Περιεχόμενα

Κεφάλαιο 2 Θεωρία απόφασης Bayes

2.1 Εισαγωγή

Η θεωρία απόφασης του Bayes είναι μια θεμελιώδης στατιστική προσέγγιση στο πρόβλημα της ταξινόμησης προτύπων. Αυτή η προσέγγιση είναι βασισμένη στον ποσοτικό προσδιορισμό των ανταλλαγών, μεταξύ διάφορων αποφάσεων ταξινόμησης, που χρησιμοποιούν την πιθανότητα, και το κόστος που συνοδεύει τέτοιες αποφάσεις. Κάνει την υπόθεση ότι το πρόβλημα απόφασης τίθεται σε πιθανολογικούς όρους, και ότι όλες οι σχετικές τιμές πιθανοτήτων είναι γνωστές. Σε αυτό το κεφάλαιο αναπτύσσουμε τις βασικές αρχές αυτής της θεωρίας, και επιδεικνύουμε πώς μπορεί να αντιμετωπισθεί ως μια απλή διαμόρφωση διαδικασιών κοινής λογικής. Στα επόμενα κεφάλαια θα εξετάσουμε τα προβλήματα που προκύπτουν όταν η πιθανολογική δομή δεν είναι εντελώς γνωστή.

Ενώ θα δώσουμε μια αρκετά γενική, αφηρημένη ανάπτυξη από την θεωρία απόφασης του Bayes στο τμήμα. ??, αρχίζουμε τη συζήτησή μας με ένα συγκεκριμένο παράδειγμα. Ας επανεξετάσουμε το υποθετικό πρόβλημα που τίθεται στο κεφ. ?? του σχεδιασμού ενός ταξινομητή για να διαχωρίζει δύο είδη ψαριών: πέρκες θάλασσας και σολομό. Υποθέστε ότι ένας παρατηρητής που παρακολουθεί ψάρια φτάνει κατά μήκος της ζώνης μεταφορέων, το βρίσκει δύσκολο να προβλέψει ποιος τύπος θα προκύψει έπειτα και ότι η ακολουθία των τύπων ψαριών εμφανίζεται να είναι τυχαία. Στην ορολογία της θεωρίας αποφάσεων θα λέγαμε ότι καθώς κάθε ψάρι εμφανίζεται φυσικά θα είναι στη μία η την άλλη από δύο πιθανές καταστάσεις: είτε το ψάρι είναι μια πέρκα θάλασσας ή το ψάρι είναι ένας σολομός. Επιλέγουμε το ω να δείχνει τη ***κατάσταση της φύσης***, με $\omega=\omega_1$ για τις πέρκες θάλασσας και $\omega=\omega_2$ για το σολομό. Επειδή η κατάσταση της φύσης είναι τόσο απρόβλεπτη, θεωρούμε το ω σα μια μεταβλητή που πρέπει να περιγραφεί πιθανολογικά..

Εάν η ψαριά είχε τόση πέρκα θάλασσας όσο και σολομό, θα λέγαμε ότι το επόμενο ψάρι είναι εξίσου πιθανό να είναι πέρκα θάλασσας ή σολομός. Γενικότερα, υποθέτουμε ότι υπάρχει κάποια a priori πιθανότητα (ή απλά εκ των προτέρων) $P(\omega_1)$ ότι το επόμενο ψάρι είναι πέρκα θάλασσας, και κάποια εκ των προτέρων πιθανότητα $P(\omega 2)$ ότι είναι σολομός. Εάν υποθέσουμε ότι δεν υπάρχει κανένας άλλος τύπος ψαριών σχετικών εδώ, τότε $P(\omega_1)$ και $P(\omega_2)$ έχουν άθροισμα 1. Αυτές οι εκ των προτέρων πιθανότητες αντανακλούν τη εκ των προτέρων γνώση μας για το πόσο πιθανό είναι να πάρουμε πέρκες θάλασσας ή σολομό πριν εμφανιστεί πραγματικά το ψάρι. Παραδείγματος χάριν, μπορεί να βασίζεται στο χρόνο του έτους ή την επιλογή της περιοχής.

Υποθέστε για μια στιγμή ότι αναγκαστήκαμε να λάβουμε μια απόφαση για τον τύπο ψαριών που θα εμφανιστεί μετά χωρίς να είναι επιτρεπτό να το δούμε. Προς το παρόν, θα υποθέσουμε ότι οποιαδήποτε ανακριβής ταξινόμηση συνεπάγεται το ίδιο κόστος ή συνέπεια, και ότι οι μόνες πληροφορίες που έχουμε την άδεια να χρησιμοποιήσουμε είναι η τιμή των εκ των προτέρων πιθανοτήτων. Εάν μια απόφαση πρέπει να ληφθεί με τόσο λίγες πληροφορίες, φαίνεται λογικό να χρησιμοποιηθεί ο ακόλουθος ***κανόνα απόφασης***:

Επιλέγουμε ω_1 εάν $P(\omega_1) >_P(\omega_2)$; Αλλιώς επιλέγουμε ω_2.

Αυτός ο κανόνας έχει νόημα εάν πρόκειται να κρίνουμε μόνο ένα ψάρι, αλλά εάν πρόκειται να κρίνουμε πολλά ψάρια, η χρησιμοποίηση αυτού του κανόνα επανειλημμένα μπορεί να φανεί λίγο παράξενη. Τελικά, θα λαμβάναμε πάντα την ίδια απόφαση ακόμα κι αν ξέρουμε ότι και οι δύο τύποι ψαριών θα εμφανιστούν. Πόσο καλά λειτουργεί εξαρτάται από τις τιμές των εκ των προτέρων πιθανοτήτων. Εάν $P(\omega_1)$ είναι πάρα πολύ μεγαλύτερο από $P(\omega_2)$, η απόφασή μας υπέρ ω_1 θα είναι σωστή τις περισσότερες φορές. Εάν $P(\omega_1)=P(\omega_2)$, έχουμε μόνο μια πενήντα-πενήντα πιθανότητα να είμαστε σωστοί. Γενικά, η πιθανότητα του λάθους είναι η μικρότερη των $P(\omega_1)$ και $P(\omega_2)$, και θα δούμε αργότερα ότι υπό αυτούς τους όρους κανένας άλλος κανόνας απόφασης δεν μπορεί να δώσει μια μεγαλύτερη πιθανότητα να είμαστε ακριβείς.

Στις περισσότερες περιστάσεις δεν καλούμαστε να λάβουμε τις αποφάσεις με τόσες λίγες πληροφορίες. Στο παράδειγμά μας, θα μπορούσαμε να χρησιμοποιήσουμε ένα μετρητή ελαφρότητας x για να βελτιώσουμε τον ταξινομητή μας. Διαφορετικά ψάρια θα δώσουν διαφορετικές τιμές ελαφρότητας και εκφράζουμε αυτήν την διαφορετικότητα με πιθανολογικούς όρους; θεωρούμε το x μια συνεχή τυχαία μεταβλητή της οποίας η κατανομή εξαρτάται από την κατάσταση της φύσης, και εκφράζεται ως $p(x|\omega_1)$.* Αυτή είναι συνάρτηση πυκνότητας πιθανότητας των κλάσεων. Για να κυριολεκτήσουμε, η συνάρτηση πυκνότητας πιθανότητας $p(x|\omega_1)$ πρέπει να γράφετε $p_X(x|\omega_1)$ για να δείξει ότι μιλάμε για μια ιδιαίτερη συνάρτηση πυκνότητας για την τυχαία μεταβλητή X. Αυτή η πιο ξεκάθαρη σήμανση καθιστά σαφές ότι $p_X(\cdot)$ και $p_Y(\cdot)$ δηλώνουν δύο διαφορετικές συναρτήσεις, ένα γεγονός που παρερμηνεύεται όταν γράφουμε p(x) και p(y). Δεδομένου ότι αυτή η πιθανή σύγχυση προκύπτει σπάνια στην πράξη, έχουμε επιλέξει να υιοθετήσουμε την απλούστερη σήμανση. Αναγνώστες που είναι αβέβαιοι της σήμανσης μας ή που θα επιθυμούσαν να θυμηθούν τη θεωρία πιθανοτήτων πρέπει να δουν το παράρτημα ??). Αυτή είναι η συνάρτηση πυκνότητας πιθανότητας για το X δεδομένου ότι η κατάσταση της φύσης είναι ω_1. (Αναφέρεται και ως πυκνότητα πιθανότητας όσο αναφορά την κατάσταση.) Κατόπιν η διαφορά μεταξύ $p(x|\omega_1)$ και $p(x|\omega_2)$ περιγράφει τη διαφορά στην ελαφρότητα μεταξύ των πληθυσμών των πέρκων θάλασσας και του σολομού (Σχ. 2.1).

Υποθέστε ότι ξέρουμε και τις δύο εκ των προτέρων πιθανότητες $P(\omega_j)$ και τις δεσμευμένες πυκνότητες $p(x|\omega_j)$. Υποθέστε περαιτέρω ότι μετράμε την ελαφρότητα ενός ψαριού και ανακαλύπτουμε ότι η τιμή της είναι x. Πως αυτό επηρεάζει τη συμπεριφορά μας όσο αναφορά την πραγματική φύση του προβλήματος, που είναι η κατηγορία των ψαριών; Παρατηρούμε ότι η (δεσμευμένη) πυκνότητα πιθανότητας της εύρεσης ενός μοτίβου που ανήκει στην κατηγορία ω_j και έχει αξία εκτιμήτριας x μπορεί να γραφτεί με δύο τρόπους: $p(\omega_j,x)=P(\omega_j|x)p(x)=p(x|\omega_j)P(\omega_j)$. Κάνοντας ανακατανομή αυτών οδηγούμαστε στην απάντηση της ερώτησης μας που καλείται εξίσωση του Bayes:

$$P(\omega_j|x)=\frac{p(x|\omega_j)P(\omega_j)}{p(x)}. \tag{1}$$

Που στην περίπτωση των δύο αυτών κατηγοριών γίνεται:

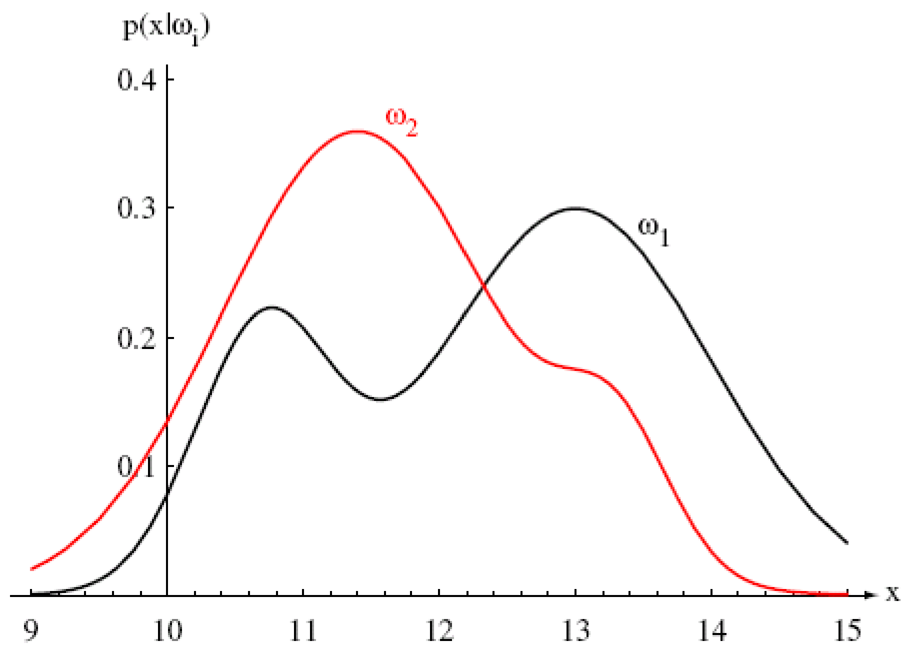
$$p(x)=\sum_{j=1}^{2} p(x|\omega_j)P(\omega_j). \tag{2}$$

* Γενικά χρησιμοποιούμε κεφαλαίο $P(\cdot)$ για να δηλώσουμε συνάρτηση πιθανότητας και μικρό $p(\cdot)$ για να δηλώσουνε συνάρτηση πυκνότητας πιθανότητας.

Η εξίσωση του Bayes μπορεί να εκφραστεί ως

$$posterior = \frac{likelihood \times prior}{evidence}. \tag{3}$$

Η εξίσωση του Bayes δείχνει ότι παρατηρώντας την τιμή του x μπορούμε να μετατρέψουμε την εκ των προτέρων πιθανότητα $P(\omega_j)$ στην εκ των υστέρων (ή μεταγενέστερη) πιθανότητα $P(\omega_j | x)$ — η πιθανότητα η κατάσταση της φύσης να είναι ω_j δεδομένου ότι η εκτιμήτρια x έχει μετρηθεί. Καλούμε $p(x|\omega_j)$ την πιθανότητα του ω_j με εκτιμήτρια το x (ένας όρος που έχει επιλεχθεί για να δείξει ότι, αν όλες οι άλλες παράμετροι είναι ίσοι η κατηγορία ω_j για την οποία η $p(x|\omega_j)$ είναι μεγαλύτερη είναι πιο πιθανό να είναι η σωστή κατηγορία). Πρέπει να προσεχθεί ότι είναι απαραίτητο να υπολογιστεί η εκ των υστέρων πιθανότητα λόγω της πιθανότητας και της εκ των προτέρων πιθανότητας. ο στοιχειώδης παράγοντας, p(x), μπορεί να είναι μόνο ένας παράγοντας κλίμακας που εγγυάται ότι οι πιθανότητες θα έχουν άθροισμα ένα, όπως όλες οι καλές πιθανότητες πρέπει να έχουν. Η γραφική παράσταση της $P(\omega_j | x)$ σε σχέση με το x φαίνεται στο Σχ. 2.2 για τις περιπτώσεις που $P(\omega_1)$=2/3 and $P(\omega_2)$=1/3.

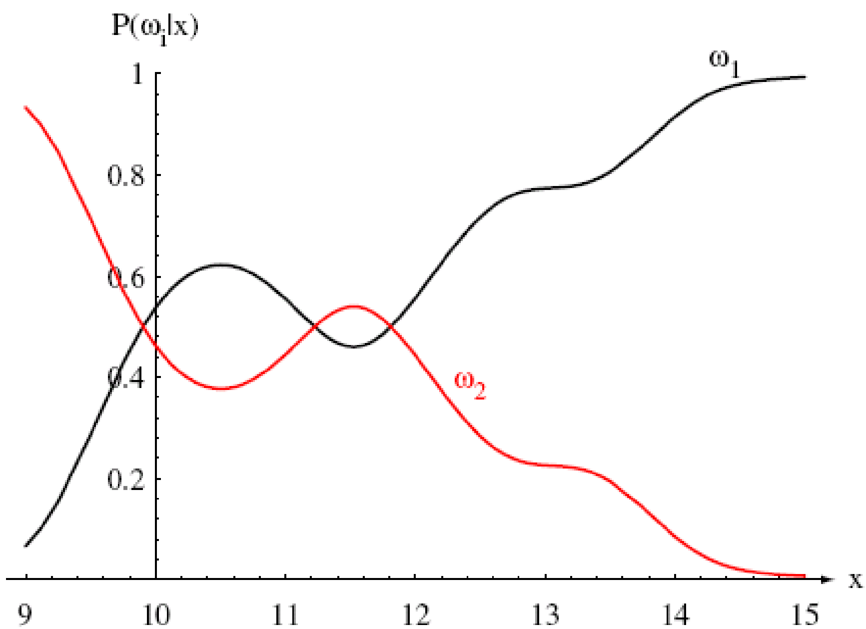


Σχήμα 2.1: Υποθετικές συναρτήσεις πυκνότητας πιθανότητας για την ταξινόμηση κλάσεων δείχνουν τη πυκνότητα πιθανότητας της μέτρησης μιας εκτιμήτριας x με δεδομένο ότι η ακολουθία ανήκει στην κατηγορία ω_i. Εάν το x αναπαριστά το μήκος του ψαριού, οι δύο καμπύλες μπορεί να περιγράφουν τη διαφορά στο μήκος των πληθυσμών δύο τύπων ψαριών. Οι συναρτήσεις πυκνότητας είναι κανονικοποιημένες και επομένως η περιοχή κάτω από κάθε καμπύλη είναι 1.0.

Εάν έχουμε μια τιμή x για την οποία $P(\omega_1|x)$ είναι μεγαλύτερο από $P(\omega_2|x)$, θα μπορούσαμε να οδηγηθούμε στην απόφαση ότι η σωστή κατάσταση της φύσης είναι το ω_1. Ομοίως, αν το $P(\omega_2|x)$ είναι μεγαλύτερο από το $P(\omega_1|x)$, θα διαλέγαμε ω_2. Για να δικαιολογήσουμε αυτή τη διαδικασία επιλογής, ας υπολογίζουμε την πιθανότητα λάθους όποτε παίρνουμε μια απόφαση. Όταν παίρνουμε μια τιμή x,

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{αν αποφασίσουμε } \omega_2 \\ P(\omega_2|x) & \text{αν αποφασίσουμε } \omega_1 \end{cases} \quad (4)$$

Προφανώς για δεδομένη τιμή του x μπορούμε να ελαχιστοποιήσουμε την πιθανότητα του λάθους επιλέγοντας ω_1 αν $P(\omega_1|x) > P(\omega_2|x)$ και ω_2 σε άλλη περίπτωση. Βέβαια, δε μπορούμε ποτέ να δούμε ακριβώς την ίδια τιμή του x δύο φορές. Θα ελαχιστοποιήσει όμως αυτός ο κανόνας την πιθανότητα λάθους; Ναι γιατί η μέση πιθανότητα σφάλματος δίνεται από:



Σχήμα 2.2: Εκ των υστέρων πιθανότητες με βάση τις εκ των προτέρων $P(\omega_1)=2/3$ και $P(\omega_2)=1/3$ για τις συναρτήσεις πυκνότητας πιθανότητας για την ταξινόμηση κλάσεων που φαίνονται στο Σχ. 2.1. Σε αυτή τη περίπτωση δεδομένου ότι μια ακολουθία μετριέται να έχει εκτιμήτρια x=14, η πιθανότητα να είναι στη κατηγορία ω_2 είναι μόλις 0.08, και στην ω_1 είναι 0.92. Για κάθε x οι εκ των υστέρων πιθανότητες έχουν σύνολο 1.0.

$$P(error) = \int_{-\infty}^{\infty} P(error,x)\,dx = \int_{-\infty}^{\infty} P(error,x)\,p(x)\,dx \quad (5)$$

και αν για κάθε x εξασφαλίζαμε ότι P(error|x) θα ήταν όσο μικρό είναι δυνατόν, τότε ο ακέραιος θα ήταν όσο μικρότερος γίνεται. Επομένως έχουμε διατυπώσει τον επόμενο κανόνα απόφασης του Bayes για την ελαχιστοποίηση της πιθανότητας σφάλματος:

Επιλέγουμε ω_1 αν $P(\omega_1|x) > P(\omega_2|x)$. Αλλιώς επιλέγουμε ω_2 (6)

Και με αυτό τον κανόνα η εξίσωση 4 γίνεται

$$P(error|\mathrm{x}) = \min\left[P(\omega_1|\mathrm{x}), P(\omega_2|\mathrm{x})\right] \quad (7)$$

Αυτή η μορφή του κανόνα απόφασης εστιάζει στο ρόλο των εκ των υστέρων πιθανοτήτων. Χρησιμοποιώντας την Εξ. 1, μπορούμε να εκφράσουμε τον κανόνα με όρους υπό συνθήκης και εκ των προτέρων πιθανοτήτων. Πρώτα παρατηρήστε ότι το *στοιχείο* p(x), στην Εξ. 1 όσο αναφορά τη λήψη μιας απόφασης, είναι απλά ένας παράγοντας που δηλώνει πόσο συχνά θα παίρνουμε τιμές σε μια ακολουθία της εκτιμήτριας x; Η ύπαρξη του στη Εξ. 1 μας διαβεβαιώνει ότι $P(\omega_1|x)+P(\omega_2|x) = 1$. Διαγράφοντας αυτό τον παράγοντα, έχουμε τον επόμενο εντελώς ισοδύναμο κανόνα:

Επιλέγουμε ω_1 αν $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$. Αλλιώς επιλέγουμε ω_2. (8)

Μερικές επιπλέον πληροφορίες μπορούν να αποκτηθούν αν αναλογιστούμε μερικές ειδικές περιπτώσεις. Αν για κάποιο x έχουμε $p(x|\omega_1)= p(x|\omega_2)$, τότε αυτή η παρατήρηση δε μας πληροφορία για τη κατάσταση της φύσης. Σε αυτή την περίπτωση η απόφαση έγκειται απόλυτα στις εκ των προτέρων πιθανότητες. Από την άλλη, αν $P(\omega_1)=P(\omega_2)$, οι καταστάσεις τις φύσης είναι εξίσου πιθανές. Σε αυτή την περίπτωση η απόφαση βασίζεται εξ' ολοκλήρου στις $p(x|\omega_j)$. Γενικά, και οι δύο αυτοί παράγοντες είναι σημαντικοί στη λήψη μιας απόφασης, και ο κανόνας απόφασης του Bayes τους συνδυάζει για να πετύχει την ελάχιστη πιθανότητα σφάλματος.

2.2. Θεωρία Απόφασης του Bayes – Συνεχείς Εκτιμήτριες

Θα τυποποιήσουμε τώρα τις ιδέες που αναφέραμε, και θα τις γενικεύσουμε με τέσσερις τρόπους:

• Επιτρέποντας τη χρήση παραπάνω από μιας εκτιμήτριας
• Επιτρέποντας δύο καταστάσεις της φύσης
• Επιτρέποντας ενέργειες διαφορετικές από την απόφαση της κατάστασης της φύσης
• Εισάγοντας μια συνάρτηση απώλειας πιο γενική από την πιθανότητα σφάλματος.

Αυτές οι γενικεύσεις και οι συνοδευτικές περιπλοκές σήμανσής τους δε πρέπει να εμποδίσουν τα κεντρικά σημεία που απεικονίζονται στο απλό παράδειγμα μας. Επιτρέποντας τη χρήση παραπάνω από μιας εκτιμήτριας απαιτεί αντικατάσταση του x με το διάνυσμα εκτιμήτριας x, όπου x είναι ένας d-διαστάσεων Ευκλείδειος χώρος R^d, που ονομάζεται διάστημα εκτιμήτριας. Επιτρέποντας περισσότερες από δύο καταστάσεις της φύσης μας δίνει μια χρήσιμη γενίκευση για ένα μικρό βάρος σήμανσης. Επιτρέποντας ενέργειες άλλες από την ταξινόμηση αρχικά επιτρέπει την πιθανότητα της απόρριψης, π.χ., την άρνηση της λήψης της απόφασης για παρόμοιες περιπτώσεις; Αυτό είναι μια χρήσιμη επιλογή αν η αναποφασιστικότητα δε είναι πολύ δαπανηρή. Τυπικά, η συνάρτηση απώλειας δηλώνει ακριβώς πόσο δαπανηρή είναι η κάθε απόφαση, και χρησιμοποιείται για να μετατρέψει έναν πιθανολογικό προσδιορισμό σε απόφαση. Οι συναρτήσεις κόστους μας αφήνουν να αντιμετωπίζουμε καταστάσεις στις οποίες κάποια λάθη ταξινόμησης κλάσεων είναι πιο δαπανηρά από άλλες, αν και συχνά συζητάμε την πιο απλή κατάσταση, όπου όλα τα λάθη έχουν το ίδιο βάρος. Με αυτά σαν πρόλογο, ας αρχίσουμε μια πιο επίσημη επεξεργασία.

Έστω $\omega_1, \ldots, \omega_c$ είναι το πεπερασμένο σύνολο των c καταστάσεων της φύσης (“κατηγορίες”) και $\alpha_1, \ldots, \alpha_a$ είναι το πεπερασμένο σύνολο των πιθανών ενεργειών. Η συνάρτηση απώλειας $\lambda(\alpha_i|\omega_j)$ περιγράφει την απώλεια που προκύπτει όταν επιλέγουμε την ενέργεια α_i όταν η κατάσταση της φύσης είναι ω_j . Έστω το διάνυσμα εκτιμήτριας x ότι είναι μία d-διαστάσεων διανυσματικά εκτιμούμενη τυχαία μεταβλητή, και έστω $p(x|\omega_j)$ είναι υπό όρους κατάστασης η συνάρτηση πυκνότητας πιθανότητας για το x — η συνάρτηση πυκνότητας πιθανότητας για το x με τη προϋπόθεση ότι ω_j είναι η σωστή κατάσταση της φύσης. Όπως προηγουμένως, το $P(\omega_j)$ την εκ των προτέρων πιθανότητα η κατάσταση της φύσης να είναι ω_j . Τότε η εκ των υστέρων πιθανότητα $P(\omega_j|x)$ μπορεί να υπολογιστεί από την $p(x|\omega_j)$ μέσω του τύπου του Bayes:

$$P\left(\omega_{\mathrm{j}}\,|\mathrm{x}\right)=\frac{p\left(\mathrm{x}|\omega_{\mathrm{j}}\right)P\left(\omega_j\right)}{p\left(x\right)} \tag{9}$$

όπου το στοιχείο είναι τώρα

$$p\left(x\right)=\sum_{j=1}^{c}p\left(\mathrm{x}|\omega_{\mathrm{j}}\right)P\left(\omega_j\right) \tag{10}$$

Υποθέστε ότι παρατηρούμε ένα συγκεκριμένο x και ότι αναλογιζόμαστε να επιλέξουμε την ενέργεια α_i. Αν η πραγματική κατάσταση της φύσης είναι ω_j , εξ’ ορισμού θα έχουμε μια απώλεια $\lambda(\alpha_i|\omega_j)$. Αφού $P(\omega_j|x)$ είναι η πιθανότητα η σωστή κατάσταση της φύσης να είναι ω_j, η αναμενόμενη απώλεια σε συνδυασμό με την ενέργεια α_i είναι μόνο:

$$R\left(\alpha_{\mathrm{i}}|\mathrm{x}\right)=\sum_{j=1}^{c}\lambda\left(\alpha_{\mathrm{i}}|\omega_{\mathrm{j}}\right)P\left(\omega_{\mathrm{j}}|\mathrm{x}\right) \tag{11}$$

Σε όρους της θεωρίας αποφάσεων, μια αναμενόμενη απώλεια καλείται ρίσκο, και $R(\alpha_i|x)$ καλείται δεσμευμένο ρίσκο. Όποτε ερχόμαστε αντιμέτωποι με μια παρατήρηση του x, μπορούμε να ελαχιστοποιήσουμε την αναμενόμενη απώλεια επιλέγοντας την ενέργεια που θα αποφέρει τη μικρότερο δεσμευμένο ρίσκο. Θα δείξουμε τώρα ότι αυτή η διαδικασία απόφασης του Bayes μας δίνει στην πραγματικότητα την βέλτιστη απόδοση στο συνολικό ρίσκο.

Τυπικά καθορισμένο, το πρόβλημα μας είναι να βρούμε ένα κανόνα απόφασης για $P(\omega_j)$ ο οποίος ελαχιστοποιεί το συνολικό ρίσκο. Ένας γενικός κανόνας απόφασης είναι μια συνάρτηση α(x) η οποία μας λέει ποια ενέργεια πρέπει να ακολουθήσουμε για κάθε δυνατή παρατήρηση. Για να είμαστε πιο συγκεκριμένοι, για κάθε x η συνάρτηση απόφασης α(x) παίρνει μια από τις a τιμές $\alpha_1, \ldots, \alpha_a$. Το συνολικό ρίσκο R είναι η αναμενόμενη απώλεια συνδυασμένη με ένα δοσμένο κανόνα απόφασης. Αφού $R(\alpha_i|x)$ είναι το δεσμευμένο ρίσκο συνδυασμένο με την ενέργεια α_i, και αφού ο κανόνας απόφασης καθορίζει την ενέργεια, το συνολικό ρίσκο δίνεται από:

$$R=\int R\left(a\left(x\right)|\mathrm{x}\right)p\left(x\right)dx \tag{12}$$

όπου dx είναι η σήμανση για ένα d-διαστάσεων στοιχείο, και όπου ο ακέραιος εκτείνεται σε όλο τον χώρο της εκτιμήτριας. Προφανώς, αν α(x) επιλεχθεί ώστε $R(\alpha_i(x))$ είναι όσο το δυνατόν μικρότερο για κάθε x, τότε το συνολικό ρίσκο θα ελαχιστοποιηθεί. Αυτό επιβεβαιώνει το επόμενη δήλωση του κανόνα απόφασης του Bayes: Για να ελαχιστοποιήσουμε το συνολικό ρίσκο, υπολογίζουμε το δεσμευμένο ρίσκο

$$R\left(\alpha_{\mathrm{i}}|\mathrm{x}\right)=\sum_{j=1}^{c}\lambda\left(\alpha_{\mathrm{i}}|\omega_{\mathrm{j}}\right)P\left(\omega_{\mathrm{j}}|\mathrm{x}\right) \tag{13}$$

για i= 1,...,a και επιλέγουμε την ενέργεια α_i για την οποία το $R(\alpha_i|x)$ είναι ελάχιστο.* Το ελάχιστο αποτέλεσμα του συνολικού ρίσκου καλείται ρίσκο του Bayes, δηλώνεται ως R^*, και είναι η καλύτερη εφικτή απόδοση.

2.2.1 Ταξινόμηση Δύο Κατηγοριών

Ας συνυπολογίσουμε αυτά τα αποτελέσματα όταν εφαρμόζονται στην ειδική περίπτωση προβλημάτων ταξινόμησης δύο κατηγοριών. Εδώ η ενέργεια α_1 αντιστοιχεί στην απόφαση ότι η σωστή κατάσταση της φύσης είναι ω_1, και η ενέργεια α_2 αντιστοιχεί στην απόφαση ότι η σωστή κατάσταση της φύσης είναι ω2. Για απλότητα σήμανσης, ας είναι $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ η απώλεια που προκύπτει για την επιλογή ω_i όταν η σωστή κατάσταση της φύσης είναι ω_j . Αν γράψουμε το δεσμευμένο ρίσκο που δίνεται από την Εξ. 13, έχουμε:

$$R\left(\alpha_1|\mathrm{x}\right)=\lambda_{11}P\left(\omega_1|\mathrm{x}\right)+\lambda_{12}P\left(\omega_2|\mathrm{x}\right) \quad \text{και}$$

$$R\left(\alpha_2|\mathrm{x}\right)=\lambda_{21}P\left(\omega_1|\mathrm{x}\right)+\lambda_{22}P\left(\omega_2|\mathrm{x}\right). \tag{14}$$

Υπάρχουν πολλοί τρόποι να εκφραστεί ο κανόνας απόφασης ελαχίστου ρίσκου, και καθένας έχει τα δικά του μικρά πλεονεκτήματα. Ο θεμελιώδης κανόνας είναι να επιλέξουμε ω_1 αν $R(\alpha_1|x)<R(\alpha_2|x)$. Σε όρους εκ των υστέρων πιθανοτήτων, επιλέγουμε ω_1 αν

$$\left(\lambda_{21}-\lambda_{11}\right)P\left(\omega_1|\mathrm{x}\right)>\left(\lambda_{12}-\lambda_{22}\right)P\left(\omega_2|\mathrm{x}\right) \tag{15}$$

* Προσέξτε ότι αν περισσότερες από μια ενέργειες ελαχιστοποιούν την R(αx), δε παίζει ρόλο ποια ενέργεια θα ληφθεί, και οποιοσδήποτε βολικός κανόνας που θα δώσει μια απόφαση για τις ισοδύναμες ενέργειες μπορεί να χρησιμοποιηθεί.

Συνήθως, η απώλεια που προκύπτει από ένα λάθος είναι μεγαλύτερη από την απώλεια για ένα σωστό, και οι δύο παράγοντες $\lambda_{21}-\lambda_{11}$ και $\lambda_{12}-\lambda_{22}$ είναι θετικοί. Κατά συνέπεια στην πράξη, η απόφασή μας καθορίζεται γενικά από την πιθανότερη κατάσταση της φύσης, αν και πρέπει να βαθμονομήσουμε τις εκ των υστέρων πιθανότητες με τις διαφορές στις απώλειες. Χρησιμοποιώντας το τύπο του Bayes , μπορούμε να αντικαταστήσουμε τις εκ των υστέρων πιθανότητες με τις εκ των προτέρων πιθανότητες και τις δεσμευμένες πυκνότητες. Αυτό οδηγεί στον ισοδύναμο κανόνα, να αποφασίσουμε ω_1 αν

$$(\lambda_{21}-\lambda_{11})\,p(x|\omega_1)\,P(\omega_1) > (\lambda_{12}-\lambda_{22})\,p(x|\omega_2)\,P(\omega_2), \qquad (16)$$

αλλιώς ω_2.

Μια άλλη εναλλακτική, η οποία ακολουθεί το λογικό συμπέρασμα ότι $\lambda_{21}>\lambda_{11}$, είναι να επιλέξουμε ω_1 αν

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(\omega_1)}{P(\omega_2)}. \qquad (17)$$

Αυτή η μορφή του κανόνα απόφασης εστιάζει στην εξάρτηση από το x των πυκνοτήτων πιθανότητας. Μπορούμε να θεωρήσουμε $p(x|\omega_j)$ μια συνάρτηση του ω_j (π.χ., η συνάρτηση πιθανότητας), και μετά να σχηματίσουμε το ***ρυθμό πιθανότητας*** $p(x|\omega_1)/p(x|\omega_2)$. Επομένως ο κανόνας απόφασης του Bayes μπορεί να ερμηνευτεί ως κλήση για επιλογή του ω_1 αν ο ρυθμός πιθανότητας υπερβαίνει μια τιμή κατωφλίου που είναι ανεξάρτητη από την παρατήρηση του x.

2.3 Ταξινόμηση Ποσοστού Ελάχιστου Σφάλματος

Σε προβλήματα ταξινόμησης, κάθε κατάσταση της φύσης συνήθως συνδέεται με μια διαφορετική από τις c κλάσεις, και η ενέργεια α_i εκφράζεται συνήθως ως η απόφαση ότι η σωστή κατάσταση της φύσης είναι ω_i. Αν επιλεχθεί η ενέργεια α_i και η πραγματική κατάσταση της φύσης είναι ω_j , τότε η απόφαση είναι σωστή αν i=j, και λάθος αν $i \neq j$. Για να αποφευχθούν τα λάθη, είναι φυσικό να αναζητήσουμε ένα κανόνα απόφασης που ελαχιστοποιεί την πιθανότητα σφάλματος, π.χ., το ποσοστό λάθους. Η συνάρτηση απώλειας για αυτή την περίπτωση είναι η καλούμενη *συμμετρική* ή *μηδέν-ένα* συνάρτηση απώλειας,

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i=j \\ 1 & i\neq j \end{cases} \qquad i,j=1,\ldots,c \qquad (18)$$

Αυτή η συνάρτηση απώλειας δεν εισάγει καθόλου απώλεια σε μια σωστή απόφαση, και εισάγει μια μοναδιαία απώλεια σε κάθε λάθος. Επομένως όλα τα λάθη έχουν το ίδιο κόστος.* Το ρίσκο που αντιστοιχεί σε αυτή τη συνάρτηση απώλειας είναι ακριβώς η μέση πιθανότητα σφάλματος, αφού το δεσμευμένο ρίσκο είναι

$$\begin{aligned} R(\alpha_i|\mathrm{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_\mathrm{j}) P(\omega_j|\mathrm{x}) \\ &= \sum_{j\neq i} P(\omega_j|\mathrm{x}) \\ &= 1 - P(\omega_i|\mathrm{x}) \end{aligned} \tag{19}$$

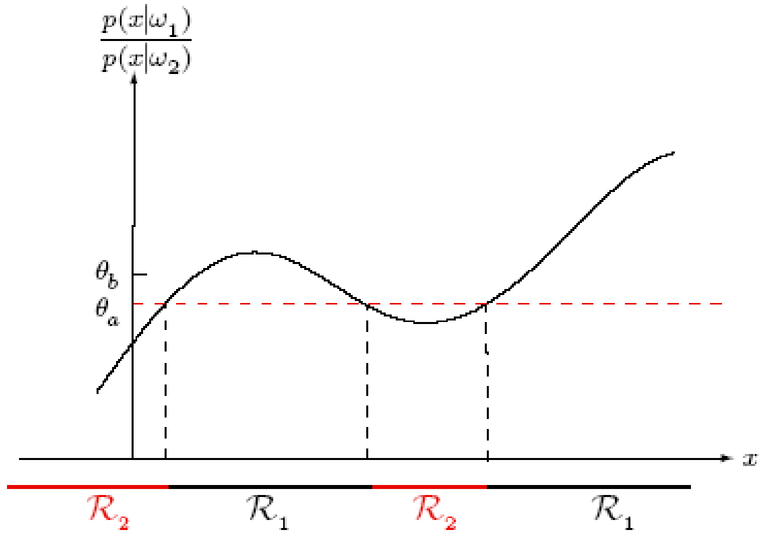
και $P(\omega_i|x)$ είναι η δεσμευμένη πιθανότητα ότι η ενέργεια α_i είναι σωστή. Ο κανόνας απόφασης του Bayes για την ελαχιστοποίηση ρίσκου θέλει την επιλογή της ενέργειας που ελαχιστοποιεί το δεσμευμένο ρίσκο. Επομένως, για να ελαχιστοποιηθεί η μέση πιθανότητα σφάλματος, θα επιλέγαμε το $_i$ που μεγιστοποιεί την εκ των υστέρων πιθανότητα $P(\omega_i|x)$. Με άλλα λόγια, για ελάχιστο ποσοστό σφάλματος:

Επιλέγουμε ω_i αν $P(\omega_i|x) > P(\omega_j|x)$ για όλα τα $_j = _i$. (20)

Αυτός είναι ο ίδιος κανόνας με την Εξ. 6.

Είδαμε στο Σχ. 2.2 μερικές πυκνότητες πιθανότητας για την ταξινόμηση κλάσεων και τις εκ των υστέρων πιθανότητες; Το Σχ. 2.3 δείχνει το ρυθμό πιθανότητας $p(x|\omega_1)/p(x\omega2)$ για την ίδια περίπτωση. Γενικά, αυτός ο ρυθμός μπορεί να πάρει τιμές μεταξύ μηδέν και άπειρο. Η τιμή κατωφλίου θα είναι οι ίδιες εκ των προτέρων πιθανότητες αλλά με συναρτήσεις μηδέν-ένα. Παρατηρήστε ότι αυτό οδηγεί σε απόφαση ίδιων ορίων όπως στο Σχ. 2.2, όπως θα έπρεπε. Αν τιμωρήσουμε τα λάθη ταξινομώντας ακολουθίες ω_1 ως ω2 περισσότερο από ότι αντίστροφα (π.χ., $\lambda_{21} > \lambda_{12}$), τότε η Εξ. 17 οδηγεί στο σημειωμένο κατώφλι θ_b . Προσέξτε ότι οι τιμές του x για τις οποίες ταξινομούμε μια ακολουθία σαν ω_1 γίνονται μικρότερες, όπως θα έπρεπε.

* Επισημαίνουμε ότι οι υπόλοιπες συναρτήσεις απώλειας, όπως η δευτέρου βαθμού και η γραμμικής διαφοράς, βρίσκουν ευρύτερη χρήση στους στόχους οπισθοδρόμησης, όπου υπάρχει μια φυσική σειρά στις προβλέψεις και μπορούμε με σημασία να τιμωρήσουμε τις προβλέψεις που είναι «πιο λάθος» από άλλες.



Σχήμα 2.3: Οι ρυθμοί πιθανότητας $p(x|\omega_1)/p(x|\omega_2)$ για τις κατανομές που φαίνονται στο Σχ. 2.1. Αν χρησιμοποιήσουμε απώλεια μηδέν-ένα ή ταξινόμησης, τα όρια της απόφασης μας καθορίζονται από το κατώφλι θ_a. Αν η συνάρτηση απώλειας τιμωρεί τις λάθος κατηγοριοποιημένες ακολουθίες ω_2 ως ω_1 περισσότερο από το αντίστροφο, (π.χ., $\lambda_{12}>\lambda_{21}$), παίρνουμε μεγαλύτερο κατώφλι θb, και επομένως το R_1 γίνεται μικρότερο.

2.3.1 *Κριτήριο Μεγίστου Ελαχίστου

Μερικές φορές πρέπει να σχεδιάσουμε ένα ταξινομητή να λειτουργεί καλά ένα εύρος εκ των προτέρων πιθανοτήτων. Για παράδειγμα, στο πρόβλημα κατηγοριοποίησης ψαριών μπορούμε να φανταστούμε ότι όσο οι φυσικές ιδιότητες της ελαφρότητας και του πλάτους κάθε είδους ψαριού παραμένουν σταθερές, οι εκ των προτέρων πιθανότητες μπορεί να διαφέρουν αρκετά και με απρόβλεπτο τρόπο, ή εναλλακτικά θέλουμε να χρησιμοποιήσουμε τον ταξινομητή σε ένα διαφορετικό φυτό που δε ξέρουμε τις εκ των προτέρων πιθανότητες. Μια λογική προσέγγιση είναι τότε να δημιουργήσουμε τον ταξινομητή μας έτσι ώστε το χειρότερο ολικά ρίσκο για οποιαδήποτε τιμή των εκ των προτέρων πιθανοτήτων να είναι όσο μικρότερη είναι δυνατόν — δηλαδή, να ελαχιστοποιήσουμε το μέγιστο πιθανό συνολικά ρίσκο.

Για να το καταλάβουμε αυτό, έχουμε το R_1 να δείχνει (άγνωστο μέχρι τώρα) περιοχή στο χώρο εκτιμήτριας όπου ο ταξινομητής αποφασίζει ω_1 και το ίδιο για τα R_2 και ω_2, και μετά γράφουμε το συνολικό ρίσκο Εξ. 12 σε όρους δεσμευμένου ρίσκου:

$$\begin{aligned} R = &\int_{R_1} \left[\lambda_{11} P(\omega_1) p(x|\omega_1) + \lambda_{12} P(\omega_2) p(x|\omega_2) \right] \\ &+ \int_{R_2} \left[\lambda_{21} P(\omega_1) p(x|\omega_1) + \lambda_{22} P(\omega_2) p(x|\omega_2) \right] \end{aligned} \tag{21}$$

Μπορούμε να χρησιμοποιήσουμε το γεγονός $P(\omega_2)=1-P(\omega_1)$ και ότι $\int_{R_1} p(x|\omega_1)dx = 1 - \int_{R_2} p(x|\omega_1)dx$ και μπορούμε να ξαναγράψουμε το ρίσκο ως:

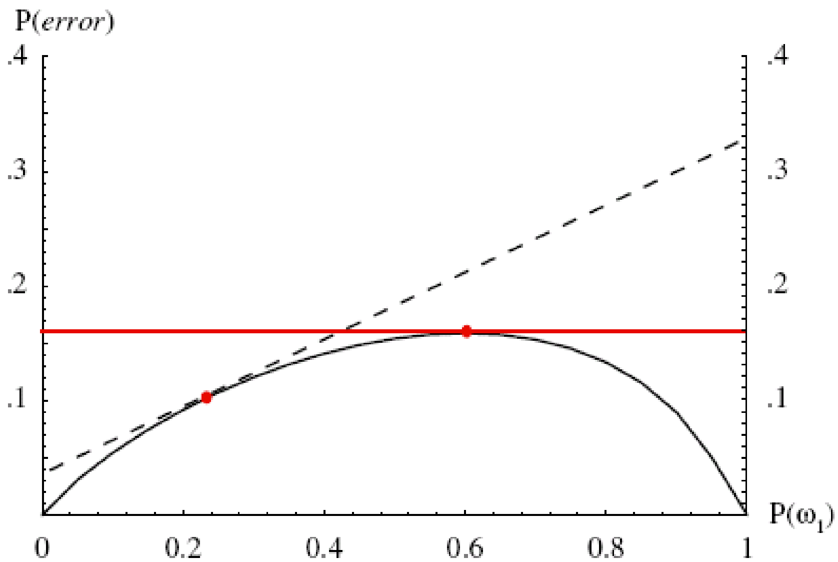
$$R(P(\omega_1)) = \overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx}^{=R_{mm},\ minimax\ \ risk}$$
$$+P(\omega_1) \underbrace{\left[(\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)\,dx - (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)\,dx \right]}_{0\ \ for\ \ minimax\ \ solution}. \quad (22)$$

Αυτή η εξίσωση δείχνει ότι μόλις όρια έχουν οριστεί (π.χ., R_1 και R_2), το συνολικό ρίσκο είναι γραμμικό της $P(\omega_1)$. Αν μπορούμε να βρούμε ένα όριο τέτοιο ώστε η σταθερά αναλογίας να είναι 0, τότε το ρίσκο είναι ανεξάρτητο των εκ των προτέρων πιθανοτήτων. Αυτή είναι η λύση Μεγίστου Ελαχίστου, και το ρίσκο μεγίστου ελαχίστου, R_{mm}, μπορεί να ληφθεί από την Εξ. 22:

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(x|\omega_1)dx. \end{aligned} \quad (23)$$

Στο σχήμα 2.4 φαίνεται αυτή η προσέγγιση. Εν συντομία, ψάχνουμε για την εκ των προτέρων για την οποία το ρίσκο του Bayes είναι μέγιστο, το αντίστοιχο όριο απόφασης δίνει τη λύση μεγίστου ελαχίστου. Η τιμή του ρίσκου μεγίστου ελαχίστου, R_{mm}, είναι σα συνέπεια ίσο με το χειρότερο ρίσκο Bayes. Πρακτικά, η εύρεση του ορίου απόφασης για ρίσκο μεγίστου ελαχίστου μπορεί να είναι δύσκολη, ειδικά όταν οι κατανομές είναι περίπλοκες. Εντούτοις, σε μερικές περιπτώσεις το όριο μπορεί να βρεθεί αναλυτικά (Πρόβλημα 3).

Το κριτήριο μεγίστου ελαχίστου βρίσκει ευρεία εφαρμογή στη θεωρεία παιχνιδιών από ότι στην παραδοσιακή αναγνώριση προτύπων. Στη θεωρία παιχνιδιών, έχετε ένα εχθρικό αντίπαλο που αναμένεται να κάνει μια ενέργεια μέγιστης καταστροφής για σας. Επομένως είναι λογικό να ενεργήσετε (π.χ. να κάνετε μια ταξινόμηση) όπου το κόστος — λόγω των επόμενων ενεργειών των αντίπαλων σας — ελαχιστοποιούνται.



Σχήμα 2.4: Η καμπύλη στο κάτω μέρος δείχνει το ελάχιστο (Bayes) λάθος σαν συνάρτηση εκ των προτέρων πιθανοτήτων $P(\omega_1)$ σε ένα πρόβλημα ταξινόμησης δύο κατηγοριών με σταθερές κατανομές. Για κάθε τιμή των εκ των προτέρων (π.χ., $P(\omega_1)=0.25$) υπάρχει ένα αντίστοιχο βέλτιστο όριο απόφασης και συνδυασμένος ρυθμός σφάλματος του Bayes. Για κάθε (προσαρμοσμένο) τέτοιο όριο, αν οι εκ των προτέρων αλλαχθούν, η πιθανότητα σφάλματος θα αλλάξει σαν γραμμική συνάρτηση του $P(\omega_1)$ (διακεκομμένη γραμμή). Το μέγιστο τέτοιο λάθος θα συμβεί σε μια ακραία τιμή των εκ των προτέρων, εδώ στο $P(\omega_1) = 1$. Για να ελαχιστοποιήσουμε το μέγιστο ενός τέτοιου λάθους, θα πρέπει να σχεδιάσουμε το όριο απόφασης για το μέγιστο σφάλμα του Bayes (εδώ $P(\omega_1)=0.6$), και επομένως το σφάλμα δε θα αλλάξει ως συνάρτηση των εκ των προτέρων, όπως φαίνεται από την συμπαγή κόκκινη γραμμή.

2.3.2 * Κριτήριο των Neyman-Pearson

Σε μερικά προβλήματα, μπορεί να θέλουμε να ελαχιστοποιήσουμε το συνολικό ρίσκο σε ένα περιορισμό; Για παράδειγμα, μπορεί να θέλουμε να ελαχιστοποιήσουμε το συνολικό ρίσκο σε ένα περιορισμό $\int R(\alpha_i|\mathrm{x})dx < constant$ για κάποιο συγκεκριμένο i. Ένας τέτοιος περιορισμός μπορεί να προκύψει όταν υπάρχει ένας σταθερός πόρος που συνοδεύει μια ιδιαίτερη ενέργεια α_i, η όταν δεν πρέπει να καταχωρήσουμε λάθος το πρότυπο από μια ιδιαίτερη κατάσταση της φύσης ω_i σε περισσότερο από κάποια περιορισμένη συχνότητα. Παραδείγματος χάριν, στο παράδειγμα ψαριών μας, να υπάρξει κάποιος κυβερνητικός κανονισμός ότι δεν πρέπει να καταχωρήσουμε λάθος περισσότερο από 1% από το σολομό ή τις πέρκες θάλασσας. Έπειτα να επιδιώξουμε μια απόφαση αυτός ελαχιστοποιεί την πιθανότητα ταξινόμησης μια πέρκα θάλασσας ως σολομό.

Ικανοποιούμε γενικά το κριτήριο Neyman-Pearson με να ρυθμίσει τα όρια απόφασης αριθμητικά. Εντούτοις, για τις γκαουσσιανές και μερικές άλλες κατανομές, λύσεις του Neyman-Pearson μπορούν να βρεθούν αναλυτικά (Προβλήματα 5 & 6). Θα έχουμε την αιτία για να αναφέρουμε τα κριτήρια Neyman-Pearson ξανά στο Κεφ. 2.8.3 στα χαρακτηριστικά λειτουργίας.

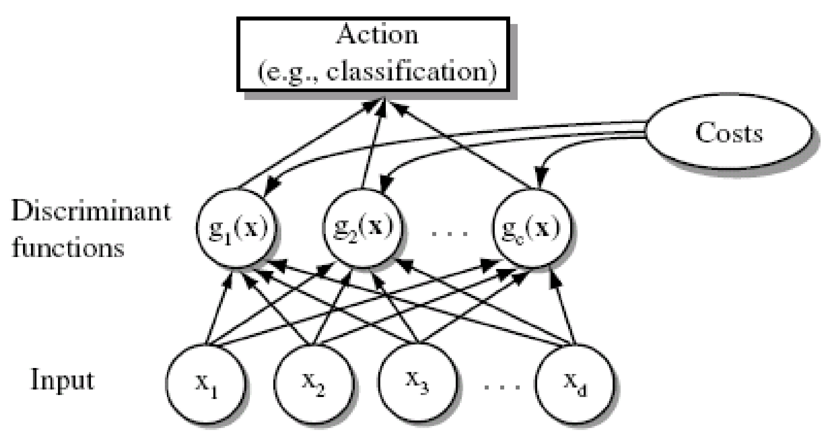
2.4 Ταξινομητές, Διακρίνουσες συναρτήσεις και Επιφάνειες αποφάσης

2.4.1 Η περίπτωση πολλών κατηγοριών

Υπάρχουν πολλοί διαφορετικοί τρόποι που αντιπροσωπεύουν ταξινομητές προτύπων. Ένας από τους πιο χρήσιμους είναι μια σειρά από *διακρίνουσες συναρτήσεις* $g_i(x)$, i =1, ..., c. Ο ταξινομητής χρησιμοποιείται για να ορίσει ένα διάνυσμα εκτιμήτριας x στη κλάση ω_i αν

$$g_i(x) > g_j(x) \qquad \text{για κάθε } j \neq I \tag{24}$$

Κατά συνέπεια, ο ταξινομητής αντιμετωπίζεται ως δίκτυο ή μηχανή που υπολογίζει c *διακρίνουσες συναρτήσεις* και επιλέγει την κατηγορία που αντιστοιχεί στη μεγαλύτερη διακρίνουσα συνάρτηση. Μια δικτυακή αναπαράσταση ενός ταξινομητή φαίνεται στο Σχ. 2.5.



Σχήμα 2.5: Η λειτουργική δομή ενός γενικού στατιστικού ταξινομητή προτύπων ο οποίος περιλαμβάνει d εισόδους και c διακρίνουσες συναρτήσεις $g_i(x)$. Ένα επόμενο βήμα καθορίζει ποια από τις τιμές διακρίνουσας συνάρτησης είναι η μέγιστη, και ταξινομεί το πρότυπο εισαγωγής αναλόγως. Τα βέλη παρουσιάζουν την κατεύθυνση της ροής των πληροφοριών, αν και συχνά τα βέλη παραλείπονται όταν η κατεύθυνση της ροής είναι αυτονόητη.

Ένας ταξινομητής Bayes εύκολα και φυσικά αναπαρίσταται κατ' αυτό τον τρόπο. Για τη γενική περίπτωση με τα ρίσκα, μπορούμε να θέσουμε $g_i(x) = -R(\alpha_i|x)$, από τη μέγιστη διακρίνουσα συνάρτηση αντιστοιχεί έπειτα στο ελάχιστο δεσμευμένο ρίσκο. Για την περίπτωση του ελάχιστου ποσοστού σφαλμάτων, μπορούμε να απλοποιήσουμε τα πράγματα περαιτέρω με τη λήψη $g_i(x)=P(\omega_i|x)$, έτσι ώστε η μέγιστη διακρίνουσα συνάρτηση αντιστοιχεί στη μέγιστη εκ των υστέρων πιθανότητα.

Σαφώς, η επιλογή διακρίνουσων συναρτήσεων δεν είναι μοναδική. Μπορούμε πάντα να πολλαπλασιάσουμε όλες τις διακρίνουσες συναρτήσεις με την ίδια θετική σταθερά ή να τους μετατοπίσουμε με την ίδια προσθετική σταθερά χωρίς επηρεασμό της απόφασης. Γενικότερα, εάν αντικαθιστάμε το κάθε $g_i(x)$ με $f(g_i(x))$, όπου $f(\cdot)$είναι μια μονότονα αυξανόμενη συνάρτηση, η προκύπτουσα ταξινόμηση είναι αμετάβλητη. Αυτή η παρατήρηση μπορεί να οδηγήσει σε σημαντικές αναλυτικές και υπολογιστικές απλοποιήσεις. Ειδικότερα, για την ταξινόμηση ελάχιστου ποσοστού σφαλμάτων, οποιαδήποτε από τις ακόλουθες επιλογές δίνει τα ίδια αποτελέσματα ταξινόμησης, αλλά μερικά μπορούν να είναι πολύ απλούστερα να κατανοηθούν ή να υπολογιστούν από άλλα:

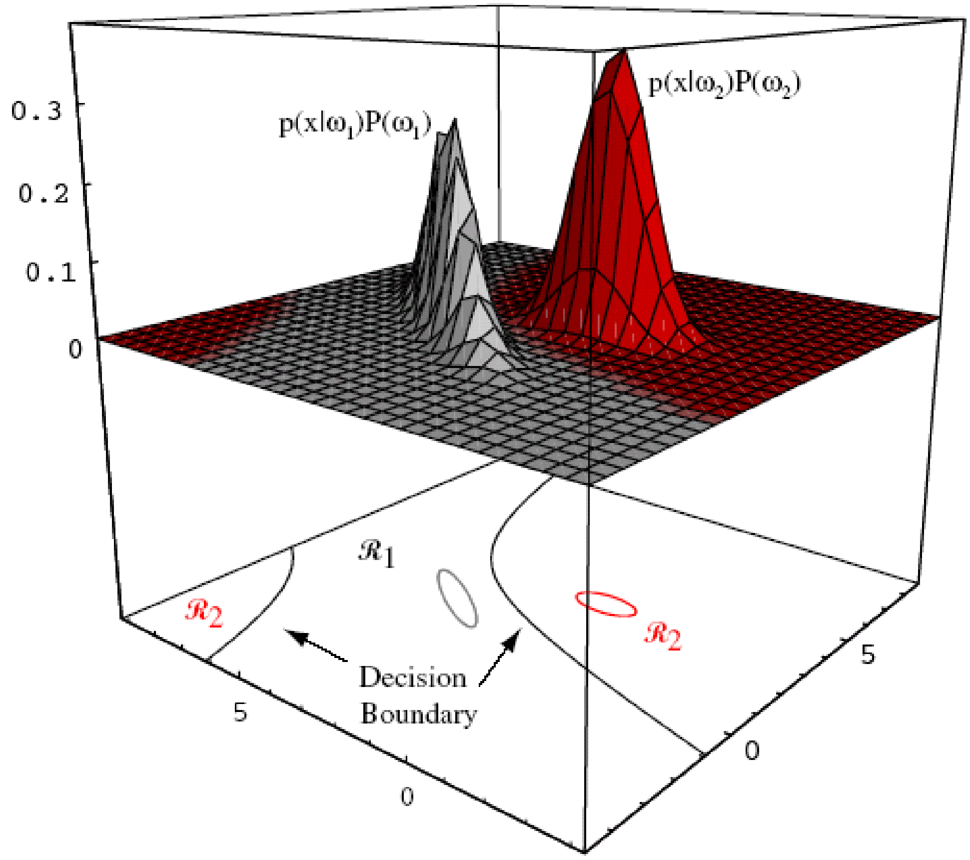
$$g_i(x) = P(\omega_i|\mathrm{x}) = \frac{p(x|\omega_\mathrm{i})P(\omega_i)}{\sum_{j=1}^{c} p(x|\omega_\mathrm{i})P(\omega_i)} \tag{25}$$

$$g_i(x) = p(x|\omega_\mathrm{i})P(\omega_i) \tag{26}$$

$$g_i(x) = \ln p(x|\omega_\mathrm{i}) + \ln P(\omega_i) \tag{27}$$

όπου το ln δηλώνει φυσικό λογάριθμο.

Ακόμα κι αν οι διακρίνουσες συναρτήσεις μπορούν να γραφτούν με ποικίλες μορφές, οι κανόνες απόφασης είναι ισοδύναμοι. Η επίδραση οποιουδήποτε κανόνα απόφασης είναι να διαιρέσει το χώρο εκτιμήτριας σε c περιοχές απόφασης, R_1,...,Rc. Αν $g_i(x)>g_j(x$ για κάθε $j \neq i$, τότε x είναι στη περιοχή R_i, και ο κανόνας απόφασης μας λέει να τοποθετήσουμε το x στο ω_i. Οι περιοχές είναι χωρισμένες από τα όρια απόφασης, επιφάνειες στο χώρο εκτιμήτριας όπου εμφανίζονται ισοπαλίες μεταξύ των μεγαλύτερων διακρίνουσων συναρτήσεων (Σχ. 2.6).



Σχήμα 2.6: Σε αυτό το δισδιάστατο ταξινομητή δύο κατηγοριών, οι πυκνότητες πιθανότητας είναι γκαουσσιανές (με 1/e ελλείψεις που παρουσιάζονται), το όριο απόφασης αποτελείται από δύο υπερβολές, και έτσι η περιοχή απόφασης R_2 δεν συνδέεται απλά.

2.4.2 Η περίπτωση δύο κατηγοριών

Ενώ η περίπτωση δύο κατηγοριών είναι ακριβώς μια ειδική περίπτωση της περίπτωσης πολλαπλών κατηγοριών, έχει λάβει παραδοσιακά χωριστή επεξεργασία. Πράγματι, ένας ταξινομητής που τοποθετεί ένα πρότυπο σε μια από μόνο δύο κατηγορίες έχει ένα ειδικό όνομα — μια *διχοτόμος*.* Αντί της χρησιμοποίησης δύο διχοτόμων διακρίνουσων συναρτήσεων g_1 και g_2 και αντιστοιχώντας x στο ω_1 αν $g_1 > g_2$, είναι πιο φυσικό να καθοριστεί μια ενιαία διακρίνουσα συνάρτηση

$$g(x) \equiv g_1(x) - g_2(x), \tag{28}$$

και για να χρησιμοποιηθεί ο ακόλουθος κανόνα απόφασης: Επιλέγουμε ω_1 αν g(x) > 0; αλλιώς επιλέγουμε ω2. Κατά συνέπεια, μια διχοτόμος μπορεί να θεωρηθεί σαν μια μηχανή ου υπολογίζει μια ενιαία διακρίνουσα συνάρτηση g(x), και ταξινομεί το x σύμφωνα με το αλγεβρικό σημάδι του αποτελέσματος. Από τις διάφορες μορφές που μπορεί να γραφτεί η διακρίνουσα συνάρτηση ελάχιστου ποσοστού λάθους, οι ακόλουθες δύο (παραγόμενες από τις Εξ. 25 & 27) είναι ιδιαίτερα βολικές:

$$g(x)=P(\omega_1|\mathrm{x})-P(\omega_2|\mathrm{x}) \tag{29}$$

$$g(x)=\ln\frac{p(\mathrm{x}|\omega_1)}{\mathrm{p}(\mathrm{x}|\omega_2)}+\ln\frac{P(\omega_1)}{P(\omega_2)}. \tag{30}$$

2.5 Η Κανονική Πυκνότητα

Η δομή ενός ταξινομητή Bayes καθορίζεται από τις δεσμευμένες πυκνότητες $p(x|\omega_i)$ καθώς και από τις εκ των προτέρων πιθανότητες. Από τις διάφορες συναρτήσεις πυκνότητας που έχουν διερευνηθεί, καμία δεν έχουμε προσέξει περισσότερο από την κανονική πολλών μεταβλητών ή την γκαουσσιανή πυκνότητα. Σε μεγάλο βαθμό αυτή η προσοχή οφείλεται στην αναλυτική ευπείθειά της. Εντούτοις η πολλών μεταβλητών κανονική πυκνότητα είναι επίσης ένα κατάλληλο πρότυπο για μια σημαντική κατάσταση, δηλαδή, την περίπτωση που τα διανύσματα εκτιμήτριας x για μια δοσμένη κλάση ω_i έχουν συνεχείς τιμές, τυχαία αλλοιωμένες εκδόσεις ενός ενιαίου χαρακτηριστικού ή ενός πρωτοτύπου διανύσματος μ_i. Σε αυτό το τμήμα παρέχουμε μια συνοπτική έκθεση της πολλών μεταβλητών κανονικής πυκνότητας, εστιάζοντας στις ιδιότητες μέγιστου ενδιαφέροντος για τα προβλήματα ταξινόμησης.

Κατ' αρχάς, θυμηθείτε τον ορισμό της *αναμενόμενης αξίας* μιας κλιμακωτής συνάρτησης f (x), καθορισμένη για κάποια πυκνότητα p(x):

$$\mathrm{E}[f(x)]\equiv\int_{-\infty}^{\infty}f(x)p(x)dx. \tag{31}$$

Αν έχουμε τιμές από ένα χώρο D από μια διακριτή κατανομή πρέπει να αθροίσουμε όλες τις τιμές μας

$$\mathrm{E}[f(x)]=\sum_{x\in D}f(x)P(x), \tag{32}$$

όπου p(x) είναι η μάζα πιθανότητας του x. Θα πρέπει συχνά να υπολογίσουμε τις αναμενόμενες τιμές — από αυτές και ανάλογες εξισώσεις που καθορίζονται σε υψηλότερες διαστάσεις (δείτε το παράρτημα Κεφ. ??, ?? & ??).*

* Ένας ταξινομητής για περισσότερες από δύο κατηγορίες καλείται πόλυδιχοτόμος.

* Θα χρησιμοποιήσουμε συχνά την κάπως χαλαρή ορολογία εφαρμοσμένης μηχανικής και θα αναφερθούμε σε ένα μοναδικό σημείο ως "δείγμα." Οι στατιστικοί, εν τούτοις, αναφέρονται πάντα σε ένα δείγμα ως συλλογή των σημείων, και το αναφέρουν ως "ένα δείγμα του μεγέθους n". Όταν λαμβάνεται υπόψη, υπάρχουν σπάνια ασάφειες σε τέτοια χρήση.

2.5.1 Μεταβλητή Πυκνότητα

Αρχίζουμε με τη συνεχή μεταβλητή κανονική ή γκαουσσιανή πυκνότητα,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right], \tag{33}$$

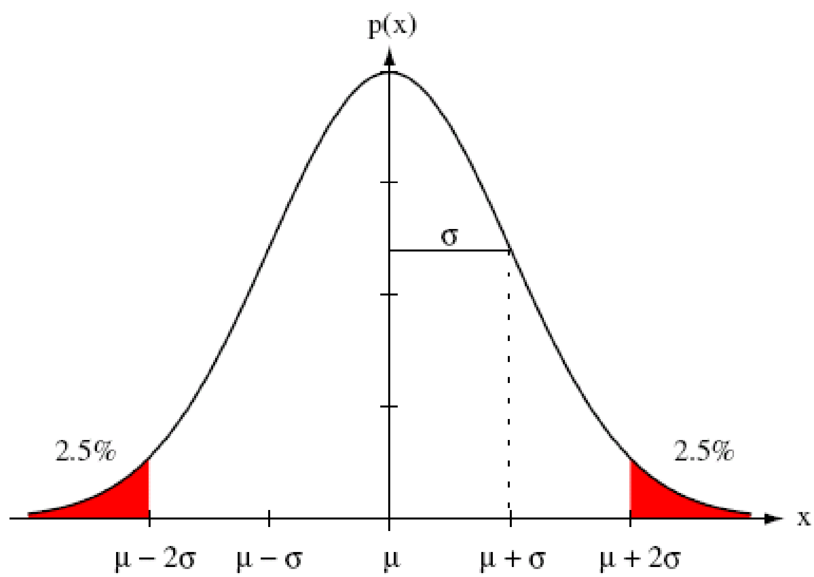
για την οποία η αναμενόμενη τιμή του x (μια μέση τιμή, εδώ έχει ληφθεί στο χώρο εκτιμήτριας) είναι

$$\mu \equiv \mathrm{E}[x] = \int_{-\infty}^{\infty} x p(x) dx, \tag{34}$$

και όπου το αναμενόμενο τετράγωνο της τυπικής απόκλισης ή *μεταβλητότητα*

$$\sigma^2 \equiv \mathrm{E}\left[(x-\mu)^2 \right] = \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x) dx. \tag{35}$$

Η μεταβλητή κανονική πυκνότητα διευκρινίζεται από δύο παραμέτρους: το μέσο μ και την μεταβλητότητα σ^2 . Για απλότητα, βλέπουμε συχνά την Εξ. 33 γράφοντας $p(x) \sim \mathrm{N}(\mu, \sigma^2)$ για να δηλώσουμε ότι x έχει κανονική κατανομή με μέσο μ μεταβλητότητα σ^2. Τα δείγματα από τις κανονικές κατανομές τείνουν να συγκεντρωθούν γύρο απο το μέσο όρο, με μια διάδοση σχετική με τη σταθερή απόκλιση σ (Σχ. 2.7).



Σχήμα 2.7: Μια μεταβλητή κανονική κατανομή έχει κατά προσέγγιση 95% της περιοχή της στο εύρος $|x-\mu| \le 2\sigma$, όπως φαίνεται. Η αιχμή της κατανομής έχει την αξία

$$p(\mu) = 1 \Big/ \sqrt{2\pi}\sigma .$$

Υπάρχει μια βαθιά σχέση μεταξύ της κανονικής κατανομής και της εντροπίας. Θα εξετάσουμε την εντροπία λεπτομερέστερα στο Κεφ. ??, αλλά για τώρα μόνο δηλώνουμε ότι η εντροπία μιας κατανομής δίνεται από

$$H\left(p(x)\right) = -\int p(x)\ln p(x)\,dx, \tag{36}$$

και μετριέται σε nats. Αν χρησιμοποιηθεί $\log_2$, η μονάδα είναι b_it. Η εντροπία είναι μια μη αρνητική ποσότητα που περιγράφει τη θεμελιώδη αβεβαιότητα στις τιμές των σημείων που επιλέγονται τυχαία από μια κατανομή. Μπορεί να αποδειχθεί ότι η κανονική κατανομή έχει τη μέγιστη εντροπία όλων των κατανομών που έχουν δεδομένο μέσο όρο και μεταβλητότητα(Πρόβλημα 20). Επιπλέον, όπως δηλώνεται από το ***Κεντρικό Οριακό Θεώρημα***, το συνολικό αποτέλεσμα ενός μεγάλου αριθμού μικρών, ανεξάρτητων τυχαίων διαταραχών θα οδηγήσει σε μια γκαουσσιανή κατανομή (Ασκηση υπολογιστών ??). Επειδή πολλά πρότυπα — στα ψάρια, στους χειρόγραφους χαρακτήρες σε μερικούς λεκτικούς ήχους — μπορούν να αντιμετωπισθούν ως κάποιο ιδανικό ή πρωτότυπο πρότυπο που αλλοιώνεται από έναν μεγάλο αριθμό τυχαίων διαδικασιών, η γκαουσσιανή είναι συχνά ένα καλό πρότυπο για την πραγματική κατανομή πιθανότητας.

2.5.2 Πυκνότητα Πολλών μεταβλητών

Η γενική πολλών μεταβλητών κανονική πυκνότητα σε d διαστάσεις γράφεται ως

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^t\,\Sigma^{-1}(x-\mu)\right], \tag{37}$$

όπου x είναι δ-συστατικών διάνυσμα στηλών, μ είναι το δ-συστατικών διάνυσμα μέσου, Σ είναι dxd *συνδιακύμαινόμενος πίνακας*, |Σ| and Σ^{-1} είναι το μέτρο και αντίστροφος του, αντίστοιχα, και $(x-\mu)^t$ είναι ο μεταθετικός του x−μ.* Η σημείωσή μας για το *εσωτερικό προϊόν* είναι

$$a^t b = \sum_{i=1}^{d} \alpha_i b_i \tag{38}$$

και αποκαλείται συχνά *προϊόν σημείων*. Για την απλότητα, βλέπουμε συχνά την Εξ. 37 ως $p(x) \sim N(\mu,\Sigma)$. Τυπικά, έχουμε

$$\mu \equiv E[x] = \int x p(x)\,dx \tag{39}$$

* Οι μαθηματικές εκφράσεις για την πολλών μεταβλητών κανονική πυκνότητα απλοποιούνται πολύ με την υιοθέτηση των εννοιών και της σήμανσης της γραμμικής άλγεβρας. Οι αναγνώστες που είναι αβέβαιοι της σήμανσης μας ή που θα επιθυμούσαν να θυμιθούν τη γραμμική άλγεβρα πρέπει να δουν το Παράρτημα ??.

και

$$\Sigma \equiv \mathrm{E}\left[(x-\mu)(x-\mu)^t\right] = \int (x-\mu)(x-\mu)^t\, p(x)\, dx, \tag{40}$$

όπου η αναμενόμενη τιμή ενός διανύσματος ή το πίνακα βρίσκεται με τη λήψη των αναμενόμενων τιμών των συστατικών του. Με άλλα λόγια, αν x_i είναι το $_i$-οστό συστατικό του x, μ_i είναι το $_i$-οστό συστατικό του μ, και σ_{ij} είναι το ij-οστό συστατικό του Σ, τότε

$$\mu_i = \mathrm{E}[x_i] \tag{41}$$

και

$$\sigma_{ij} = \mathrm{E}\left[(x_i-\mu_i)(x_j-\mu_j)\right]. \tag{42}$$

Ο πίνακας συνδιακύμανσης Σ είναι πάντα συμμετρικός και θετικός ημιάπειρος. Θα περιορίσουμε την προσοχή μας στην περίπτωση στην οποία Σ είναι θετικός, έτσι ώστε ο καθοριστικός παράγοντας του Σ είναι αυστηρά θετικός.* Τα διαγώνια στοιχεία σ_{ii} είναι οι διαφορές του αντίστοιχου x_i (π.χ.., σ_i 2), και τα μη-διαγώνια στοιχεία σ_{ij} είναι οι συνδιακυμάνσεις x_i και x_j . Θα αναμέναμε μια θετική συνδιακύμανση για τα χαρακτηριστικά γνωρίσματα μήκους και βάρους ενός πληθυσμού των ψαριών, για παράδειγμα. Αν x_i και x_j είναι στατιστικά ανεξάρτητα, σ_{ij} =0. Εάν όλα τα μη-διαγώνια στοιχεία είναι μηδέν, p(x) μειώνεται στο προϊόν των μεταβλητών κανονικών πυκνοτήτων για τα συστατικά x.

Οι γραμμικοί συνδυασμοί από κανονικά διανεμημένων τυχαίων δεσμευμένων μεταβλητών, ανεξάρτητων ή όχι, διανέμονται κανονικά. Ειδικότερα, εάν Α είναι ένας dxk πίνακας και $y=A^t x$ είναι ένα διάνυσμα k-συστατικών, τότε $p(y) \sim \mathrm{N}\left(\mathrm{A}^t\mu, \mathrm{A}^t\Sigma\mathrm{A}\right)$, όπως φαίνεται στο Σχ. 2.8. Στην ειδική περίπτωση που k = 1 και Α είναι ένα μοναδιαίου μήκους διάνυσμα a, $y = a^t x$ είναι ένας κλιμακωτός όρος που αντιπροσωπεύει την προβολή του x σε μια γραμμή στη διεύθυνση του a; Σε αυτή την περίπτωση $a^t\Sigma a$ είναι η μεταβλητότητα της προβολής του x στην a. Γενικά έπειτα, η γνώση ου πίνακα συνδιακύμανσης μας επιτρέπει να υπολογίσουμε τη διασπορά των στοιχείων σε οποιαδήποτε κατεύθυνση, ή σε οποιοδήποτε υποχώρο.

* Εάν τα διανύσματα δειγμάτων προέρχονται από γραμμικό υποχώρο, |Σ|= 0 και p(x) είναι εκφυλισμένα. Αυτό εμφανίζεται, παραδείγματος χάριν, όταν ένα συστατικό του x έχει μηδενική μεταβλητότητα, ή όταν δύο συστατικά είναι ίδια ή πολλαπλάσια το ένα του άλλου.

Είναι μερικές φορές βολικό να εκτελεσθεί ένας ισότιμος μετασχηματισμός που μετατρέπει μια αυθαίρετη πολλών μεταβλητών κανονική κατανομή σε μια σφαιρική, π.χ., μία που έχει το πίνακα συνδιακύμανσης ανάλογο προς τον πίνακα ταυτότητας I. Αν θέσουμε το Φ να είναι ο πίνακας του οποίου οι στήλες είναι τα ορθοκανονικά διανύσματα του Σ, και Λ οι διαγώνιοι πίνακες των αντίστοιχων τιμών, τότε ο μετασχηματισμός $A_w = \Phi\Lambda^{-1/2}$ όταν εφαρμοστεί στις συντεταγμένες καθιστά σαφές ότι η μετάσχηματισμένη κατανομή έχει πίνακα συνδιακύμανσης ίσο με τον πίνακα ταυτότητας. Στην επεξεργασία σήματος, ο μετασχηματισμός A_w καλείται μετασχηματισμός *λεύκανσης* (Whitening Transformation), δεδομένου ότι καθιστά το φάσμα των διανυσμάτων της μετασχηματισμένης κατανομής ομοιόμορφο.

Η πολλών μεταβλητών κανονική πυκνότητα διευκρινίζεται πλήρως από $d+d(d+1)/2$ παραμέτρους — τα στοιχεία του μέσου διανύσματος μ και τα ανεξάρτητα στοιχεία του πίνακα συνδιακύμανσης Σ. Τα δείγματα που προέρχονται από έναν κανονικό πληθυσμό τείνουν να εμπέσουν σε ένα ενιαίο σύννεφο ή μια συστάδα (Σχ. 2.9); το κέντρο της συστάδας καθορίζεται από το μέσο διάνυσμα, και η μορφή της συστάδας καθορίζεται από τη πίνακας συνδιακύμανσης. Προκύπτει από την Εξ. 37 ότι οι γεωμετρικοί τόποι των σημείων της σταθερής πυκνότητας είναι υπερελλιψοειδή για τα οποία η τετραγωνική μορφή $(x-\mu)^t\Sigma^{-1}(x-\mu)$ είναι σταθερή. Οι κύριοι άξονες αυτού του υπερελλιψοειδή δίνονται από διανύσματα Σ (που περιγράφεται από την Φ); οι τιμές (που περιγράφονται από το Λ) καθορίζουν τα μήκη αυτών των αξόνων. Η ποσότητα

$$r^2 = \left(x-\mu\right)^t \Sigma^{-1} \left(x-\mu\right) \tag{43}$$

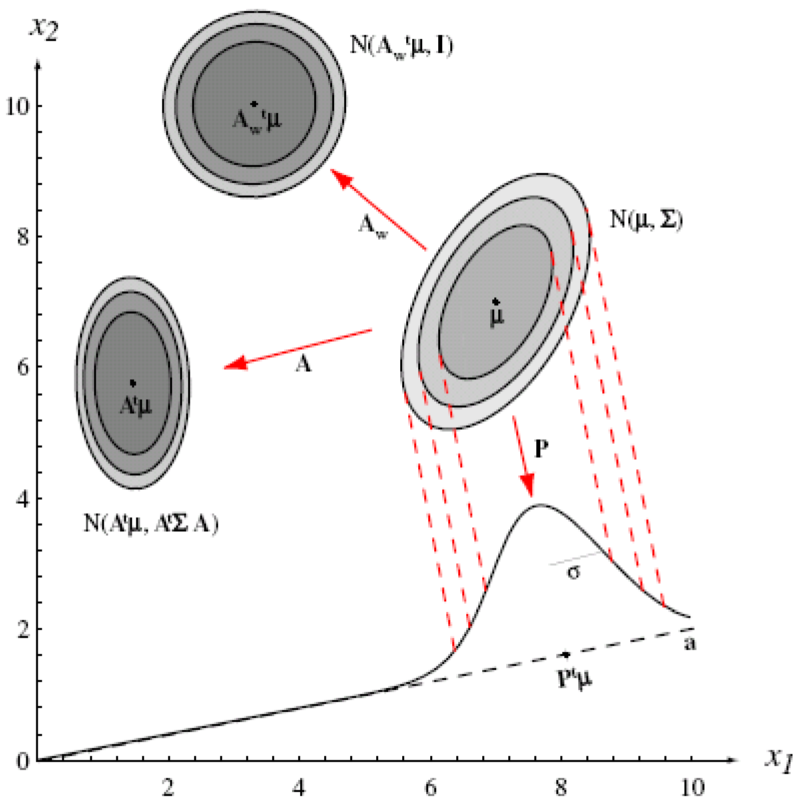
καλείται μερικές φορές τετραγωνική απόσταση Mahalanobis από το x στο μ. Κατά συνέπεια, τα περιγράμματα της σταθερής πυκνότητας είναι υπερελλιψοειδή της σταθερής απόστασης Mahalanobis στο μ και ο όγκος αυτών των υπερελλιψοειδών μετρά τη διασπορά των δειγμάτων κοντά στο μέσο όρο. Μπορεί να αποδειχθεί (Προβλήματα 15 & 16) ότι ο όγκος των υπερελλιψοειδών που αντιστοιχεί σε μια απόσταση Mahalanobis r δίνεται από

$$V = V_d \left|\Sigma\right|^{1/2} r^d, \tag{44}$$

Όπου Vd είναι ο όγκος μιας d-διάστατης μονάδας Υπερσφάιρας:

$$V_d = \begin{cases} \pi^{d/2} \Big/ \left(d/2\right)! & d \quad \pi\varepsilon\rho\iota\tau\tau o \\ 2^d\, \pi^{(d-1)/2} \left(\dfrac{d-1}{2}\right)! \Big/ (d)! & d \quad \alpha\rho\tau\iota o \end{cases} \tag{45}$$

Κατά συνέπεια, για μια δεδομένη διαστατικότητα, η διασπορά των δειγμάτων ποικίλλει άμεσα με $|\Sigma|^{1/2}$ (Πρόβλημα 17).



Σχήμα 2.8: Η εκτέλεση ενός γραμμικού μετασχηματισμού στο διάστημα χαρακτηριστικών γνωρισμάτων θα μετατρέψει μια αυθαίρετη κανονική κατανομή σε μια άλλη κανονική κατανομή. Ένας μετασχηματισμός, A, παίρνει τη πηγαία κατανομή στη κατανομή N ($A^t\mu$, $A^t\Sigma A$). Ένας άλλος γραμμικός μετασχηματισμός — μια προβολή Pεπάνω στη γραμμή a — οδηγεί στο N (μ, σ^2) μετρούμενο πάνω στην a. Ενώ οι μετατροπές παράγουν τις κατανομές σε διαφορετικό χώρο, τους παρουσιάζουμε στον αρχικό x_1−x_2 χώρο. Ένας μετασχηματισμός λεύκανσης οδηγεί σε μια κυκλικά συμμετρική γκαουσσιανή, εδώ παρουσιάζεται μετατοπισμένη.

2.6 Διακρίνουσες Συναρτήσεις για την Κανονική Πυκνότητα

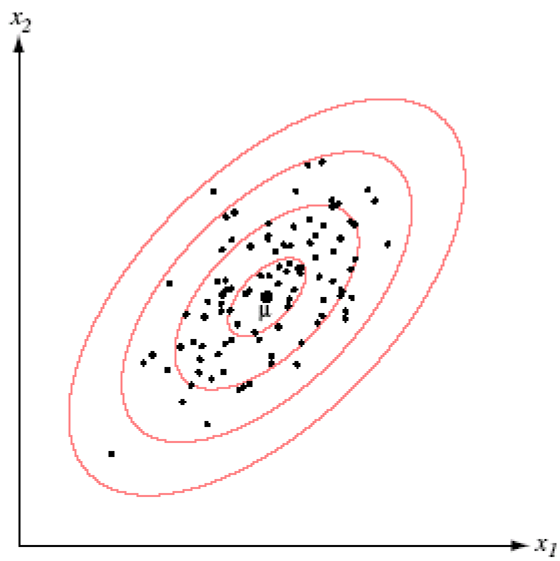
Στο Κεφ. 2.4.1 είδαμε ότι η ταξινόμηση ελάχιστου ποσοστού λάθους μπορεί να επιτευχθεί μέσω των διακρινουσών συναρτήσεων

$$g_i(x) = \ln p(\mathrm{x}|\omega_i) + \ln P(\omega_i). \tag{46}$$

Αυτή η έκφραση μπορεί να αξιολογηθεί εύκολα εάν οι πυκνότητες p(x|ω_i) είναι κανονικές πολλών μεταβλητών, π.χ., αν $p(\mathrm{x}|\omega_i) \sim \mathrm{N}(\mu_i, \Sigma_i)$. Σε αυτή την περίπτωση από την Εξ. 37 έχουμε

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i). \qquad (47)$$

Ας εξετάσουμε τη διακρίνουσα συνάρτηση και την προκύπτουσα ταξινόμηση για διάφορες ειδικές περιπτώσεις.



Σχήμα 2.9: Δείγματα από μια δισδιάστατη γκαουσσιανή κατανομή που βρίσκονται σε ένα σύννεφο κενταρισμένο στο μέσο μ. Οι κόκκινες εκλείψεις δείχνουν γραμμές ίσης πυκνότητας πιθανότητας της γκαουσσιανής.

2.6.1 Περίπτωση 1: $\Sigma_i = \sigma^2 I$

Η απλούστερη περίπτωση εμφανίζεται όταν τα χαρακτηριστικά γνωρίσματα είναι στατιστικά ανεξάρτητα, και όταν έχει κάθε εκτιμήτρια την ίδια μεταβλητότητα, σ^2. Σε αυτήν την περίπτωση η πίνακας συνδιακύμανσης είναι διαγώνια, όντας μόνο σ^2 φορές η πίνακας ταυτότητας Ι. Γεωμετρικά, αυτό αντιστοιχεί στην κατάσταση στην οποία τα δείγματα εμπίπτουν στις υπερσφαιρικές συστάδες ίσου μεγέθους, και η συστάδα για την i-οστή κατηγορία που κεντροθετείται γύρω από μέσο διάνυσμα μ_i. Ο υπολογισμός του μέτρου και του αντιστρόφου Σ_i είναι ιδιαίτερα εύκολος: $|\Sigma_i| = \sigma^{2d}$ και $\Sigma^{-1} = (1/\sigma^2)I$.

Αφού και το $|\Sigma_i|$ και το $(d/2)\ln 2\pi$ είναι ανεξάρτητοι του Ι στη Εξ. 47, είναι ασήμαντες πρόσθετες σταθερές που μπορούν να αγνοηθούν. Κατά συνέπεια λαμβάνουμε τις απλές διακρίνουσες συναρτήσεις

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i), \qquad (48)$$

Όπου $\|\,.\,\|$ είναι η Ευκλείδια νόρμα, η οποία είναι,

$$\|x-\mu_i\|^2 = (x-\mu_i)^t (x-\mu_i). \tag{49}$$

Εάν οι εκ των προτέρων πιθανότητες δεν είναι ίσες, τότε η Εξ. 48 δείχνει ότι η τετραγωνική απόσταση $||x-\mu||^2$ πρέπει να κανονικοποιηθεί από τη μεταβλητότητα σ^2 και να αλλαχθεί με την προσθήκη ln $P(\omega_i)$. Κατά συνέπεια, εάν το x είναι εξίσου κοντά σε δύο διαφορετικά μέσα διανύσματα, τη βέλτιστη απόφαση θα ευνοήσει την εκ των προτέρων πιθανότερη κατηγορία.

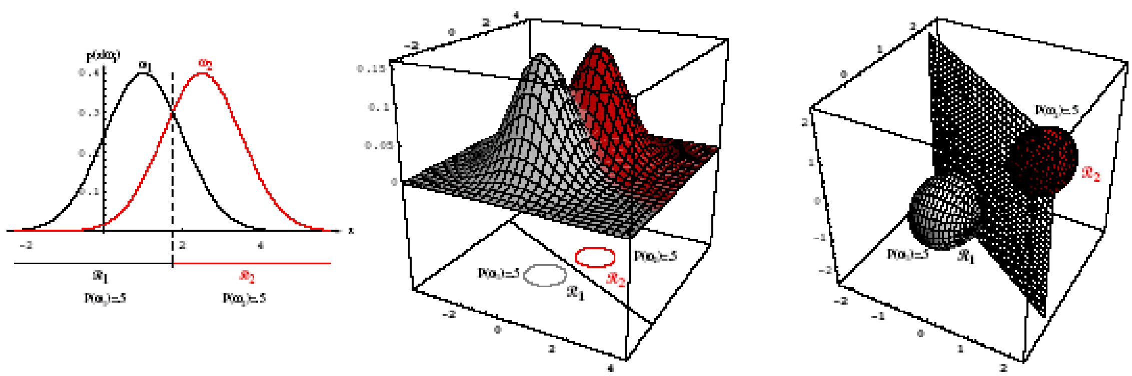
Ανεξάρτητα από εάν οι εκ των προτέρων πιθανότητες είναι ίσες ή όχι, δεν είναι πραγματικά απαραίτητο να υπολογιστούν οι αποστάσεις. Επέκταση της τετραγωνικής μορφής $(x-\mu_i)^t(x-\mu_i)$ δίνει

$$g_i(x) = -\frac{1}{2\sigma^2}\left[x^t x - 2\mu_i^2 x + \mu_i^t \mu_i\right] + \ln P(\omega_i), \tag{50}$$

η όποια εμφανίζεται να είναι μια τετραγωνική συνάρτηση x. Εντούτοις, ο τετραγωνικός όρος $x^t x$ είναι ίδιος για όλα τα $_i$, κάνοντας το μια παραβλέψιμη πρόσθετη σταθερά. Κατά συνέπεια, λαμβάνουμε τις ισοδύναμες γραμμικές διακρίνουσες συναρτήσεις

$$g_i(x) = w_i^t x + \omega_{i0}, \tag{51}$$

όπου



Σχήμα 2.10: Εάν οι συνδιακυμάνσεις δύο κατανομών είναι ίσες και ανάλογες προς η πίνακας ταυτότητας, έπειτα οι διανομές είναι σφαιρικές σε d διαστάσεις, και το όριο είναι γενικευμένο υπερευθείες d-1 διαστάσεων, κάθετες στις γραμμές που χωρίζουν τα μέσα. Σε αυτά τα μονό,δις,τρις-διάστατα παραδείγματα τονίζουμε το $p(x|\omega_i)$ και τα όρια για την περίπτωση $P(\omega_1) = P(\omega_2)$.Στον 3-διάσταστη περίπτωση, το πλέγμα χωρίζει το R_1 από το R_2.

$$w_i = \frac{1}{\sigma^2}\mu_i \tag{52}$$

και

$$\omega_{i0} = \frac{-1}{2\sigma^2}\mu_i^t\mu_i + \ln P\left(\omega_i\right). \tag{53}$$

Καλούμε w_{i0} το κατώφλι ή το bias στην $_i$-οστή κατεύθυνση.

Ένας ταξινομητής που χρησιμοποιεί τις γραμμικές διακρίνουσες συναρτήσεις καλείται γραμμική μηχανή. Αυτό το είδος ταξινομητή έχει πολλές ενδιαφέρουσες θεωρητικές ιδιότητες, μερικές από τις οποίες θα συζητήσουμε λεπτομερώς στο κεφ. ??. Σε αυτό το σημείο μόνο σημειώνουμε ότι οι επιφάνειες απόφασης για μια γραμμική μηχανή είναι κομμάτια υπερευθειών που καθορίζονται από τις γραμμικές εξισώσεις $g_i(x) = g_j(x)$ για τις δύο κατηγορίες με τις υψηλότερες εκ των υστέρων πιθανότητες. Για τη συγκεκριμένη περίπτωση, αυτή η εξίσωση μπορεί να γραφτεί ως

$$w^t\left(x - x_0\right) = 0, \tag{54}$$

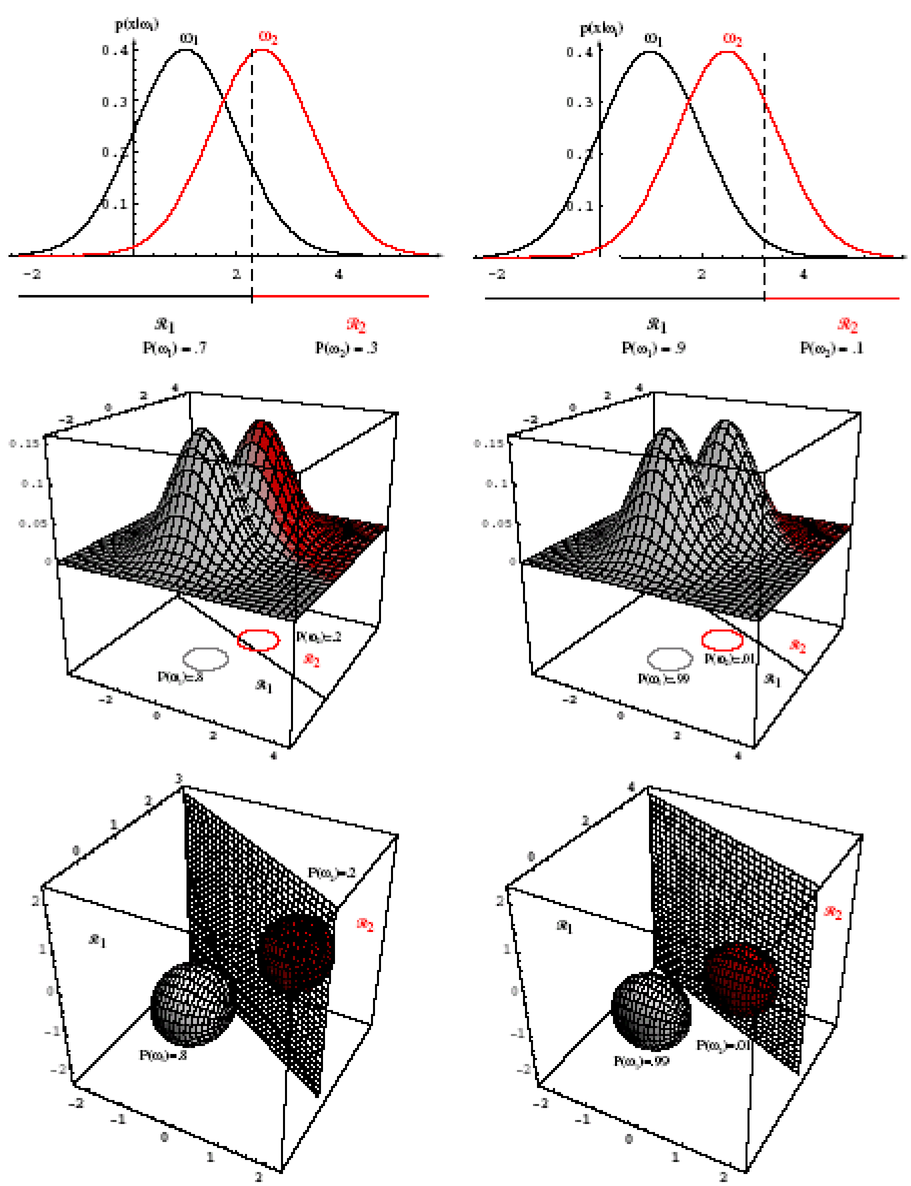
όπου

$$w = \mu_i - \mu_j \tag{55}$$

και

$$x_0 = \frac{1}{2}\left(\mu_i + \mu_j\right) - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2}\ln\frac{P\left(\omega_i\right)}{P\left(\omega_j\right)}\left(\mu_i - \mu_j\right). \tag{56}$$

Αυτή η εξίσωση καθορίζει ένα υπερευθείες μέσω του σημείου x_0 και ορθογώνιο στο διάστημα w. Αφού $w = \mu_i - \mu_j$, το υπερευθείες που χωρίζει τις R_i and R_j είναι ορθογώνιο στη γραμμή που ενώνει τα μέσα. Αν $P(\omega_i) = P(\omega_j)$, ο δεύτερος όρος στα δεξιά της Εξ. 56 διαγράφεται, και επομένως το σημείο x_0 είναι στη μέση των μέσων, και το υπερευθείες είναι η κάθετος διαχωριστική της γραμμής μεταξύ των μέσων (Σχ. 2.11). Αν $P(\omega_i) \neq P(\omega_j)$, το σημείο x_0 απομακρύνεται από τον πιο πιθανό μεσο. Προσέξτε ότι αν η μεταβλητότητα σ^2 είναι μικρή σε σχέση με την τετραγωνική απόσταση $\|\mu_i - \mu_j\|$, κατόπιν η θέση της του ορίου απόφασης είναι σχετικά ανεξάρτητο από τις ακριβείς τιμές των εκ των προτέρων πιθανοτήτων. Εάν οι εκ των προτέρων πιθανότητες $P(\omega_i)$ είναι ίδιες για όλες τις c κλάσεις, τότε ο όρος $\ln P(\omega_i)$ γίνεται μια άλλη ασήμαντη πρόσθετη σταθερά που μπορεί να αγνοηθεί. Όταν αυτό συμβαίνει, ο βέλτιστος κανόνας απόφασης μπορεί να δηλωθεί πολύ απλά: Για να ταξινομήσουμε ένα διάστημα εκτιμήτριας x, μετράμε την Ευκλείδεια απόσταση $\|x - \mu_i\|$ για κάθε x σε κάθε c διάστημα μέσου, και θέτουμε x στην κατηγορία του κοντινότερου μέσου όρου. Ένας τέτοιος ταξινομητής καλείται ταξινομητής ελάχιστης απόστασης. Εάν κάθε μέσο διάνυσμα θεωρείται ως ένα ιδανικό πρωτότυπο ή ένα πρότυπο για τα σχέδια στην κατηγορία του, αυτό είναι ουσιαστικά μια διαδικασία ταιριάσματος προτύπων (Σχ. 2.10), μια τεχνική που θα εξετάσουμε πάλι στο Κεφ ?? κεφ. ??στον αλγόριθμο κοντινότερης γειτονιάς(nearest neighborhood)



Σχήμα 2.11: Αν αλλάζουν οι εκ των προτέρων, το όριο απόφασης μετακινείται; για αρκετά ανόμοιες εκ των προτέρων το όριο δεν θα βρεθεί μεταξύ των μέσων των 1-, 2- και 3-διάστατων σφαιρικών γκαουσσιανών κατανομών.

2.6.2 Περίπτωση 2: $\Sigma_i = \Sigma$

Μια άλλη απλή περίπτωση προκύπτει όταν οι πίνακες συνδιακύμανσης για όλες τις κατηγορίες είναι ίδιες αλλά ειδάλλως αυθαίρετες. Γεωμετρικά, αυτό αντιστοιχεί στην κατάσταση στην οποία τα δείγματα εμπίπτουν στις υπερελλιψοειδείς συστάδες ίσου μεγέθους και μορφής, η συστάδα για την i-οστή κατηγορία που κεντροθετείται γύρω από το μέσο διάνυσμα μ_i. Αφού και οι δύο $|\Sigma_i|$ και $(d/2)\ln 2\pi$ όροι στην Εξ. 47 είναι ανεξάρτητοι του i, μπορούν να αγνοηθούν ως περιττές πρόσθετες σταθερές. Αυτή η απλοποίηση οδηγεί στις διακρίνουσες συναρτήσεις

$$g_i\left(x\right) = -\frac{1}{2}\left(x-\mu_i\right)^t \sum^{-1}\left(x-\mu_i\right) + \ln P\left(\omega_i\right). \tag{57}$$

Αν οι εκ των προτέρων πιθανότητες $P(\omega_i)$ είναι ίδιες για όλες τις c κλάσεις, τότε ο όρος ln $P(\omega_i)$ μπορεί να αγνοηθεί. Σε αυτήν την περίπτωση, ο βέλτιστος κανόνας απόφασης μπορεί άλλη μια φορά να δηλωθεί πολύ απλά: για να ταξινομήσετε ένα διάνυσμα x χαρακτηριστικών γνωρισμάτων, μετρήστε την τετραγωνική απόσταση Mahalanob$_i$s $(x-\mu_i)^t\Sigma^{-1}(x-\mu_i)$ από το x στο κάθε ένα διάνυσμα μέσου c, και ταξινομήστε το x στην κατηγορία του κοντινότερου μέσου όρου. Όπως πριν, οι άνισες εκ των προτέρων πιθανότητες προκαταλαμβάνουν την απόφαση υπέρ της εκ των προτέρων πιθανότερης κατηγορίας.

Επέκταση της τετραγωνικής μορφής $(x-\mu_i)^t\Sigma^{-1}(x-\mu_i)$ έχει ως αποτέλεσμα ένα ποσό που περιλαμβάνει έναν τετραγωνικό όρο xtΣ−1x ο όποιος είναι εδώ ανεξάρτητος από το i. Αφου αφαιρεθεί αυτός ο όρος από την Εξ. 57, οι προκύπτουσες διακρίνουσες συναρτήσεις είναι πάλι γραμμικές:

$$g_i\left(x\right) = w_i^t x + \omega_{i0}, \tag{58}$$

όπου

$$w_i = \sum^{-1}\mu_i \tag{59}$$

και

$$\omega_{i0} = -\frac{1}{2}\mu_i^t \sum^{-1}\mu_i + \ln P\left(\omega_i\right). \tag{60}$$

Δεδομένου ότι οι διακρίνουσες είναι γραμμικές, τα προκύπτοντα όρια απόφασης είναι πάλι υπερευθείες (Σχ. 2.10). Αν R_i και R_j είναι παρακείμενος, το όριο μεταξύ τους έχει την εξίσωση

$$w^t\left(x - x_0\right) = 0, \tag{61}$$

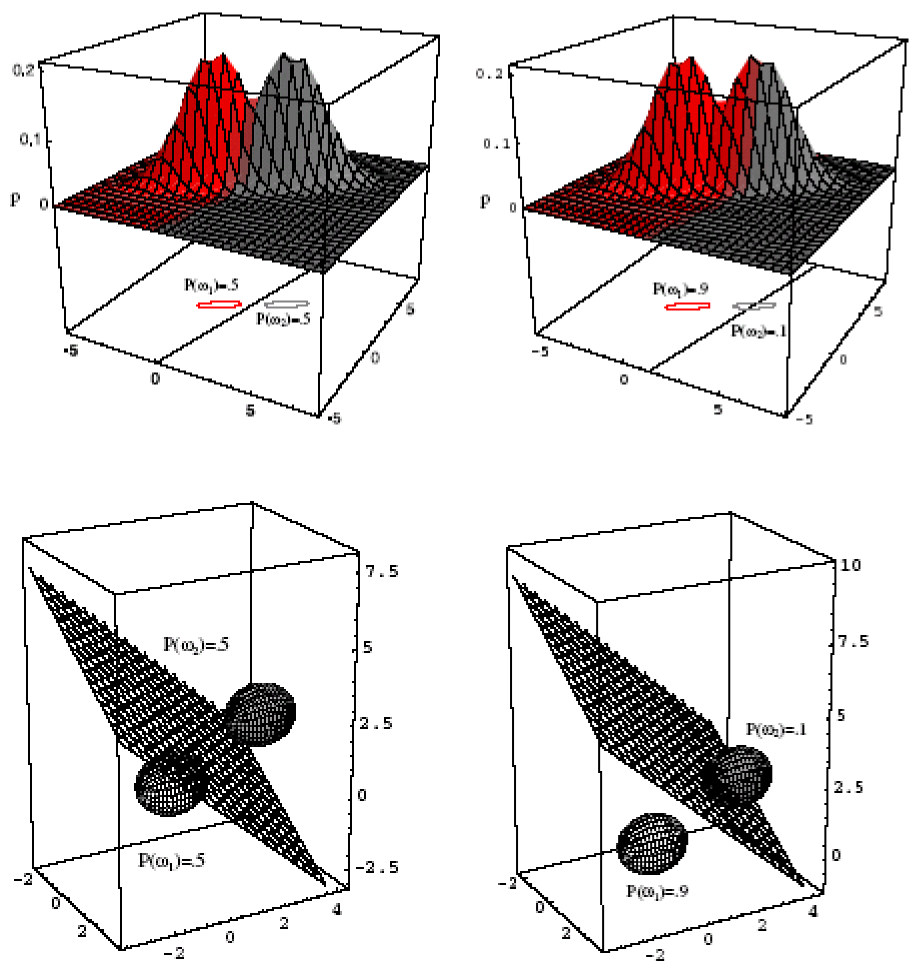
όπου

$$w = \sum^{-1}\left(\mu_i - \mu_j\right) \tag{62}$$

και

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \sum^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j). \tag{63}$$

Αφού w= $\Sigma^{-1}(\mu_i - \mu_j)$ δεν είναι γενικώς στη διεύθυνση του $\mu_i - \mu_j$, η υπερευθεία που διαχωρίζει τα R_i και R_j δεν είναι γενικά ορθογώνια στη γραμμή μεταξύ των μέσων. Εντούτοις, κόβει εκείνη την γραμμή στο σημείο x_0 που είναι στη μέση της απόστασης μεταξύ των μέσων εάν οι εκ των προτέρων πιθανότητες είναι ίσες. Εάν οι εκ των προτέρων πιθανότητες δεν είναι ίσες, η βέλτιστη υπερευθεία ορίου μετατοπίζεται μακριά από τον πιθανότερο μέσο όρο (Σχ. 2.12). Όπως πριν από, με επαρκές bias το πλέγμα απόφασης δεν χρειάζεται να βρίσκεται μεταξύ των δύο διανύσματα μέσων.



Σχήμα 2.12: Πυκνότητες πιθανότητας (υποδειγμένες από τις επιφάνειες σε δύο διαστάσεις και ελλειψοειδείς επιφάνειες σε τρεις διαστάσεις) και περιοχές απόφασης για ίσες αλλά ασύμμετρες Γκαουσσιανές κατανομές. Τα υπερευθείες απόφασης δεν χρειάζεται να είναι κάθετα στη γραμμή που συνδέει τα μέσα.

2.6.3 Περίπτωση 3: Σ_i = αυθαίρετο

Στη γενική πολλών μεταβλητών κανονική περίπτωση, οι πίνακες συνδιακύμανσης είναι διαφορετικοί για κάθε μια κατηγορία. Ο μόνος όρος που μπορεί να απαλειφθεί από την Εξ. 47 είναι ο (d/2) ln 2π, και οι προκύπτουσες διακρίνουσες συναρτήσεις είναι εγγενώς τετραγωνικές:

$$g_i(x) = x^t W_i x + w_i^t x + \omega_{i0}, \tag{64}$$

όπου

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \ , \tag{65}$$

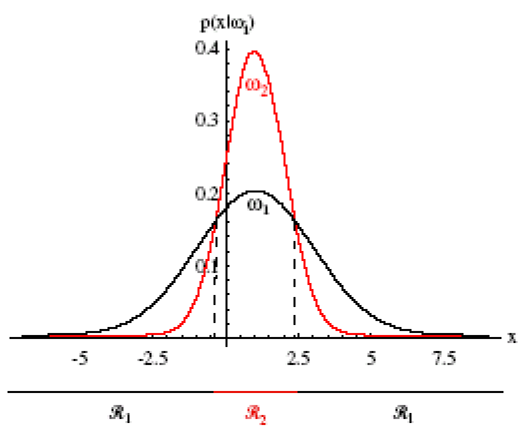
$$w_i = \Sigma_i^{-1}\mu_i \tag{66}$$

και

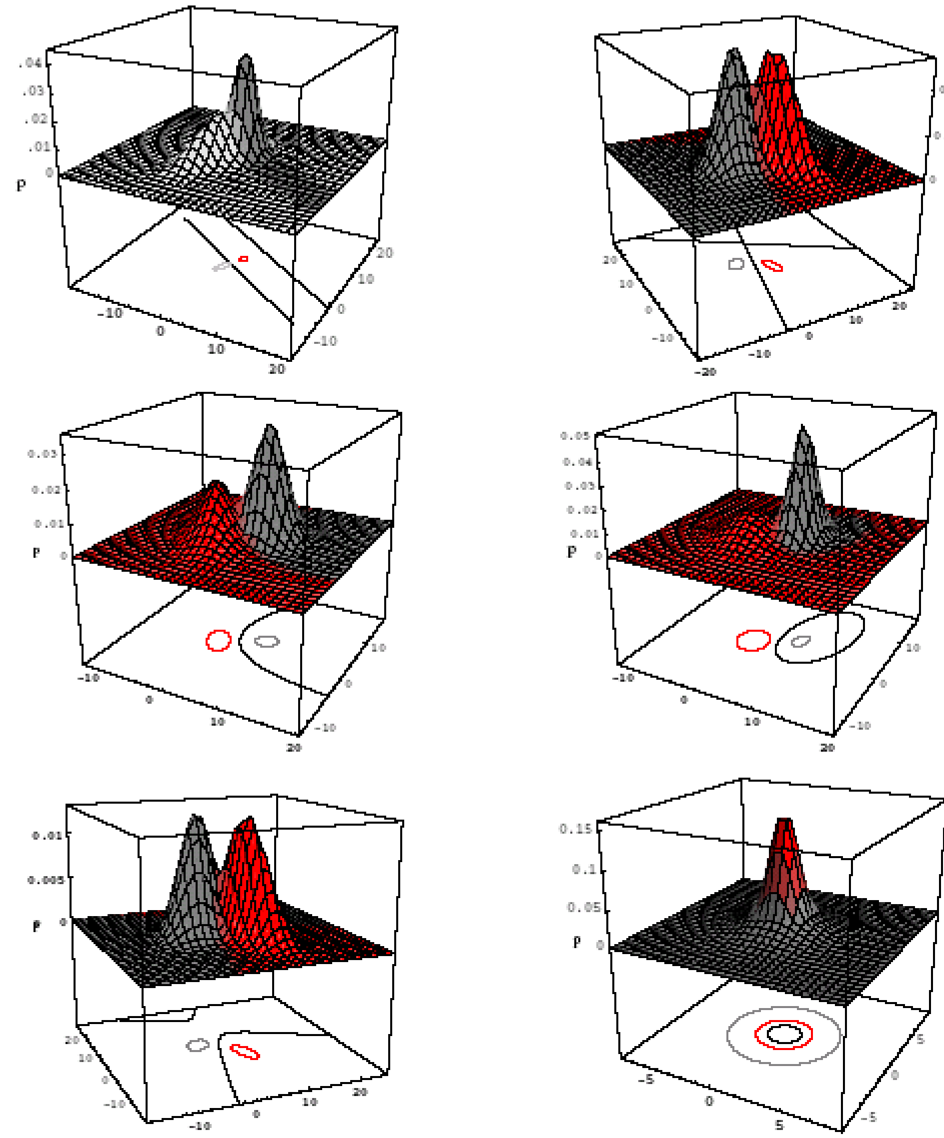
$$\omega_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i). \tag{67}$$

Οι επιφάνειες απόφασης είναι υπερτετραγωνικές και μπορουν να παρουν οποιαδήποτε από τις γενικές μορφές — ύπερευθείες, ζεύγη υπερευθειών, υπερσφαίρες, υπερελλιψοειδή, υπερπαραβολικές, και υπερυπερβολικές διαφόρων τύπων (Πρόβλημα 29). Ακόμη και σε μια διάσταση, για την αυθαίρετη συνδιακύμανση οι περιοχές απόφασης δεν χρειάζονται να συνδεθούν απλά (Σχ. 2.13). Τα δισδιάστατα και τρισδιάστατα παραδείγματα στο Σχ. 2.14 και 2.15 δείχνουν πως αυτές οι διαφορετικές μορφές μπορούν να προκύψουν. Αυτές οι διαφορές υποδεικνύονται από τα περιγράμματα της σταθερής πυκνότητας πιθανότητας.

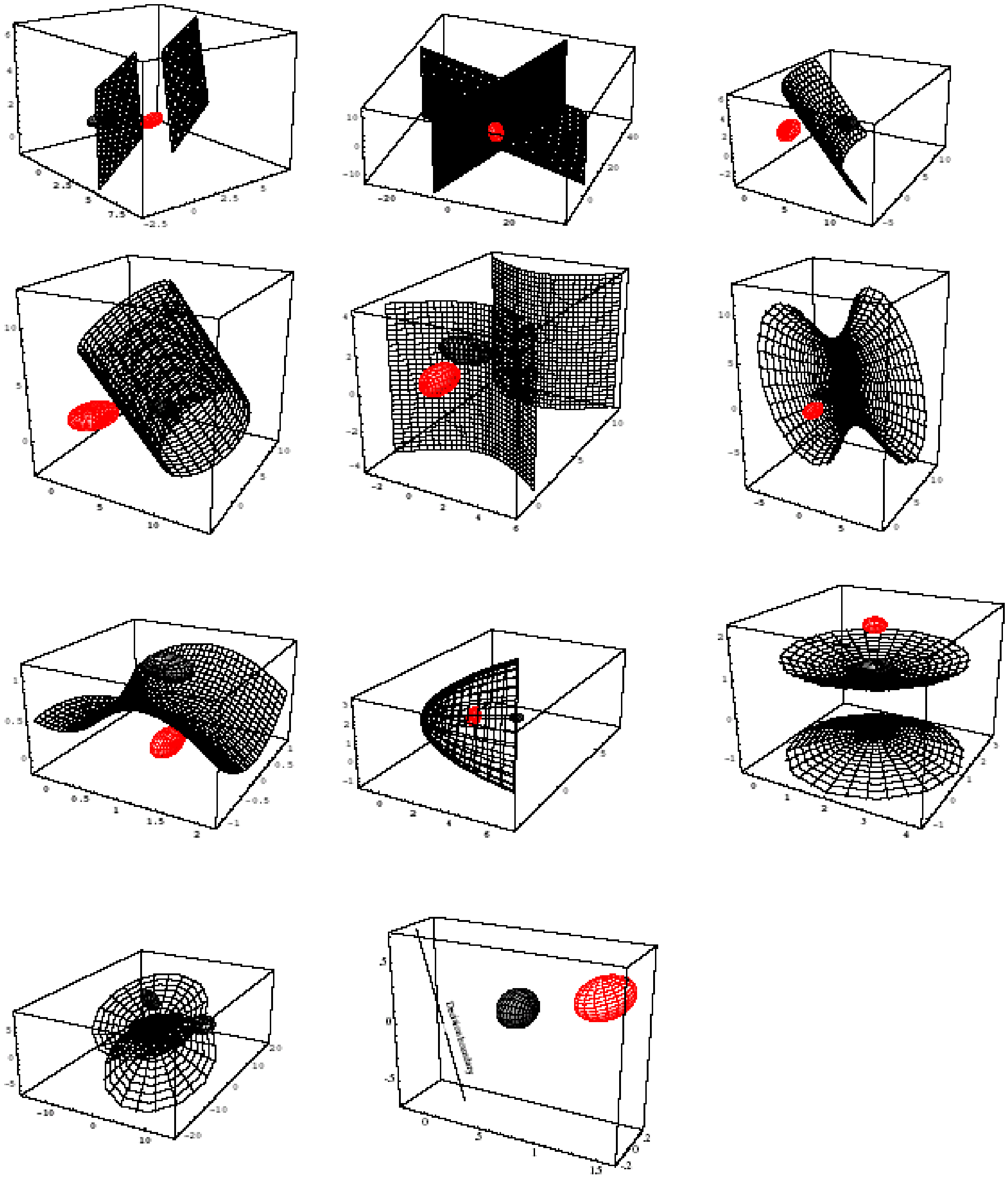
Η επέκταση αυτών οδηγεί σε παραπάνω από δύο κατηγορίες είναι ξεκάθαρο όμως ότι χρειάζεται να ξεκαθαρήσουμε ποιες δύο από τις c οι κατηγορίες είναι αρμόδιες για οποιοδήποτε τμήμα ορίου. Το Σχ. 2.16 παρουσιάζει επιφάνειες απόφασης για μια περίπτωση τεσσάρων κατηγοριών φτιαγμένη από γκαουσσιανές κατανομές. Φυσικά, εάν οι κατανομές είναι πιο περίπλοκες, οι περιοχές απόφασης μπορούν να είναι ακόμα πιό σύνθετες, αν και η ίδια θεωρία ισχύει και εκεί επίσης.



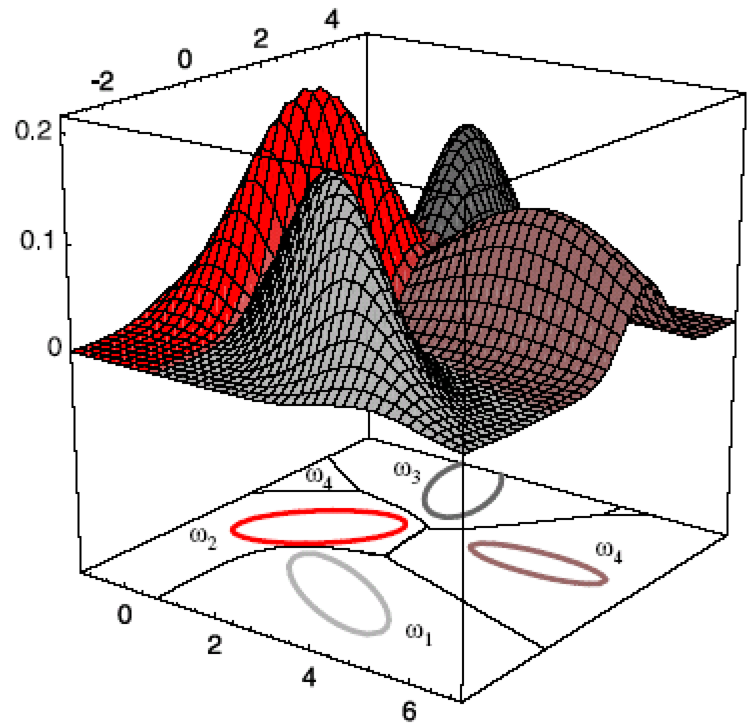
Σχήμα 2.13: Οι μη-απλά συνδεδεμένες περιοχές απόφασης μπορούν να προκύψουν σε μια διάσταση για γκαουσιανές που έχει άνιση μεταβλητότητα.



Σχήμα 2.14: Οι αυθαίρετες γκαουσσιανές κατανομές οδηγούν στα όρια απόφασης Bayes που είναι σε γενικές γραμμές υπερτετραγωνικές. Αντιθέτως, λαμβάνοντας μια υπερτετραγωνική, κάποιος μπορεί να βρει δύο Γκαουσσιανές κατανομές των οποίων το όριο απόφασης του Bayes είναι και αυτό υπερτετραγωνικό.



Σχήμα 2.15: Αυθαίρετες τρισδιάστατές γκαουσσιανές κατανομές που αφορούν την απόφαση ορίων του Bayes τα οποία είναι δισδιάστατα υπερτετραγωνικά. Υπάρχουν ακόμη και εκφυλισμένες περιπτώσεις στις οποίες το όριο απόφασης είναι μια γραμμή.



Σχήμα 2.16: Οι περιοχές απόφασης για τέσσερις κανονικές κατανομές. Ακόμη και με έναν τέτοιο χαμηλό αριθμός κατηγοριών, οι μορφές των περιοχών ορίου μπορεί να είναι αρκετά σύνθετές.

Παράδειγμα 1: Περιοχές απόφασης για δισδιάστατα γκαουσσιανά δεδομένα

Για να διευκρινίσουν αυτές τις ιδέες, υπολογίζουμε ρητά τα όρια απόφασης για τα δισδιάστατα δεδομένα δύο κατηγοριών στο σχήμα του παραδείγματος. Έστω ω1 είναι το σύνολο των τεσσάρων μαύρων σημείων, και ω2 τα κόκκινα σημεία. Αν και θα ξοδέψουμε το επόμενο κεφάλαιο για να καταλάβουμε πώς να εκτιμήσουμε τις παραμέτρους των κατανομών, για τώρα υποθέτουμε ότι χρειάζεται μόνο να υπολογίσουμε τους μέσους όρους και συνδιακυμάνσεις από τις ιδιαίτερες μορφές των Εξ. 39 & 40; Βρίσκουμε:

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

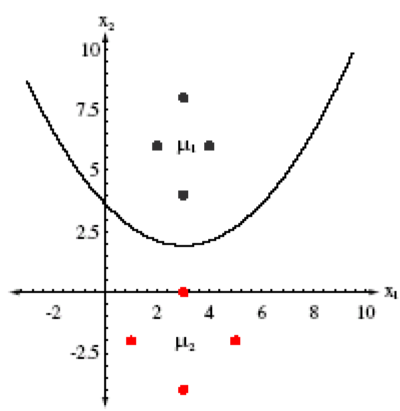
Οι αντίστροφοι πίνακες είναι έπειτα,

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \text{ and } \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

Θεωρούμε ίσες εκ των προτέρων πιθανότητες, $P(\omega_1)=P(\omega_2) = 0.5$, και τις αντικαθιστούμε στη γενική μορφή για την κατανομή, Εξ. 64 – 67, θέτοντας $g_1(x) = g_2(x)$ για να πάρουμε το όριο απόφασης¨

$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$. x2 = 3.514 − 1.125x1 + 0.1875x21

Αυτή η εξίσωση περιγράφει μια παραβολή με κορυφή στο $\begin{pmatrix} 3 \\ 1.83 \end{pmatrix}$. Προσέξτε ότι παρά το γεγονός ότι η μεταβλητότητα στα δεδομένα στη διεύθυνση x_2 και για τις δύο κατανομές είναι η ίδια, το όριο απόφασης δε περνά από το σημείο $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$, στο μέσο των δύο μέσων όρων, όπως θα είχαμε υποθέσει. Αυτό συμβαίνει γιατί για την κατανομή ω_1, η πιθανοτική κατανομή «πιέζεται» στη διεύθυνση x_1 περισσότερο από ότι για την ω_2. Επειδή οι συνολικές εκ των προτέρων πιθανότητες είναι οι ίδιες (π.χ., το ολοκλήρωμα στο διάστημα της πυκνότητας πιθανότητας), η κατανομή αυξάνεται στη διεύθυνση x2 (τη σχετική με τη κατανομή ω2). Επομένως το όριο απόφασης βρίσκεται ελάχιστα χαμηλότερα από το μέσο των μέσων όρων όπως φαίνεται στο όριο απόφασης



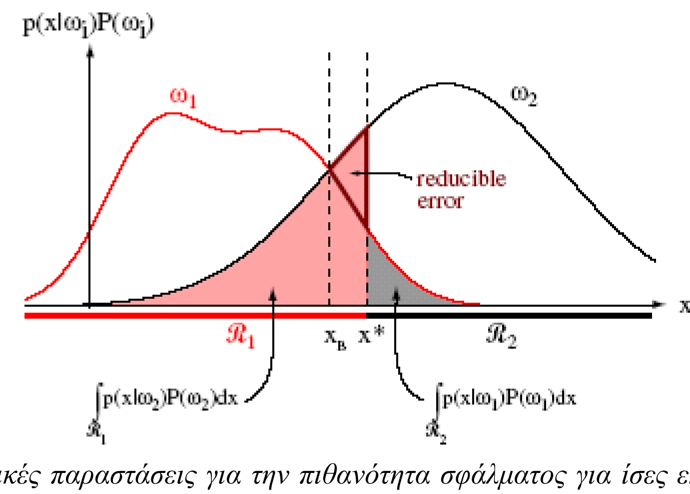
Το υπολογισμένο όριο απόφασης του Bayes για δύο γκαουσσιανές κατανομές που η κάθε μια βασίζεται σε τέσσερα σημεία δεδομένων.

2.7 *Πιθανότητες σφάλματος και ολοκληρώματα

Μπορούμε να αποκτήσουμε περαιτέρω άποψη για τη λειτουργία ενός γενικού ταξινομητή – του Bayes ή άλλου — αν λάβουμε υπόψη τις πηγές σφαλμάτων του. Εξετάστε πρώτα την περίπτωση δύο κατηγοριών, και υποθέστε ότι η διχοτόμος έχει χωρίσει το χώρο σε δύο περιοχές R_1 και R_2 με ένα πιθανό όχι ιδανικό τρόπο, Υπάρχουν δύο τρόποι με τους οποίους ένα λάθος ταξινόμησης μπορεί να προκύψει, είτε μια τιμή x πέφτει στην περιοχή R_2 και η πραγματική κατάσταση της φύσης είναι ω_1, ή το x πέφτει στην περιοχή R_1 και η πραγματική κατάσταση της φύσης είναι ω_2. Δεδομένου ότι αυτά τα γεγονότα είναι αμοιβαία αποκλειστικά και εξαντλητικά, η πιθανότητα του λάθους είναι

$$\begin{aligned} P(error) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2|\omega_1)P(\omega_1) + P(x \in R_1|\omega_2)P(\omega_2) \\ &= \int_{R_2} p(x|\omega_1)P(\omega_1)\,dx + \int_{R_1} p(x|\omega_2)P(\omega_2)\,dx. \end{aligned} \tag{68}$$

Το αποτέλεσμα απεικονίζεται για την περίπτωση της μια διάστασης στο Σχ. 2.17. Τα δύο ολοκληρώματα στην Εξ. 68 αναπαριστούν τις ροζ και τις γκρι περιοχές στις ουρές των συναρτήσεων $p(x|\omega_i)P(\omega_i)$. Επειδή το σημείο απόφασης x^* (και επομένως οι περιοχές R_1 και R_2) επιλέχθηκαν αυθαίρετα για το σχήμα, η πιθανότητα σφάλματος δεν είναι όσο μικρή όσο θα έπρεπε. Συγκεκριμένα, η τριγωνική περιοχή που έχει υπογραμμιστεί σαν «με δυνατότητα μείωσης λάθους» μπορεί να εξαλειφθεί αν το όριο απόφασης μετακινηθεί στο x_B. Αυτό είναι το ιδεατό όριο απόφασης του και δίνει την ελάχιστη πιθανότητα σφάλματος. Γενικά, αν $p(x|\omega_1)P(\omega_1)>p(x|\omega_2)P(\omega_2)$, είναι πλεονεκτικό να ταξινομήσουμε το x στην περιοχή R_1 έτσι ώστε η μικρότερη ποσότητα να συμβάλει στο ολοκλήρωμα λάθους. Αυτό είναι ακριβώς αυτό που επιτυγχάνει ο κανόνας του Bayes.



Σχήμα 2.17: Γραφικές παραστάσεις για την πιθανότητα σφάλματος για ίσες εκ των προτέρων και (όχι ιδεατό) όριο απόφαση x. Η ροζ περιοχή ανταποκρίνεται στην πιθανότητα σφάλματος για την απόφαση ω1 όταν η κατάσταση της φύσης είναι ω_2, και η γκρι περιοχή αναπαριστά το αντίθετο, όπως δίνεται από την Εξ. 68. Αν αντίθετα το όριο απόφασης είναι στο σημείο ποτ οι εκ των υστέρων πιθανότητας είναι ίσες, x_B, τότε το «με δυνατότητα μείωσης λάθος» και η ολική σκιασμένη περιοχή είναι τα ελάχιστα δυνατά — Αυτή είναι η απόφαση του Bayes και δίνει το ρυθμό σφάλματος του Bayes.*

Στην περίπτωση πολλών κατηγοριών, υπάρχουν περισσότεροι τρόποι να είναι λάθος από σωστό, και είναι ευκολότερο να υπολογίσουμε την πιθανότητα να είναι σωστό. Ξεκάθαρα

$$
\begin{aligned}
P(correct) &= \sum_{i=1}^{c} P(x \in R_i, \omega_i) \\
&= \sum_{i=1}^{c} P(x \in R_i | \omega_i) P(\omega_i) \\
&= \sum_{i=1}^{c} \int_{R_i} p(x|\omega_i) P(\omega_i) dx.
\end{aligned} \tag{69}
$$

Το γενικό αποτέλεσμα της Εξ. 69 δεν εξαρτάται ούτε από το πώς ο χώρος εκτιμήτριας είναι καταμερισμένος σε περιοχές απόφασης ούτε στη μορφή των ελλοχεύουσων κατανομών. Ο ταξινομητής του Bayes μεγιστοποιεί αυτή την πιθανότητα επιλέγοντας τις περιοχές έτσι ώστε η προς ολοκλήρωση παράσταση να είναι μέγιστη για όλα τα x; κανένας άλλος καταμερισμός δεν μπορεί να παραγάγει μια μικρότερη πιθανότητα σφάλματος.

2.8 *Όρια λάθους για κανονικές πυκνότητες

Ο κανόνας απόφασης του Bayes εγγυάται το χαμηλότερο μέσο ποσοστό λάθους, και έχουμε δει πώς να υπολογίσει τα όρια απόφασης για τις κανονικές πυκνότητες. Εντούτοις, αυτά τα αποτελέσματα δεν μας λένε η πιθανότητα του λάθους ποια θα είναι πραγματικά. Ο πλήρης υπολογισμός του λάθους για την γκαουσσιανή περίπτωση θα ήταν αρκετά δύσκολος, ειδικά στις υψηλές διαστάσεις, λόγω της ασυνεχής φύση των περιοχών απόφασης στο ολοκλήρωμα στην Εξ. 69. Εντούτοις, στην περίπτωση δύο κατηγορίων το γενικό ολοκλήρωμα λάθους της Εξ. 5 μπορεί να προσεγγιστεί αναλυτικά για να μας δώσει έναν ανώτερο όριο λάθους.

2.8.1 Όριο Chernoff

Για να παραγάγουμε ένα όριο για το λάθος, χρειαζόμαστε την ακόλουθη ανισότητα:

$$\min[a,b] \le a^{\beta} b^{1-\beta} \quad for \quad a,b \ge 0 \quad and \quad 0 \le \beta \le 1. \tag{70}$$

Για να καταλάβει αυτή η ανισότητα μπορούμε, χωρίς απώλεια γενικότητας, να υποθέσουμε $a \ge b$. Κατά συνέπεια χρειαζόμαστε μόνο να δείξουμε ότι $b \le a^{\beta} b^{1-\beta} = (a/b)^{\beta} b$. Αλλά αυτή η ανισότητα ισχύει προφανώς, αφού $(a/b)^{\beta} \ge 1$. Χρησιμοποιώντας τις Εξ. 7 & 1, εφαρμόζουμε την ανισότητα στη Εξ. 5 και παίρνουμε το όριο:

$$P(error) \le P^{\beta}(\omega_1) P^{1-\beta}(\omega_2) \int p^{\beta}(x|\omega_1) p^{1-\beta}(x|\omega_2) dx \quad for \quad 0 \le \beta \le 1. \tag{71}$$

Σημειώστε ειδικά ότι αυτό το ολοκλήρωμα είναι σε όλο το διάστημα εκτιμήτριας — δεν πρέπει να επιβάλουμε όρια ολοκλήρωσης που αντιστοιχούν στα όρια απόφασης.

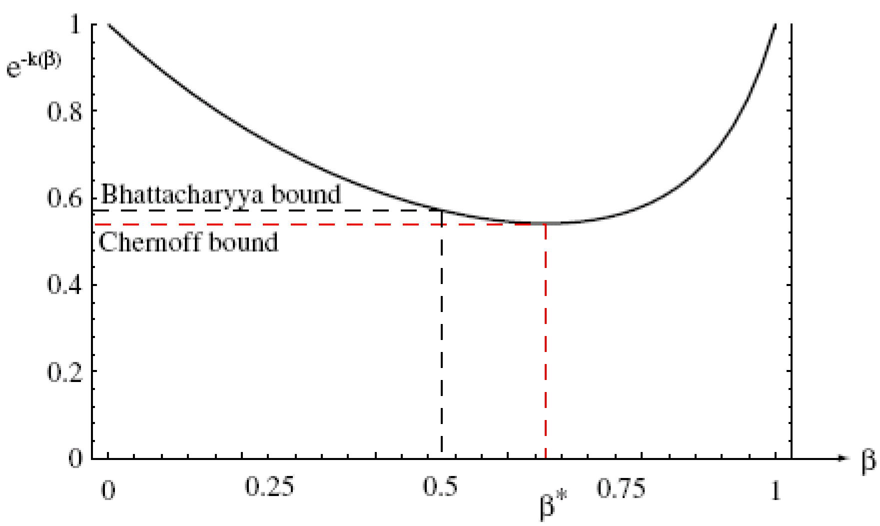
Εάν οι υπό όρους πιθανότητες είναι κανονικές, τα ολοκληρώματα στην Εξ. 71 μπορούν να αξιολογηθούν αναλυτικά (Πρόβλημα 35), δίνοντας:

$$\int p^{\beta}\left(x|\omega_1\right) p^{1-\beta}\left(x|\omega_2\right) dx = e^{-k(\beta)} \quad (72)$$

όπου

$$k\left(\beta\right) = \frac{\beta\left(1-\beta\right)}{2}\left(\mu_2-\mu_1\right)^t\left[\beta\sum_1 + \left(1-\beta\right)\sum_2\right]^{-1}\left(\mu_2-\mu_1\right) + \frac{1}{2}\ln\frac{\left|\beta\sum_1 + \left(1-\beta\right)\sum_2\right|}{\left|\sum_1\right|^{\beta}\left|\sum_2\right|^{1-\beta}}. \quad (73)$$

Το γράφημα στο Σχ. 2.18 παρουσιάζει ένα χαρακτηριστικό παράδειγμα για το πώς $e^{-k(\beta)}$ ποικίλει ανάλογα με το β. Το όριο Chernoff , του P(error) βρίσκεται με αναλυτική ή αριθμητική εύρεση της τιμής του β που ελαχιστοποιεί το $e^{-k(\beta)}$, και αντικαθιστώντας τα αποτελέσματα στην Εξ. 71. Το βασικό όφελος εδώ είναι ότι αυτή η βελτιστοποίηση είναι στο μονοδιάστατο χώρο β, παρά το γεγονός ότι οι κατανομές οι ίδιες μπορεί να είναι σε ένα χώρο αυθαίρετα υψηλής διάστασης.



Σχήμα 2.18: Το λάθος του ορίουChernoff δεν είναι ποτέ χαλαρότερος από το όριο Bhattacharyya. Για αυτό το παράδειγμα, το όριο Chernoff τυχαίνει να είναιστο β = 0.66, και είναι ελαφρώς αυστηρότερο από το όριο Bhattacharyya (β = 0.5).*

2.8.2 Όριο Bhattacharyya

Η γενική εξάρτηση του ορίου Chernoff από το β που φαίνεται στο Σχ. 2.18 είναι χαρακτηριστική για ευρύ φάσμα προβλημάτων — Το όριο είναι χαλαρό για ακραίες τιμές (π.χ., β → 1 και β → 0), και πιο σφιχτό για ενδιάμεσες. Ενώ η ακριβής αξία του βέλτιστου β εξαρτάται από τις παραμέτρους των κατανομών και των εκ των προτέρων πιθανοτήτων, ένα υπολογιστικά απλούστερο, αλλά ελαφρώς λιγότερο σφιχτό όριο μπορεί να παραχθεί για β = 1/2. Αυτό το αποτέλεσμα είναι το αποκαλούμενο όριο λάθους του Bhattacharyya, όπου η Εξ. 71 έπειτα έχει τη μορφή

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)}\int \sqrt{p(x|\omega_1)\,p(x|\omega_2)}\,dx = \sqrt{P(\omega_1)P(\omega_2)}\,e^{-k(1/2)}, \quad (74)$$

όπου από την Εξ. 73έχουμε την γκαουσσιανή περίπτωση:

$$k(1/2) = 1/8(\mu_2-\mu_1)^t\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}}. \quad (75)$$

Τα όρια Chernoff και Bhatacharyya μπορούν ακόμα να χρησιμοποιηθούν ακόμα κι αν οι κατανομές δεν είναι γκαουσσιανές. Εντούτοις, για τις κατανομές που παρεκκλίνουν εμφανώς από γκαουσσιανές, τα όρια δεν θα δώσουν πληροφορίες (Πρόβλημα 32).

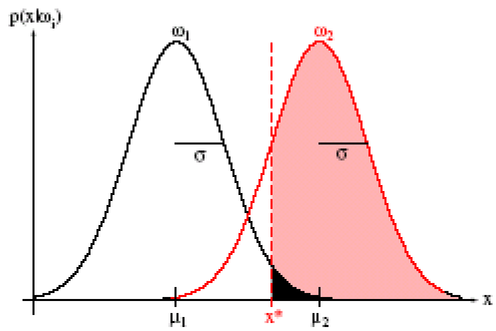
Παράδειγμα 2: Όρια λάθους για γκαουσσιανές κατανομές.

Είναι ένα απλό θέμα για να υπολογίσει το όριο Bhattacharyya για τα δισδιάστατα σύνολα στοιχείων του παραδείγματος 1. Αντικατάσταση των μέσων και των συνδιακυμάνσεων του παραδείγματος 1 στην Εξ. 75 μας δίνει k(1/2) = 4.11 και επομένως από τις Εξ. 74 & 75 το όριο Bhattacharyya στο λάθος P(error) ≤ 0.016382.

Ένας σφιχτότερο όριο στο λάθος μπορεί να προσεγγιστεί με το να ψάξει κανείς αριθμητικά για το όριο Chernoff της Εξ. 73, το οποίο για αυτό το πρόβλημα δίνει 0.016380. Κάποιος μπορεί να πάρει την καλύτερη εκτίμηση υπολογίζοντας αριθμητικά απευθείας το ποσοστό λάθους από την Εξ.. 5, το οποίο δίνει 0.0021, και έτσι τα όρια δεν είναι εδώ ιδιαίτερα σφιχτά. Τέτοια αριθμητική ολοκλήρωση είναι συχνά μη πρακτική για γκαουσσιανές κατανομές σε υψηλότερο από δύο ή τρεις διαστάσεις.

2.8.3 Θεωρία ανίχνευσης σημάτων και χαρακτηριστικά λειτουργίας

Ένα άλλο μέτρο της απόστασης μεταξύ δύο γκαουσσιανών διανομών έχει βρεί μεγάλη χρήση στην πειραματική ψυχολογία, την ανίχνευση ραντάρ και άλλους τομείς. Υποθέστε ότι ενδιαφερόμαστε για την ανίχνευση ενός αδύνατου παλμού, όπως μια αμυδρή λάμψη του φωτός ή ενός αδύνατου αντανάκλαση ραντάρ. Το πρότυπό μας είναι, κατόπιν, ότι σε κάποιο σημείο στον ανιχνευτή υπάρχει εσωτερικό σήμα (όπως μια τάση) x, του οποίου η τιμή έχει μέσο όρο μ2 όταν το εξωτερικό σήμα (παλμός) είναι παρόν, και μ1 όταν δεν υπάρχει. Λόγω του τυχαίου θορύβου —μέσα και έξω από τον ίδιο τον ανιχνευτή — η πραγματική αξία είναι μια τυχαία μεταβλητή. Υποθέτουμε ότι οι διανομές είναι κανονικές με διαφορετικά μέσα αλλά την ίδια μεταβλητότητα, π.χ. $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$ όπως φαίνεται στο Σχ. 2.19.



Σχήμα 2.19: Κατά τη διάρκεια μιας στιγμής όπου κανένας εξωτερικός παλμός δεν είναι παρών, η πυκνότητα πιθανότητας για ένα εσωτερικό σήμα είναι κανονική, π.χ., $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$. *όταν το εξωτερικό σήμα είναι παρόν, η πυκνότητα είναι* $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. *Οποιοδήποτε κατώτατο όριο απόφασης x* θα καθορίσει την πιθανότητα ενός χτυπήματος (η κόκκινη περιοχή κάτω από την καμπύλη* ω_2*, επάνω από το x*) και ενός ψεύτικου συναγερμού (η μαύρη περιοχή κάτω από την καμπύλη* ω_1*, πάνω από x*).*

Ο ανιχνευτής (ταξινομητής) υιοθετεί μια αξία κατώτατων ορίων x* για τον καθορισμό εάν ο εξωτερικός παλμός είναι παρών, αλλά ας υποθέσουμε ότι εμείς, ως πειραματιστές, δεν έχουμε πρόσβαση σε αυτή την τιμή (ούτε στα μέσα και τις σταθερές αποκλίσεις των κατανομών). Επιδιώκουμε να βρούμε κάποια μέτρηση εάν ο παλμός είναι παρών ή όχι, σε μιας μορφής ανεξάρτητα από την επιλογή του x*. Μια τέτοιο μέτρηση είναι η *ιδιότητα διάκρισης*, η οποία περιγράφει τις έμφυτες και αμετάβλητες ιδιότητες λόγω του θορύβου και της δύναμης του εξωτερικού σήματος, αλλά όχι λόγω της στρατηγικής της απόφασης (δηλ., η πραγματική επιλογή του x*)

.

Η ιδιότητα διάκρισης ορίζεται ως

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}. \tag{76}$$

Ένα μεγάλο d’ είναι βεβαίως επιθυμητό.

Ενώ δεν γνωρίζουμε τα μ_1, μ_2, σ και x*, υποθέτουμε εδώ ότι ξέρουμε την κατάσταση της φύσης και την απόφαση του συστήματος. Αυτές οι πληροφορίες μας επιτρέπουν να βρούμε το d'. Για αυτόν τον λόγο, εξετάζουμε τις ακόλουθες τέσσερις πιθανότητες:

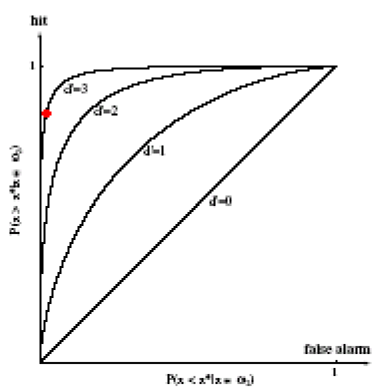
• $P(x > x^* \mid x \in \omega_2)$: Χτύπημα — η πιθανότητα ότι το εσωτερικό σήμα είναι επάνω από το x*
δεδομένου ότι το εξωτερικό σήμα είναι παρόν

• $P(x > x^* \mid x \in \omega_1)$: Ψευτικος συναγερμός — η πιθανότητα ότι το εσωτερικό σήμα είναι επάνω από το x* η μοχθηρία απουσία του εξωτερικού σήματος είναι παρούσα

• $P(x < x^* \mid x \in \omega_2)$: αστοχία — η πιθανότητα ότι το εσωτερικό σήμα είναι κάτω από το x*
δεδομένου ότι το εξωτερικό σήμα είναι παρόν

• $P(x < x^* \mid x \in \omega_1)$: μια απόρριψη του σωστού — η πιθανότητα που το εσωτερικό σήμα είναι κάτω από το x* δεδομένου ότι το εξωτερικό σήμα δεν είναι παρόν.

Εάν πραγματοποιούμε έναν μεγάλο αριθμό δοκιμών (και μπορούμε να υποθέσουμε ότι το x* καθορίζεται, αν και σε άγνωστη τιμή), μπορούμε να καθορίσουμε αυτές τις πιθανότητες πειραματικά, ειδικότερα τα ποσοστά χτυπήματος και ψεύτικων συναγερμών. Σχεδιάζουμε ένα σημείο που αντιπροσωπεύει αυτά τα ποσοστά σε μια δισδιάστατη γραφική παράσταση. Εάν οι πυκνότητες είναι καθορισμένες αλλά το κατώφλι x* αλλάζει, κατόπιν τα ποσοστά χτυπήματος και ψεύτικων συναγερμών θα αλλάξουν επίσης. Κατά συνέπεια βλέπουμε ότι για δεδομένη ιδιότητα διάκρισης δ', το σημείο μας θα κινηθεί κατά μήκος μιας ομαλής καμπύλης — ένα χαρακτηριστικό λειτουργίας δεκτών ή Καμπύλη Roc (Σχ. 2.20)



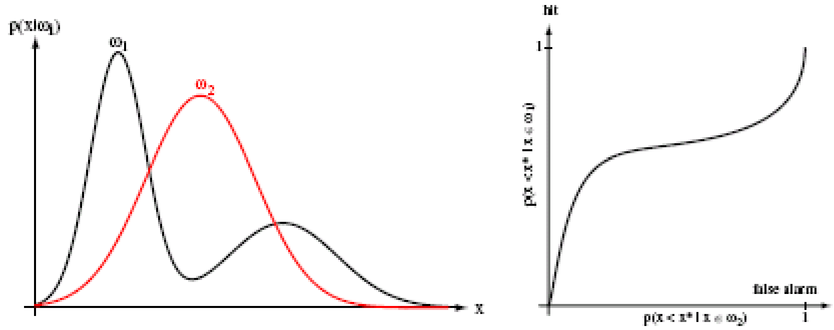
Σχήμα 2.20: Σε μια καμπύλη χαρακτηριστικών λειτουργίας δεκτών (ROC), η τετμημένη είναι η πιθανότητα του ψεύτικου συναγερμού, $P(x>x^*|x \in \omega_1)$, και η τεταγμένη η πιθανότητα του χτυπήματος, $P(x>x^*|x \in \omega_2)$. Από τα μετρημένα ποσοστά χτυπήματος και ψεύτικων συναγερμών (που αντιστοιχούν στο x^* στο σχέδιο 2.19, και παρουσιάζονται ως κόκκινα σημείο), μπορούμε να συναγάγουμε ότι d'= 3.

Το μεγάλο όφελος αυτού του πλαισίου ανίχνευσης σημάτων είναι ότι μπορούμε να διακρίνουμε λειτουργικά μεταξύ ιδιότητας διάκρισης και της απόφασης bias — ενώ η πρώτη είναι η έμφυτη ιδιοκτησία του συστήματος ανιχνευτών, η τελευταία οφείλεται Στο πίνακα απωλειών του δέκτη ο οποίος είναι μεταβλητός. Αν και μέσω οποιουδήποτε ζευγαριού χτυπήματος και ψεύτικου συναγερμού τα ποσοστά περνούν από τη μια και μόνο μια ROC καμπύλη, όμως, εφ' όσον το ποσοστό δεν είναι ακριβώς 0 ή 1, μπορούμε να καθορίσουμε ιδιότητα διάκρισης από αυτά τα ποσοστά (Πρόβλημα 38). Επιπλέον, εάν η γκαουσσιανή υπόθεση περιέχει ένα προσδιορισμός της ιδιότητας διάκρισης (από έναν αυθαίρετο x*) μας επιτρέπει να υπολογίσουμε το ποσοστό λάθους Bayes — τη σημαντικότερη ιδιότητα οποιουδήποτε ταξινομητή. Εάν το πραγματικό ποσοστό λάθους διαφέρει από το ποσοστό Bayes που προκύπτει κατ' αυτό τον τρόπο, πρέπει να αλλάξουμε το κατώφλι x* αναλόγως.

Είναι απλό θέμα να γενικευτεί η ανωτέρω συζήτηση και να εφαρμοστεί σε δύο κατηγορίες που έχουν αυθαίρετες πολυδιάστατες κατανομές, γκαουσσιανές ή όχι. Υποθέστε ότι έχουμε δύο κατανομές $p(x|\omega_1)$ and $p(x|\omega_2)$ οι όποιες επικαλύπτονται, και έχουμε έτσι μη - μηδενικό λάθος ταξινόμησης Bayes. Ακριβώς όπως είδαμε ανωτέρω, οποιοδήποτε πρότυπο από το ω_2 θα μπορούσε να ταξινομηθεί κατάλληλα ως ω_2 (ένα "χτύπημα") ή να καταχωρηθεί λάθος ως ω1 (ένας "ψεύτικος συναγερμός"). Αντίθετα από τη μονοδιάστατη περίπτωση ανωτέρω, εντούτοις, μπορούν να υπάρξουν πολλά όρια απόφασης που δίνουν ένα συγκεκριμένο ποσοστό χτυπημάτων, το κάθε ένα με ένα διαφορετικό ποσοστό ψεύτικων συναγερμών. Σαφώς, εδώ δεν μπορούμε να καθορίσουμε μια θεμελιώδη μέτρηση ιδιότητας διάκρισης χωρίς γνώση περισσότερων για τον ελλοχεύοντα κανόνα απόφασης εκτός από τα ποσοστά χτυπήματος και ψεύτικων συναγερμών.

Σε ένα σπάνια εφικτό ιδανικό, μπορούμε να φανταστούμε ότι τα μετρημένα ποσοστά χτυπήματος και ψεύτικου συναγερμού είναι βέλτιστα, παραδείγματος χάριν από όλους τους κανόνες απόφασης που δίνουν τη μέτρηση του ποσοστού χτυπήματος, ο κανόνας που χρησιμοποιείται πραγματικά είναι αυτό που έχει το ελάχιστο ποσοστό ψεύτικων συναγερμών.

Εάν κατασκευάζαμε έναν πολυδιάστατο ταξινομητή — ανεξάρτητα από τις κατανομές που θα χρησιμοποιούσαμε — μπορούμε να προσπαθήσουμε να χαρακτηρίσουμε το πρόβλημα κατ' αυτό τον τρόπο, αν και πιθανώς απαιτούνται μεγάλοι υπολογιστικοί πόροι για να ψάξουμε για τέτοια βέλτιστο ποσοστά χτυπήματος και ψεύτικου συναγερμού. Στην πράξη, αντ' αυτού αποφεύγουμε το ιδεατό, και χρησιμοποιούμε απλά μια ενιαία παράμετρο ελέγχοντας τον κανόνα απόφασης και σχεδιάζοντας τα προκύπτοντα ποσοστά χτυπημάτων και ψεύτικων συναγερμών – μια καμπύλη που καλείτε ένα *χαρακτηριστικό λειτουργίας*. Μια τέτοια παράμετρος ελέγχου μπορεί να είναι το b$_i$as ή η μη γραμμικότητα σε μια διακρίνουσα συνάρτηση. Είναι παραδοσιακό να επιλεχτεί μια παράμετρος ελέγχου που να μπορεί να παραγάγει, στις ακραίες τιμές, είτε ένα εξαφανισμένο ποσοστό ψεύτικου συναγερμού ή εξαφανισμένο ποσοστό χτυπήματος, ακριβώς όπως μπορεί να επιτευχθεί με ένα πολύ μεγάλο ή πολύ μικρό x* σε μια ROC καμπύλη. Πρέπει να σημειώσουμε ότι αφού οι κατανομές μπορούν να είναι αυθαίρετες, το χαρακτηριστικό λειτουργίας δεν χρειάζεται να είναι συμμετρικό (σχέδιο 2.21) σε σπάνιες περιπτώσεις δεν χρειάζεται ακόμη και να είναι κοίλο προς τα κάτω σε όλα τα σημεία.



Σχήμα 2.21: Σε μια γενική καμπύλη χαρακτηριστικής λειτουργίας, η τετμημένη είναι η πιθανότητα του ψεύτικο συναγερμό, $P(x \in R_2|x \in \omega_1)$, και η τεταγμένη η πιθανότητα του χτυπήματος, $P(x \in R_2|x \in \omega_2)$. Όπως διευκρινίζεται εδώ, οι καμπύλες χαρακτηριστικής λειτουργίας δεν είναι γενικά συμμετρικές, όπως παρουσιάζονται δεξιά. Οι καμπύλες λειτουργίας ταξινομητών είναι σημαντικές για προβλήματα όπου ο πίνακας απωλειών λ_{ij} μπορεί να αλλάξει. Εάν το χαρακτηριστικό λειτουργίας έχει καθοριστεί ως συνάρτηση της παραμέτρου ελέγχου μπροστά από το χρόνο, είναι απλό θέμα, όταν βρίσκεται αντιμέτωπο με μια νέα συνάρτηση απώλειας, να συναχθεί η παράμετρος ελέγχου που θέτει ότι θα ελαχιστοποιήσει το αναμενόμενο ρίσκο (Πρόβλημα 38).

2.9 Θεωρία απόφασης Bayes — Διακριτές εκτιμήτριες

Μέχρι τώρα υποθέταμε ότι το διάνυσμα εκτιμήτριας x μπορούσε να είναι οποιοδήποτε σημείο στο d-διάστατο Ευκλείδειο χώρο, R^d. Εντούτοις, σε πολλές πρακτικές εφαρμογές τα συστατικά του χ είναι δυαδικά τετραδικά η μεγαλύτερης αξίας ακεραίου έτσι ώστε το x να μπορεί να πάρει m τιμές μόνο $v_1, ..., v_m$. Σε τέτοιες περιπτώσεις, η λειτουργία πυκνότητας πιθανότητας $p(x|\omega_j)$ γίνεται μοναδική. Ολοκληρώματα της μορφής

$$\int p\left(x|\omega_{\mathrm{j}}\right) dx \qquad (77)$$

πρέπει έπειτα να αντικατασταθούν από τα αντίστοιχα αθροίσματα, όπως

$$\sum_x P\left(x|\omega_{\mathrm{j}}\right), \qquad (78)$$

όπου καταλαβαίνουμε ότι το άθροισμα είναι σε όλες τις τιμές του x στην διακριτή κατανομή.* Ο εξίσωση του Bayes περιλαμβάνει έπειτα τις πιθανότητες, παρά τις πυκνότητες πιθανότητας:

$$\boxed{P\left(x|\omega_{\mathrm{j}}\right) = \frac{P\left(x|\omega_{\mathrm{j}}\right)P\left(\omega_{\mathrm{j}}\right)}{P(x)},} \qquad (79)$$

* Από τεχνικής άποψης, η Εξ. 78 πρέπει να γραφεί $\sum_k P\left(v_k \mid \omega_j\right)$ όπου $P\left(v_k \mid \omega_j\right)$ όπου είναι η δεσμευμένη πιθανότητα του $x = v_k$ δεδομένου ότι η κατάσταση της φύσης είναι ω_j

όπου

$$P(x) = \sum_{j=1}^{c} P(x|\omega_{\text{j}}) P(\omega_{\text{j}}). \tag{80}$$

Ο ορισμός του δεσμευμένου ρίσκου R(α|x) μένει αναλλοίωτος, και ο θεμελιώδης κανόνας απόφασης του Bayes παραμένει ο ίδιος: Για να ελαχιστοποιήσουμε το συνολικό ρίσκο, επιλέγουμε την ενέργεια α_i για την οποία $R(\alpha_i|x)$ είναι ελάχιστη, ή δίνεται,

$$a^* = \arg\max_{i} R(a_i|\text{x}). \tag{81}$$

Ο βασικός κανόνας για να ελαχιστοποιηθεί το ποσοστό σφάλματος με τη μεγιστοποίηση της εκ των υστέρων πιθανότητας είναι επίσης αμετάβλητος όπως είναι και οι διακρίνουσες συναρτήσεις των Εξ. 25 – 27, αντικαθιστώντας προφανώς τις πυκνότητες p(·) με τις πιθανότητες P(·).

2.9.1 Ανεξάρτητες δυαδικές εκτιμήτριες

Σαν παράδειγμα μιας ταξινόμησης που περιλαμβάνει διακριτές εκτιμήτριες, θεωρήστε το πρόβλημα δύο κατηγοριών στο οποίο τα συστατικά του διανύσματος εκτιμήτριας είναι δυαδικά και υπό όρους ανεξάρτητα. Για να είμαστε πιο συγκεκριμένοι θέτουμε $x = (x_1, ..., x_d)^t$, όπου τα συστατικά x_i είναι είτε 0 ή 1, με

$$p_i = Prob(x_i = 1|\omega_1) \tag{82}$$

και

$$q_i = Prob(x_i = 1|\omega_2) \tag{83}$$

Αυτό είναι ένα πρότυπο ενός προβλήματος ταξινόμησης στο οποίο κάθε εκτιμήτρια μας δίνει μια ναι/όχι απάντηση για το πρότυπο. Αν $p_i > q_i$, περιμένουμε ότι η i-οστή εκτιμήτρια να δώσει “ναι” σαν απάντηση ποιο συχνά όταν η κατάσταση της φύσης είναι ω_1 από ότι όταν είναι ω_2. (Για παράδειγμα, θεωρήστε δύο εργοστάσια που κάθε ένα κατασκευάζει το ίδιο αυτοκίνητο, κάθε ενός εκ των οποίων τα d τμήματα θα μπορούσαν να είναι λειτουργικά ή ελαττωματικά. Εάν ήταν γνωστό πώς τα εργοστάσια διέφεραν στην αξιοπιστία τους για την κατασκευή κάθε συστατικού, κατόπιν αυτό το μοντέλο θα μπορούσε να χρησιμοποιηθεί για να κρίνει ποιο εργοστάσιο κατασκεύασε ένα αυτοκίνητο βασισμένο στη γνώση ποιες εκτιμήτριες είναι λειτουργικές και ποιες ελαττωματικές.) Με να υποθέσουμε την υπό όρους ανεξαρτησία μπορούμε να γράψουμε $P(x|\omega_i)$ σαν το προϊόν των πιθανοτήτων για τα συστατικά x.

Λαμβάνοντας υπόψη αυτήν την υπόθεση, ένας ιδιαίτερα κατάλληλος τρόπος για να γράψουμε τις σχετιζόμενες με τη κλάση πιθανότητες είναι ο ακόλουθος:

$$P(x|\omega_1) = \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1-x_i} \tag{84}$$

και

$$P\left(x|\omega_2\right)=\prod_{i=1}^{d} q_i^{x_i}\left(1-q_i\right)^{1-x_i} \tag{85}$$

Κατόπιν ο ρυθμός πιθανότητας δίνεται από

$$\frac{P\left(x|\omega_1\right)}{P\left(x|\omega_2\right)}=\prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i} \tag{86}$$

και συνεπώς η Εξ. 30 παράγει τη διακρίνουσα συνάρτηση

$$g\left(x\right)=\sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i}+\left(1-x_i\right)\ln\frac{1-p_i}{1-q_i}\right]+\ln\frac{P\left(\omega_1\right)}{P\left(\omega_2\right)}. \tag{87}$$

Σημειώνουμε ειδικά ότι αυτή η διακρίνουσα συνάρτηση είναι γραμμική στο x_i και επομένως μπορούμε να γράψουμε

$$g\left(x\right)=\sum_{i=1}^{d} w_i x_i + w_0, \tag{88}$$

όπου

$$w_i=\ln\frac{p_i\left(1-p_i\right)}{q_i\left(1-q_i\right)} \qquad i=1,\ldots,d \tag{89}$$

και

$$w_0=\sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i}+\ln\frac{P\left(\omega_1\right)}{P\left(\omega_2\right)}. \tag{90}$$

Εξετάστε αυτά τα αποτελέσματα για να δούμε τι αποτελέσματα μπορούν να δώσουν. Θυμηθείτε πρώτα ότι αποφασίζουμε ω_1 αν $g(x) > 0$ και ω_2 αν $g(x) \le 0$. Είδαμε ότι g(x) είναι ένας υπολογισμένος συνδυασμός των συστατικών του x. Το πλάτος του βάρους w_i δείχνει τη σχετικότητα με την απάντηση "ναι " για το x_i ώστε να αποφασίσουμε την ταξινόμηση. Αν $p_i = q_i$, x_i δε μας δίνει πληροφορία για την κατάσταση της φύσης, και $w_i = 0$, όπως θα περιμέναμε.

Αν $p_i>q_i$, τότε $1-p_i<1-q_i$ και w_i θετικό. Επομένως σε αυτή την περίπτωση μια απάντηση "ναι" για το x_i συνεισφέρει w_i ψήφους για το ω1. Επιπλέον, για κάθε σταθερό $q_i < 1$, w_i μεγαλώνει όσο το p_i μεγαλώνει. Αφ' ετέρου, αν $p_i<q_i$, w_i είναι αρνητικό και μια απάντηση "ναι" συνεισφέρει $|w_i|$ ψήφους για ω2.

Η κατάσταση της ανεξαρτησίας εκτιμητριών οδηγεί σε έναν πολύ απλό (γραμμικό) ταξινομητή. φυσικά εάν οι εκτιμήτριες δεν ήταν ανεξάρτητες, θα χρειαζόταν ένας πιό περίπλοκος ταξινομητής. Θα το συναντήσουμε αυτό πάλι για τα συστήματα με τις συνεχής εκτιμήτριες στο κεφ. ??, αλλά σημειώνουμε εδώ ότι όσο πιο ανεξάρτητες μπορούμε να κάνουμε τις εκτιμήτριες, τόσο απλούστερος ο ταξινομητής θα είναι.

Οι εκ των προτέρων πιθανότητες $P(\omega_i)$ εμφανίζονται στη διακρίνουσα συνάρτηση μόνο μέσω του κατωφλίου βάρος w_0. Αύξηση του $P(\omega_1)$ αυξάνει το w_0 και επηρεάζει την απόφαση προς το μέρος του ω_1, ενώ η μείωση του $P(\omega_1)$ έχει το αντίθετο αποτέλεσμα. Γεωμετρικά, οι πιθανές τιμές για x εμφανίζονται ως κορυφές ενός d-διαστατού υπερκύβου; η επιφάνεια απόφασης που καθορίζεται από το g(x) = 0 είναι μια υπερευθεία που χωρίζει τις κορυφές του ω_1 από τις κορυφές τις ω_2.

Παράδειγμα 3: Αποφάσεις του Bayes για τρισδιάστατες δυαδικές εκτιμήτριες

Υποθέστε ότι δύο κατηγορίες αποτελούνται από ανεξάρτητες δυαδικές εκτιμήτριες σε τρεις διαστάσεις με γνωστές τις πιθανότητες εκτιμητριών. Θα κατασκευάσουμε το όριο απόφασης του Bayes εάν P(ω1) = P(ω2) = 0.5 και τα επιμέρους συστατικά υπακούνε:

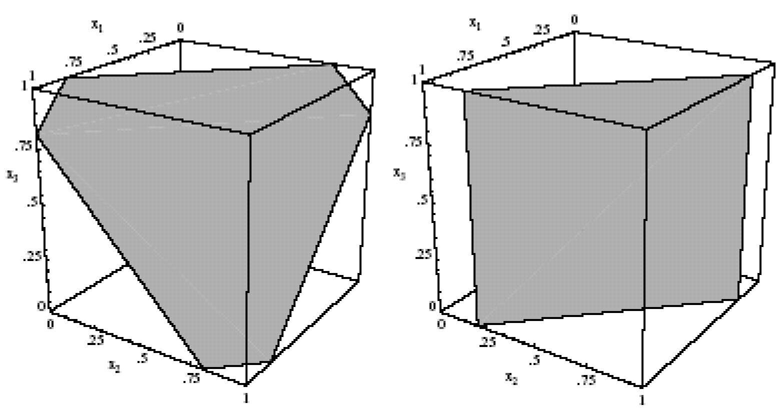
$$\begin{cases} p_i = 0.8 \\ q_i = 0.5 \end{cases} \qquad i = 1, 2, 3.$$

Από τις Εξ. 89 & 90 έχουμε ότι τα βάρη είναι

$$w_i = \ln \frac{0.8(1-0.5)}{0.5(1-0.8)} = 1.3863$$

Και η τιμή του bias

$$w_0 = \sum_1^3 \ln \frac{1-0.8}{1-0.5} + \ln \frac{0.5}{0.5} = 1.2.$$



Το όριο απόφασης για το παράδειγμα που περιλαμβάνει τα τρισδιάστατα δυαδικά χαρακτηριστικά γνωρίσματα. Στο αριστερό παρουσιάζουμε την περίπτωση p_i=0.8 και q_i=0.5. Στο δεξί χρησιμοποιούμε τις ίδιες τιμές εκτός από $p_3 = q_3$, το οποίο οδηγεί σε w_3=0 και μια επιφάνεια απόφασης παράλληλη στο άξονα x_3.

Η επιφάνεια g(x) = 0 της Εξ. 88 φαίνεται στα αριστερά του σχήματος. Πράγματι, όπως περιμέναμε, το όριο βάζει τα σημεία με δύο ή πιο πολλές απαντήσεις " ναι" στη κατηγορία ω1, δεδομένου ότι εκείνη η κατηγορία έχει μια υψηλότερη πιθανότητα της κατοχής οποιασδήποτε εκτιμήτριας με τιμή 1.

Υποθέστε άντ' αυτού ότι ενώ οι εκ των προτέρων πιθανότητες παρέμειναν οι ίδιες, ατομικά συστατικά υπακούνε:

$$\begin{cases} p_1 = p_2 = 0.8, p_3 = 0.5 \\ q_1 = q_2 = q_3 = 0.5 \end{cases}$$

Σε αυτήν την περίπτωση η εκτιμήτρια x3 δεν μας δίνει καμία πληροφορία για τις κατηγορίες, και ως εκ τούτου το όριο απόφασης είναι παράλληλο στο x3 άξονα. Σημειώστε ότι σε αυτήν την ιδιαίτερη περίπτωση υπάρχει ένα μεγάλο εύρος στις θέσεις του ορίου απόφασης που αφήνει την κατηγοριοποίηση αμετάβλητη, όπως είναι ιδιαίτερα σαφές στο δεξί σχήμα.

2.10 *Αγνοούμενες Εκτιμήτριες και Εκτιμήτριες με θόρυβο

Εάν ξέρουμε την πλήρη πιθανοτική δομή ενός προβλήματος, μπορούμε να κατασκευάσουμε (βελτιστοποιήσουμε) τον Κανόνας απόφασης του Bayes. Υποθέστε ότι αναπτύσσουμε έναν ταξινομητή Bayes χρησιμοποιώντας τα αδιάφθορα στοιχεία, αλλά τα στοιχεία εισαγωγής μας (δοκιμή) είναι έπειτα αλλοιωμένα με γνωστό τρόπο. Πώς μπορούμε να ταξινομήσουμε τέτοιες αλλοιωμένες εισαγωγές για να λάβουμε ένα ελάχιστο λάθος;

Υπάρχουν δύο περιπτώσεις που επιλύονται αναλυτικά ιδιαίτερου ενδιαφέροντος: όταν μερικά από τα χαρακτηριστικά γνωρίσματα λείπουν, και όταν αλλοιώνονται από μια πηγή θορύβου με γνωστές ιδιότητες. Σε κάθε περίπτωση η βασική προσέγγισή μας είναι να ανακτήσουμε όσες πληροφορίες για την ελλοχεύουσα κατανομή μπορούμε και να χρησιμοποιήσουμε τον κανόνα απόφασης του Bayes.

2.10.1 Αγνοούμενες Εκτιμήτριες

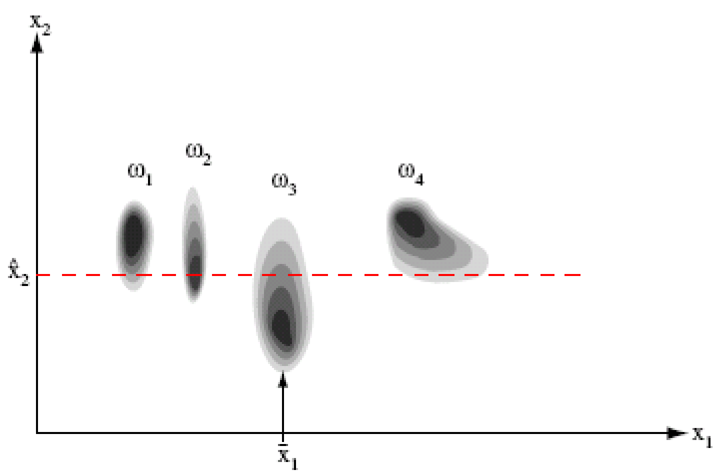
Υποθέστε ότι έχουμε ένα σύστημα αναγνώρισης του Bayes (ή άλλο) για ένα πρόβλημα που χρησιμοποιεί δύο εκτιμήτριες, αλλά για ένα ιδιαίτερη ακολουθία που ταξινομείται, μια από τις εκτιμήτριες λείπει.* Παραδείγματος χάριν, μπορούμε εύκολα να φανταστούμε ότι η ελαφρότητα μπορεί να μετρηθεί από μια πλευρά ενός ψαριού, αλλά το πλάτος δεν μπορεί λόγω επαφής με ένα άλλο ψάρι.

* Στην πράξη, η ανακάλυψη ότι η εκτιμήτρια αγνοείται είναι χειρότερο από το να έχει τιμή μηδέν (η του μέσου όρου)

Μπορούμε να επεξηγήσουμε με τέσσερις κατηγορίες μια κάπως γενικότερη περίπτωση (σχέδιο 2.22). Υποθέστε ότι για ένα ιδιαίτερο σχέδιο δοκιμής το χαρακτηριστικό γνώρισμα x_1 λείπει, και η μετρημένη αξία από x_2 είναι $\hat{x}_2$. Σαφώς εάν υποθέτουμε η άγνωστη τιμή είναι ο μέσος όρος όλων των x_0 τιμών, δηλ., $\bar{x}_1$, θα ταξινομήσουμε το σχέδιο ως ω_3. Εντούτοις, εάν οι εκ των προτέρων είναι ίσες, η ω_2 θα ήταν καλύτερη απόφαση , δεδομένου ότι ο αριθμός υπονοεί ότι $p\left(\hat{x}_2 \mid \omega_2\right)$είναι το μεγαλύτερο των τεσσάρων πιθανοτήτων.

Για να διευκρινίσουμε την παραδοχή μας αφήνουμε $X=[x_g,x_b]$, όπου x_g αντιπροσωπεύει γνωστή ή τα "καλά" χαρακτηριστικά γνωρίσματα και x_b αντιπροσωπεύουν τους "κακούς" , δηλ., είτε άγνωστους είτε που λείπουν. Εμείς Επιδιώκουμε τον κανόνα Bayes δεδομένου των καλών χαρακτηριστικών γνωρισμάτων, και για αυτό οι εκ των υστέρων πιθανότητες απαιτούνται. Από την άποψη των καλών χαρακτηριστικών γνωρισμάτων οι εκ των υστέρων είναι

$$\begin{aligned} P(\omega_i|x_g) &= \frac{p(\omega_i|x_g)}{p(x_g)} = \frac{\int p(\omega_i, x_g, x_b)dx_b}{p(x_g)} \\ &= \frac{\int P(\omega_i|x_g, x_b)p(x_g, x_b)dx_b}{p(x_g)} \\ &= \frac{\int g_i(x)p(x)dx_b}{\int p(x)dx_b}, \end{aligned} \tag{91}$$



Σχήμα 2.22: Four categories have equal priors and the class-conditional distributionsshown. Εάν ένα σημείο ελέγχου είναι παρών σε μια αγνοούμενη εκτιμήτρια (εδώ, x1) και η άλλη μετρείται ότι έχει τιμή $\hat{x}_2$ (κόκκινη διακεκομμένη γραμμή), θέλουμε ο ταξινομητής μας να ταξινόμηση το πρότυπο ως ω_2, επειδή $p\left(\hat{x}_2 \mid \omega_2\right)$είναι η μεγαλύτερη από τις πιθανότητες.

όπου $g_i(x) = g_i(x_g,x_b) = P(\omega_i|x_g,x_b)$ είναι μια μορφή της διακρίνουσας συνάρτησης.

Αναφερόμαστε στο $\int p\left(\omega_i, x_g, x_b\right) dx_b$, σα μια οριακή κατανομή, και λέμε ότι η πλήρης δεσμευμένη οριακή κατανομή έχει όρια γύρω από τη μεταβλητή xb. Σε συντομία, η Εξ. 91 δείχνει ότι πρέπει να ενσωματώσουμε (οριοθετήσουμε) τις εκ των υστέρων πιθανότητες γύρω από τις «κακές» εκτιμήτριες. Τελικά χρησιμοποιούμε τον κανόνα απόφασης του Bayes για τις προκύπτουσες εκ των υστέρων πιθανότητες, π.χ., επιλέγουμε ω_i αν $P(\omega_i|x_g) > P(\omega_j|x_g)$ για κάθε i και j. Θα εξετάσουμε τον αλγόριθμο Μεγιστοποίησης της προσδοκίας(ΕΜ)στο κεφ . ??, που αναφέρεται σε ένα σχετικό πρόβλημα αγνοούμενης εκτιμήτριας.

2.10.2 Εκτιμήτριες με θόρυβο

Είναι εύκολο να γενικεύσουμε τα αποτελέσματα της Εξ. 91 στην περίπτωση που μια συγκεκριμένη εκτιμήτρια έχει αλλοιωθεί από στατιστικά ανεξάρτητο θόρυβο*. Για παράδειγμα, στο πρώτο παράδειγμα ταξινόμησης μας, μπορεί να έχουμε μια αξιόπιστη μέτρηση του μήκους ενώ η μεταβλητότητα της πηγής φωτισμού να επηρεάσει τη μέτρηση φωτεινότητας. Υποθέτουμε ότι έχουμε μη αναλλοίωτες (καλές) εκτιμήτριες x_g, όπως και πριν, και ένα πρότυπο θορύβου, το οποίο εκφράζεται ως $p(x_b|x_t)$. Έστω x_t η πραγματική τιμή των εκτιμητριών x_b, π.χ χωρίς την παρουσία θορύβου, αν οι xb παρακολουθούνταν αντί για τις x_t. Υποθέτουμε ότι αν οι x_t ήταν γνωστές, x_b θα ήταν ανεξάρτητες από τα ω_i και x_g. Από μια τέτοια υπόθεση παίρνουμε:

$$P(\omega_i|x_g,x_b) = \frac{\int p(\omega_i,x_g,x_b,x_t)dx_t}{p(x_g,x_b)}. \tag{92}$$

Τώρα που $p(\omega_i,x_g,x_b,x_t) = P(\omega_i|x_g,x_b,x_t)p(x_g,x_b,x_t)$ από την υπόθεση ανεξαρτησίας που δεχτήκαμε, αν γνωρίζουμε το x_t, τότε το x_b δε δίνει επιπλέον πληροφορίες για το ω_i. Επομένως έχουμε $P(\omega_i|x_g,x_b,x_t)=P(\omega_i|x_g,x_t)$. Ομοίως, έχουμε $p(x_g,x_b,x_t)=p(x_b|x_g,x_t)p(x_g,x_t)$, και $p(x_b|x_g,x_t) = p(x_b|x_t)$. Βάζοντας τα μαζί έχουμε

$$\begin{aligned} P(\omega_i|x_g,x_b) &= \frac{\int P(\omega_i|x_g,x_t)p(x_g,x_t)p(x_b|x_t)dx_t}{\int p(x_g,x_t)p(x_b|x_t)dx_t} \\ &= \frac{\int g_i(x)p(x)p(x_b|x_t)dx_t}{\int p(x)p(x_b|x_t)dx_t}, \end{aligned} \tag{93}$$

τις οποίες χρησιμοποιούμε σαν διακρίνουσες συναρτήσεις όπως απαιτεί ο κανόνας του Bayes.

Η εξίσωση 93 διαφέρει από την Εξ. 91 απλώς από το γεγονός ότι ο ακέραιος μετριέται από το πρότυπο θορύβου. Στην ακραία περίπτωση που $p(x_b|x_t)$ είναι ομοιόμορφο σε όλο το χώρο (και επομένως δε παρέχει καμία πληροφορία πρόβλεψης για την ταξινόμηση), η εξίσωση γίνεται ίδια με αυτή για αγνοούμενες εκτιμήτριες — ένα ικανοποιητικό αποτέλεσμα.

* Φυσικά, για να δώσουμε στο ταξινομητή να καταλάβει ότι μια εκτιμήτρια λείπει, πρέπει αυτός να έχει σχεδιαστεί ώστε να παίρνει και τιμές που δεν είναι αριθμητικές

2.11 *Σύνθετη θεωρία του Bayes και Πλαίσιο

Ας επανεξετάσουμε το εισαγωγικό παράδειγμά σχεδίασης ενός ταξινομητή για να κατηγοριοποιεί δύο τύπους από ψάρια. Η αρχική μας υπόθεσή ήταν ότι η ακολουθία τύπων ψαριών ήταν τόσο απρόβλεπτη που η κατάσταση της φύσης έμοιαζε με μια τυχαία μεταβλητή. Χωρίς να εγκαταλείψουμε αυτή τη τοποθέτηση, ας εξετάσουμε τη δυνατότητα ότι οι διαδοχικές καταστάσεις της φύσης να μην είναι στατιστικά ανεξάρτητες. Πρέπει να είμαστε σε θέση να εκμεταλλευτούμε τέτοια στατιστική εξάρτηση για να κερδίσουμε βελτιωμένη απόδοση. Αυτό είναι ένα παράδειγμα της χρήσης του πλαισίου στην ενίσχυση λήψη απόφασης.

Ο τρόπος με τον οποίο εκμεταλλευόμαστε τέτοιες πληροφορίες πλαισίου είναι κάπως διαφορετικός όταν μπορούμε να περιμένουμε τα n ψάρια να προκύψουν και να λάβουμε έπειτα όλες τις n αποφάσεις από κοινού από όταν πρέπει να αποφασίσουμε για κάθε ψάρι όταν εμφανίζεται. Το πρώτο πρόβλημα είναι ένα σύνθετο πρόβλημα απόφασης, και το δεύτερο είναι ένα διαδοχικό σύνθετο πρόβλημα απόφασης. Η προηγούμενη περίπτωση είναι εννοιολογικά απλούστερη, και είναι αυτή που θα εξετάσουμε εδώ.

Για να ορίσουμε το γενικό πρόβλημα, έστω $\omega=(\omega(1)..., \omega(n))^t$ είναι ένα διάνυσμα που δείχνει τις n καταστάσεις της φύσης, με $\omega(_i)$ να παίρνει μια από τις τιμές $\omega_1, ..., \omega_c$ Έστω P(ω) είναι η εκ των προτέρων πιθανότητα για τις n καταστάσεις τις φύσης. Έστω $X=(x_1..., x_n)$ να είναι ένας πίνακας που δίνει τα n διανύσματα όπου x_i είναι το διάνυσμα εκτιμήτριας όταν η κατάσταση της φύσης είναι $\omega(_i)$. Τέλος, έστω p(X|ω) είναι η συνάρτηση δεσμευμένης πυκνότητας πιθανότητας για X δεδομένου ότι η κατάσταση της φύσης είναι ω. Χρησιμοποιώντας αυτήν την σήμανση βλέπουμε
ότι η εκ των υστέρων πιθανότητα ω δίνεται από

$$\mathrm{P}(\omega|\mathrm{X})=\frac{\mathrm{P}(\mathrm{X}|\omega)\mathrm{P}(\omega)}{\mathrm{p}(\mathrm{X})}=\frac{\mathrm{P}(\mathrm{X}|\omega)\mathrm{P}(\omega)}{\sum_{\omega}\mathrm{p}(\mathrm{X}|\omega)\mathrm{P}(\omega)}. \tag{94}$$

Γενικά, κάποιος μπορεί να καθορίσει ένα πίνακα απώλειας για το σύνθετο πρόβλημα απόφασης και να επιδιώξει έναν κανόνα απόφασης που ελαχιστοποιεί το σύνθετο κίνδυνο. Η ανάπτυξη αυτής της θεωρίας παραλληλίζει τη συζήτησή μας για το απλό πρόβλημα απόφασης, και καταλήγει στο συμπέρασμα ότι η βέλτιστη διαδικασία είναι να ελαχιστοποιηθεί ο σύνθετος δεσμευμένος κίνδυνος. Ειδικότερα, δεν υπάρχει καμία απώλεια για την ύπαρξη σωστού, και εάν όλα τα λάθη είναι εξίσου δαπανηρά, κατόπιν η διαδικασία μειώνεται στον υπολογισμό του P(ω|X) για κάθε ω και στην επιλογή του ω για το οποίο η εκ των υστέρων πιθανότητα να είναι μέγιστη.

Αν και αυτό παρέχει τη θεωρητική λύση στην πράξη ο υπολογισμός του P(ω|X) μπορεί εύκολα να αποδειχθεί χρονοβόρα διαδικασία. Αν κάθε $\omega(_i)$ μπορεί να πάρει c τιμές, υπάρχουν c^n πιθανές τιμές του ω που πρέπει να ληφθούν υπόψη. Κάποια απλοποίηση μπορεί να ληφθεί εάν η κατανομή του διανύσματος εκτιμήτριας x_i εξαρτάται μόνο από την αντίστοιχη κατάσταση της φύσης ω (ι), και όχι από τις τιμές των άλλων διανυσμάτων εκτιμητριών ή τις άλλες καταστάσεις τις φύσης. Σε αυτήν την περίπτωση η δεσμευμένη πυκνότητα p(X|ω) είναι μόνο το προϊόν των πυκνοτήτων των συστατικών $p(x_i|\omega(_i))$:

$$p\left(X|\omega\right)=\prod_{\mathrm{i=1}}^{\mathrm{n}}p\left(x_i|\omega\left(i\right)\right). \tag{95}$$

Ενώ αυτό απλοποιεί το πρόβλημα του υπολογισμού p(X|ω), υπάρχει ακόμα το πρόβλημα από τον υπολογισμό των εκ των προτέρων πιθανοτήτων P(ω). Αυτή η δεσμευμένη πιθανότητα είναι κεντρική στο σύνθετο πρόβλημα απόφασης Bayes, δεδομένου ότι απεικονίζει την αλληλεξάρτηση των καταστάσεων της φύσης. Κατά συνέπεια είναι απαράδεκτο να απλοποιηθεί το πρόβλημα του υπολογισμού P(ω) με να υποθέσει κανείς ότι οι καταστάσεις της φύσης είναι ανεξάρτητες. Επιπλέον, οι πρακτικές εφαρμογές απαιτούν συνήθως κάποια μέθοδο υπολογισμού του P(ω|X) για όλες τις c^n πιθανές τιμές ω. Θα βρούμε κάποια λύση σε αυτό το πρόβλημα στο κεφάλαιο. ??.

Περίληψη

Οι βασικές ιδέες που απορρέουν από τον κανόνα απόφασης του Bayes είναι πολύ απλές. Για να ελαχιστοποιήσει το γενικό κίνδυνο, κάποιος πρέπει πάντα να επιλέξει τη ενέργεια που ελαχιστοποιεί τον δεσμευμένο κίνδυνο, $R(\alpha|x)$. Ειδικότερα, για να ελαχιστοποιήσει την πιθανότητα του λάθους σε ένα πρόβλημα ταξινόμησης, κάποιος πρέπει πάντα να επιλέξει την κατάσταση της φύσης που μεγιστοποιεί τη εκ των υστέρων πιθανότητα $P(\omega_j\ |x)$. Ο εξίσωση του Bayes μας επιτρέπει να υπολογίσουμε τέτοιες πιθανότητες από τις εκ των προτέρων πιθανότητες $P(\omega_j)$ και τις δεσμευμένες πυκνότητες $p(x|\omega_j)$. Εάν υπάρχουν διαφορετικές ποινές για τη λανθασμένη καταχώρηση των σχεδίων από ω_i σαν από ω_j, οι εκ των υστέρων πρέπει να ζυγιστούν πρώτα σύμφωνα με τέτοιες ποινες πρίν ληφθούν μέτρα.

Εάν οι ελλοχεύουσες κατανομές είναι πολλών μεταβλητών γκαουσσιανές, τα όρια απόφασης θα είναι υπερτετραγωνικά, των οποίων η μορφή και η θέση εξαρτάται από τις εκ των προτέρων πιθανότητες, τους μέσους όρους και συνδιακυμάνσεις των κατανομών. Το πραγματικό αναμενόμενο λάθος μπορεί να βρεθεί από τα όρια Chernoff και το υπολογιστικά ευκολότερο όριο Bhattacharyya. Αν η ακολουθία εισόδου έχει αγνοούμενες ή αλλοιωμένες εκτιμήτριες, πρέπει να διαμορφώσουμε τις οριακές κατανομές με την ενσωμάτωση τέτοιων εκτιμητριών, και έπειτα τη χρησιμοποίηση της διαδικασίας απόφασης του Bayes στις προκύπτουσες κατανομές. Οι καμπύλες χαρακτηριστικής λειτουργίας δεκτών περιγράφουν τις έμφυτες και αμετάβλητες ιδιότητες ενός ταξινομητή και μπορούν να χρησιμοποιηθούν, παραδείγματος χάριν, για να καθορίσουν το ποσοστό Bayes.

Για πολλές εφαρμογές ταξινόμησης σχεδίων, το κύριο πρόβλημα στην εφαρμογή αυτών των αποτελεσμάτων είναι ότι οι υπό όρους πυκνότητες $p(x|\omega_j)$ δεν είναι γνωστές. Σε μερικές περιπτώσεις μπορούμε να ξέρουμε τη μορφή που αυτές οι πυκνότητες έχουν, αλλά μπορεί να μην ξέρουμε τις χαρακτηριστικές τιμές των παραμέτρων. Η κλασική περίπτωση εμφανίζεται όταν είναι γνωστές οι πυκνότητες, ή μπορούμε να υποθέσουμε ότι είναι πολλών μεταβλητών κανονικές, αλλά οι τιμές των μέσων διανυσμάτων και πινάκων συνδιακύμανσης δεν είναι γνωστές. Συχνότερα ακόμα λιγότερα είναι γνωστά για τις δεσμευμένες πυκνότητες, και διαδικασίες που είναι λιγότερο ευαίσθητες στις συγκεκριμένες υποθέσεις για τις πυκνότητες πρέπει να χρησιμοποιηθούν. Το μεγαλύτερο μέρος του υπολοίπου αυτού του βιβλίου θα αφιερωθεί στις διάφορες διαδικασίες που έχουν αναπτυχθεί για να αντιμετωπιστούν σε τέτοια προβλήματα.