ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ (PATTERN RECOGNITION)

Σέργιος Θεοδωρίδης

Κωνσταντίνος Κουτρούμπας

Version 2

Στη συνέχεια θα μιλήσουμε μόνο για στατιστικές μεθόδους ταξινόμησης προτύπων (statistical pattern classification).

Υπόθεση: Σε μια μεγάλη αίθουσα γυμναστηρίου βρίσκονται *100* χορευτές (**D**) και *200* καλαθοσφαιριστές (**B**)

Ερώτηση 1: *«Αν επιλέξουμε κάποιον από την αίθουσα, μαντέψτε αν είναι D ή B.»*

Απάντηση: Για χορευτή: $P(D)=100/(100+200)=1/3$.

Για καλαθοσφαιριστή: $P(B)=200/(100+200)=2/3$

Άρα επιλέγεται η απάντηση **B**

Επιπλέον δεδομένα: Οι κατανομές πυκνότητας πιθανότητας των βαρών των δύο κατηγοριών:

$$p(x/D) = \frac{1}{\sqrt{300\pi}} e^{-\frac{(x-50)^2}{300}} \qquad p(x/B) = \frac{1}{\sqrt{300\pi}} e^{-\frac{(x-90)^2}{300}}.$$

Ερώτηση 2: *«Αν επιλέξουμε κάποιον από την αίθουσα, μαντέψτε αν είναι D ή B.»*

Απάντηση: Ίδια με την ερώτηση 1.

Ερώτηση 3: *««Αν επιλέξουμε κάποιον από την αίθουσα, μαντέψτε αν είναι D ή B, δοθέντος ότι το ύψος του είναι 155 εκ;»*

Απάντηση: Ίδια με την ερώτηση 1.

Ερώτηση 4: *««Αν επιλέξουμε κάποιον από την αίθουσα, μαντέψτε αν είναι D ή B, δοθέντος ότι το βάρος του είναι 60 κιλά;»*

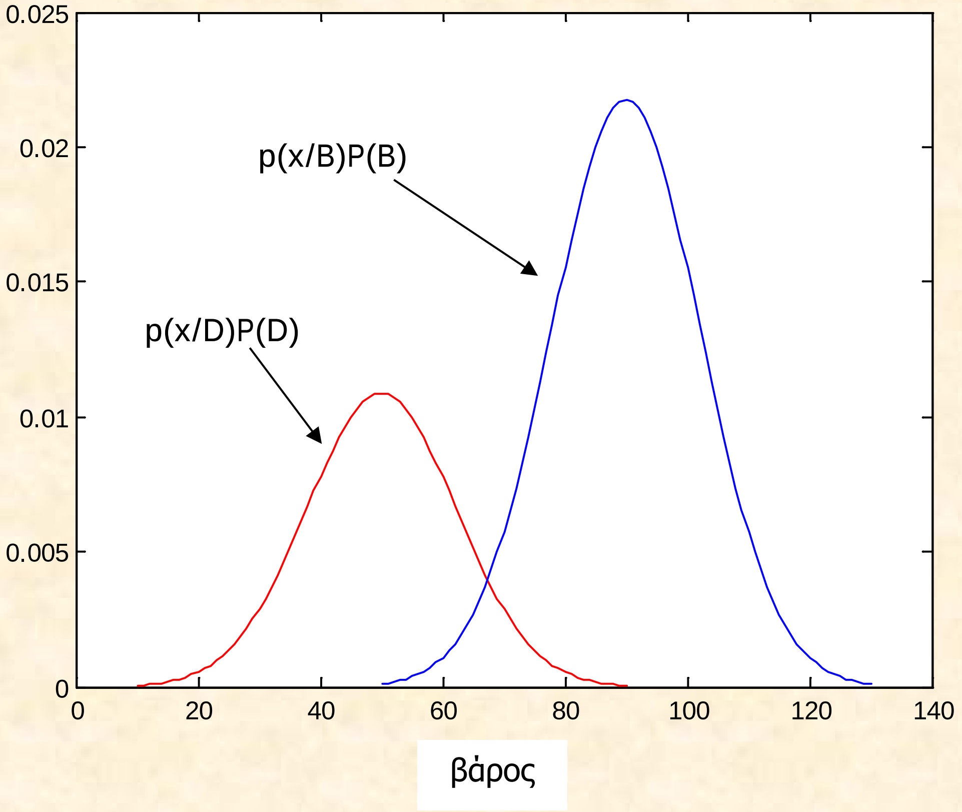
Απάντηση:

$$P(D/\beta=60)=\frac{p(60/D)P(D)}{p(60/D)P(D)+p(60/B)P(B)} \quad = \quad 0.8780$$

$$P(\kappa/\beta=60)=\frac{p(60/\kappa)P(\kappa)}{p(60/\alpha)P(\alpha)+p(60/\kappa)P(\kappa)}. \quad = \quad 0.1220$$

Προβληματισμός: Γιατί η κατάσταση άλλαξε τόσο δραματικά με την επιπλέον πληροφορία;

Εξήγηση: Οι τιμές του βάρους διαφοροποιούνται πολύ στις δύο κατηγορίες.



Υπόθεση: Ας θεωρήσουμε ότι το επιπλέον χαρακτηριστικό είναι η απόχρωση των μαλλιών (η ίδια ομοιόμορφη κατανομή και για τις δύο κατηγορίες)

$$p(x/D)=\begin{cases}\dfrac{1}{\chi_2-\chi_1}, & x\in[\chi_1,\chi_2]\\ 0 & x\notin[\chi_1,\chi_2]\end{cases} \qquad p(x/B)=\begin{cases}\dfrac{1}{\chi_2-\chi_1}, & x\in[\chi_1,\chi_2]\\ 0 & x\notin[\chi_1,\chi_2]\end{cases}.$$

Ερώτηση 5: *««Αν επιλέξουμε κάποιον από την αίθουσα, μαντέψτε αν είναι D ή B, δοθέντος ότι η απόχρωση των μαλλιών του είναι y∈[χ1,χ2]».*

Απάντηση:

$$P(D/\chi = y) = \frac{p(y/D)P(D)}{p(y/D)P(D) + p(y/B)P(B)} = \frac{\frac{1}{\chi_2 - \chi_1}\frac{1}{3}}{\frac{1}{\chi_2 - \chi_1}\frac{1}{3} + \frac{1}{\chi_2 - \chi_1}\frac{2}{3}} = \frac{1}{3} = P(D)$$

$$P(B/\chi = y) = \frac{p(y/B)P(B)}{p(y/D)P(D) + p(y/B)P(B)} = \frac{\frac{1}{\chi_2 - \chi_1}\frac{2}{3}}{\frac{1}{\chi_2 - \chi_1}\frac{1}{3} + \frac{1}{\chi_2 - \chi_1}\frac{2}{3}} = \frac{2}{3} = P(B).$$

Συμπέρασμα: Αν έχουμε τη δυνατότητα να επιλέξουμε χαρακτηριστικά ας διαλέγουμε εκείνα που διαφοροποιούνται όσο το δυνατόν περισσότερο στις υπό εξέταση κατηγορίες (τα καταλληλότερα χαρακτηριστικά επιλέγονται με τη βοήθεια ενός ειδικού στην υπό μελέτη εφαρμογή).

Έστω και πάλι ότι το επιπλέον χαρακτηριστικό είναι το βάρος

Ερώτηση 6: *«Αν επιλέξω κάποιον αθλητή από την αίθουσα, τι είδους αθλητής είναι αυτός, δεδομένου ότι το βάρος του είναι x κιλά;»*

Καλύτερη δυνατή απάντηση: Είναι αθλητής από την κατηγορία με τη **μεγαλύτερη εκ των υστέρων** (**a posteriori**) πιθανότητα.

Κανόνας του Bayes (1):

Αν *P(D/x)>P(B/x)* τότε καταχώρησε τον αθλητή στην κατηγορία **D**.

Διαφορετικά καταχώρησε τον αθλητή στην κατηγορία **B**.

Λαμβανομένου υπόψη ότι

$$P(D/x) = \frac{p(x/D)P(D)}{p(x/D)P(D) + p(x/B)P(B)} \qquad P(B/x) = \frac{p(x/B)P(B)}{p(x/D)P(D) + p(x/B)P(B)}$$

Κανόνας του Bayes (2):

Αν *P(D)p(x/D)>P(B)p(x/B)* τότε καταχώρησε τον αθλητή στην κατηγορία «χορευτής»

Διαφορετικά καταχώρησε τον αθλητή στην κατηγορία «καλαθοσφαιριστής».

Υπόθεση: Οι a priori πιθανότητες των κατηγοριών είναι ίσες.

Κανόνας του Bayes (ειδ):

Αν *p(x/D)>p(x/B)* τότε καταχώρησε τον αθλητή στην κατηγορία «χορευτής»

Διαφορετικά καταχώρησε τον αθλητή στην κατηγορία «καλαθοσφαιριστής».

Σημαντική παρατήρηση: ΠΑΝΤΑ υπάρχει και η πιθανότητα λάθους (ένας «αφύσικα» αδύνατος καλαθοσφαιριστής η ένας «αφύσικα» παχουλός χορευτής μπορούν να ταξινομηθούν λάθος).

Ταξινόμηση κατά Bayes και γεωμετρική ταξινόμηση

• Τα σημεία όπου

$$p(x/D)P(D) = p(x/B)P(B)$$

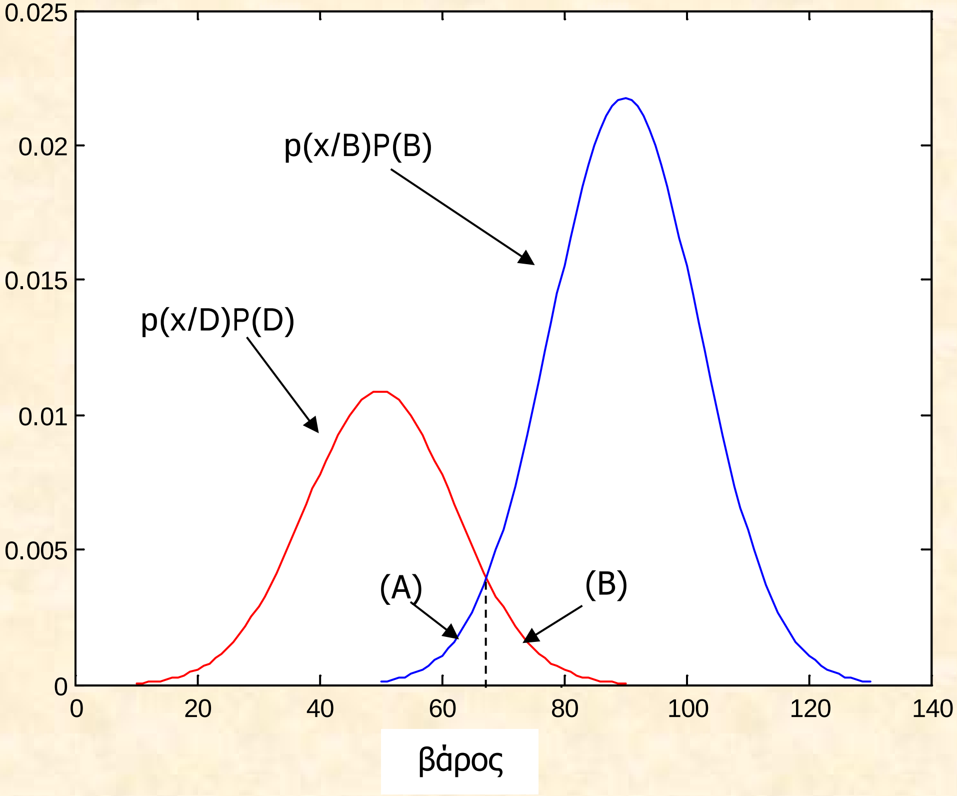
οριοθετούν τις περιοχές του χώρου των χαρακτηριστικών (βάρος) που αντιστοιχούν στις δύο κατηγορίες

•Ταξινόμηση οντότητας με συγκεκριμένη τιμή βάρους ανάλογα με την περιοχή στην οποία αυτή ανήκει.

Ορισμός περιοχών: Λύνοντας την εξίσωση

$$p(x/D)P(D) = p(x/B)P(B)$$

$$\frac{1}{3}\frac{1}{\sqrt{300\pi}}e^{-\frac{(x-50)^2}{300}} = \frac{2}{3}\frac{1}{\sqrt{300\pi}}e^{-\frac{(x-90)^2}{300}} \Leftrightarrow -(x-50)^2 = -(x-90)^2 + 300\ln 2 \Leftrightarrow x \approx 67.4 \equiv x_0$$



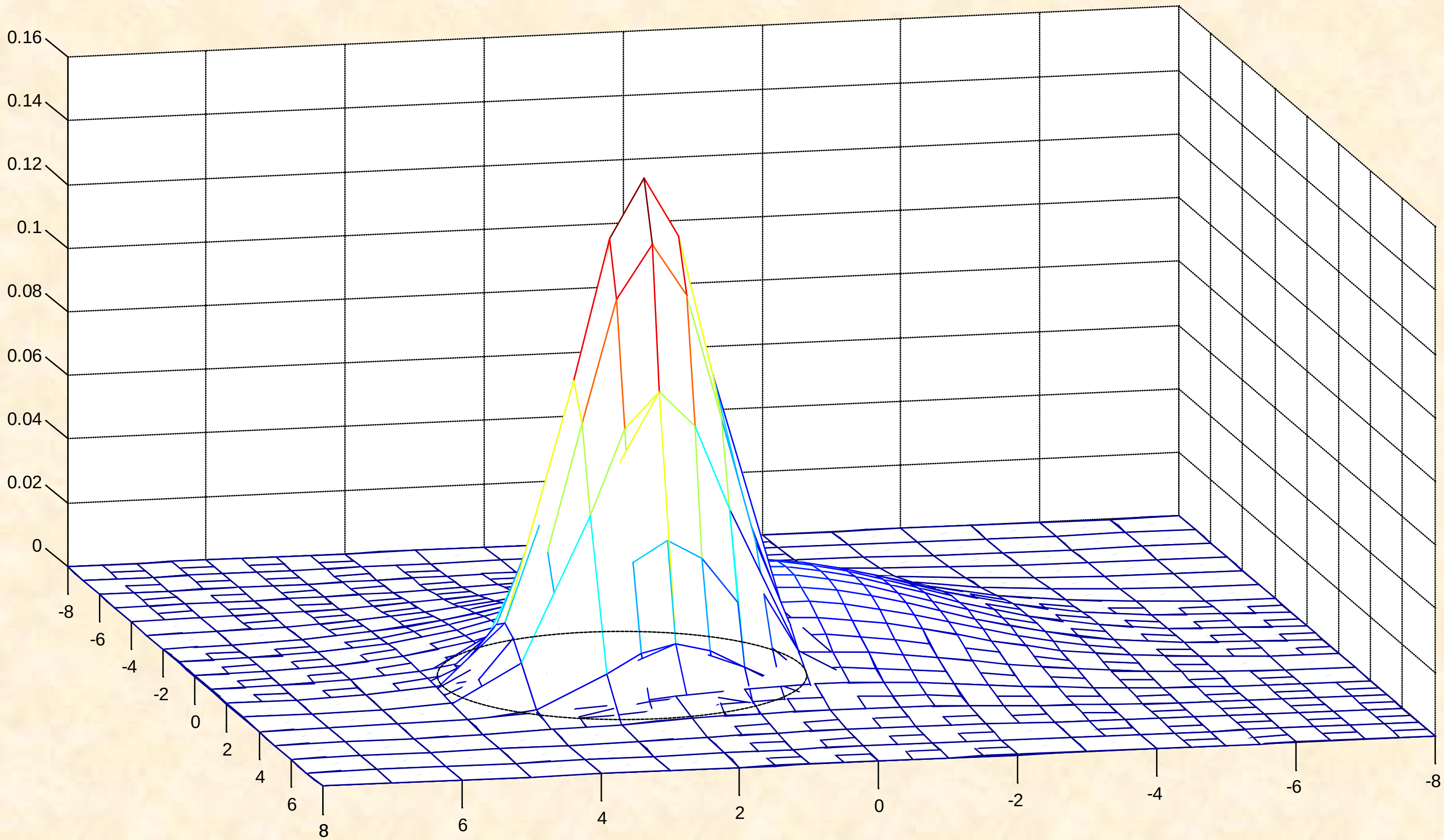
Χορευτής: $R_D=\{x: x<67.4\}$

Καλαθοσφ.: $R_B=\{x: x>67.4\}$

Ερώτηση: Ποιά είναι η πιθανότητα λάθους;

Απάντηση: $$P_\lambda = \mathrm{E}\mu\beta(\mathrm{A}) + \mathrm{E}\mu\beta(\mathrm{B}) = \int_{x\in R_B} P(D)\,p(x/D)dx + \int_{x\in R_D} P(B)\,p(x/B)dx.$$

Για το παράδειγμά μας: $$P_\lambda = \frac{1}{3}0.0778 + \frac{2}{3}0.0322 = 0.0474.$$



Παράδειγμα δύο κατηγοριών στον 2-διάστατο χώρο.

Σχεδιασμός ταξινομητή

Συμβολισμός:

- $\omega_1, \omega_2, \ldots, \omega_c$ οι κατηγορίες στις οποίες θα ταξινομηθεί μια οντότητα
- l το πλήθος των χαρακτηριστικών που αναπαριστούν μια οντότητα (διάσταση του χώρου των χαρακτηριστικών)
- $\mathbf{x}=[x_1,x_2,\ldots x_l]^T$ Διάνυσμα αναπαράστασης οντότητας (x_i η τιμή του i χαρακτηριστικού γι' αυτήν την οντότητα)
- $P(\omega_i)$ η *a priori* πιθανότητα η υπό εξέταση οντότητα να ανήκει στην i κατηγορία
- $P(\omega_i/\mathbf{x})$ η *a posteriori* πιθανότητα η υπό εξέταση οντότητα να ανήκει στην i κατηγορία δοθέντος του διανύσματος μετρήσεων $\mathbf{x}$
- $p(\mathbf{x}/\omega_i)$ η συνάρτηση πυκνότητας πιθανότητας που περιγράφει την κατανομή των χαρακτηριστικών διανυσμάτων $\mathbf{x}$ για την κατηγορία ω_i.

Ερώτηση 7: *«Δοθέντος ενός διανύσματος μετρήσεων* **x** *μιας υπό εξέταση οντότητας να προσδιοριστεί η κατηγορία* ω_i *στην οποία ανήκει η οντότητα»*

Bayes 1: *«καταχώρησε την υπό εξέταση οντότητα στην κατηγορία* ω_i *για την οποία*

$$P(\omega_i/\mathbf{x}) = \max_{j=1,\ldots,c} P(\omega_j/\mathbf{x}).»$$

Δεδομένου ότι:

$$P(\omega_i / x) = \frac{p(x/\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(x/\omega_j)P(\omega_j)}$$

Bayes 2: *«καταχώρησε την υπό εξέταση οντότητα στην κατηγορία* ω_i *για την οποία*

$$p(\mathbf{x}/\omega_i)P(\omega_i) = \max_{j=1,\ldots,c} p(\mathbf{x}/\omega_j)P(\omega_j).»$$

Bayes (ειδ): *«Αν οι κατηγορίες είναι ισοπίθανες, καταχώρησε την υπό εξέταση οντότητα στην κατηγορία ωi για την οποία*

$$p(\mathbf{x}/\omega_i) = \max_{j=1,\ldots,c} p(\mathbf{x}/\omega_j).»$$

Παράδειγμα: Σ’ ένα πρόβλημα τριών κλάσεων, χρησιμοποιείται μόνο ένα χαρακτηριστικό για την ταξινόμηση των δειγμάτων. Οι αντίστοιχες πυκνότητες πιθανότητας και οι a priori πιθανότητες για τις κλάσεις είναι:

ω1: $P(\omega 1)=1/2$, $p(x|\omega 1)=(1/\sqrt{2\pi})*\exp(-x^2/2)$

ω2: $P(\omega 2)=1/3$, $p(x|\omega 2)=(1/\sqrt{2\pi})*\exp(-(x-1)^2/2)$

ω3: $P(\omega 3)=1/6$, $p(x|\omega 3)=(1/\sqrt{2\pi})*\exp(-(x-2)^2/2)$

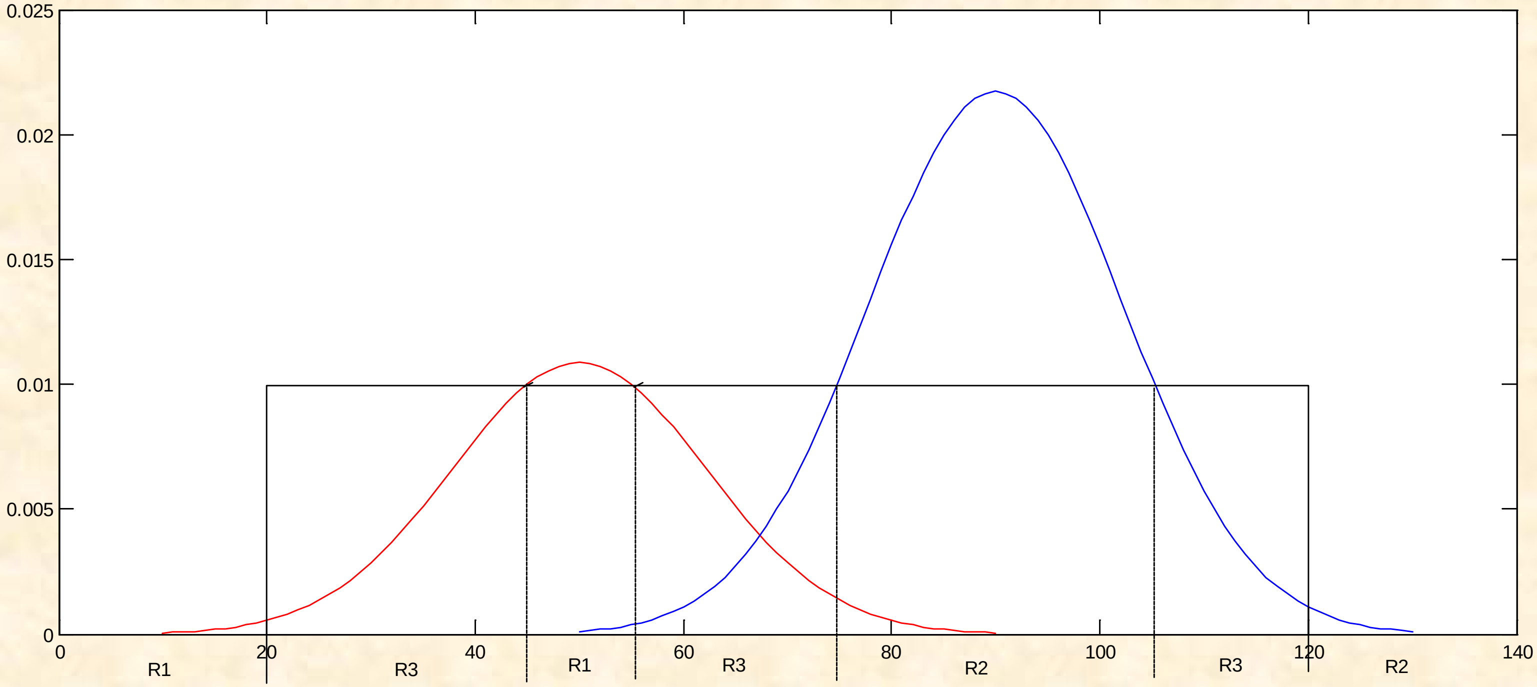
Ταξινομήστε σε μία από τις παραπάνω κατηγορίες άγνωστο δείγμα με τιμή χαρακτηριστικού **x=1.6**.

Είναι:

$P(\omega 1)p(x|\ \omega 1)=$ **0.0555**, $P(\omega 2)p(x|\ \omega 2)=$ **0.1111**, $P(\omega 3)p(x|\ \omega 3)=$ **0.0614**

Συνεπώς το δείγμα καταχωρείται στην κατηγορία **ω2.**

Παράδειγμα 3 τριών κατανομών



Πιθανότητα λάθους:

$$P_\lambda = \sum_{i=1}^{c} \int_{R_i} \left(\sum_{k=1, k\neq i}^{c} p(x/\omega_k) P(\omega_k) \right) dx$$

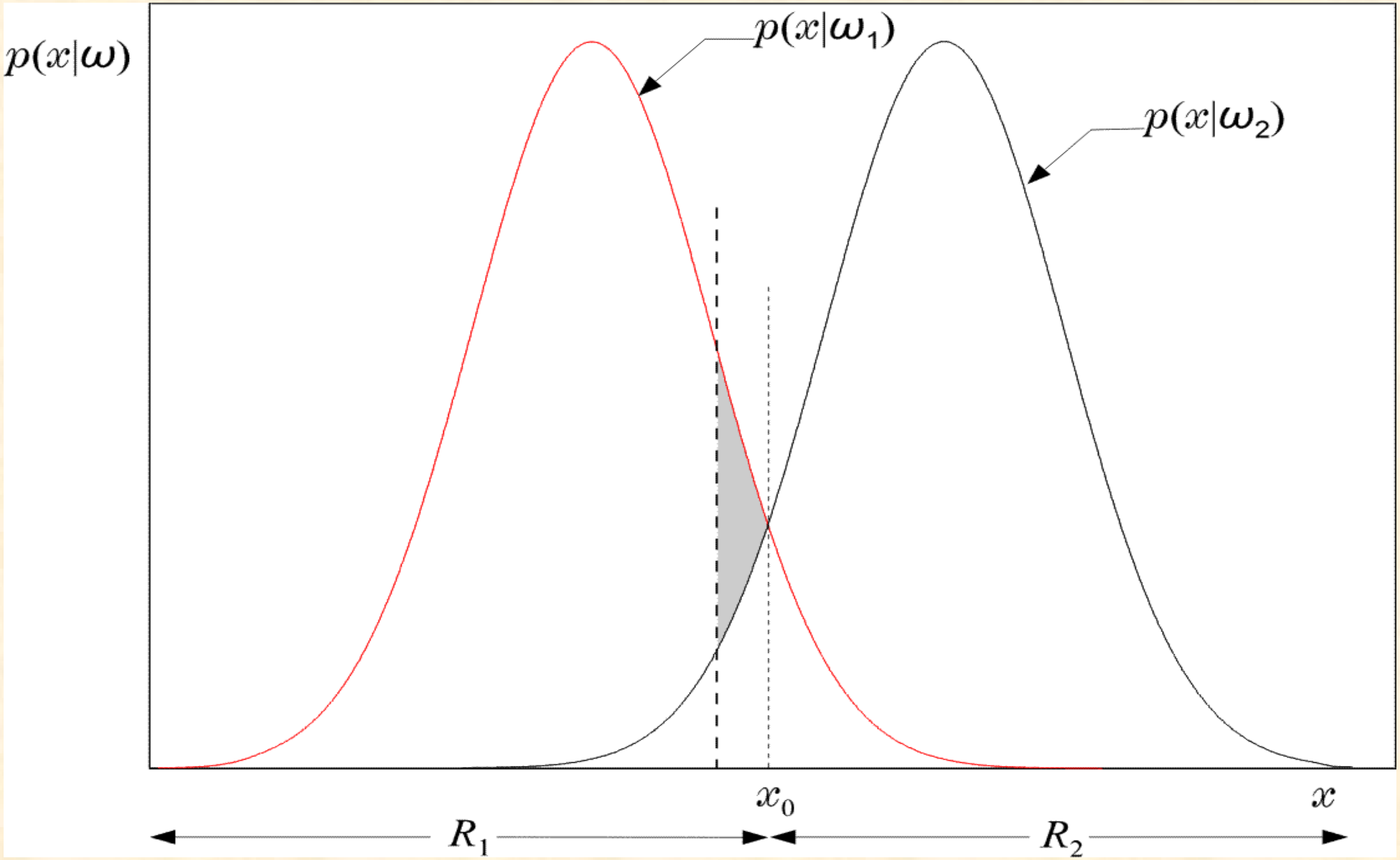
Ερώτηση: Γιατί τόση επιμονή στον ταξινομητή Bayes;

Απάντηση:

•*γιατί η λογική του δεν έρχεται σε αντίθεση με την διαίσθησή μας.*

•*Γιατί περιλαμβάνει ως ειδικές περιπτώσεις πολλούς ταξινομητές που χρησιμοποιούνται στην πράξη.*

•*Ο κανόνας ταξινόμησης κατά Bayes είναι βέλτιστος με την έννοια ότι* ***ελαχιστοποιεί την πιθανότητα λάθους***.

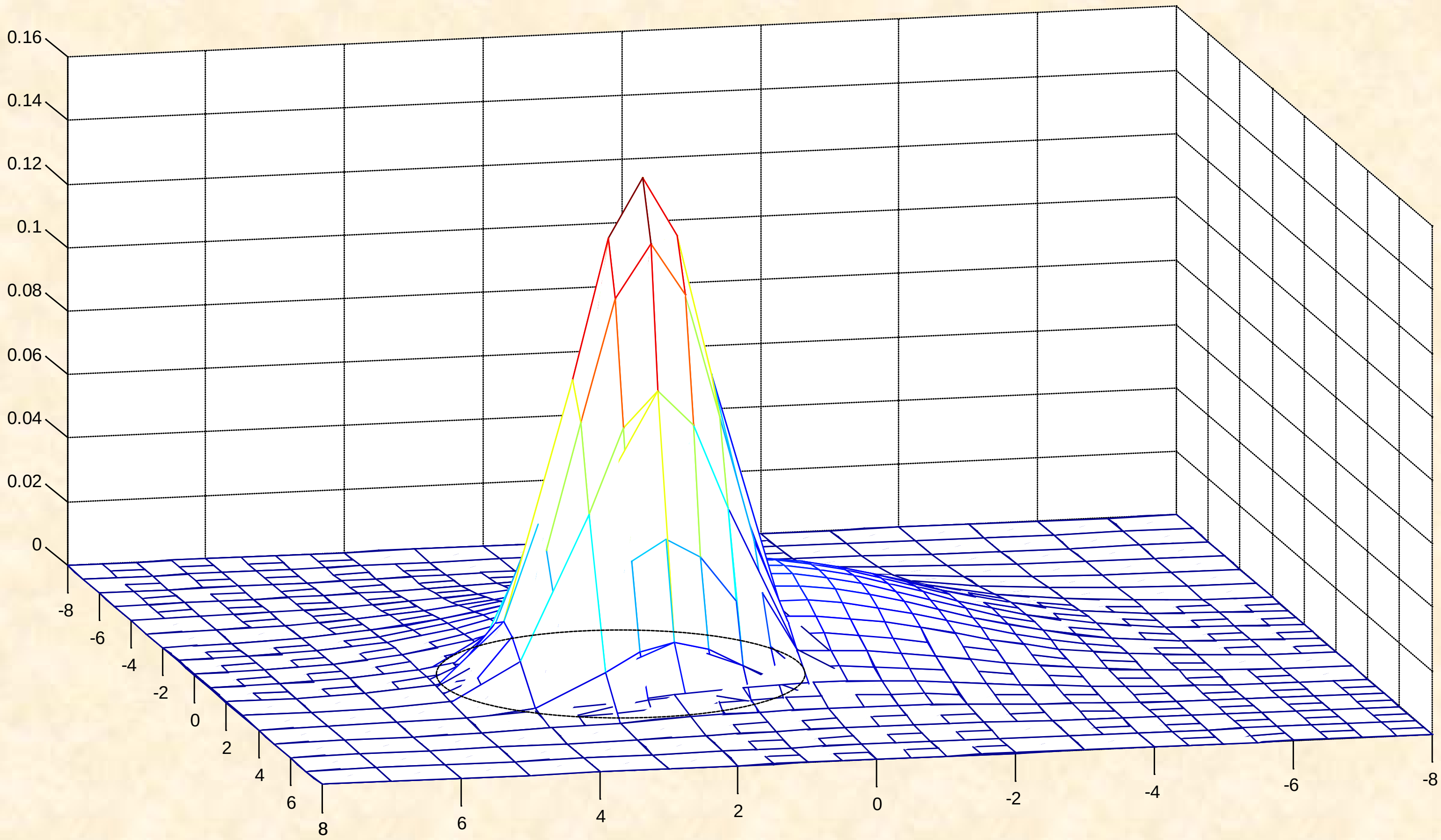


Παρατήρηση: οι πυκνότητες πιθανότητας των κατηγοριών μαζί με τις a priori πιθανότητές τους ορίζουν **μονοσήμαντα** μια διαμέριση του χώρου του *l*-διάστατου χώρου των χαρακτηριστικών, σε (όχι απαραίτητα ενιαίες) περιοχές (R_i) όπου κάθε μια αντιστοιχεί και σε μια κατηγορία (ω_i).

Έτσι, η ταξινόμηση ενός διανύσματος μπορεί να γίνει

(α) είτε με χρήση του κανόνα του Bayes όπως αυτός διατυπώνεται στον *Bayes 1* ή *Bayes 2*

(β) είτε με προσδιορισμό των περιοχών των χώρου των χαρακτηριστικών που αντιστοιχούν στις υπό εξέταση κατηγορίες και για κάθε νέο διάνυσμα χαρακτηριστικών να εξετάζει απλώς σε ποια περιοχή ανήκει.



Παράδειγμα δύο κατηγοριών στον 2-διάστατο χώρο.