΡΙΤΗΡΙΑ

Υπάρχουν και άλλα κριτήρια που χρησιμοποιούνται για τη δημιουργία του βέλτιστου κανόνα κατάταξης.

1)ΤΡΜ=P(λανθασμένα ταξινομημένη στην π_1 ή λανθασμένα ταξινομημένη στην π_2)=

P(η παρατήρηση να προέρχεται από την π1 και να είναι λανθασμένα ταξινομημένη) + P(η παρατήρηση να προέρχεται από την π_2 και να είναι λανθασμένα ταξινομημένη)=

$$p_1 \int_{R_2} f_1(x)\,dx \quad + \quad p_2 \int_{R_1} f_2(x)\,dx$$

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

Η ελαχιστοποίηση του TPM ισοδυναμεί με την ελαχιστοποίηση του ECM όταν τα κόστη των λανθασμένων κατατάξεων είναι ίσα. Άρα ο κανόνας κατάταξης δίνεται από τον τύπο:

$R_1: f_1(x)/f_2(x) \geq p_2/p_1$ $\quad$ $R_2: f_1(x)/f_2(x) < p_2/p_1$

2)Οι εκ των υστέρων πιθανότητες κατά Bayes:

$P(\pi_1|x_0)$=P(Η παρατήρηση x_0 να ανήκει στο π_1)/ P (να παρατηρήσω x_0)=

P(να παρατηρήσουμε $x_0|\pi_1)P(\pi_1)$ / {P(να παρατηρήσουμε $x_0|\pi_1)P(\pi_1)$ +

P(να παρατηρήσουμε $x_0 \mid \pi_2)\, P(\pi_2)\}$ = $p_1f_1(x_0)/\{p_1f_1(x_0)/\{p_1f_1(x_0)+p_2f_2(x_0)\}$ και

ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ ΓΙΑ 2 ΠΛΗΘΥΣΜΟΥΣ

$P(\pi_2|x_0) = 1-P(\pi_1|x_0) = \ldots = p_2f_2(x_0)/\{p_1f_1(x_0)+p_2f_2(x_0)\}$

- Ταξινομούμε μια παρατήρηση x_0 στο π_1 όταν $P(\pi_1|x_0)>P(\pi_2|x_0)$ που ισοδυναμεί με το να θεωρήσουμε ότι τα κόστη λανθασμένων κατατάξεων είναι ίσα.

Α : Όταν $\Sigma_1=\Sigma_2=\Sigma$

Θεωρούμε τις από κοινού πυκνότητες του $X'=[X_1,X_2,\ldots,X_p]$για του πληθυσμούς π_1 και π_2 που δίνονται από τον τύπο:

$F_i(x)=\{(2\pi)^{p/2}|\Sigma|^{1/2}\}^{-1}\exp\{-(x-\mu_i)'\Sigma^{-1}(x-\mu_i)\}$ για i=1,2, όπου μ_1,μ_2 γνωστά.

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

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

Έτσι καταλήγουμε στον κανόνα κατάταξης:

x_0 ανήκει στο π_1 αν $(\mu_1-\mu_2)'\Sigma^{-1}X_0-(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2)/2$

$\geq \ln[\frac{C(1|2)}{C(2|1)}(p_2/p_1)]$

Διαφορετικά x_0 ανήκει στο π_2.

Στην πράξη

Έστω n_1, n_2 παρατηρήσεις από τους πληθυσμούς π_1 και π_2 αντίστοιχα με τους πίνακες δεδομένων τους:

$X_{1(p\times n1)}=[X_{11},X_{12},\ldots,X_{1n1}]$ και $X_{2(p\times n2)}=[X_{11},X_{12},\ldots,X_{1n}]$

Υπολογίζουμε τα $\bar{x}_1=\Sigma X_{1j}/n_1$ $j=1,2,...,n_1$, $\bar{x}_2=\Sigma X_{2j}/n_2$

$j=1,2,\ldots,n_2$, $S_{1(p\times p)}=\Sigma(X_{1j}-\bar{x}_1)(X_{1j}-\bar{x}_1)'/(n_1-1)$

$S_{2(p\times p)=}\Sigma(X_{2j}-\bar{x}_2)(X_{2j}-\bar{x}_2)'/(n_2-1)$. Συνδυάζοντας τα S_1 και S_2 για τον υπολογισμό ενός αμερόληπτου εκτιμητή(Α.Ε) του Σ:

ΤΑΞΙΝΟΜΗΣΗ ΜΕ 2 ΠΟΛΥΔΙΑΣΤΑΤΟΥΣ ΚΑΝΟΝΙΚΟΥΣ ΠΛΗΘΥΣΜΟΥΣ

$$S_{pooled}=\left[\frac{n_1-1}{(n_1-1)+(n_2-1)}\right]S_1+\left[\frac{n_2-1}{(n_1-1)+(n_2-1)}\right]S_2=\frac{(n_1-1)S_1+(n_2-1)S_2}{(n_1+n_2-2)}$$

και παίρνουμε τον εξής κανόνα ταξινόμησης:

- Ταξινομούμε το x_0 στον π_1 αν :

$$(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}x_0-\frac{1}{2}(\bar{x}_1-\bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1+\bar{x}_2)\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

- διαφορετικά στον π_2 .

- Έστω $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)=1$ τότε ln(1)=0 και ο

παραπάνω κανόνας για δύο κανονικούς πληθυσμούς ισοδυναμεί με τη σύγκριση της μεταβλητής $y=\hat{l}'x=(\bar{x}_1-\bar{x}_2)\,S_{pooled}^{-1}x$ στην παρατήρηση x_0 με τον αριθμό

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

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

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

- Δηλαδή, ο κανόνας που προέκυψε από την ελαχιστοποίηση του ECM για δύο κανονικούς πληθυσμούς είναι ισοδύναμος με τη δημιουργία δύο μονοδιάστατων πληθυσμών π_1 και π_2.

- Ταξινομούμε μια νέα παρατήρηση x_0, στον π_1 ή στον π_2, ανάλογα με το αν το $y_0 = \hat{l}' x_0$ πέφτει δεξιά ή αριστερά του μέσου ($\hat{m}$) μεταξύ των μέσων τιμών $\bar{y}_1$, $\bar{y}_2$.

Β : Όταν $\Sigma_1 \neq \Sigma_2$

Κατά όμοιο τρόπο τα R_1, R_2 ορίζονται ως εξής:

R_1: $X'(\Sigma_1^{-1}-\Sigma_2^{-1})X/2 + (\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1})X-k \geq \ln\{\frac{C(1|2)}{C(2|1)}(p_2/p_1)\}$ (1)

R_2: $X'(\Sigma_1^{-1}-\Sigma_2^{-1})X/2 + (\mu_1'\Sigma_1^{-1}-\mu 2'\Sigma_2^{-1})X-k < \ln\{\frac{C(1|2)}{C(2|1)}(p_2/p_1)\}$ (2)

Όπου $k=\ln\{|\Sigma_1/\Sigma_2|\}+\mu_1'\Sigma_1^{-1}\mu_1-\mu 2'\Sigma_2^{-1}\mu_2)$

Και έτσι ο κανόνας κατάταξης διαμορφώνεται ως εξής:

Κατατάσσουμε το x_0 στο π_1 αν ισχύει η σχέση (1) όπου $X=x_0$ διαφορετικά στο π_2.

ΤΑΞΙΝΟΜΗΣΗ ΜΕ 2 ΠΟΛΥΔΙΑΣΤΑΤΟΥΣ ΚΑΝΟΝΙΚΟΥΣ ΠΛΗΘΥΣΜΟΥΣ

<u>ΑΔΥΝΑΜΙΑ ΤΟΥ ΚΑΝΟΝΑ ΚΑΤΑΤΑΞΗΣ:</u>

Η αδυναμία του κανόνα κατάταξης έγκειται στο ότι είναι ευαίσθητος στις αποκλίσεις από την κανονική κατανομή.

ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΤΑΞΙΝΟΜΗΣΗΣ:

Οι δειγματικές συναρτήσεις ταξινόμησης μπορούν να αξιολογηθούν από την τιμή του AER(actual error rate):

$$AER=p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

Όπου R_1,R_2 αντιπροσωπεύουν τις περιοχές ταξινόμησης που ορίζονται από τα δείγματα μεγέθους n_1 και n_2 αντίστοιχα.

Ο AER δείχνει πως η δειγματική συνάρτηση ταξινόμησης θα αποδώσει σε μελλοντικά δείγματα.

Apparent error rate(APER)

Υπάρχει ένα μέτρο απόδοσης το οποίο δεν εξαρτάται από την μορφή των αρχικών δειγμάτων το οποίο καλείται apparent error rate(APER)

Για n_1 και n_2 παρατηρήσεις από τους πληθυσμούς π_1 και π_2 αντίστοιχα παίρνουμε τον ακόλουθο confusion matrix και $APER=(n_{1M}+n_{2M}) / (n_1+n_2)$

		Predicted membership		
		π_1	π_2	
Actual membership	π_1	n_{1c}	$n_{1M}=n_1-n_{1c}$	n_1
	π_2	$n_{2M}=n_2-n_{2C}$	n_{2c}	n_2

Αδυναμία του APER

- Ο APER υποεκτιμά τον AER κι αυτό αντιμετωπίζεται μόνο με τη χρήση μεγάλων δειγμάτων. Αυτό συμβαίνει διότι τα δεδομένα που έχουμε χρησιμοποιούνται για τη δημιουργία αλλά και την αξιολόγηση της συνάρτησης ταξινόμησης.
- Μπορούν να κατασκευαστούν άλλοι error-rate εκτιμητές που είναι καλύτεροι από τον APER, υπολογίζονται εύκολα και δεν απαιτούν υποθέσεις για τις κατανομές.

Μέθοδοι Αντιμετώπισης

1. Μια μέθοδος είναι να χωρίσουμε τα δεδομένα σε ένα training sample κι ένα validation sample. Έχει όμως δυο μειονεκτήματα:

a. Απαιτεί μεγάλα δείγματα και

b. Η συνάρτηση που αξιολογείται δεν είναι αυτή που μας ενδιαφέρει.

2. Μια δεύτερη μέθοδος είναι η Lachenbruch “holdout”.

1. Παραλείπουμε μια παρατήρηση από τον πληθυσμό π_1 και δημιουργούμε τη συνάρτηση ταξινόμησης που προκύπτει από αυτές που υπολείπονται $(n_1\text{-}1, n_2)$.
2. Ταξινομούμε τη holdout παρατήρηση με βάση την παραπάνω συνάρτηση.
3. Επαναλαμβάνουμε τα βήματα 1 και 2 ώσπου να ταξινομηθούν όλες οι παρατηρήσεις του π_1.($n_{1M}^{(H)}$=ο αριθμός των παρατηρήσεων που έχουν ταξινομηθεί λανθασμένα)

4. Επαναλαμβάνουμε τα βήματα 1 έως 3 για τις παρατηρήσεις του πληθυσμού π_2.($n_{2M}^{(H)}$=ο αριθμός των holdout παρατηρήσεων που έχουν ταξινομηθεί λανθασμένα)

- Εκτιμούμε τις πιθανότητες P(2|1), P(1|2):

$$\hat{P}(2|1)=\frac{n_{1M}^{(H)}}{n_1} \qquad \hat{P}(2|1)=\frac{n_{2M}^{(H)}}{n_2}$$

$$\hat{E}(AER)=\frac{n_{1M}^{(H)}+n_{2M}^{(H)}}{n_1+n_2}$$

Παράδειγμα

- Εφαρμόζουμε τη μέθοδο Lachenbruch's "holdout" για τον υπολογισμό των error-rates εκτιμητών για ίσα κόστη και ίσες εκ των προτέρων πιθανότητες.

$$X_1 = \begin{bmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{bmatrix} \qquad \bar{x}_1 = \begin{bmatrix} 3 \\ 10 \end{bmatrix} \qquad 2S_1 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

$$X_2 = \begin{bmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{bmatrix} \qquad \bar{x}_2 = \begin{bmatrix} 4 \\ 7 \end{bmatrix} \qquad 2S_2 = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix}$$

$$S_{pooled} = \frac{1}{4}(2S_1 + 2S_2) = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$$

- Με τη χρήση του κατάλληλου κανόνα κατάταξης προκύπτει ο confusion matrix :

True population:	Classify as: π_1	Classify as: π_2
π_1	2	1
π_2	1	2

$$\text{APER} = \frac{2}{6} = .33$$

- Αφαιρώντας μια παρατήρηση $\bar{x}_H' = [2 \quad 12]$ από τον X_1 έχουμε :

$$X_{1H} = \begin{bmatrix} 4 & 3 \\ 10 & 8 \end{bmatrix} \qquad \bar{x}_{1H} = \begin{bmatrix} 3.5 \\ 9 \end{bmatrix} \qquad 1S_{1H} = \begin{bmatrix} .5 & 1 \\ 1 & 2 \end{bmatrix}$$

$$S_{H,pooled} = \frac{1}{3}\begin{bmatrix} 2.5 & -1 \\ -1 & 10 \end{bmatrix} \qquad S_{H,pooled}^{-1} = \frac{1}{8}\begin{bmatrix} 10 & 1 \\ 1 & 2.5 \end{bmatrix}$$

Τετραγωνική Απόσταση της x_H από τη $\bar{x}_{1H}$ =

$$\left(x_H - \bar{x}_{1H}\right)' S_{H,pooled}^{-1} \left(x_H - \bar{x}_{1H}\right) = 4.5$$

Τετραγωνική Απόσταση της x_H από τη $\bar{x}_{2H}$ =

$$\left(x_H - \bar{x}_{2H}\right)' S_{H,pooled}^{-1} \left(x_H - \bar{x}_{2H}\right) = 10.3$$

- Επειδή 10.3 > 4.5 κατατάσσουμε την παρατήρηση x_H στον πληθυσμό π_1.

- Ακολουθώντας τα βήματα της μεθόδου καταλήγουμε στα αποτελέσματα:

  $n_{1M}^{(H)} = 2$ και $n_{2M}^{(H)} = 1$
- Άρα :

$$\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} = \frac{2+1}{3+3} = .5$$

Παράδειγμα

- X_1=διάμετρος δακτυλίων για το 1ο έτος που αναπτύχθηκε ο σολομός σε γλυκό νερό
- X_2= διάμετρος δακτυλίων για το 1ο έτος που αναπτύχθηκε ο σολομός σε θαλασσινό νερό

- Gender= $\begin{cases} 1, \theta\eta\lambda\upsilon\kappa\acute{o} \\ 2, \alpha\rho\sigma\varepsilon\nu\iota\kappa\acute{o} \end{cases}$

TABLE 11.2 SALMON DATA (GROWTH-RING DIAMETERS)

Alaskan			Canadian		
Gender	Freshwater	Marine	Gender	Freshwater	Marine
2	108	368	1	129	420
1	131	355	1	148	371
1	105	469	1	179	407
2	86	506	2	152	381
1	99	402	2	166	377
2	87	423	2	124	389
1	94	440	1	156	419
2	117	489	2	131	345
2	79	432	1	140	362
1	99	403	2	144	345
1	114	428	2	149	393
2	123	372	1	108	330
1	123	372	1	135	355
2	109	420	2	170	386
2	112	394	1	152	301
1	104	407	1	153	397
2	111	422	1	152	301
2	126	423	2	136	438
2	105	434	2	122	306
1	119	474	1	148	383
1	114	396	2	90	385
2	100	470	1	145	337
2	84	399	1	123	364
2	102	429	2	145	376
2	101	469	2	115	354
2	85	444	2	134	383
1	109	397	1	117	355
2	106	442	2	126	345
1	82	431	1	118	379
2	118	381	2	120	369
1	105	388	1	153	403
1	121	403	2	150	354
1	85	451	1	154	390
1	83	453	1	155	349
1	53	427	2	109	325
1	95	411	2	117	344
1	76	442	1	128	400
1	95	426	1	144	403
2	87	402	2	163	370
1	70	397	2	145	355
2	84	511	1	133	375
2	91	469	1	128	383
1	74	451	2	123	349
2	101	474	1	144	373
1	80	398	2	140	388
1	95	433	2	150	339
2	92	404	2	124	341
1	99	481	1	125	346
2	94	491	1	153	352
1	87	480	1	108	339

SOURCE: Data courtesy of K. A. Jensen and B. Van Alen of the State of Alaska Department of Fish and Game.

- Στο training sample με $n_1=n_2=50$ έχουμε:

$$\bar{x}_1 = \begin{bmatrix} 98.38 \\ 429.66 \end{bmatrix} \quad S_1 = \begin{bmatrix} 260.608 & -188.093 \\ -188.093 & 1399.086 \end{bmatrix}$$

$$\bar{x}_2 = \begin{bmatrix} 137.46 \\ 366.62 \end{bmatrix} \quad S_2 = \begin{bmatrix} 326.090 & 133.505 \\ 133.505 & 893.261 \end{bmatrix}$$

Με τον κανόνα κατάταξης για ίσα κόστη και ίσες εκ των προτέρων πιθανότητες προκύπτει ο confusion matrix:

Actual membership	Predicted membership: π_1: Alaskan	Predicted membership: π_2: Canadian
π_1: Alaskan	44	6
π_2: Canadian	1	49

- Η συνάρτηση ταξινόμησης είναι :
  $\hat{y} = -5.54121 - .12839x_1 + .05194x_2$
- Ο APER= 0.07 είναι αρκετά μικρός.

- Υπάρχει διαφορά στις διασπορές του $\hat{y}$ για τους 2 πληθυσμούς :

	n	Sample Mean	Sample Standard Deviation
Alaskan	50	4.144	3.253
Canadian	50	-4.147	2.450

- P(Canadian|Alaskan)> P(Alaskan|Canadian)

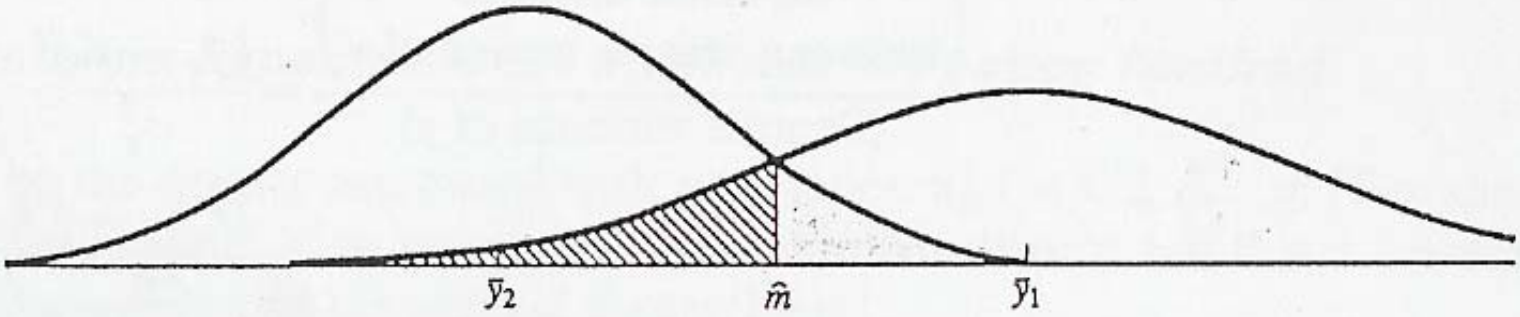

Figure 11.7 Schematic of normal densities for the linear discriminant—salmon data.

Συνάρτηση διαχωρισμού κατά Fisher

- Ο Fisher μετέτρεψε τις πολυδιάστατες παρατηρήσεις x σε μονοδιάστατες y τέτοιες ώστε οι y που προέρχονται από τους πληθυσμούς π_1 και π_2 να διαχωρίζονται όσο το δυνατόν περισσότερο.
- Δημιούργησε τις y σαν γραμμικό συνδυασμό των x.
- Η προσέγγισή του δεν προϋποθέτει την κανονικότητα των δεδομένων αλλά την ισότητα των πινάκων συνδιασπορών.

- Η μεταβλητή y παίρνει τις τιμές y_{11}, $y_{12},\ldots, y_{1n1}$ για τον πρώτο πληθυσμό και $y_{21}, y_{22},\ldots, y_{2n2}$ για το δεύτερο πληθυσμό. Ο διαχωρισμός των δύο συνόλων εκφράζεται με τη μορφή: $\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y}$ όπου $s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$

- Σκοπός είναι να επιλέξουμε το γραμμικό μετασχηματισμό των x που επιτυγχάνει το μέγιστο διαχωρισμό των δειγματικών μέσων τιμών $\bar{y}_1$, $\bar{y}_2$.

- Ο γραμμικός μετασχηματισμός :

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

Μεγιστοποιεί το λόγο :

$$\frac{\left(\begin{array}{c}\text{Τετραγωνική Απόσταση}\\ \text{μεταξύ των δειγματικών}\\ \text{μέσων τιμών της y}\end{array}\right)}{\left(\text{Δειγματική Διασπορά της y}\right)} = \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{s_y^2} = \frac{\left(\hat{l}'\bar{x}_1 - \hat{l}'\bar{x}_2\right)^2}{\hat{l}' S_{pooled}\hat{l}} = \frac{\left(\hat{l}' d\right)^2}{\hat{l}' S_{pooled}\hat{l}}$$

για όλα τα πιθανά διανύσματα $\hat{l}$ όπου $d=\left(\bar{x}_1-\bar{x}_2\right)$

- Η μέγιστη τιμή του λόγου είναι:

$$D^2 = \left(\bar{x}_1 - \bar{x}_2\right)' S_{pooled}^{-1}\left(\bar{x}_1 - \bar{x}_2\right)$$

ΚΑΝΟΝΑΣ ΚΑΤΑΤΑΞΗΣ ΚΑΤΑ FISHER

- Κατατάσσουμε το x_0 στον πληθυσμό π_1 αν:

$$y_0 = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 \geq \hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$$

ή $$y_0 - \hat{m} \geq 0$$

- Κατατάσσουμε το x_0 στον πληθυσμό π_2 αν:

$$y_0 < \hat{m}$$

ή $$y_0 - \hat{m} < 0$$

- Η εύρεση του κατάλληλου $\hat{l}$ απεικονίζεται γραφικά για p=2 στο παρακάτω σχήμα:

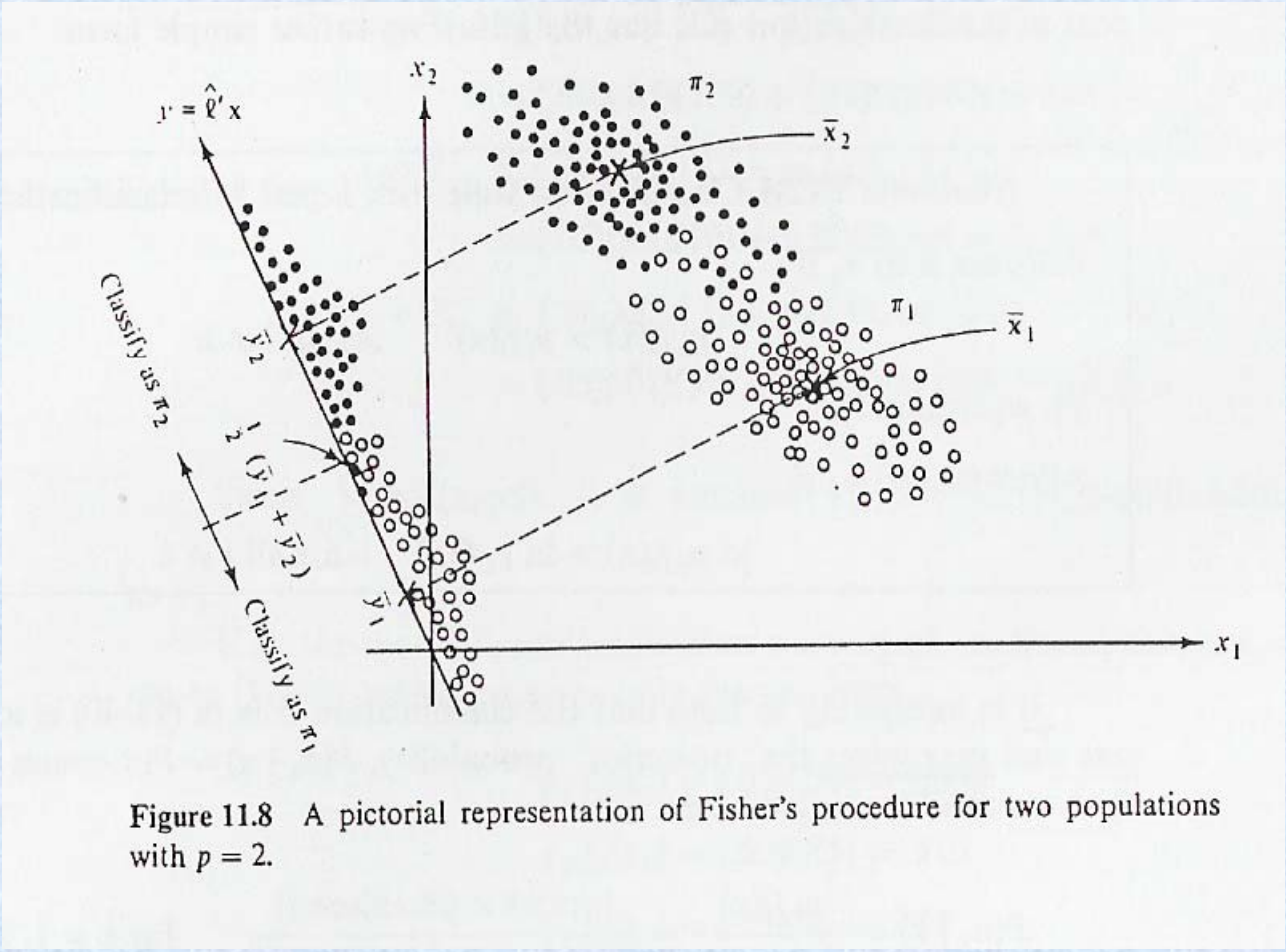

Figure 11.8 A pictorial representation of Fisher's procedure for two populations with $p = 2$.

Έλεγχος Διαχωρισμού

- Υποθέτουμε ότι έχουμε δύο πολυδιάστατους κανονικούς πληθυσμούς π_1 και π_2 με κοινό πίνακα διασπορών Σ. Τότε ο έλεγχος $H_0:\mu_1=\mu_2$ με $H_1:\mu_1\neq\mu_2$ γίνεται με το στατιστικό:

$$\left(\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\right)\left(\frac{n_1 n_2}{n_1+n_2}\right)D^2 \sim F_{u,v}$$

  όπου $u=p$ και $v = n_1+n_2-p-1$.

- Αν η H_0 απορριφθεί συμπεραίνουμε ότι ο διαχωρισμός μεταξύ των πληθυσμών είναι σημαντικός.

Σχόλιο

- Ο σημαντικός διαχωρισμός δεν υποδηλώνει καλή ταξινόμηση.
- Κάθε κανόνας ταξινόμησης μπορεί να αξιολογηθεί ανεξάρτητα από κάθε έλεγχο διαχωρισμού.