Κεφάλαιο 2: Θεωρία Απόφασης του Bayes

2.1 Εισαγωγή

Η θεωρία απόφασης του Bayes αποτελεί μια από τις σημαντικότερες στατιστικές προσεγγίσεις για το πρόβλημα της ταξινόμησης προτύπων. Βασίζεται στη σύγκριση μεταξύ διαφόρων αποφάσεων ταξινόμησης με βάση τις πιθανότητες και τα κόστη που σχετίζονται με τις αποφάσεις αυτές. Θεωρεί ότι το πρόβλημα απόφασης ορίζεται με πιθανοθεωρητικούς όρους και ότι όλες οι σχετικές πιθανότητες είναι γνωστές. Στο κεφάλαιο αυτό παρουσιάζονται οι βασικές αρχές της θεωρίας του Bayes.

Έστω (βλέπε κεφάλαιο 1), ότι πρέπει να σχεδιαστεί ένας ταξινομητής για το διαχωρισμό δύο ειδών ψαριού: πέρκας και σολομού. Έστω, επίσης, ότι ένας παρατηρητής, ο οποίος παρατηρεί τα ψάρια που μεταφέρονται στον κινούμενο ιμάντα, δεν είναι σε θέση να προβλέπει το είδος ψαριού που εμφανίζεται κάθε φορά και το ότι η ακολουθία με την οποία εμφανίζονται τα ψάρια είναι εντελώς τυχαία. Προφανώς οι περιπτώσεις που μπορούν να συμβαίνουν είναι δύο: κάθε ψάρι που έρχεται μπορεί να είναι είτε πέρκα είτε σολομός. Έστω ότι με ω συμβολίζεται η κατάσταση της φύσης, με $\omega=\omega_1$ για την πέρκα και $\omega=\omega_2$ για τον σολομό. Επειδή προφανώς η κατάσταση της φύσης είναι εντελώς απρόβλεπτη, το ω θεωρείται ως μία μεταβλητή που πρέπει να περιγραφεί πιθανοτικά.

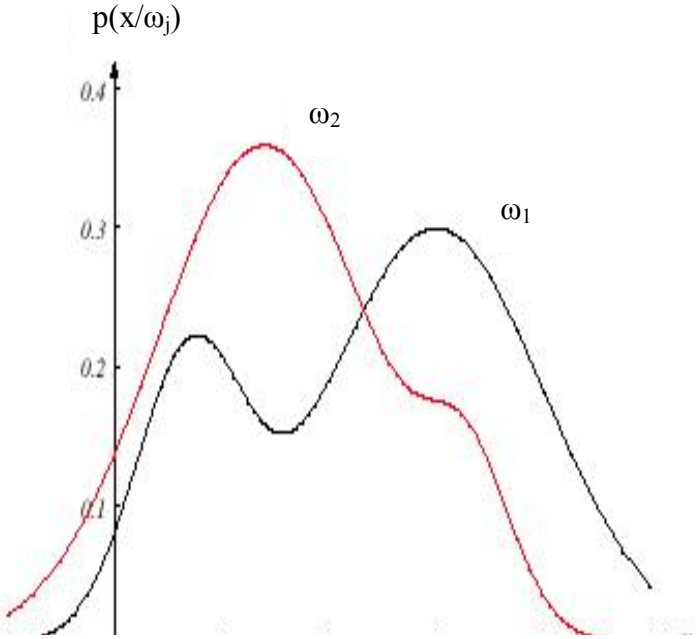
Εάν ο αριθμός των πέρκων, συνολικά, είναι ίδιος με τον αριθμό των σολομών, θα μπορούσε κάποιος να ισχυριστεί ότι το επόμενο ψάρι έχει την ίδια πιθανότητα να είναι είτε πέρκα είτε σολομός. Πιο γενικά, μπορούμε να ισχυριστούμε ότι υπάρχει κάποια εκ των προτέρων (a priori) πιθανότητα $P(\omega_1)$ ότι το επόμενο ψάρι θα είναι πέρκα, και κάποια εκ των προτέρων πιθανότητα $P(\omega_2)$ ότι το επόμενο ψάρι θα είναι σολομός. Θεωρώντας ότι δεν υπάρχουν άλλα είδη ψαριών στην συγκεκριμένη περίπτωση, το άθροισμα των πιθανοτήτων $P(\omega_1)$ και $P(\omega_2)$ ισούται προφανώς με το ένα. Αυτές οι εκ των προτέρων πιθανότητες δείχνουν τη εκ των προτέρων γνώση που έχουμε για το τι ψάρι είναι πιθανόν να εμφανιστεί προτού την πραγματική εμφάνισή του. Αυτή η γνώση, για παράδειγμα, επηρεάζεται από την εποχή του χρόνου που γίνεται η παρατήρηση ή το θαλάσσιο χώρο που έγινε το ψάρεμα.

Έστω λοιπόν ότι σε κάποια χρονική στιγμή πρέπει να πάρουμε απόφαση για το τι ψάρι πρόκειται να εμφανιστεί χωρίς φυσικά να είμαστε σε θέση να το δούμε. Αρχικά, θα μπορούσε κάποιος να ισχυριστεί ότι οποιαδήποτε λανθασμένη απόφαση εμπεριέχει το ίδιο κόστος και ότι η μοναδική πληροφορία που μπορεί να χρησιμοποιηθεί για την απόφαση είναι οι εκ των προτέρων πιθανότητες. Εάν επομένως η απόφαση θα έπρεπε να παρθεί υπό αυτές τις συνθήκες, το πιο λογικό θα ήταν να χρησιμοποιηθεί ο παρακάτω κανόνας απόφασης:

Αποφάσισε ω_1 εάν $P(\omega_1)>P(\omega_2)$, διαφορετικά αποφάσισε ω_2.

Ο κανόνας αυτός έχει νόημα εάν πρέπει να παρθεί απόφαση για ένα μόνο ψάρι. Εάν όμως πρέπει να αποφασίσουμε για πολλά ψάρια, ο κανόνας αυτός δεν θα είναι λογικός. Η απόφαση κάθε φορά θα είναι η ίδια εάν και από πριν είναι γνωστό ότι τα είδη των ψαριών που μπορεί να εμφανιστούν είναι δύο. Η απόδοση αυτού του κανόνα απόφασης εξαρτάται αποκλειστικά από τις τιμές των εκ των προτέρων πιθανοτήτων. Εάν η $P(\omega_1)$ είναι πολύ μεγαλύτερη από την $P(\omega_2)$, τότε η απόφαση υπέρ της ω_1 θα είναι σχεδόν πάντα σωστή. Εάν $P(\omega_1)=P(\omega_2)$ τότε η πιθανότητα σωστής απόφασης θα είναι 50%. Γενικά, η πιθανότητα λάθους είναι ίση με την μικρότερη πιθανότητα από τις $P(\omega_1)$ και $P(\omega_2)$ και όπως θα δούμε και παρακάτω υπό αυτές τις προϋποθέσεις

κανένας άλλος κανόνας δεν μπορεί να επιτύχει μεγαλύτερη πιθανότητα σωστής απόφασης.
Ευτυχώς, στις περισσότερες περιπτώσεις η διαθέσιμη πληροφορία είναι πολύ μεγαλύτερη. Στο παράδειγμά ταξινόμησης ψαριών, για τη βελτίωση του ταξινομητή, θα μπορούσε να χρησιμοποιηθεί η φωτεινότητα του δέρματος των ψαριών x. Διαφορετικό είδος ψαριού εμφανίζει και διαφορετική φωτεινότητα στο δέρμα. Θεωρούμε ότι το x είναι μια συνεχής τυχαία μεταβλητή της οποίας η συνάρτηση κατανομής εξαρτάται από την κατάσταση της φύσης και ισούται με $p(x/\omega)$. Αυτή είναι η υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας, δηλαδή η συνάρτηση πυκνότητας πιθανότητας για το x δεδομένου ότι η κατάσταση της φύσης είναι ω. Επομένως, η διαφορά μεταξύ της $p(x/\omega_1)$ και $p(x/\omega_2)$ περιγράφει τη διαφορά στη φωτεινότητα ανάμεσα στις πέρκες και τους σολομούς (εικόνα 2.1).



Εικόνα 2.1: Οι υποθετικές υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας δείχνουν την πυκνότητα πιθανότητας της μέτρησης μιας συγκεκριμένης τιμής ενός χαρακτηριστικού x, δεδομένου ότι το πρότυπο ανήκει στην κατηγορία ω_i. Εάν το x αντιπροσωπεύει τη φωτεινότητα ενός ψαριού, οι δύο καμπύλες περιγράφουν τη διαφορά στη φωτεινότητα μεταξύ των πληθυσμών δύο διαφορετικών τύπων ψαριών. Οι συναρτήσεις είναι κανονικοποιημένες και επομένως το εμβαδόν που βρίσκεται κάτω από κάθε καμπύλη ισούται με τη μονάδα.

Έστω ότι είναι γνωστές τόσο οι εκ των προτέρων πιθανότητες $P(\omega_j)$ όσο και οι υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_j)$, $j=1,2$. Έστω επίσης ότι μετά από τη μέτρηση της φωτεινότητας του δέρματος ενός ψαριού αυτή βρίσκεται ίση με x. Πως μπορεί να επηρεάσει η μέτρηση αυτή την απόφασή μας για το είδος στο οποίο ανήκει στην πραγματικότητα το ψάρι; Η συνάρτηση πυκνότητας πιθανότητας να βρεθεί ένα ψάρι το οποίο να ανήκει στην κατηγορία ω_j και να έχει τιμή x για το χαρακτηριστικό της φωτεινότητας ισούται με $p(\omega_j/x)=P(\omega_j/x)\cdot p(x)=p(x/\omega_j)\cdot P(\omega_j)$. Έτσι, μπορούμε να καταλήξουμε στον παρακάτω τύπο απόφασης του Bayes:

$$P(\omega_j / x) = \frac{p(x/\omega_j)\cdot P(\omega_j)}{p(x)} \quad (2.1)$$

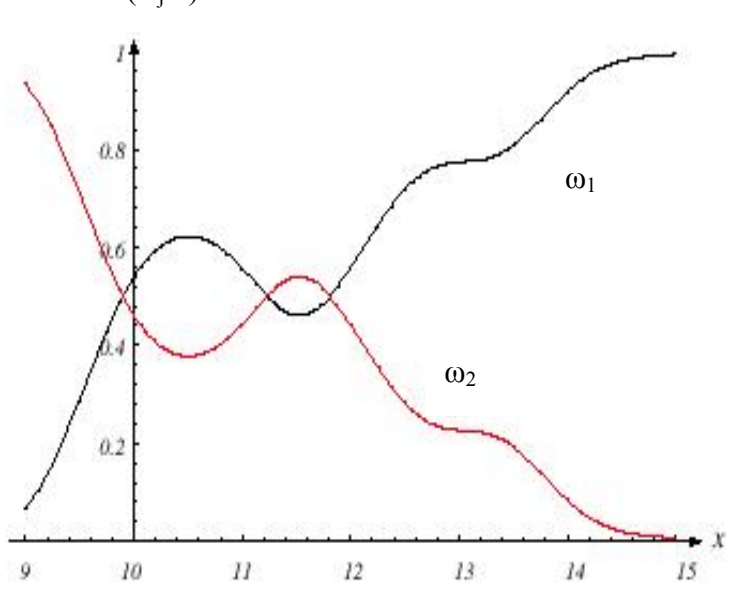
όπου στην περίπτωση των δύο κατηγοριών

$$p(x) = \sum_{j=1}^{2} p(x/\omega_j)\cdot P(\omega_j) \quad (2.2)$$

Ο τύπος απόφασης του Bayes μπορεί να περιγραφεί με λόγια ως εξής:

$$\text{εκ των υστέρων πιθανότητα} = \frac{\text{πιθανοφάνεια} \times \text{εκ των προτέρων πιθανότητα}}{\text{γεγονός}} \quad (2.3)$$

Ο τύπος απόφασης του Bayes δηλώνει ότι με την βοήθεια της παρατήρησης της τιμής του x είναι να δυνατόν να μετατραπεί η εκ των προτέρων πιθανότητα $P(\omega_j)$ στην εκ των υστέρων πιθανότητα $P(\omega_j/x)$, δηλαδή την πιθανότητα η κατάσταση της φύσης να είναι η ω_j δεδομένου ότι έχει μετρηθεί η τιμή x για το χαρακτηριστικό. Η $p(x/\omega_j)$ καλείται συνάρτηση πιθανοφάνειας της ω_j σε σχέση με το x και χρησιμοποιείται για να δηλώσει ότι, εάν όλες οι υπόλοιποι παράμετροι είναι ίσες, η κατηγορία ω_j για την οποία η $p(x/\omega_j)$ έχει μεγάλη τιμή έχει μεγαλύτερη πιθανότητα να είναι η σωστή κατηγορία. Να σημειωθεί ότι το γινόμενο της πιθανοφάνειας και της εκ των προτέρων πιθανότητας είναι αυτό που καθορίζει την τιμή της εκ των υστέρων πιθανότητας. Ο παράγοντας p(x), μπορεί να θεωρηθεί περισσότερο ως ένας παράγοντας κανονικοποίησης που εγγυάται ότι το άθροισμα των εκ των υστέρων πιθανοτήτων θα ισούται με τη μονάδα. Η διακύμανση της $P(\omega_j/x)$ σε σχέση με το x φαίνεται στην εικόνα 2.2 για την περίπτωση όπου $P(\omega_1)=2/3$ και $P(\omega_2)=1/3$.



Εικόνα 2.2: Οι εκ των υστέρων πιθανότητες για τις συγκεκριμένες τιμές των εκ των προτέρων πιθανοτήτων $P(\omega_1) = 2/3$ και $P(\omega_2) = 1/3$ για τις υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας που φαίνονται στην εικόνα 2.1. Στην παραπάνω περίπτωση, δεδομένου ότι ένα πρότυπο έχει μετρηθεί να έχει τιμή χαρακτηριστικού x = 14, η πιθανότητα να ανήκει στην κατηγορία ω_2 είναι 0.08, ενώ η πιθανότητα να ανήκει στην ω_1 είναι 0.92. Για κάθε x οι τιμές των εκ των υστέρων πιθανοτήτων έχουν άθροισμα τη μονάδα.

Εάν υπάρχει μια παρατήρηση x για την οποία η $P(\omega_1/x)$ είναι μεγαλύτερη από την $P(\omega_2/x)$, αυτόματα υπονοείται ότι η πραγματική κατάσταση της φύσης είναι η ω_1. Αντίστοιχα, εάν η $P(\omega_2/x)$ είναι μεγαλύτερη από την $P(\omega_1/x)$, θα πρέπει να επιλεχτεί η ω_2 ως η πραγματική κατάσταση της φύσης. Στη συνέχεια θα προσπαθήσουμε να αποδείξουμε την ορθότητα αυτής της διαδικασίας απόφασης υπολογίζοντας την πιθανότητα λάθους κάθε φορά που παίρνεται μία απόφαση. Οποτεδήποτε μετριέται μία συγκεκριμένη τιμή του x, η πιθανότητα λάθους ισούται με:

$$P(\text{λάθος}/x) = \begin{cases} P(\omega_1/x), & \text{εάν αποφασίζουμε } \omega_2 \\ P(\omega_2/x), & \text{εάν αποφασίζουμε } \omega_1 \end{cases} \tag{2.4}$$

Προφανώς, για μία δεδομένη τιμή του x μπορούμε να ελαχιστοποιήσουμε την πιθανότητα λάθους αποφασίζοντας ω_1 εάν $P(\omega_1/x) > P(\omega_2/x)$ και ω_2 διαφορετικά. Ο κανόνας αυτός όπως θα δούμε και στη συνέχεια πραγματικά ελαχιστοποιεί την μέση πιθανότητα λάθους. Η μέση πιθανότητα λάθους ισούται με:

$$P(\text{λάθος}) = \int_{-\infty}^{\infty} P(\text{λάθος}, x)\,dx = \int_{-\infty}^{\infty} P(\text{λάθος}/x)\cdot p(x)\,dx \tag{2.5}$$

Εάν για κάθε x εγγυηθούμε ότι η P(λάθος/x) είναι μικρότερη δυνατή, τότε το διάστημα ολοκλήρωσης θα είναι όσο το δυνατόν μικρότερο. Επομένως, αποδείχθηκε ο ακόλουθος κανόνας απόφασης του Bayes ο οποίος ελαχιστοποιεί την πιθανότητα λάθους:

Αποφάσισε ω_1 εάν $P(\omega_1/x) > P(\omega_2/x)$ διαφορετικά αποφάσισε ω_2 (2.6)

Με βάση αυτόν τον κανόνα η εξίσωση 2.4 γράφεται ως εξής:

$$P(\text{λάθος}/x) = \min[P(\omega_1/x), P(\omega_2/x)] \tag{2.7}$$

Αυτή η μορφή του κανόνα απόφασης δίνει ιδιαίτερη έμφαση στο ρόλο των εκ των υστέρων πιθανοτήτων. Χρησιμοποιώντας την εξίσωση 2.1 και απαλείφοντας τον όρο p(x), ο οποίος δεν παίζει κανένα ρόλο στην λήψη απόφασης, ο κανόνας απόφασης μπορεί να εκφραστεί με βάση τις υπό συνθήκη και εκ των προτέρων πιθανότητες:

Αποφάσισε ω_1 εάν $p(x/\omega_1)P(\omega_1) > p(x/\omega_2)P(\omega_2)$ διαφορετικά αποφάσισε ω_2 (2.8)

Παρατηρήσεις

1. Εάν για κάποια τιμή του x οι υπό συνθήκη πιθανότητες είναι ίσες ($p(x/\omega_1) = p(x/\omega_2)$), τότε η συγκεκριμένη παρατήρηση δεν παρέχει κάποια χρήσιμη πληροφορία για την πραγματική κατάσταση της φύσης. Σε αυτήν την περίπτωση η απόφαση εξαρτάται αποκλειστικά από τις εκ των προτέρων πιθανότητες.
2. Εάν οι εκ των προτέρων πιθανότητες είναι ίσες ($P(\omega_1) = P(\omega_2)$), τότε οι καταστάσεις τις φύσης είναι ισοπίθανες. Σε αυτήν την περίπτωση η απόφαση εξαρτάται αποκλειστικά από τις συναρτήσεις πιθανοφάνειας.

2.2 Θεωρία Απόφασης του Bayes – Συνεχή Χαρακτηριστικά

Στην ενότητα αυτή γίνεται μια γενίκευση των ιδεών που παρουσιάστηκαν στην προηγούμενη με βάση τα παρακάτω σημεία:

- τη χρήση περισσότερων του ενός χαρακτηριστικών γνωρισμάτων.
- την ύπαρξη περισσότερων των δύο καταστάσεων της φύσης.
- τη δυνατότητα για ενέργειες διαφορετικές από την απόφαση κατηγοριοποίησης σε κάποια κατάσταση της φύσης (π.χ. απόρριψη).

- την εισαγωγή μιας συνάρτησης κόστους η οποία είναι γενικότερη από τη συνάρτηση της πιθανότητας λάθους.

Η δυνατότητα χρήσης περισσότερων του ενός χαρακτηριστικών γνωρισμάτων απαιτεί την αντικατάσταση του βαθμωτού x από ένα διάνυσμα χαρακτηριστικών x, όπου το x ανήκει σε έναν d-διάστατο Ευκλείδειο χώρο R^d, ο οποίος καλείται χώρος χαρακτηριστικών. Από την άλλη, η ύπαρξη περισσότερων από δύο καταστάσεων της φύσης μας δίνει τη δυνατότητα να γενικεύσουμε τα συμπεράσματά μας. Η δυνατότητα για επιλογή ενέργειας διαφορετικής από την επιλογή κατάστασης της φύσης (ταξινόμηση-κατηγοριοποίηση δειγμάτων) επιτρέπει τη δυνατότητα της απόρριψης, δηλαδή τη δυνατότητα να μην γίνει κατηγοριοποίηση για κάποια δείγματα. Αυτή η επιλογή είναι αρκετά χρήσιμη στην περίπτωση που η μη κατηγοριοποίηση έχει σχετικά χαμηλό κόστος. Η συνάρτηση κόστους δηλώνει ξεκάθαρα πόσο κοστίζει η κάθε ενέργεια και χρησιμοποιείται ως επί το πλείστον για την μετατροπή του καθορισμού μιας πιθανότητας σε απόφαση για ενέργεια. Οι συναρτήσεις κόστους επιτρέπουν την επιτυχή αντιμετώπιση περιπτώσεων όπου κάποια λάθη ταξινόμησης έχουν μεγαλύτερο κόστος από κάποια άλλα.
Έστω $\{\omega_1,...,\omega_c\}$ το πεπερασμένο σύνολο c διαφορετικών καταστάσεων της φύσης («κατηγορίες») και $\alpha_1,...,\alpha_a$ το πεπερασμένο σύνολο των α πιθανών ενεργειών. Η συνάρτηση κόστους $\lambda(\alpha_i/\omega_j)$ περιγράφει το κόστος που αντιστοιχεί στην ενέργεια a_i όταν η κατάσταση της φύσης είναι η ω_j. Έστω ότι το διάνυσμα των χαρακτηριστικών γνωρισμάτων x είναι μια d-διάστατη τυχαία μεταβλητή και $p(x/\omega_j)$ είναι υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας για το x, δηλαδή η συνάρτηση πυκνότητας πιθανότητας για το x υπό τη συνθήκη ότι η ω_j είναι η πραγματική κατάσταση της φύσης. Φυσικά, με $P(\omega_j)$ παριστάνεται η εκ των προτέρων πιθανότητα ότι η κατάσταση της φύσης είναι η ω_j. Επομένως, η εκ των υστέρων πιθανότητα $P(\omega_j/x)$ μπορεί να υπολογιστεί από την $p(x/\omega_j)$ με βάση τον τύπο του Bayes:

$$P(\omega_j/x)=\frac{p(x/\omega_j)\cdot P(\omega_j)}{p(x)} \tag{2.9}$$

όπου

$$p(x)=\sum_{j=1}^{c} p(x/\omega_j)\cdot P(\omega_j) \tag{2.10}$$

Έστω ότι παρατηρείται ένα συγκεκριμένο x και λαμβάνεται η ενέργεια α_i. Εάν η πραγματική κατάσταση της φύσης είναι η ω_j εξ ορισμού θα έχουμε κόστος ίσο με $\lambda(\alpha_i/\omega_j)$. Αφού η $P(\omega_j/x)$ είναι η πιθανότητα η πραγματική κατάσταση της φύσης να είναι η ω_j, το αναμενόμενο κόστος που σχετίζεται με την ενέργεια α_i θα είναι:

$$R(\alpha_i/x)=\sum_{j=1}^{c}\lambda(\alpha_i/\omega_j)\cdot P(\omega_j/x) \tag{2.11}$$

Στην ορολογία της Θεωρίας Αποφάσεων το αναμενόμενο κόστος καλείται ρίσκο και το $R(\alpha_i/x)$ υπό συνθήκη ρίσκο. Για κάθε ένα δείγμα x το αναμενόμενο κόστος (ρίσκο) είναι το μικρότερο αν επιλεγεί η ενέργεια (α_i) εκείνη που ελαχιστοποιεί το υπό συνθήκη ρίσκο. Στις επόμενες παραγράφους θα δείξουμε πως η βέλτιστη απόδοση μπορεί να επιτευχθεί με τη χρήση του κανόνα απόφασης του Bayes.

Το πρόβλημα που πρέπει να αντιμετωπιστεί είναι η εύρεση ενός κανόνα απόφασης ως προς το $P(\omega_j)$ ο οποίος να ελαχιστοποιεί το συνολικό ρίσκο. Ένας γενικός κανόνας απόφασης είναι μία συνάρτηση α(x) η οποία δηλώνει ξεκάθαρα ποια ενέργεια πρέπει να γίνει για κάθε παρατήρηση. Πιο συγκεκριμένα, για κάθε x, η συνάρτηση απόφασης α(x) επιλέγει μία από τις α τιμές $\alpha_1,...,\alpha_\alpha$. Το συνολικό ρίσκο R είναι το αναμενόμενο κόστος που σχετίζεται με ένα δεδομένο κανόνα απόφασης. Αφού το $R(\alpha_i / x)$ είναι το υπό συνθήκη ρίσκο που σχετίζεται με την ενέργεια α_i και αφού ο κανόνας απόφασης καθορίζει την ενέργεια, το συνολικό ρίσκο δίνεται από τη σχέση:

$$R = \int R(\alpha(x)/x)\cdot p(x)dx \tag{2.12}$$

όπου το dx αναφέρεται σε d-διάστατο δείγμα χαρακτηριστικών και το διάστημα ολοκλήρωσης περιλαμβάνει ολόκληρο το χώρο των χαρακτηριστικών. Προφανώς, εάν το α(x) έχει επιλεγεί με τέτοιο τρόπο ώστε το $R(\alpha_i(x))$ να είναι όσο το δυνατόν μικρότερο για κάθε x, τότε το συνολικό ρίσκο θα ελαχιστοποιείται. Αυτό βρίσκεται σε πλήρη συμφωνία με την ακόλουθη πρόταση του κανόνα απόφασης του Bayes:
Για να ελαχιστοποιηθεί το συνολικό ρίσκο, υπολόγισε το υπό συνθήκη ρίσκο

$$R(\alpha_i / x) = \sum_{j=1}^{c} \lambda(\alpha_i / \omega_j)\cdot P(\omega_j / x) \tag{2.13}$$

για i = 1,…, α και στη συνέχεια επέλεξε την ενέργεια α_i για την οποία το $R(\alpha_i/x)$ είναι ελάχιστο. Το ελάχιστο συνολικό ρίσκο που προκύπτει καλείται ρίσκο του Bayes, συμβολίζεται με R^*, και είναι η βέλτιστη απόδοση που μπορεί να επιτευχθεί.

2.2.1 Ταξινόμηση Δύο Κατηγοριών

Στην ενότητα αυτή θα παρουσιαστεί η εφαρμογή των παραπάνω κανόνων στο πρόβλημα της ταξινόμησης σε δύο κατηγορίες. Στην περίπτωση αυτή η ενέργεια α_1 αντιστοιχεί στην απόφαση ότι η πραγματική κατάσταση της φύσης είναι η ω_1 και η ενέργεια α_2 στην απόφαση ότι είναι η ω_2. Έστω ότι $\lambda_{ij} = \lambda(a_i / \omega_j)$ είναι το κόστος που υπάρχει όταν αποφασίζουμε υπέρ της ω_i ενώ η πραγματική κατάσταση της φύσης είναι η ω_j. Χρησιμοποιώντας την εξίσωση 2.13 για το υπό συνθήκη ρίσκο έχουμε:

$$R(\alpha_1 / x) = \lambda_{11}\cdot P(\omega_1 / x) + \lambda_{12}\cdot P(\omega_2 / x) \tag{2.14}$$

$$R(\alpha_2 / x) = \lambda_{21}\cdot P(\omega_1 / x) + \lambda_{22}\cdot P(\omega_2 / x) \tag{2.15}$$

Υπάρχουν πολλοί διαφορετικοί τρόποι για τη διατύπωση του κανόνα απόφασης ελάχιστου ρίσκου, καθένας από τους οποίους έχει τα δικά του πλεονεκτήματα. Πάντως, ο βασικός κανόνας είναι να αποφασίσουμε ω_1 εάν $R(\alpha_1 / x) < R(\alpha_2 / x)$. Όσον αφορά τις εκ των υστέρων πιθανότητες, αποφασίζουμε ω_1 εάν

$$(\lambda_{21} - \lambda_{11})\cdot P(\omega_1 / x) > (\lambda_{12} - \lambda_{22})\cdot P(\omega_2 / x) \tag{2.16}$$

Προφανώς, το κόστος που υπάρχει όταν γίνεται λάθος είναι μεγαλύτερο από το κόστος που συμβαίνει όταν είμαστε σωστοί και οι παράγοντες λ_{21}-λ_{11} και λ_{12}-λ_{22} είναι θετικοί. Χρησιμοποιώντας τον τύπο του Bayes, οι εκ των υστέρων πιθανότητες μπορούν να αντικατασταθούν από τις εκ των προτέρων πιθανότητες και της υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας. Έτσι, προκύπτει ο ισοδύναμος κανόνας απόφασης: αποφάσισε ω_1 εάν

$$(\lambda_{21} - \lambda_{11})\cdot p(x / \omega_1)\cdot P(\omega_1) > (\lambda_{12} - \lambda_{22})\cdot p(x / \omega_2)\cdot P(\omega_2) \tag{2.17}$$

διαφορετικά αποφάσισε ω_2.
Μια άλλη εναλλακτική μορφή, η οποία προκύπτει από το λογικό συλλογισμό ότι $\lambda_{21} > \lambda_{11}$ είναι να αποφασίζουμε ω_1 εάν

$$\frac{p(x / \omega_1)}{p(x / \omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}\frac{P(\omega_2)}{P(\omega_1)} \tag{2.18}$$

Η τελευταία μορφή του κανόνα απόφασης επικεντρώνεται στην εξάρτηση των συναρτήσεων πυκνότητας πιθανότητας από το x. Μπορούμε να θεωρήσουμε την $p(x/\omega_j)$ ως μία συνάρτηση του ω_j (δηλαδή ως μια συνάρτηση πιθανοφάνειας) και στη συνέχεια να υπολογίσουμε το λόγο $p(x/\omega_1)/p(x/\omega_2)$. Επομένως, ο κανόνας απόφασης του Bayes μπορεί να ερμηνευθεί ως απόφαση για ω_1 εάν ο λόγος πιθανοφάνειας είναι μεγαλύτερος από μία τιμή κατωφλίου, η οποία είναι ανεξάρτητη από το διάνυσμα παρατήρησης x.

2.3 Ταξινόμηση Ελάχιστου Ρυθμού Λάθους

Στα προβλήματα ταξινόμησης, κάθε κατάσταση της φύσης συσχετίζεται συνήθως με μία από τις c διαφορετικές κατηγορίες, και κάθε ενέργεια α_i ερμηνεύεται συνήθως ως η απόφαση ότι η πραγματική κατάσταση της φύσης είναι η ω_i. Εάν εκτελεστεί η ενέργεια α_i και η πραγματική κατάσταση της φύσης είναι η ω_j, τότε η απόφαση είναι σωστή εάν $i = j$ και λάθος εάν $i \neq j$. Εάν, όπως είναι το φυσιολογικό, επιθυμούμε να αποφεύγονται τα λάθη, πρέπει να βρεθεί ένας κανόνας απόφασης ο οποίος να ελαχιστοποιεί την πιθανότητα λάθους, δηλαδή το ρυθμό λάθους.

Η συνάρτηση κόστους για αυτήν την περίπτωση είναι η, όπως καλείται και στη βιβλιογραφία, συμμετρική ή μηδέν-ένα συνάρτηση κόστους:

$$\lambda(\alpha_i/\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1,\ldots,c \tag{2.19}$$

Αυτή η συνάρτηση κόστους δεν αντιστοιχεί κανένα κόστος στις σωστές αποφάσεις ενώ αντιστοιχεί μοναδιαίο κόστος σε κάθε λανθασμένη απόφαση. Έτσι, όλα τα λάθη είναι ισοδύναμα από πλευράς κόστους. Το ρίσκο που αντιστοιχεί σε αυτή τη συνάρτηση κόστους είναι ακριβώς ίδιο με τη μέση πιθανότητα λάθους διότι το υπό συνθήκη ρίσκο ισούται με

$$\begin{aligned} R(\alpha_i/x) &= \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j)\cdot P(\omega_j/x) \\ &= \sum_{j\neq i} P(\omega_j/x) \\ &= 1 - P(\omega_i/x) \end{aligned} \tag{2.20}$$

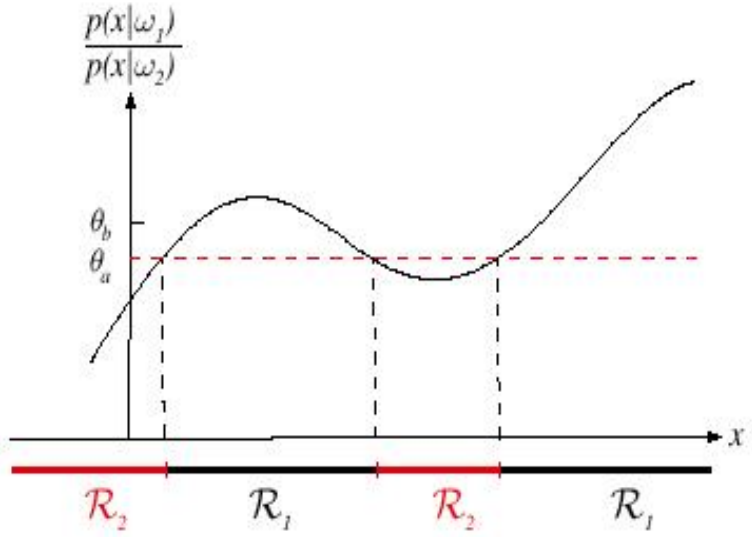
και η $P(\omega_j/x)$ είναι η υπό συνθήκη πιθανότητα ότι η ενέργεια α_i είναι σωστή. Ο κανόνας απόφασης του Bayes για την ελαχιστοποίηση του ρίσκου αναφέρεται στην επιλογή της ενέργειας που ελαχιστοποιεί το υπό συνθήκη ρίσκο. Έτσι, για να ελαχιστοποιηθεί η μέση πιθανότητα λάθους πρέπει να επιλεγεί το i το οποίο μεγιστοποιεί την εκ των υστέρων πιθανότητα $P(\omega_i/x)$. Με άλλα λόγια, ο κανόνας απόφασης για ελάχιστο ρυθμό λάθους είναι ο εξής:

Αποφάσισε ω_i εάν $P(\omega_i/x) > P(\omega_j/x)$ για κάθε $j \neq i$ (2.21)

Αυτός ο κανόνας είναι ο ίδιος με τον κανόνα της εξίσωσης 2.6. Η περιοχή απόφασης στο χώρο των δειγμάτων για την οποία αποφασίζουμε ω_i δηλώνεται ως $\mathcal{R}_i$. Μια τέτοια περιοχή προφανώς δεν απαιτείται να είναι συνεχής.

Στην εικόνα 2.2 παρουσιάστηκαν κάποιες συναρτήσεις πυκνότητας πιθανότητας και οι αντίστοιχες εκ των υστέρων πιθανότητες. Στην εικόνα 2.3 παρουσιάζεται ο λόγος πιθανοφάνειας $p(x/\omega_1) / p(x/\omega_2)$ για την ίδια περίπτωση, ο οποίος γενικά μπορεί να κυμαίνεται από το μηδέν έως το άπειρο. Όταν χρησιμοποιείται η μηδέν-ένα συνάρτηση κόστους, τα όρια απόφασης καθορίζονται από τιμή κατωφλίου ίση με θ_α. Παρατηρείται ότι αυτό οδηγεί στα ίδια όρια απόφασης με την εικόνα 2.2. Εάν το λάθος της ταξινόμησης στην κλάση ω_1, δειγμάτων που ανήκουν στην κλάση ω_2,

τιμωρείται περισσότερο από το αντίστροφο, δηλαδή εάν $\lambda_{12} > \lambda_{21}$), τότε η εξίσωση 2.18 οδηγεί στο κατώφλι θ_b. Σημειώνεται ότι το εύρος των τιμών του x για τις οποίες ένα πρότυπο ταξινομείται στην κλάση ω_1 μικραίνει.



Εικόνα 2.3: Ο λόγος πιθανοφάνειας $p(x/\omega_1)$ / $p(x/\omega_2)$ για τις κατανομές που παρουσιάστηκαν στο σχήμα 2.2. Εάν χρησιμοποιηθεί μία μηδέν-ένα συνάρτηση κόστους, τα όρια απόφασης καθορίζονται από το κατώφλι θ_a. Εάν η συνάρτηση κόστους τιμωρεί τη λάθος κατηγοριοποίηση στην ω_1 δειγμάτων που ανήκουν στην ω_2 περισσότερο από το αντίστροφο, προκύπτει το κατώφλι θ_b, και ως αποτέλεσμα η περιοχή απόφασης R_1 γίνεται μικρότερη.

2.3.1 Το Κριτήριο Minimax

Πολλές φορές πρέπει να σχεδιαστεί ένας καινούριος ταξινομητής για να έχουμε καλή απόδοση σε περιπτώσεις που οι τιμές για τις εκ των προτέρων πιθανότητες αλλάζουν ή είναι άγνωστες. Για παράδειγμα, στο πρόβλημα ταξινόμησης ψαριών που παρουσιάστηκε στο προηγούμενο κεφάλαιο, αν και οι φυσικές ιδιότητες της φωτεινότητας και του πλάτους κάθε ψαριού παραμένουν σταθερές, οι τιμές των εκ των προτέρων πιθανοτήτων μπορεί να αλλάζουν σημαντικά και με απρόβλεπτο τρόπο. Επίσης, μπορεί να θέλουμε να χρησιμοποιήσουμε τον ταξινομητή σε ένα άλλο εργοστάσιο όπου δε γνωρίζουμε τις τιμές των εκ των προτέρων πιθανοτήτων. Μια λογική προσέγγιση στο πρόβλημα αυτό είναι να σχεδιαστεί ο ταξινομητής με τέτοιο τρόπο ώστε το μέγιστο συνολικό ρίσκο για οποιαδήποτε τιμή των εκ των προτέρων πιθανοτήτων να είναι όσο το δυνατόν μικρότερο, δηλαδή να ελαχιστοποιηθεί το μέγιστο πιθανό συνολικό ρίσκο.

Έστω ότι με $\mathcal{R}_1$ συμβολίζεται η (προς το παρόν άγνωστη) περιοχή στο χώρο των χαρακτηριστικών για την οποία ο ταξινομητής αποφασίζει ω_1 και αντίστοιχα με $\mathcal{R}_2$ συμβολίζεται η περιοχή για την οποία αποφασίζει ω_2. Το συνολικό ρίσκο (εξίσωση 2.12) μπορεί να γραφεί με χρήση των υπό συνθήκη ρίσκων ως εξής:

$$R = \int_{\mathcal{R}_1} [\lambda_{11} P(\omega_1) p(x/\omega_1) + \lambda_{12} P(\omega_2) p(x/\omega_2)] dx + \int_{\mathcal{R}_2} [\lambda_{21} P(\omega_1) p(x/\omega_1) + \lambda_{22} P(\omega_2) p(x/\omega_2)] dx \quad (2.22)$$

Χρησιμοποιώντας τις σχέσεις $P(\omega_2) = 1 - P(\omega_1)$ και $\int_{\Re_1} p(x/\omega_1)dx = 1 - \int_{\Re_2} p(x/\omega_1)dx$ η εξίσωση του ρίσκου μπορεί να γραφεί ως εξής:

R_{mm}, minimax ρίσκο

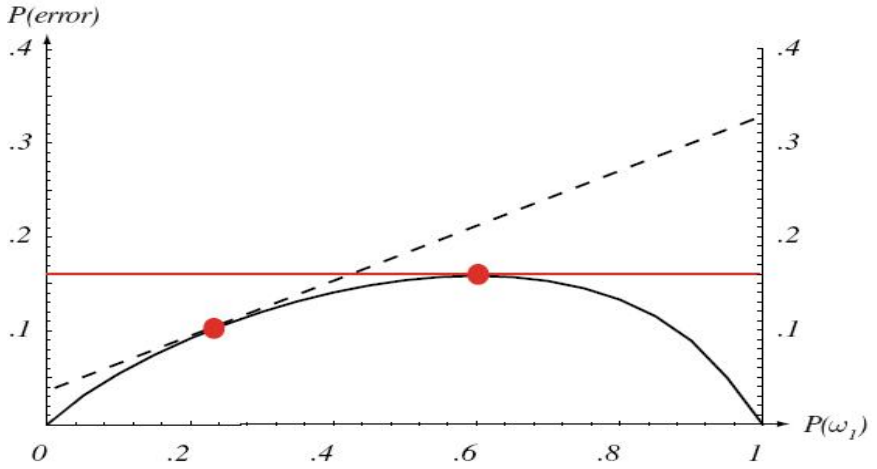
$$R(P(\omega_1)) = \overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22})\int_{\Re_1} p(x/\omega_2)dx}$$
$$+ P(\omega_1)\underbrace{\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11})\int_{\Re_2} p(x/\omega_1)dx - (\lambda_{12} - \lambda_{22})\int_{\Re_1} p(x/\omega_2)dx\right]}_{} \quad (2.23)$$

= 0 για τη minimax λύση

Η παραπάνω εξίσωση δείχνει ότι από τη στιγμή που αποφασιστεί το όριο απόφασης (καθοριστούν οι περιοχές απόφασης $\Re_1$ και $\Re_2$) το συνολικό ρίσκο είναι ανάλογο με την $P(\omega_1)$. Εάν μπορεί να βρεθεί ένα όριο απόφασης τέτοιο ώστε ο συντελεστής της αναλογικότητας να ισούται με το μηδέν, το ρίσκο θα είναι ανεξάρτητο από τις εκ των προτέρων πιθανότητες. Αυτή είναι η minimax λύση, ενώ η τιμή του minimax ρίσκου, R_{mm}, μπορεί να υπολογιστεί από την εξίσωση 2.23:

$$R_{mm} = \lambda_{22} + (\lambda_{12} - \lambda_{22})\int_{\Re_1} p(x/\omega_2)dx = \lambda_{11} + (\lambda_{21} - \lambda_{11})\int_{\Re_2} p(x/\omega_1)dx \quad (2.24)$$

Στην εικόνα 2.4 παρουσιάζεται γραφικά η προσέγγιση. Αρχικά, ψάχνουμε για την εκ των προτέρων πιθανότητα για την οποία το ρίσκο του Bayes είναι μέγιστο. Το αντίστοιχο όριο απόφασης παρέχει τη minimax λύση. Η τιμή του minimax ρίσκου, R_{mm}, είναι στην περίπτωση αυτή ίση με το χειρότερο Bayes ρίσκο. Πρακτικά, η εύρεση του ορίου απόφασης για το minimax ρίσκο είναι συνήθως αρκετά δύσκολη, ιδιαίτερα σε περιπτώσεις που οι κατανομές είναι πολύπλοκες. Παρ' όλα αυτά, σε κάποιες περιπτώσεις το όριο μπορεί να βρεθεί αναλυτικά.
Το minimax κριτήριο βρίσκει μεγαλύτερη χρήση στη Θεωρία Παιγνίων από την παραδοσιακή Αναγνώριση Προτύπων. Στη Θεωρία Παιγνίων υπάρχει συνήθως ένας αντίπαλος ο οποίος αναμένεται να επιλέξει μια ενέργεια μέγιστα εχθρική προς εμάς. Έτσι, έχει μεγάλη σημασία να μπορούμε να επιλέγουμε μία ενέργεια (π.χ. μια ταξινόμηση) η οποία να ελαχιστοποιεί το κόστος σε σχέση με τις μέγιστα εχθρικές ενέργειες του αντιπάλου μας.



Εικόνα 2.4: Η καμπύλη στο κάτω μέρος δείχνει το ελάχιστο λάθος κατά Bayes ως συνάρτηση της εκ των προτέρων πιθανότητας $P(\omega_1)$ σε ένα πρόβλημα ταξινόμησης δύο κατηγοριών με καθορισμένες κατανομές. Για κάθε τιμή των εκ των προτέρων πιθανοτήτων (π.χ. $P(\omega_1) = 0.25$) υπάρχει ένα αντίστοιχο βέλτιστο όριο απόφασης και ένας αντίστοιχος ρυθμός λάθους κατά Bayes. Για κάθε (καθορισμένο) τέτοιο όριο, εάν αλλάξουν οι τιμές των εκ των προτέρων πιθανοτήτων, η πιθανότητα του λάθους θα μεταβληθεί ως μία γραμμική συνάρτηση του $P(\omega_1)$, όπως φαίνεται από τη διακεκομμένη γραμμή. Η μέγιστη τιμή ενός τέτοιου λάθους θα συμβεί σε μία ακραία τιμή της εκ των προτέρων πιθανότητας και πιο συγκεκριμένα για $P(\omega_1) = 1$. Για να ελαχιστοποιηθεί το μέγιστο ενός τέτοιου λάθους, πρέπει να σχεδιαστεί το όριο απόφασης για το μέγιστο λάθος κατά Bayes ($P(\omega_1) = 0.6$) και επομένως το λάθος να μη μεταβάλλεται ως συνάρτηση της τιμής της εκ των προτέρων πιθανότητας (όπως φαίνεται από την οριζόντια γραμμή).

2.3.2 Το Κριτήριο Neyman-Pearson

Σε μερικά προβλήματα, ο στόχος είναι να ελαχιστοποιηθεί το συνολικό ρίσκο σε σχέση με ένα συγκεκριμένο περιορισμό. Για παράδειγμα, μπορεί να θέλαμε να ελαχιστοποιήσουμε το συνολικό ρίσκο σε σχέση με τον περιορισμό $\int R(\alpha_i / x)dx < \varepsilon$, όπου ε = σταθερά, για κάποιο συγκεκριμένο i. Ένας τέτοιος περιορισμός μπορεί να εμφανιστεί όταν υπάρχει μια καθορισμένη πηγή που συνοδεύει μια συγκεκριμένη ενέργεια α_i ή όταν δεν πρέπει να ταξινομούμε λανθασμένα ένα πρότυπο που ανήκει σε κάποια συγκεκριμένη κατάσταση της φύσης ω_i με συχνότητα μεγαλύτερη από κάποια συγκεκριμένη τιμή. Για παράδειγμα, στην περίπτωση της ταξινόμησης ψαριών που αναφέραμε στο προηγούμενο κεφάλαιο, θα μπορούσε να υπάρχει ένας κανονισμός ο οποίος να απαγορεύει τη λανθασμένη κατηγοριοποίηση περισσοτέρων από 1% σολομών ως πέρκες. Στην περίπτωση αυτή θα ψάχναμε να βρούμε μία απόφαση η οποία ελαχιστοποιεί την πιθανότητα της ταξινόμησης μιας πέρκας ως σολομό με βάση αυτή τη συνθήκη. Γενικά, ένα τέτοιο Neyman-Pearson κριτήριο ικανοποιείται με την αριθμητική προσαρμογή των ορίων απόφασης. Φυσικά, για τη Gaussian και μερικές άλλες κατανομές, οι Neyman-Pearson λύσεις μπορούν να βρεθούν αναλυτικά.

2.4 Ταξινομητές, Διακρίνουσες Συναρτήσεις και Επιφάνειες Απόφασης

2.4.1 Η Περίπτωση Πολλών Κατηγοριών

Υπάρχουν πολλοί διαφορετικοί τρόποι για την αναπαράσταση ταξινομητών προτύπων. Ένας από τους πιο χρήσιμους είναι η χρήση ενός συνόλου από διακρίνουσες συναρτήσεις $g_i(x)$, $i = 1,\ldots, c$. Ο ταξινομητής αναθέτει ένα διάνυσμα χαρακτηριστικών x στην κατηγορία ω_i εάν

$$g_i(x) > g_j(x) \text{ για όλα τα } j \neq i \tag{2.25}$$

Έτσι, ο ταξινομητής αντιμετωπίζεται ως ένα δίκτυο (ή μία μηχανή) το οποίο υπολογίζει τις τιμές για c διαφορετικές διακρίνουσες συναρτήσεις και επιλέγει την κατηγορία που αντιστοιχεί στη συνάρτηση που έχει τη μεγαλύτερη τιμή. Η αναπαράσταση ενός ταξινομητή με τη μορφή δικτύου φαίνεται στην εικόνα 2.5.

Ένας ταξινομητής Bayes μπορεί με εύκολο και φυσικό τρόπο να αναπαρασταθεί με αυτόν τον τρόπο. Για την γενική περίπτωση, όπου υπολογίζεται το ρίσκο, θέτουμε $g_i(x) = -R(a_i / x)$, επειδή η διακρίνουσα συνάρτηση με τη μέγιστη τιμή αντιστοιχεί στο ελάχιστο υπό συνθήκη ρίσκο. Για την περίπτωση του ελάχιστου ρυθμού λάθους, τα πράγματα μπορούν να απλοποιηθούν ακόμα περισσότερο θέτοντας $g_i(x) = P(\omega_i / x)$. Στην περίπτωση αυτή η διακρίνουσα συνάρτηση με τη μέγιστη τιμή αντιστοιχεί στη μέγιστη εκ των υστέρων πιθανότητα.

Προφανώς, η επιλογή διακρινουσών συναρτήσεων δεν είναι μοναδική. Εάν κάθε διακρίνουσα συνάρτηση $g_i(x)$ αντικατασταθεί από την $f(g_i(x))$, όπου η $f(\cdot)$ είναι μια μονότονη αύξουσα συνάρτηση, η ταξινόμηση θα παραμείνει ανεπηρέαστη. Η παρατήρηση αυτή οδηγεί συνήθως σε σημαντικές αναλυτικές και υπολογιστικές απλοποιήσεις. Πιο συγκεκριμένα, για την περίπτωση του ρυθμού ελάχιστου λάθους, όλες οι παρακάτω επιλογές έχουν τα ίδια αποτελέσματα ταξινόμησης, αν και κάποιες από αυτές, ανάλογα με την περίπτωση, είναι απλούστερες και πιο εύκολα υπολογίσιμες:

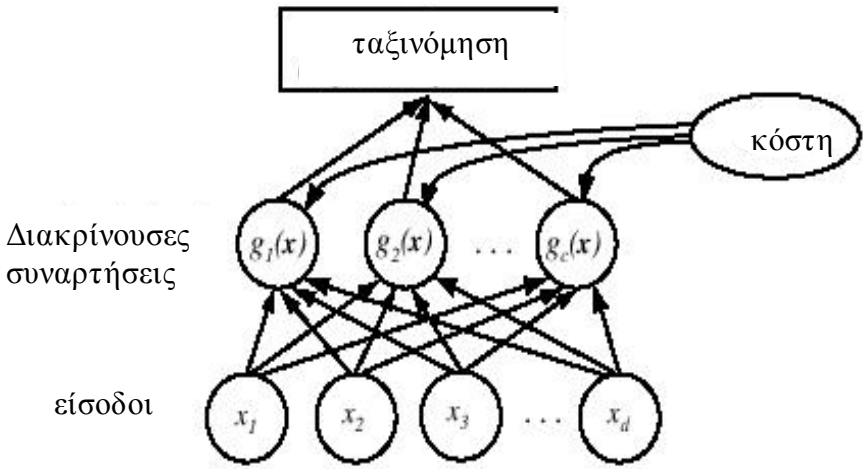
$$g_i(x) = P(\omega_i / x) = \frac{p(x / \omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(x / \omega_j)P(\omega_j)} \tag{2.26}$$

$$g_i(x) = p(x / \omega_i)P(\omega_i) \tag{2.27}$$

$$g_i(x) = \ln p(x / \omega_i) + \ln P(\omega_i) \tag{2.28}$$

όπου το ln υποδηλώνει το φυσικό λογάριθμο.

Εάν και οι διακρίνουσες συναρτήσεις μπορούν να γραφούν με πολλούς διαφορετικούς τρόπους, οι κανόνες απόφασης είναι ισοδύναμοι. Το αποτέλεσμα της εφαρμογής όλων των κανόνων απόφασης είναι η διαίρεση του χώρου χαρακτηριστικών σε c περιοχές απόφασης, $R_1, \ldots, R_c$. Εάν $g_i(x) > g_j(x)$ για κάθε $i \neq j$, τότε το x ανήκει στην περιοχή απόφασης R_i και ο κανόνας απόφασης μας οδηγεί να ταξινομήσουμε το x στην κατηγορία ω_i. Οι περιοχές διαχωρίζονται από όρια απόφασης, επιφάνειες δηλαδή στο χώρο των χαρακτηριστικών όπου οι επικρατούσες διακρίνουσες συναρτήσεις έχουν την ίδια τιμή (εικόνα 2.6).



Εικόνα 2.5: Η δομή ενός γενικού στατιστικού ταξινομητή προτύπων που περιλαμβάνει d εισόδους και c διακρίνουσες συναρτήσεις $g_i(x)$. Το επόμενο βήμα καθορίζει ποια από τις διακρίνουσες συναρτήσεις έχει μέγιστη τιμή και κατηγοριοποιεί αντίστοιχα το πρότυπο εισόδου. Τα βέλη δείχνουν την κατεύθυνση της ροής της πληροφορίας, εάν και συνήθως παραλείπονται σε περιπτώσεις όπου η κατεύθυνση της ροής της πληροφορίας είναι προφανής.

2.4.2 Η Περίπτωση Δύο Κατηγοριών

Ο ταξινομητής που τοποθετεί ένα δείγμα σε μία από δύο πιθανές κατηγορίες καλείται διχοτόμος. Αντί να χρησιμοποιηθούν δύο διακρίνουσες συναρτήσεις g_1 και g_2 και να ταξινομείται το x στην ω_1 εάν $g_1 > g_2$ συνήθως ορίζεται μία μόνο διακρίνουσα συνάρτηση

$$g(x) \equiv g_1(x) - g_2(x) \tag{2.29}$$

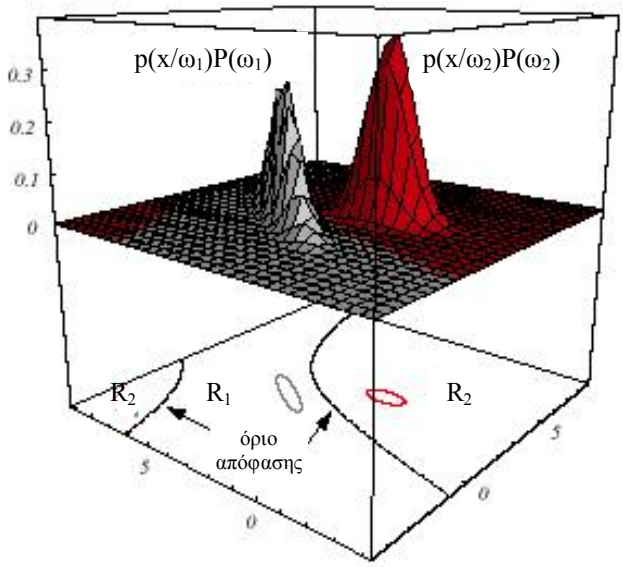
και χρησιμοποιείται ο ακόλουθος κανόνας απόφασης:

Αποφάσισε ω_1 εάν $g(x) > 0$, διαφορετικά αποφάσισε ω_2.

Έτσι, ένας διχοτόμος μπορεί να θεωρηθεί ως μία μηχανή που υπολογίζει μία διακρίνουσα συνάρτηση g(x) και ταξινομεί το x σύμφωνα με το αλγεβρικό πρόσημο του αποτελέσματος. Από τις διάφορες μορφές με τις οποίες μπορούν να γραφούν οι διακρίνουσες συναρτήσεις ελάχιστου ρυθμού λάθους, οι ακόλουθες δύο είναι ιδιαίτερα χρήσιμες:

$$g(x) = P(\omega_1 / x) - P(\omega_2 / x) \tag{2.30}$$

$$g(x) = \ln \frac{p(x/\omega_1)}{p(x/\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \tag{2.31}$$



Εικόνα 2.6: Σε αυτόν το δυσδιάστατο δύο κατηγοριών ταξινομητή οι συναρτήσεις πυκνότητας πιθανότητας είναι Gaussian, το όριο απόφασης αποτελείται από δύο υπερβολές και επομένως η περιοχή απόφασης R_2 δεν είναι απλώς συνεκτική. Οι ελλείψεις αντιστοιχούν στα σημεία όπου η συνάρτηση πυκνότητας πιθανότητας έχει τιμή ίση με 1 / e φορές την τιμή της κορυφής της κατανομής.

2.5 Η κανονική συνάρτηση πυκνότητας πιθανότητας

Η δομή ενός ταξινομητή Bayes καθορίζεται από τις υπό συνθήκη συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_i)$ όπως επίσης και από τις εκ των προτέρων πιθανότητες $P(\omega_i)$. Από τις διάφορες συναρτήσεις πυκνότητας πιθανότητας που έχουν ερευνηθεί, το μεγαλύτερο ενδιαφέρον παρουσιάζεται για την πολυδιάστατη κανονική ή Gaussian συνάρτηση πυκνότητας πιθανότητας. Στην ενότητα αυτή γίνεται μια σύντομη παρουσίαση της πολυδιάστατης κανονικής συνάρτησης πυκνότητας πιθανότητας, τονίζοντας τις ιδιότητές της οι οποίες παρουσιάζουν μεγαλύτερο ενδιαφέρον για προβλήματα ταξινόμησης.

Αρχικά, ας θυμηθούμε τον ορισμό της αναμενόμενης τιμής μιας βαθμωτής συνάρτησης f(x), που ορίζεται για κάποια συνάρτηση πυκνότητας πιθανότητας p(x):

$$E[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)dx \tag{2.32}$$

Εάν οι τιμές του χαρακτηριστικού x περιορίζονται σε σημεία που ανήκουν σε ένα διακριτό σύνολο D, πρέπει να αθροίσουμε πάνω σε όλα τα δείγματα ως εξής:

$$E[f(x)] = \sum_{x \in D} f(x)P(x) \tag{2.33}$$

Όπου η P(x) είναι η πυκνότητα πιθανότητας στο x.

2.5.1 Συνάρτηση Πυκνότητας Πιθανότητας Μιας Μεταβλητής.

Ο τύπος που δίνει τη συνεχή κανονική (ή Gaussian) συνάρτηση πυκνότητας πιθανότητας μιας μεταβλητής είναι ο εξής:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right] \tag{2.34}$$

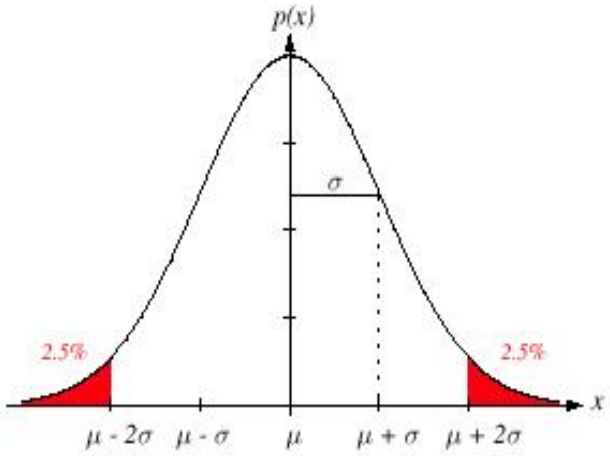
για την οποία η αναμενόμενη τιμή του x (ένας μέσος όρος πάνω σε όλο το χώρο των χαρακτηριστικών) ισούται με:

$$\mu \equiv E[x] = \int_{-\infty}^{\infty} x p(x) dx \tag{2.35}$$

και όπου η αναμενόμενη τετραγωνική απόκλιση (διασπορά) δίνεται από τον τύπο:

$$\sigma^2 \equiv E\left[(x-\mu)^2 \right] = \int_{-\infty}^{\infty} (x-\mu)^2 \, p(x) dx \tag{2.36}$$

Η κανονική συνάρτηση πυκνότητας πιθανότητας μιας μεταβλητής καθορίζεται πλήρως από δύο παραμέτρους: τη μέση τιμή της μ και την διασπορά σ^2. Στη βιβλιογραφία χρησιμοποιείται συνήθως η συντομογραφία $p(x) \sim N(\mu, \sigma^2)$ για να δηλώσει ότι η μεταβλητή x έχει κανονική κατανομή με μέση τιμή μ και διασπορά σ^2. Οι μεταβλητές που έχουν κανονικές κατανομές τείνουν να μαζεύονται γύρω από τη μέση τιμή τους, σε διάστημα που εξαρτάται από την τιμή της τυπικής τους απόκλισης σ (εικόνα 2.7).



Εικόνα 2.7: Μία κανονική κατανομή μιας μεταβλητής έχει το 95% των τιμών της στο διάστημα $|x - \mu| \le 2\sigma$. Η κορυφή της κατανομής έχει τιμή $p(\mu) = 1/\sqrt{2\pi\sigma}$.

2.5.2 Συνάρτηση Πυκνότητας Πιθανότητας Πολλών Μεταβλητών

Η γενικής μορφής συνάρτηση πυκνότητας πιθανότητας πολλών μεταβλητών σε d διαστάσεις δίνεται από τον τύπο:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (x-\mu)^t \Sigma^{-1} (x-\mu) \right] \tag{2.37}$$

όπου το x είναι ένα d-διάστατο (με d στοιχεία) διάνυσμα στήλη, το μ είναι το d-διάστατο διάνυσμα της μέσης τιμής, το Σ είναι ο d × d πίνακας συνδιασποράς, και $|\Sigma|$, Σ^{-1} είναι αντίστοιχα η ορίζουσα και ο αντίστροφος πίνακας του Σ. Επιπλέον, έστω ότι με $(x - \mu)^t$ συμβολίζεται το ανάστροφο του $x - \mu$. Ο συμβολισμός που θα χρησιμοποιηθεί για το εσωτερικό γινόμενο είναι ο εξής:

$$a^t b = \sum_{i=1}^{d} a_i b_i \tag{2.38}$$

Για ευκολία, η σχέση 2.37 γράφεται συνήθως ως $p(x) \sim N(\mu, \Sigma)$.
Γενικά, ισχύουν τα εξής:

$$\mu \equiv E[x] = \int x\, p(x)\, dx \tag{2.39}$$

και

$$\Sigma \equiv E\left[(x-\mu)(x-\mu)^t\right] = \int (x-\mu)(x-\mu)^t\, p(x) dx \tag{2.40}$$

όπου η αναμενόμενη τιμή ενός διανύσματος ή ενός πίνακα υπολογίζεται από τις αναμενόμενες τιμές των στοιχείων του. Με άλλα λόγια, εάν το x_i είναι το i-οστό στοιχείο του x, το μ_i θα είναι το i-οστό στοιχείο του μ και το σ_{ij} θα είναι ij-οστό στοιχείο του Σ. Επομένως:

$$\mu_i = E[x_i] \tag{2.41}$$

και

$$\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \tag{2.42}$$

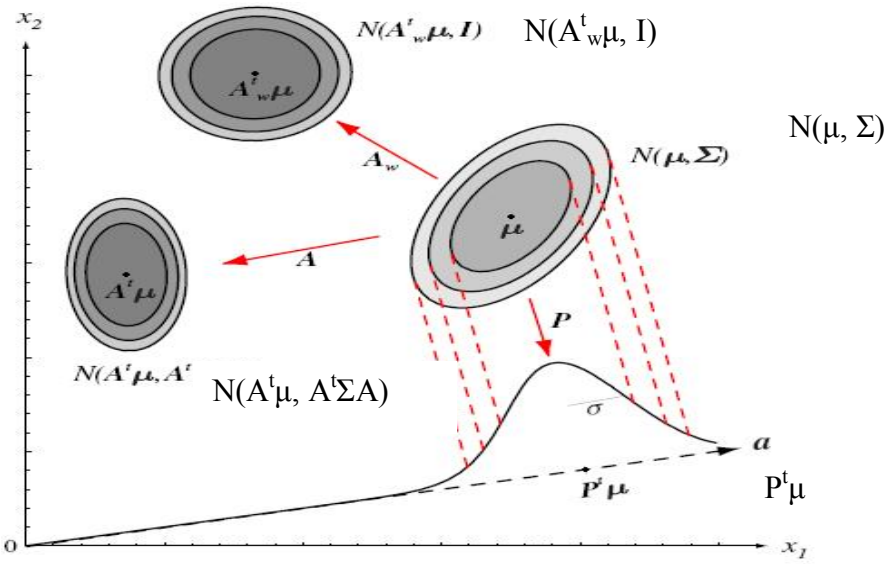
Ο πίνακας συνδιασποράς Σ είναι πάντοτε συμμετρικός και θετικά ημι-ορισμένος. Η ανάλυσή μας θα περιοριστεί στην περίπτωση όπου ο Σ είναι θετικά ορισμένος έτσι ώστε η ορίζουσα να είναι αυστηρά θετική. Τα διαγώνια στοιχεία σ_{ij} είναι οι διασπορές των αντίστοιχων στοιχείων x_i (δηλαδή τα σ_i^2) και τα μη διαγώνια στοιχεία σ_{ij} είναι οι συνδιασπορές των στοιχείων x_i και x_j. Εάν τα x_i και x_j είναι στατιστικά ανεξάρτητα, τότε $\sigma_{ij} = 0$. Εάν όλα τα μη διαγώνια στοιχεία είναι ίσα με το μηδέν, το p(x) ανάγεται στο γινόμενο των συναρτήσεων πυκνότητας πιθανότητας μιας μεταβλητής για κάθε στοιχείο του x.

Γραμμικοί συνδυασμοί μεταξύ τυχαίων μεταβλητών με κανονικές κατανομές, ανεξάρτητες μεταξύ τους ή μη, έχουν επίσης κανονικές κατανομές. Πιο συγκεκριμένα, εάν $p(x) \sim N(\mu, \Sigma)$, ο A είναι ένας $d \times k$ πίνακας και το $y = A^t x$ είναι ένα διάνυσμα k στοιχείων, τότε $p(y) \sim N(A^t\mu, A^t\Sigma A)$, όπως φαίνεται και από την εικόνα 2.8. Στην ειδική περίπτωση όπου το k ισούται με 1 και ο A είναι ένα μοναδιαίου μήκους διάνυσμα a, το $y = a^t x$ είναι ένας αριθμός που αντιπροσωπεύει την προβολή του x πάνω σε μια γραμμή στη διεύθυνση του a. Γενικότερα, η γνώση της μορφής του πίνακα συνδιασποράς επιτρέπει τον υπολογισμό της διασποράς των δεδομένων προς κάθε κατεύθυνση (ή προς κάθε υποχώρο).

Είναι πολλές φορές χρήσιμο να χρησιμοποιηθεί ένας μετασχηματισμός ο οποίος μετατρέπει μια αυθαίρετη κανονική κατανομή πολλών μεταβλητών σε μία σφαιρική, δηλαδή σε μία κατανομή που έχει πίνακα συνδιασποράς ανάλογο με τον ταυτοτικό πίνακα I. Εάν ορίσουμε με Φ τον πίνακα του οποίου οι στήλες είναι ίσες με τα ορθοκανονικά ιδιοδιανύσματα του Σ και με Λ τον διαγώνιο πίνακα των αντίστοιχων ιδιοτιμών, τότε η εφαρμογή του μετασχηματισμού

$$A_w = \Phi \Lambda^{-1/2} \tag{2.43}$$

στις συντεταγμένες εγγυάται ότι η μετασχηματισμένη κατανομή έχει πίνακα συνδιασποράς ίσο με τον ταυτοτικό πίνακα. Στην ορολογία της επεξεργασίας σημάτων, ο A_w παράγει έναν «λευκό» μετασχηματισμό επειδή κάνει το φάσμα των ιδιοτιμών της μετασχηματισμένης κατανομής ομοιόμορφο.



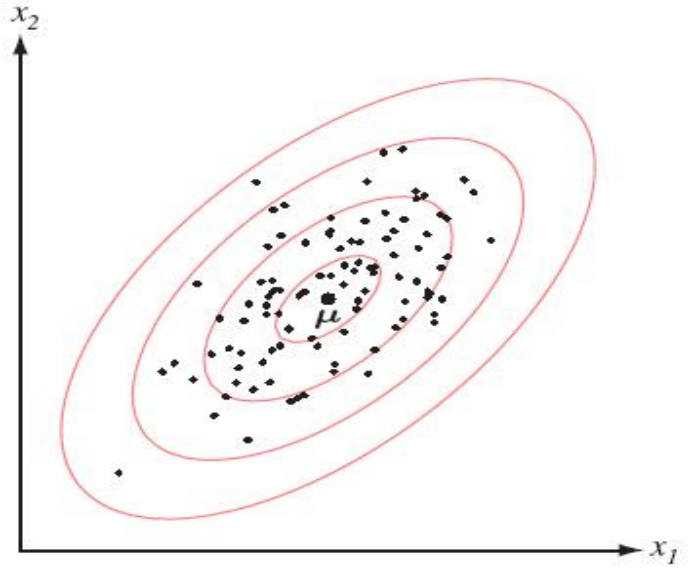
Εικόνα 2.8: Ένας γραμμικός μετασχηματισμός του χώρου των χαρακτηριστικών μετατρέπει μία τυχαία κανονική κατανομή σε μία άλλη κανονική κατανομή. Ο μετασχηματισμός , A, μετατρέπει την αρχική κατανομή στην $N(A^t\mu, A^t\Sigma A)$. Ένας άλλος – η προβολή P πάνω στη γραμμή που ορίζεται από το διάνυσμα a – οδηγεί στην $N(\mu, \sigma^2)$. Εάν και οι μετασχηματισμοί δίνουν κατανομές σε διαφορετικούς χώρους, στο σχήμα παρουσιάζονται στον αρχικό x_1x_2-χώρο. Ένας «λευκός» μετασχηματισμός, A_w, οδηγεί σε μία κυκλική συμμετρική Gaussian, η οποία στο σχήμα φαίνεται μετατοπισμένη.

Η συνάρτηση πυκνότητας πιθανότητας πολλών μεταβλητών καθορίζεται πλήρως από $d+d(d+1)/2$ παραμέτρους, δηλαδή από τα στοιχεία του διανύσματος των μέσων τιμών μ και από τα ανεξάρτητα στοιχεία του πίνακα συνδιασποράς Σ. Τα δείγματα που προέρχονται από έναν πληθυσμό κανονικής κατανομής τείνουν να περιοριστούν σε ένα σύννεφο ή τομέα (εικόνα 2.9). Το κέντρο του τομέα καθορίζεται από το διάνυσμα των μέσων τιμών, ενώ το σχήμα του καθορίζεται από τον πίνακα συνδιασποράς. Από την εξίσωση 2.38 προκύπτει ότι οι περιοχές των σημείων σταθερής πυκνότητας πιθανότητας είναι υπερελλείψεις για τις οποίες η δευτέρου βαθμού μορφή $(x-\mu)^t\Sigma^{-1}(x-\mu)$ είναι σταθερή. Οι κύριοι άξονες αυτών των υπερελλείψεων καθορίζονται από τα ιδιοδιανύσματα του πίνακα Σ (δηλαδή, τις στήλες του πίνακα Φ). Οι ιδιοτιμές (δηλαδή τα στοιχεία της διαγωνίου του πίνακα Λ) καθορίζουν τα μήκη αυτών των αξόνων. Η ποσότητα

$$r^2 = (x-\mu)^t\Sigma^{-1}(x-\mu) \qquad (2.44)$$

καλείται απόσταση Mahalanobis του x από το μ. Έτσι, οι καμπύλες σταθερής πυκνότητας πιθανότητας είναι υπερελλείψεις με σταθερή απόσταση Mahalanobis από το μ των οποίων το μέγεθος φανερώνει τη διασπορά των δειγμάτων γύρω από τη μέση τιμή τους. Το μέγεθος των υπερελλείψεων σε σχέση με την απόσταση Mahalanobis r δίνεται από τον τύπο:

$$V = V_d|\Sigma|^{1/2}r^d \qquad (2.45)$$



Εικόνα 2.9: Τα δείγματα που προέρχονται από μία δυσδιάστατη Gaussian κατανομή περιορίζονται σε ένα σύννεφο γύρω από το διάνυσμα μέσης τιμής μ. Οι ελλείψεις δείχνουν τις γραμμές (περιοχές) που έχουν ίση τιμή στη συνάρτηση πυκνότητας πιθανότητας της Gaussian κατανομής.

όπου V_d είναι το μέγεθος ενός μιας d-διάστατης μοναδιαίας υπερσφαίρας:

$$V_d = \begin{cases} \pi^{d/2} / (d/2)! & d \text{ άρτιο} \\ 2^d \pi^{(d-1)/2} \left(\frac{d-1}{2} \right)! / d! & d \text{ περιττό} \end{cases} \quad (2.46)$$

Έτσι, για δεδομένη διάσταση, η διασπορά των δειγμάτων εξαρτάται αποκλειστικά από την ποσότητα $|\Sigma|^{1/2}$.

2.6 Διακρίνουσες Συναρτήσεις για την Κανονική Συνάρτηση Πυκνότητας Πιθανότητας

Στην ενότητα 2.4.1 αποδείχθηκε ότι ταξινόμηση ελάχιστου ρυθμού λάθους μπορεί να επιτευχθεί με χρήση των ακόλουθων διακρινουσών συναρτήσεων:

$$g_i(x) = \ln p(x/\omega_i) + \ln P(\omega_i) \quad (2.47)$$

Εάν οι συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_i)$ είναι κανονικής κατανομής συναρτήσεις πολλών μεταβλητών, δηλαδή $p(x/\omega_i) \sim N(\mu_i, \Sigma_i$, από την εξίσωση 2.37 προκύπτει το εξής:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) \quad (2.48)$$

Στη συνέχεια θα εξεταστεί η μορφή της παραπάνω διακρίνουσας συνάρτησης και της αντίστοιχης διαδικασίας ταξινόμησης για κάποιες χαρακτηριστικές περιπτώσεις.

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

1. Ο πιο απλή περίπτωση είναι αυτή στην οποία όλα τα δείγματα είναι στατιστικά ανεξάρτητα μεταξύ τους και έχουν την ίδια διασπορά, σ^2. Ο πίνακας συνδιασποράς είναι διαγώνιος και ίσος με $\sigma^2 I$ (όπου I είναι ο ταυτοτικός πίνακας). Από γεωμετρική πλευρά, στην περίπτωση αυτή τα δείγματα κατανέμονται σε ισομεγέθεις υπερσφαιρικές ομάδες, όπου η ομάδα της i-οστής κατηγορίας βρίσκεται γύρω από το διάνυσμα μέσων τιμών μ_i. Ο υπολογισμός της ορίζουσας $|\Sigma_i| = \sigma^{2d}$ και του αντίστροφου $\Sigma_i^{-1} = (1/\sigma^2)I$ του

πίνακα Σ_i είναι σχετικά απλός: Επειδή οι όροι $|\Sigma_i|$ και $(d/2)\ln(2\pi)$ είναι ανεξάρτητοι του i, μπορούν να αγνοηθούν στην εξίσωση 2.48. Έτσι, προκύπτουν οι απλούστερες διακρίνουσες εξισώσεις

$$g_i(x) = -\frac{\|x - \mu_i\|^2}{2\sigma^2} + \ln P(\omega_i) \tag{2.49}$$

όπου το $\|\cdot\|$ δηλώνει την Ευκλείδεια νόρμα, η οποία ισούται με:

$$\|x - \mu_i\|^2 = (x - \mu_i)^t (x - \mu_i) \tag{2.50}$$

Εάν οι εκ των προτέρων πιθανότητες δεν είναι ίσες, σύμφωνα με την εξίσωση 2.49 η τετραγωνική απόσταση $\|x - \mu\|^2$ πρέπει να κανονικοποιηθεί από τη διασπορά σ^2 και να μετακινηθεί κατά την ποσότητα $\ln P(\omega_i)$. Έτσι, εάν ένα δείγμα x βρίσκεται το ίδιο κοντά σε δύο διαφορετικά διανύσματα μέσης τιμής, η βέλτιστη απόφαση θα είναι υπέρ της κατηγορίας που έχει τη μεγαλύτερη εκ των προτέρων πιθανότητα. Στην πραγματικότητα, δεν είναι απαραίτητο να υπολογιστούν αποστάσεις, ανεξάρτητα από το εάν οι εκ των προτέρων πιθανότητες είναι ίσες ή όχι. Από την εξίσωση 2.48 και αναπτύσσοντας τον όρο προκύπτει το εξής:

$$g_i(x) = -\frac{1}{2\sigma^2}\left[x^t x - 2\mu_i^t x + \mu_i^t \mu_i\right] + \ln P(\omega_i) \tag{2.51}$$

το οποίο φαίνεται να είναι μία τετραγωνική συνάρτηση του x. Ο όρος $x^t x$ είναι ίδιος για όλα τα i και επομένως μπορεί να παραληφθεί. Έτσι, προκύπτουν οι παρακάτω ισοδύναμες γραμμικές διακρίνουσες εξισώσεις

$$g_i(x) = w_i^t x + w_{i0} \tag{2.52}$$

όπου

$$w_i = \frac{1}{\sigma^2}\mu_i \tag{2.53}$$

και

$$w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i) \tag{2.54}$$

Το w_{i0} καλείται κατώφλι ή bias της i-οστής κατηγορίας.

Ένας ταξινομητής που χρησιμοποιεί γραμμικές διακρίνουσες συναρτήσεις καλείται γραμμική μηχανή. Οι περιοχές απόφασης για μία γραμμική μηχανή είναι κομμάτια από υπερεπίπεδα τα οποία ορίζονται από τις γραμμικές εξισώσεις $g_i(x) = g_j(x)$ για τις δύο κατηγορίες με τις μεγαλύτερες εκ των υστέρων πιθανότητες. Για τη συγκεκριμένη περίπτωση, η εξίσωση αυτή μπορεί να γραφεί ως εξής:

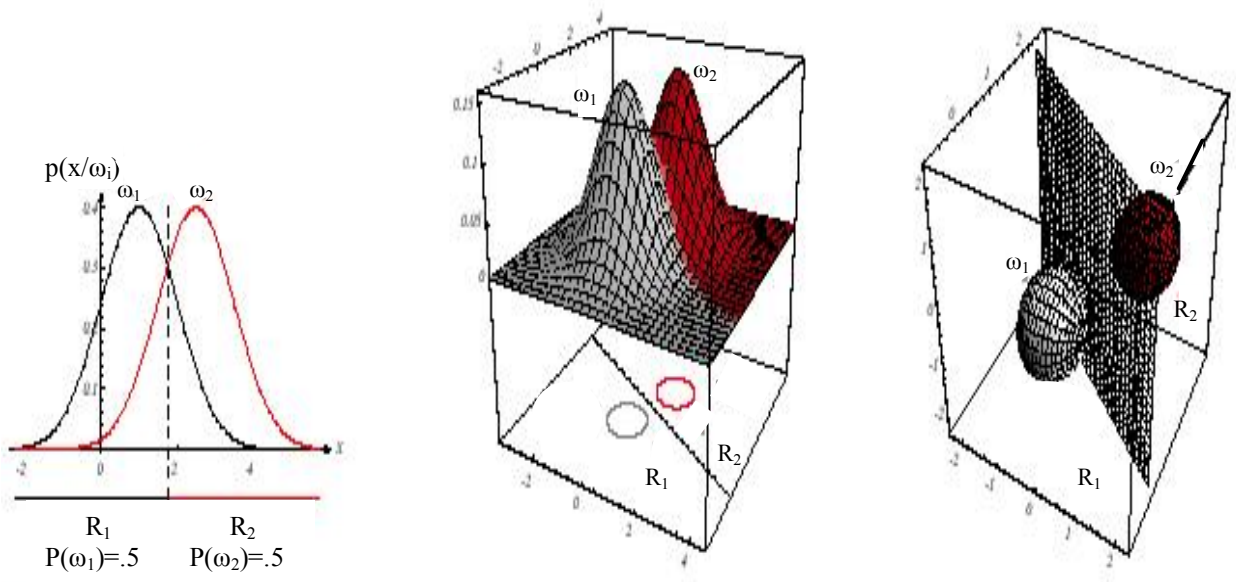
$$w^t (x - x_0) = 0 \tag{2.55}$$

όπου

$$w = \mu_i - \mu_j \tag{2.56}$$

και

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j) \tag{2.57}$$



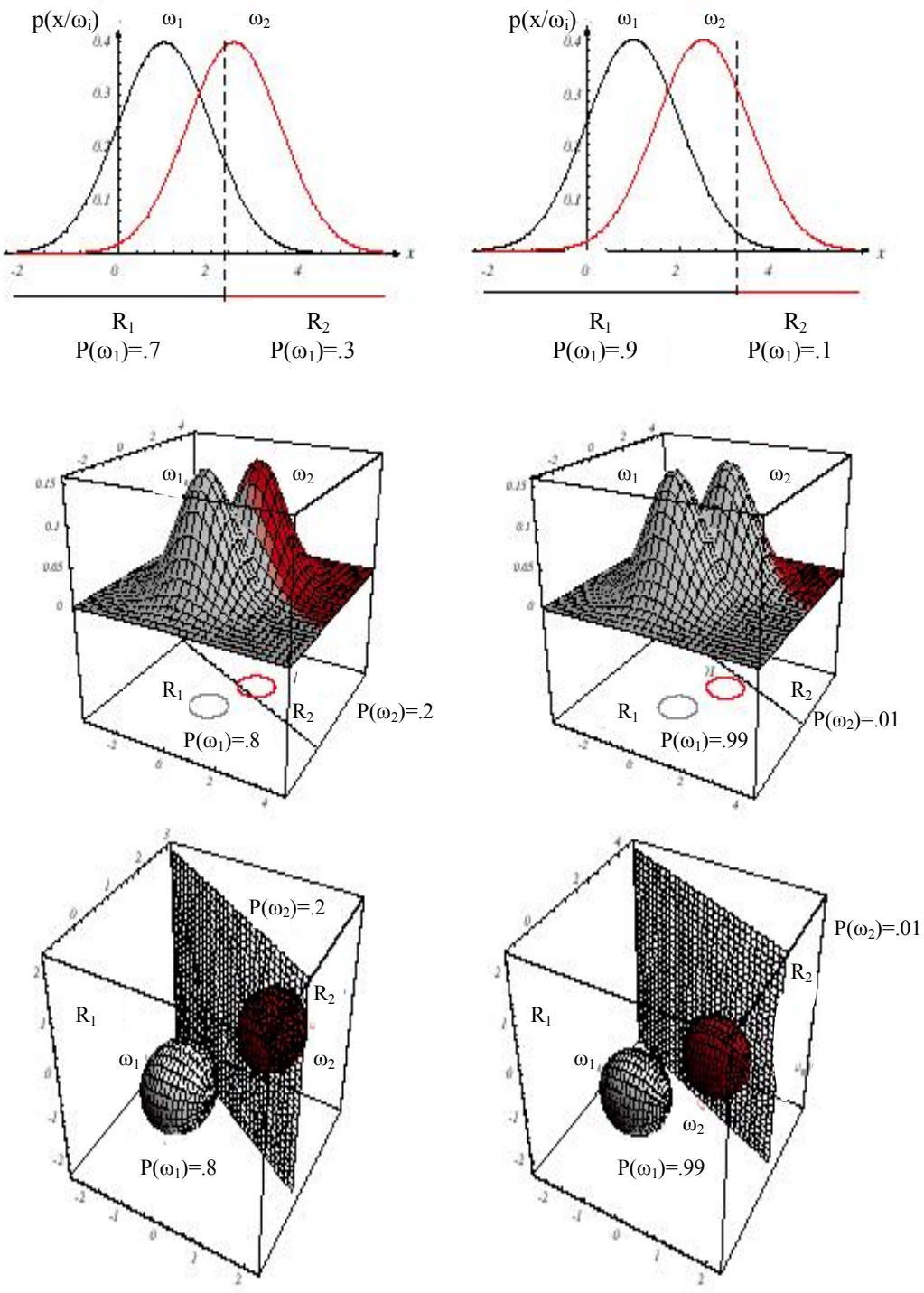
Εικόνα 2.10: Εάν οι πίνακες συνδιασποράς για δύο κατανομές είναι ίσοι και ανάλογοι του ταυτοτικού πίνακα, τότε οι κατανομές είναι σφαιρικές στις d διαστάσεις και το όριο είναι ένα γενικευμένο υπερεπίπεδο των d – 1 διαστάσεων κάθετο στη γραμμή που χωρίζει τα μέσα των κατανομών. Στο σχήμα παρουσιάζεται η $p(x/\omega_i)$ και τα όρια για την περίπτωση όπου $P(\omega_1) = P(\omega_2)$ για τη μονοδιάστατη, τη δυσδιάστατη και την τρισδιάστατη περίπτωση.

Αυτές οι εξισώσεις ορίζουν ένα υπερεπίπεδο που περνάει από το σημείο x_0 και είναι ορθογώνιο στο διάνυσμα w. Επειδή $w = \mu_i - \mu_j$, το υπερεπίπεδο που χωρίζει τις R_i και R_j είναι ορθογώνιο στη γραμμή που συνδέει τα μέσα τους. Εάν $P(\omega_i) = P(\omega_j)$, ο δεύτερος όρος στο δεξιό μέρος της εξίσωσης 2.57 μηδενίζεται και επομένως το σημείο x_0 βρίσκεται στο μέσο του τμήματος που ενώνει τα μέσα των κατηγοριών και το υπερεπίπεδο αποτελεί ένα κάθετο διχοτομητή αυτού του τμήματος (εικόνα 2.10). Εάν $P(\omega_i) \neq P(\omega_j)$, το σημείο x_0 απομακρύνεται από το μέσο της πιο πιθανής κατηγορίας και πλησιάζει το μέσο της λιγότερο πιθανής (εικόνα 2.11). Να σημειωθεί πάντως ότι, εάν η διασπορά σ^2 είναι μικρή σε σχέση με την τετραγωνική απόσταση $\|\mu_i - \mu_j\|^2$, η θέση του ορίου απόφασης δεν επηρεάζεται σημαντικά από τις τιμές των εκ των προτέρων πιθανοτήτων.

Εάν οι εκ των προτέρων πιθανότητες $P(\omega_i)$ είναι ίδιες για όλες τις κατηγορίες c, ο όρος $\ln P(\omega_i)$ μπορεί επίσης να απαλειφθεί. Σε αυτήν την περίπτωση, ο βέλτιστος κανόνας απόφασης παίρνει την εξής απλή μορφή:

Για την ταξινόμηση ενός διανύσματος χαρακτηριστικών x, πρέπει να υπολογιστεί η Ευκλείδεια απόσταση του x από καθένα από τα c διανύσματα μέσων τιμών (που αντιστοιχούν στις c διαφορετικές κατηγορίες) και να τοποθετηθεί το x στην κατηγορία του κοντινότερου διανύσματος μέσης τιμής.

Ένας τέτοιος ταξινομητής καλείται ταξινομητής ελάχιστης απόστασης.



Εικόνα 2.11: Εάν αλλάξουν οι τιμές των εκ των προτέρων πιθανοτήτων, το όριο απόφασης μετακινείται. Στην περίπτωση όπου οι τιμές των εκ των προτέρων πιθανοτήτων διαφέρουν σημαντικά, το όριο μπορεί να μην βρίσκεται καν ανάμεσα στα μέσα των παραπάνω σφαιρικών μονοδιάστατων, δυσδιάστατων και τρισδιάστατων Gaussian κατανομών.

2.6.2 Περίπτωση 2^{η}: $\Sigma_i=\Sigma$.

Μία άλλη απλή περίπτωση είναι αυτή στην οποία οι πίνακες συνδιασποράς είναι ίδιοι για όλες τις κατηγορίες αλλά τυχαίας μορφής. Από γεωμετρική άποψη, η περίπτωση

αυτή αντιστοιχεί στην κατάσταση όπου τα δείγματα κατανέμονται σε υπερελλειψοειδείς ομάδες ίδιου σχήματος και μεγέθους. Το cluster που αντιστοιχεί στην i-οστή κατηγορία συγκεντρώνεται γύρω από το διάνυσμα μέσων τιμών μ_i. Επειδή οι όροι $|\Sigma_i|$ και $(d/2)\ln(2\pi)$ στην εξίσωση 2.48 είναι ανεξάρτητοι του i, μπορούν κάλλιστα να απαλειφθούν. Αυτή η απαλοιφή οδηγεί στις ακόλουθες διακρίνουσες συναρτήσεις:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(\omega_i) \tag{2.58}$$

Εάν οι εκ των προτέρων πιθανότητες $P(\omega_i)$ είναι ίδιες για όλες τις c κατηγορίες, ο όρος $\ln P(\omega_i)$ μπορεί προφανώς να απαλειφθεί. Στην περίπτωση αυτή ο βέλτιστος κανόνας απόφασης παίρνει την ακόλουθη επίσης απλή μορφή:
Για την ταξινόμηση ενός διανύσματος χαρακτηριστικών x, πρέπει να μετρηθεί η τετραγωνική Mahalanobis απόσταση $(x-\mu_i)^t \Sigma^{-1}(x-\mu_i)$ από το x προς καθένα από τα c διανύσματα μέσων τιμών (που αντιστοιχούν στις c διαφορετικές κατηγορίες) και να τοποθετηθεί το x στην κατηγορία του κοντινότερου διανύσματος μέσης τιμής. Όπως και στην προηγούμενη περίπτωση άνισες εκ των προτέρων πιθανότητες bias την απόφαση υπέρ της κατηγορίας με την μεγαλύτερη εκ των προτέρων πιθανότητα.
Η ανάπτυξη του όρου $(x-\mu_i)^t \Sigma^{-1}(x-\mu_i)$ καταλήγει σε ένα άθροισμα που περιέχει τον όρο $x^t\Sigma^{-1}x$ ο οποίος είναι ανεξάρτητος του i. Μετά την αφαίρεση του όρου αυτού από την εξίσωση 2.58, οι διακρίνουσες εξισώσεις που προκύπτουν είναι και πάλι γραμμικές:

$$g_i(x) = w_i^t x + w_{i0} \tag{2.59}$$

όπου

$$w_i = \Sigma^{-1}\mu_i \tag{2.60}$$

και

$$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i) \tag{2.61}$$

Επειδή οι διακρίνουσες συναρτήσεις είναι γραμμικές, οι περιοχές απόφασης που προκύπτουν είναι ξανά υπερεπίπεδα. Εάν οι περιοχές απόφασης R_1 και R_2 είναι γειτονικές, το όριο απόφασης μεταξύ τους έχει την εξίσωση

$$w^t(x-x_0) = 0 \tag{2.62}$$

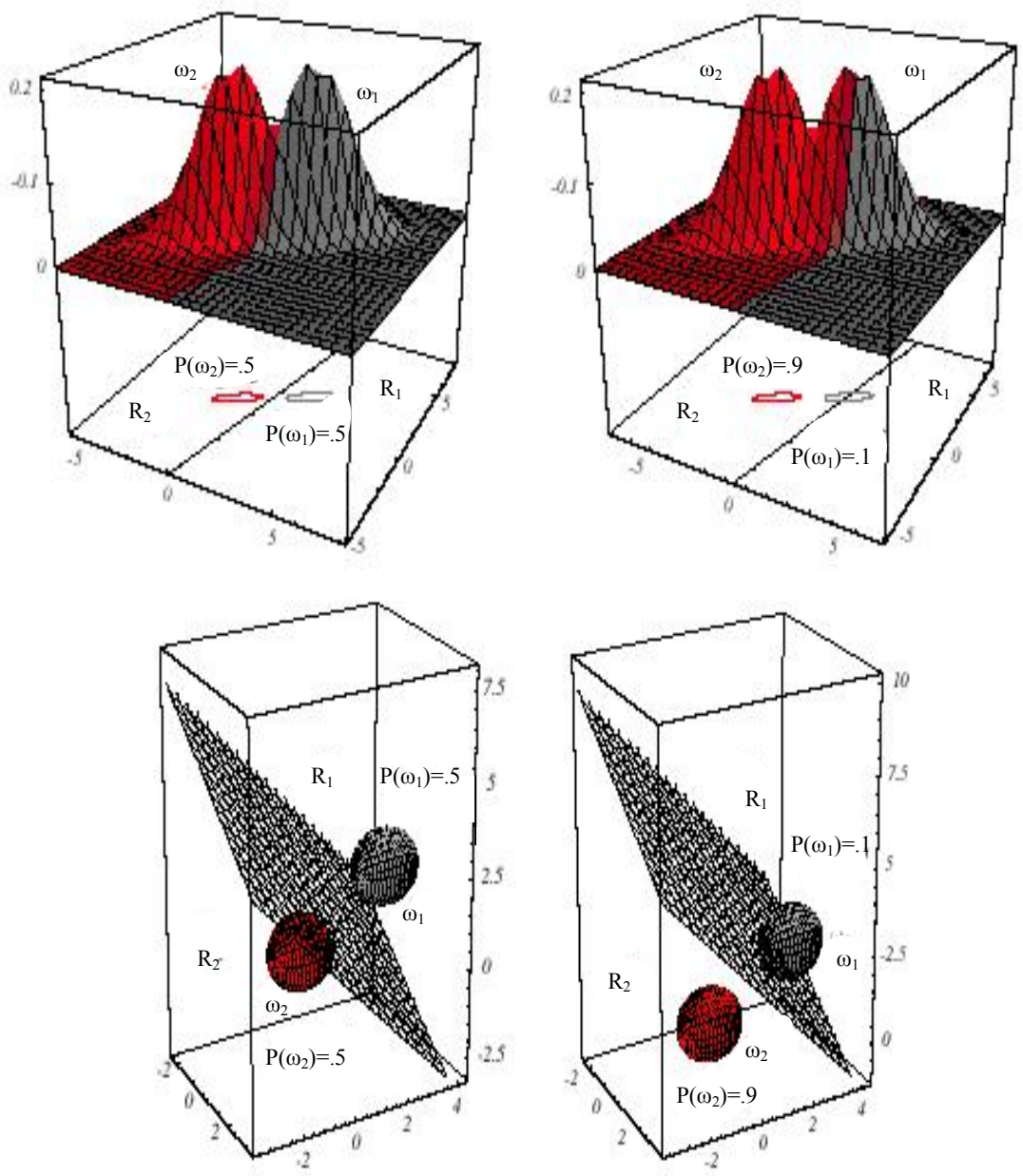
όπου

$$w = \Sigma^{-1}(\mu_i - \mu_j) \tag{2.63}$$

και

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i-\mu_j)\Sigma^{-1}(\mu_i-\mu_j)}(\mu_i - \mu_j) \tag{2.64}$$

Επειδή, το $w = \Sigma^{-1}(\mu_i - \mu_j)$ δεν είναι γενικά στην ίδια κατεύθυνση με το $\mu_i - \mu_j$, το υπερεπίπεδο που χωρίζει τις R_i και R_j δεν είναι γενικά ορθογώνιο στο τμήμα που ενώνει τα μέσα των κατηγοριών. Παρ' όλ' αυτά, τέμνει το τμήμα αυτό στο σημείο x_0. Εάν οι εκ των προτέρων πιθανότητες είναι ίσες, το x_0 βρίσκεται στο μέσο του τμήματος. Εάν οι εκ των προτέρων πιθανότητες δεν είναι ίσες, το υπερεπίπεδο του βέλτιστου ορίου απόφασης μετακινείται μακρύτερα από το πιο πιθανό μέσο (Εικόνα 2.12). Όπως και στην προηγούμενη περίπτωση ανάλογα με το κατώφλι, το επίπεδο απόφασης δεν είναι απαραίτητο να βρίσκεται ανάμεσα στα δύο διανύσματα μέσων τιμών.



Εικόνα 2.12: Οι συναρτήσεις πυκνότητας πιθανότητας (επιφάνειες στις δύο διαστάσεις και ελλειψοειδείς επιφάνειες στις τρεις διαστάσεις) και οι περιοχές απόφασης για ίσες άλλα μη συμμετρικές Gaussian κατανομές. Τα υπερεπίπεδα απόφασης δεν απαιτείται να είναι κάθετα στη γραμμή που συνδέει τα μέσα.

2.6.3 Περίπτωση 3^{η}: Σ_i = αυθαίρετης μορφής

Στην περίπτωση της γενικής κανονικής κατανομής πολλών μεταβλητών, οι πίνακες συνδιασποράς είναι διαφορετικοί για κάθε κατηγορία. Ο μοναδικός όρος που μπορεί να απαλειφθεί από την εξίσωση 2.49 είναι ο $(d/2)\ln(2\pi)$. Οι διακρίνουσες συναρτήσεις που προκύπτουν είναι εκ φύσεως τετραγωνικές:

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0} \qquad (2.65)$$

όπου

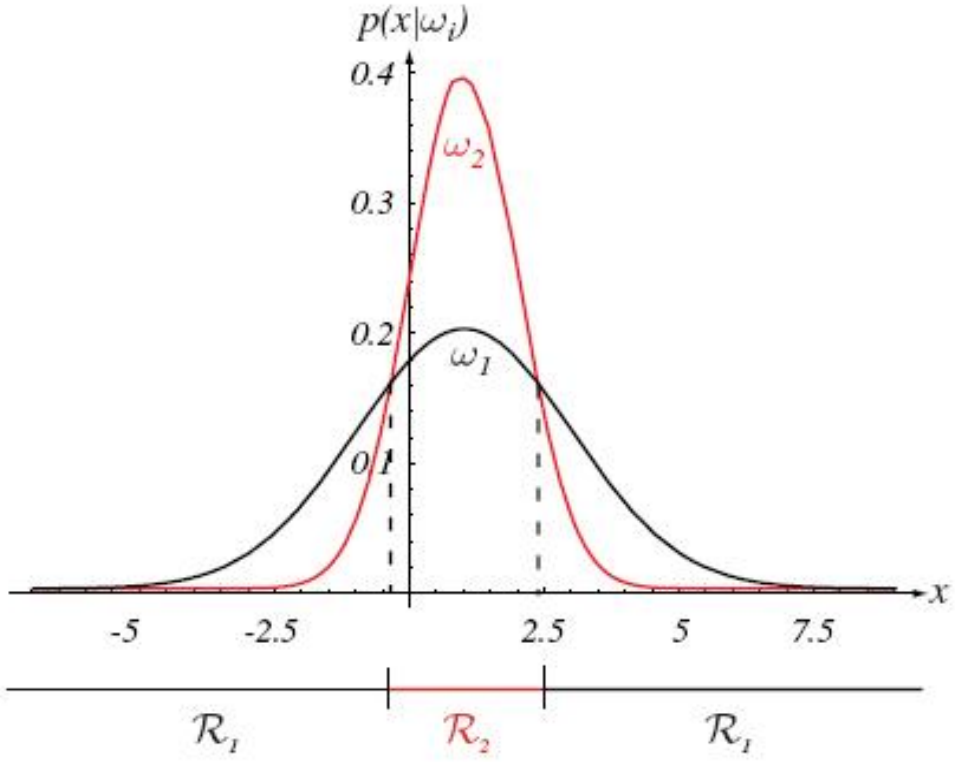
$$W_i = -\frac{1}{2}\Sigma_i^{-1} \tag{2.66}$$

$$w_i = \Sigma_i^{-1}\mu_i \tag{2.67}$$

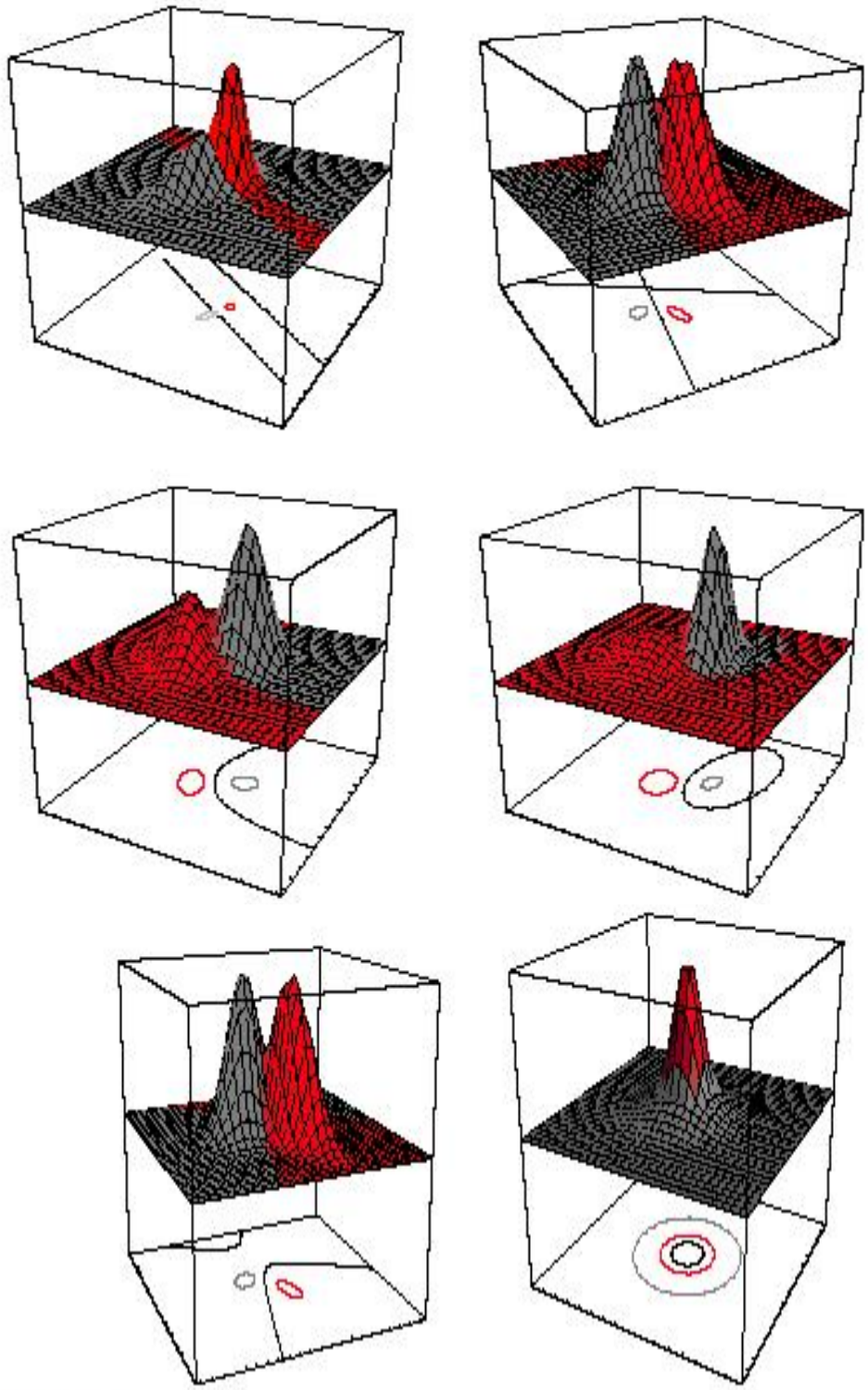
και

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) \tag{2.68}$$

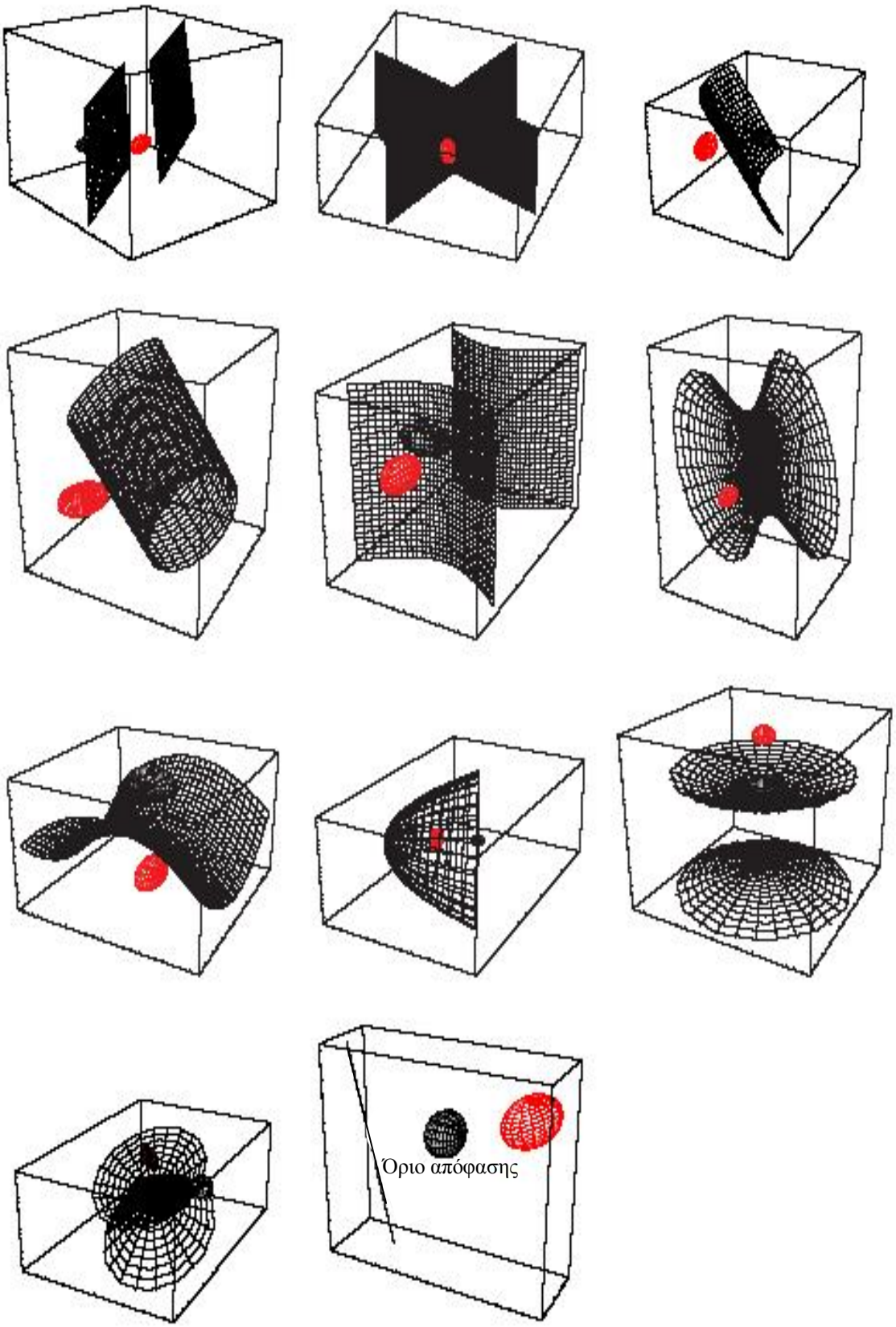
Στην περίπτωση των δύο κατηγοριών, οι επιφάνειες απόφασης είναι υπέρ-τετραγωνικές και μπορούν να έχουν οποιαδήποτε από τις επόμενες μορφές: υπερεπίπεδα, ζεύγη υπερεπιπέδων, υπερσφαίρες, υπερελλειψοειδή, υπερπαραβολοειδή και υπερυπερβολοειδή διαφόρων τύπων. Ακόμα και στη μία διάσταση, για τυχαίας μορφής διασπορά οι περιοχές απόφασης δεν απαιτείται να είναι απλώς συνεκτικές (εικόνα 2.13). Στις εικόνες 2.14 και 2.15 παρουσιάζονται διάφορες από τις μορφές που μπορεί να έχουν τα όρια και οι περιοχές απόφασης για δύο και τρεις διαστάσεις αντίστοιχα. Η επέκταση αυτών των αποτελεσμάτων σε περισσότερες από δύο κατηγορίες είναι άμεση. Το μόνο που πρέπει να αποφασιστεί κάθε φορά είναι ποιες δύο από τις συνολικά c κατηγορίες είναι υπεύθυνες για κάθε συγκεκριμένο όριο. Στην εικόνα 2.16 παρουσιάζονται οι περιοχές απόφασης για την περίπτωση τεσσάρων κατηγοριών που ακολουθούν Gaussian κατανομές. Προφανώς, εάν οι κατανομές είναι πιο πολύπλοκες, οι περιοχές απόφασης μπορεί να είναι ακόμα περισσότερο πολύπλοκες.



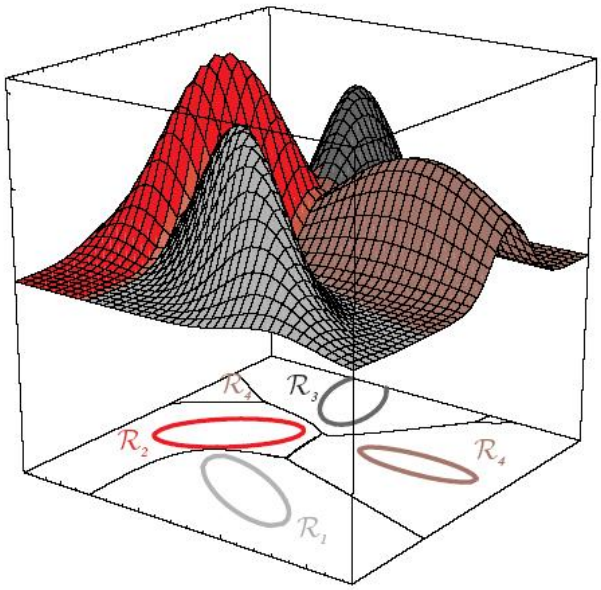
Εικόνα 2.13: Μη συνεκτικές περιοχές απόφασης μπορούν να εμφανιστούν στη μονοδιάστατη περίπτωση για Gaussian κατανομές που έχουν άνισες διασπορές, όπως φαίνεται στο παραπάνω σχήμα όπου $P(\omega_1) = P(\omega_2)$.



Εικόνα 2.14: Τυχαίας μορφής Gaussian κατανομές που οδηγούν σε όρια απόφασης κατά Bayes τα οποία είναι γενικής μορφής υπέρ-τετραγωνικά. Αντίστροφα, για οποιαδήποτε τετραγωνική μορφή, μπορεί να βρεθούν δύο Gaussian κατανομές των οποίων το όριο απόφασης κατά Bayes να είναι αυτό το quadratic. Οι διασπορές τους υποδηλώνονται από τις καμπύλες σταθερής πυκνότητας πιθανότητας.



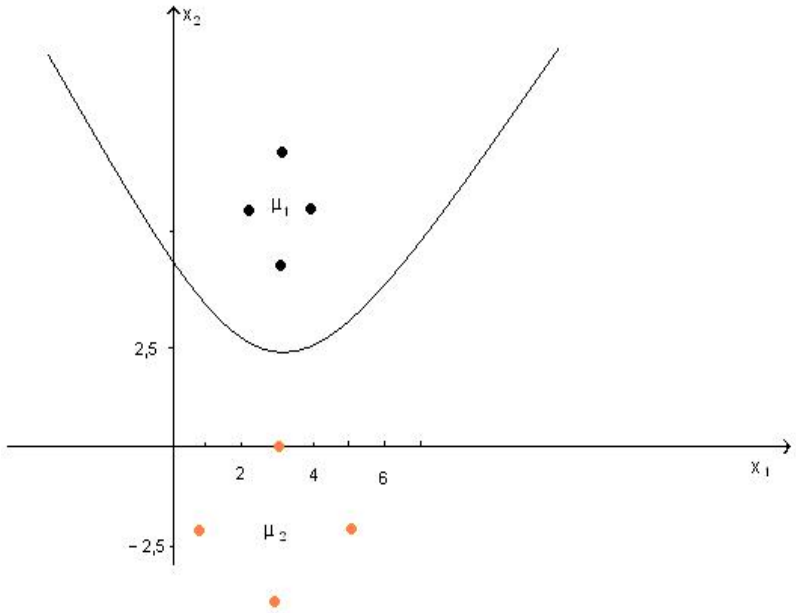
Εικόνα 2.15: Τυχαίας μορφής τρισδιάστατες Gaussian κατανομές οδηγούν σε όρια απόφασης κατά Bayes που αποτελούν δυσδιάστατα υπέρ-τετραγωνικά. Υπάρχουν ακόμα και ακραίες περιπτώσεις όπου το όριο απόφασης είναι μία ευθεία.



Εικόνα 2.16: Οι περιοχές απόφασης για τέσσερις κανονικές κατανομές. Όπως φαίνεται από το σχήμα, ακόμα και με μικρό αριθμό κατηγοριών τα σχήματα των περιοχών απόφασης μπορεί να είναι ιδιαίτερα πολύπλοκα.

Παράδειγμα 1. Περιοχές απόφασης για Gaussian Δεδομένα 2 διαστάσεων

Θα υπολογίσουμε το όριο απόφασης για τα δεδομένα (2 διαστάσεων, 2 κατηγοριών) του παρακάτω σχήματος.



Εικόνα : Το υπολογισμένο όριο απόφασης του Bayes για δύο Gaussian κατανομές, η κάθε μια από τις οποίες βασίζεται σε 4 σημεία.

Ας είναι ω_1 το σύνολο των τεσσάρων τονισμένων σημείων και ω_2 τα λιγότερο τονισμένα σημεία. Υπολογίζουμε τις μέσες τιμές και συνδιασπορές με βάση τους τύπους 2.39 και 2.40. Είναι :

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}, \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \kappa\alpha\iota \quad \mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

και

$$\Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \kappa\alpha\iota \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

Υποθέτουμε ότι οι a priori πιθανότητες είναι ίσες, δηλαδή $P(\omega_1)= P(\omega_2)=0.5$ και τις αντικαθιστούμε στους τύπους 2.65 έως 2.68, για εύρεση των διακρίνουσων συναρτήσεων. Υποθέτοντας $g_1(x)=g_2(x)$ λαμβάνουμε το παρακάτω όριο απόφασης:

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

Η παραπάνω εξίσωση περιγράφει μια παραβολή με κορυφή στο $\begin{pmatrix} 3 \\ 1.83 \end{pmatrix}$.

Παρατηρούμε ότι, παρά το γεγονός ότι η διασπορά στα δεδομένα κατά τον άξονα x_2 και για τις 2 κατανομές είναι η ίδια, το όριο απόφασης δεν περνάει από το σημείο $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$, το μέσο σημείο ανάμεσα στις μέσες τιμές, όπως θα υποθέταμε. Αυτό συμβαίνει γιατί, για την κατανομή ω_1, η κατανομή πιθανότητας «συμπιέζεται» κατά την φορά του άξονα x_1, περισσότερο από ότι αυτή της κατανομής ω_2. Η κατανομή ω_1 αυξάνεται κατά την φορά του άξονα x_2 (σε σχέση με την κατανομή ω_2). Έτσι το όριο απόφασης βρίσκεται λίγο χαμηλότερα από το μέσο σημείο μεταξύ των μέσων τιμών, κάτι που φαίνεται και στο παραπάνω σχήμα

2.7 Πιθανότητες Λάθους και Διαστήματα

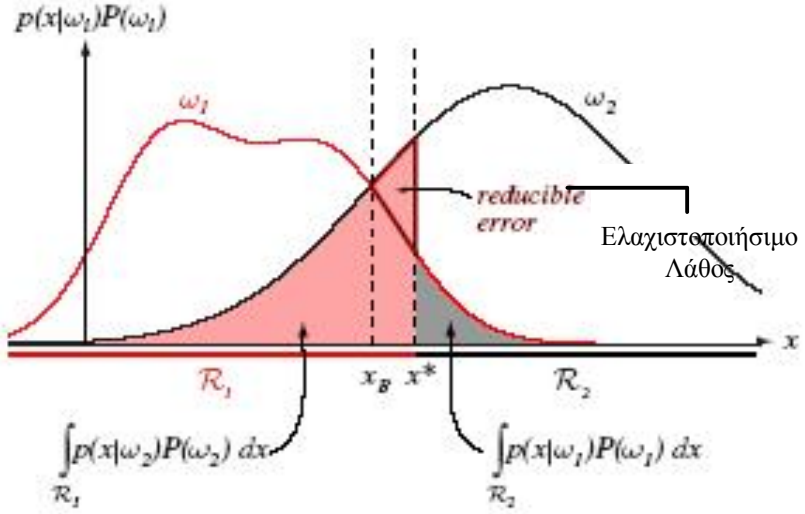
Για να γίνει πιο κατανοητή η λειτουργία ενός γενικού ταξινομητή (Bayes ή άλλου είδους) χρήσιμο θα ήταν να ερευνηθούν οι πηγές που δημιουργούν το λάθος του. Έστω αρχικά, η περίπτωση των δύο κατηγοριών όπου ο ταξινομητής (διχοτόμος) έχει χωρίσει το χώρο των χαρακτηριστικών σε δύο περιοχές απόφασης R_1 και R_2 με έναν πιθανό μη βέλτιστο τρόπο. Οι περιπτώσεις για τις οποίες μπορεί να συμβεί κάποιο λάθος ταξινόμησης είναι δύο. Είτε ένα σύνολο x βρίσκεται στην περιοχή R_2 ενώ η πραγματική κατάσταση της φύσης είναι το ω_1, είτε βρίσκεται στην περιοχή R_1 ενώ η πραγματική κατάσταση της φύσης είναι το ω_2. Επειδή οι δύο αυτές περιπτώσεις είναι αμοιβαία αποκλειόμενες και εξουδετερωμένες, η πιθανότητα του λάθους ισούται με:

$$\begin{aligned} P(\text{λάθους}) &= P(x \in R_2, \omega_1) + P(x \in R_1, \omega_2) \\ &= P(x \in R_2 / \omega_1)P(\omega_1) + P(x \in R_1 / \omega_2)P(\omega_2) \\ &= \int_{R_2} p(x/\omega_1)P(\omega_1)dx + \int_{R_1} p(x/\omega_2)P(\omega_2)dx \end{aligned} \qquad (2.69)$$

Το αποτέλεσμα αυτό παρουσιάζεται γραφικά για τη μονοδιάστατη περίπτωση στην εικόνα 2.17. Τα δύο διαστήματα στην εξίσωση 2.69 αντιπροσωπεύουν αντίστοιχα τις σκιασμένες περιοχές στις ουρές των συναρτήσεων $p(x/\omega_i)P(\omega_i)$. Επειδή το σημείο απόφασης x^* (και επομένως οι περιοχές απόφασης R_1 και R_2) έχουν επιλεχθεί τυχαία για τη συγκεκριμένη περίπτωση, η πιθανότητα του λάθους δεν είναι τόσο μικρή όσο θα μπορούσε να είναι. Πιο συγκεκριμένα, η τριγωνική περιοχή που ονομάζεται λάθος που μπορεί να ελαττωθεί (reducible error) μπορεί να εξαλειφθεί εάν το όριο απόφασης μετατοπιστεί στο σημείο x_B. Αυτό το σημείο αποτελεί το βέλτιστο όριο απόφασης κατά Bayes και δίνει την ελάχιστη πιθανότητα λάθους. Γενικά, εάν $p(x/\omega_1)P(\omega_1) > p(x/\omega_2)P(\omega_2)$ είναι προτιμότερο να ταξινομηθεί το x σαν να ήταν στην περιοχή απόφασης R_1 έτσι ώστε η μικρότερη ποσότητα από τις δύο να συνεισφέρει στο διάστημα που μετριέται το λάθος. Αυτό ακριβώς είναι που επιτυγχάνει ο κανόνας απόφασης του Bayes. Όταν οι κατηγορίες είναι περισσότερες από δύο, υπάρχουν περισσότερες περιπτώσεις να συμβεί λάθος από σωστό και επομένως είναι ευκολότερο να υπολογιστεί η πιθανότητα του σωστού. Προφανώς,

$$\begin{aligned} P(\text{σωστού}) &= \sum_{i=1}^{c} P(x \in R_i, \omega_i) \\ &= \sum_{i=1}^{c} P(x \in R_i / \omega_i) P(\omega_i) \\ &= \sum_{i=1}^{c} \int_{R_i} p(x / \omega_i) P(\omega_i) dx \end{aligned} \tag{2.70}$$

Το γενικευμένο αποτέλεσμα της εξίσωσης 2.70 δεν εξαρτάται ούτε από τον τρόπο από τον οποίο έχει διαχωριστεί ο χώρος των χαρακτηριστικών σε περιοχές απόφασης ούτε από τη μορφή των αντίστοιχων κατανομών. Ο ταξινομητής κατά Bayes μεγιστοποιεί την πιθανότητα επιλέγοντας τις περιοχές με τέτοιο τρόπο ώστε να μεγιστοποιείται η τιμή του ολοκληρώματος για όλα τα x. Κανένας άλλος διαχωρισμός του χώρου των χαρακτηριστικών δε δίνει μικρότερη πιθανότητα λάθους.



Εικόνα 2.17: Διαστήματα της πιθανότητας του λάθους για ίσες εκ των προτέρων πιθανότητες και για το (μη βέλτιστο) σημείο απόφασης x^*. Η αριστερή γραμμοσκιασμένη περιοχή αντιστοιχεί στην πιθανότητα του λάθους για απόφαση ω_1 όταν στην πραγματικότητα η κατάσταση της φύσης είναι η ω_2. Η δεξιά γραμμοσκιασμένη περιοχή αντιστοιχεί στο αντίστροφο, όπως δίνεται από την εξίσωση 2.69. Εάν το σημείο απόφασης βρίσκεται πάνω στο σημείο x_B, για το οποίο οι εκ των υστέρων πιθανότητες είναι ίσες, η περιοχή του «ελαχιστοποιήσιμου λάθους» μηδενίζεται και το συνολικό γραμμοσκιασμένο εμβαδόν είναι το ελάχιστο δυνατό. Αυτό αποτελεί το όριο απόφασης του Bayes και δίνει ως αποτέλεσμα τον αντίστοιχο ρυθμό λάθους.

2.8 Όρια Λάθους Για Κανονικές Συναρτήσεις Πυκνότητας Πιθανότητας

Όπως αναφέρθηκε σε προηγούμενες ενότητες, ο κανόνας απόφασης του Bayes εγγυάται το χαμηλότερο μέσο ρυθμό λάθους. Επίσης, σε προηγούμενη ενότητα περιγράφηκε ο τρόπος με τον οποίο υπολογίζονται τα όρια απόφασης για συναρτήσεις πυκνότητας πιθανότητας που ακολουθούν κανονική κατανομή. Όμως, τα αποτελέσματα αυτά δεν δίνουν συγκεκριμένη τιμή για την πιθανότητα λάθους. Ο πλήρης υπολογισμός του λάθους για τη Gaussian περίπτωση μπορεί να είναι πολύ δύσκολος, ιδιαίτερα στην περίπτωση μεγάλων διαστάσεων, λόγω κυρίως της ασυνεχούς φύσης των περιοχών απόφασης στο ολοκλήρωμα της εξίσωσης 2.70. Παρ' όλα αυτά, στην περίπτωση των δύο κατηγοριών, η τιμή του ολοκληρώματος του λάθους στην εξίσωση 2.5 μπορεί να υπολογιστεί αναλυτικά και να δώσει το άνω όριο του λάθους.

2.8.1 Όριο Chernoff

Για να υπολογιστεί ένα όριο για το λάθος, χρησιμοποιείται η ακόλουθη ανίσωση:

$$\min[a, b] \le a^{\beta}b^{1-\beta} \quad \text{για } a, b \ge 0 \text{ και } 0 \le \beta \le 1 \tag{2.71}$$

Για καλύτερη κατανόηση αυτής της ανίσωσης, και χωρίς χάσιμο της γενικότητας, θεωρείται ότι $a \ge b$. Έτσι, αρκεί να αποδειχθεί μόνο ότι $b \le a^{\beta}b^{1-\beta}=(a/b)^{\beta}b$. Όμως, η σχέση αυτή προφανώς ισχύει, αφού $(a/b)^{\beta} \ge 1$. Χρησιμοποιώντας τις εξισώσεις 2.7 και 2.1 και εφαρμόζοντας την παραπάνω ανίσωση στην εξίσωση 2.5, παίρνουμε το ακόλουθο όριο:

$$P(\text{λάθους}) = P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\int p^{\beta}(x/\omega_1)p^{1-\beta}(x/\omega_2)dx \quad \text{για } 0 \le \beta \le 1 \tag{2.72}$$

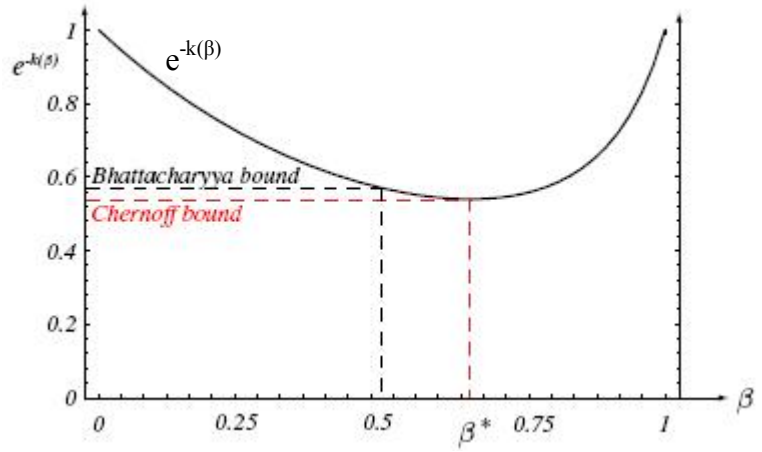
Σημειώνεται ότι το παραπάνω ολοκλήρωμα γίνεται πάνω σε ολόκληρο το χώρο των χαρακτηριστικών – δε χρειάζεται να καθοριστούν τα όρια ολοκλήρωσης που αντιστοιχούν στα όρια απόφασης. Εάν οι υπό συνθήκη πιθανότητες είναι κανονικής κατανομής, το ολοκλήρωμα της εξίσωσης 2.72 μπορεί να υπολογιστεί αναλυτικά καταλήγοντας στον παρακάτω τύπο:

$$\int p^{\beta}(x/\omega_1)p^{1-\beta}(x/\omega_2)dx = e^{-k(\beta)} \tag{2.73}$$

όπου

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2 - \mu_2)^t[\beta\,\Sigma_1 + (1-\beta)\Sigma_2]^{-1}(\mu_2 - \mu_1) + \frac{1}{2}\ln\frac{|\beta\,\Sigma_1 + (1-\beta)\Sigma_2|}{|\Sigma_1|^{\beta}|\Sigma_2|^{1-\beta}} \tag{2.74}$$

Η γραφική παράσταση της εικόνας 2.18 παρουσιάζει ένα τυπικό παράδειγμα της μεταβολής του $e^{-k(\beta)}$ σε σχέση με το β.
Το όριο Chernoff για την πιθανότητα λάθους P(λάθους) βρίσκεται από την αναλυτική ή αριθμητική εύρεση της τιμής του β που ελαχιστοποιεί το $e^{-k(\beta)}$ και την αντικατάσταση της τιμής αυτής στην εξίσωση 2.72. Το πλεονέκτημα σε αυτήν την περίπτωση είναι ότι η βελτιστοποίηση γίνεται στο μονοδιάστατο χώρο του β, αν και οι ίδιες οι κατανομές μπορεί να βρίσκονται σε χώρο πολύ μεγαλύτερης διάστασης.



Εικόνα 2.18: Το όριο λάθους Chernoff δεν είναι ποτέ πιο «χαλαρό» από το όριο λάθους Bhattacharrya. Για το παραπάνω παράδειγμα, το όριο Chernoff συμβαίνει για $\beta^* = 0.66$ και είναι ελαφρώς αυστηρότερο από το όριο Bhattacharrya (β = 0.5).

2.8.2 Όριο Bhattacharrya

Η εξάρτηση του ορίου Chernoff από την τιμή του β, που φαίνεται στην εικόνα 2.18, είναι τυπική για μία ευρεία περιοχή προβλημάτων. Το όριο είναι χαλαρό για οριακές τιμές του β (δηλαδή β→1 και β→0) ενώ είναι αυστηρότερο για ενδιάμεσες τιμές. Εάν και η ακριβής βέλτιστη τιμή του β εξαρτάται από τις παραμέτρους των κατανομών και τις εκ των προτέρων πιθανότητες, ένα υπολογιστικά απλούστερο αλλά ελάχιστα

λιγότερο αυστηρό όριο μπορεί να προκύψει θέτοντας το β ίσο με ½. Έτσι, προκύπτει το όριο Bhattacharrya για το λάθος και η εξίσωση 2.72 παίρνει τη μορφή

$$P(\text{λάθους}) \le \sqrt{P(\omega_1)P(\omega_2)}\int \sqrt{p(x/\omega_1)p(x/\omega_2)}dx = \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)} \qquad (2.75)$$

όπου από την εξίσωση 2.74 προκύπτει το εξής για τη Gaussian περίπτωση:

$$k(1/2) = \frac{1}{8}(\mu_2-\mu_1)^t\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1)+\frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \qquad (2.76)$$

Τα όρια Chernoff και Bhattacharrya μπορούν να χρησιμοποιηθούν ακόμα και σε περιπτώσεις όπου οι κατανομές δεν είναι Gaussian. Παρ' όλ' αυτά, για κατανομές που διαφέρουν σημαντικά από μία Gaussian, τα όρια δεν θα είναι informative.

Παράδειγμα 2. Όρια λάθους για Gaussian κατανομές

Στο παράδειγμα αυτό θα υπολογιστεί το όριο Bhattacharyya για τα δεδομένα του προβλήματος δύο διαστάσεων του παραδείγματος 1. Αντικαθιστώντας τις μέσες τιμές και τις συνδιασπορές του παραδείγματος 1 στην εξίσωση (2.76) βρίσκουμε k(1/2)=4.06 και έτσι από τις εξισώσεις (2.75) και (2.76) το όριο λάθους του Bhattacharyya είναι $P(\text{error}) \le 0.0087$.

Ένα πιο σφιχτό όριο για το λάθος μπορεί να υπολογιστεί βρίσκοντας το όριο Chernoff της εξίσωσης (2.74), η οποία, για το συγκεκριμένο πρόβλημα, ισούται με 0.016380. Από το ολοκλήρωμα της εξίσωσης (2.5) προκύπτει ρυθμός λάθους ίσος με error rate= 0.0021. Δηλαδή, τα όρια σε αυτήν την περίπτωση δεν είναι ιδιαίτερα σφιχτά. Η αριθμητική ολοκλήρωση αυτής της μορφής είναι συχνά μη πρακτική για Gaussian κατανομές μεγαλύτερες των δύο ή των τριών διαστάσεων.

2.8.3 Θεωρία Ανίχνευσης Σημάτων και Χαρακτηριστικές Λειτουργίας

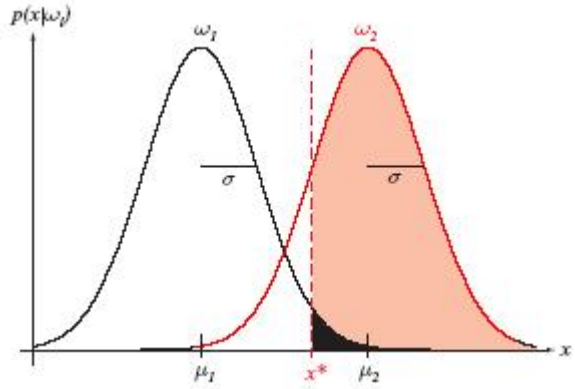
Ένα άλλο μέτρο της απόστασης μεταξύ δύο Gaussian κατανομών έχει βρει μεγάλη απόκριση και χρήση στην πειραματική ψυχολογία, την ανίχνευση σημάτων ραντάρ και άλλους τομείς. Έστω ότι πρέπει να ανιχνευθεί ένας αδύναμος παλμός, όπως μια ακτίνα φακού ή φωτός ή μία αδύναμη ανάκλαση ενός σήματος ραντάρ. Το μοντέλο που ακολουθείται είναι το εξής: σε κάποιο σημείο στον ανιχνευτή υπάρχει ένα εσωτερικό σήμα, (όπως μια τάση) x, του οποίου η μέση τιμή είναι μ_2, όταν το εξωτερικό σήμα (παλμός) είναι παρόν, και μ_1, όταν δεν είναι παρόν. Επειδή ο θόρυβος είναι τυχαίος – τόσο στο εσωτερικό όσο και στο εξωτερικό του ανιχνευτή – η πραγματική τιμή είναι μία τυχαία μεταβλητή. Οι κατανομές θεωρούνται κανονικές με διαφορετικές μέσες τιμές αλλά την ίδια διασπορά, δηλαδή $p(x/\omega_i) \sim N(\mu_i, \sigma^2)$, όπως φαίνεται στην εικόνα 2.19.

Ο ανιχνευτής (ταξινομητής) χρησιμοποιεί μία τιμή κατωφλίου x^* για να καθορίσει κάθε φορά εάν ο εξωτερικός παλμός είναι παρών. Υποτίθεται ότι ο ερευνητής δε γνωρίζει αυτή την τιμή ούτε επίσης τις μέσες τιμές και τις τυπικές αποκλίσεις των κατανομών. Υπό αυτές τις συνθήκες προσπαθεί να βρει κάποιο μέτρο για να ξεχωρίζει εάν ο παλμός είναι παρών ή όχι χρησιμοποιώντας μια μορφή ανεξάρτητη από την τιμή του x^*. Ένα τέτοιο μέτρο είναι η διαχωρισιμότητα, η οποία περιγράφει τις εσωτερικές και σταθερές ιδιότητες που εξαρτώνται από το θόρυβο και την ένταση

του εξωτερικού σήματος, και όχι από τη στρατηγική απόφασης (δηλαδή την πραγματική τιμή του x^*). Η διαχωρισιμότητα ορίζεται ως εξής:

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \tag{2.77}$$

Προφανώς είναι επιθυμητή μια υψηλή τιμή για το d΄.



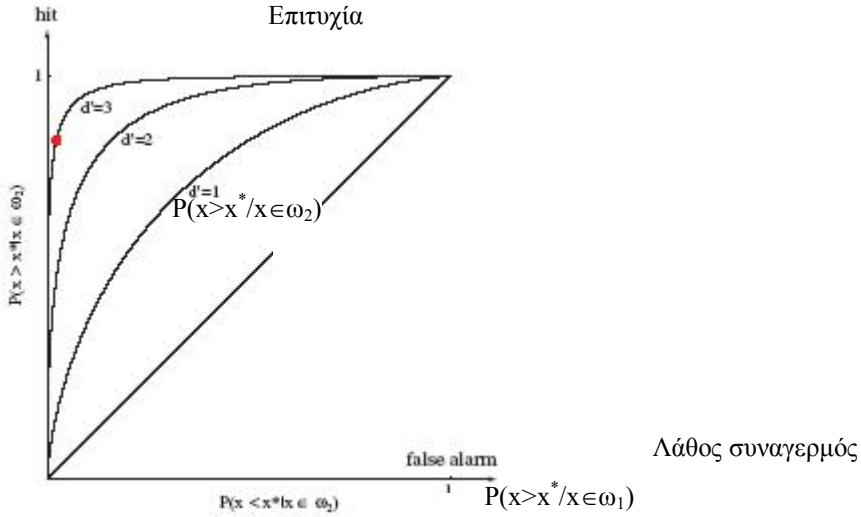
Εικόνα 2.19: Όταν δεν υπάρχει κανένας εξωτερικός παλμός, η συνάρτηση πυκνότητας πιθανότητας για ένα εσωτερικό σήμα είναι κανονική, δηλαδή $p(x/\omega_1) \sim N(\mu_1, \sigma^2)$. Όταν υπάρχει παρουσία εξωτερικού σήματος, η συνάρτηση πυκνότητας πιθανότητας ισούται με $p(x/\omega_2) \sim N(\mu_2, \sigma^2)$. Κάθε κατώφλι απόφασης x^* καθορίζει την πιθανότητα μιας επιτυχίας (το εμβαδόν αριστερά κάτω από την ω_2 και πάνω από το x^*) και ενός λάθος συναγερμού (το μαύρο εμβαδόν κάτω από την ω1 και πάνω από το x^*).

Αν και οι τιμές των μ_1, μ_2, σ και x^* είναι άγνωστες, θεωρείται ότι η κατάσταση της φύσης και η απόφαση του συστήματος είναι γνωστές. Αυτές οι πληροφορίες επιτρέπουν την εύρεση της τιμής του d΄. Χρησιμοποιούνται οι τέσσερις παρακάτω πιθανότητες:

- ✓ $P(x > x^* / x \in \omega_2)$: Μία επιτυχία – η πιθανότητα ότι το εσωτερικό σήμα είναι μεγαλύτερο από το κατώφλι x^*, δεδομένου ότι το εξωτερικό σήμα είναι παρόν.
- ✓ $P(x > x^* / x \in \omega_1)$: Ένας λάθος συναγερμός – η πιθανότητα ότι το εσωτερικό σήμα είναι μεγαλύτερο από το κατώφλι x^*, αν και κανένα εξωτερικό σήμα δεν είναι παρόν.
- ✓ $P(x < x^* / x \in \omega_2)$: Μία αποτυχία – η πιθανότητα ότι το εσωτερικό σήμα είναι μικρότερο από το κατώφλι x^*, δεδομένου ότι το εξωτερικό σήμα είναι παρόν.
- ✓ $P(x < x^* / x \in \omega_1)$: Μια επιτυχημένη απόρριψη – η πιθανότητα ότι το εσωτερικό σήμα είναι μικρότερο από το κατώφλι x^*, δεδομένου ότι το εξωτερικό σήμα δεν είναι παρόν.

Εάν υπάρχει ένας μεγάλος αριθμός δοκιμών και η τιμή του x^* εάν και άγνωστη θεωρηθεί σταθερή, οι τιμές των πιθανοτήτων αυτών μπορούν να καθοριστούν πειραματικά – κυρίως οι τιμές της πιθανότητας επιτυχίας και της πιθανότητας λάθος συναγερμού. Αρχικά απεικονίζεται γραφικά ένα σημείο που αντιπροσωπεύει αυτές τις πιθανότητες σε μία γραφικά παράσταση δύο διαστάσεων. Εάν οι συναρτήσεις πυκνότητας πιθανότητας είναι σταθερές αλλά η τιμή του κατωφλίου x^* αλλάζει, τότε οι πιθανότητες επιτυχίας και λάθος συναγερμού θα αλλάζουν και αυτές. Έτσι (εικόνα

2.20), προκύπτει ότι για μία δεδομένη διακρισιμότητα d΄, το σημείο μετακινείται κατά μήκος μιας λείας καμπύλης, η οποία καλείται χαρακτηριστική καμπύλη λειτουργίας δέκτη (receiver operating characteristic – ROC).

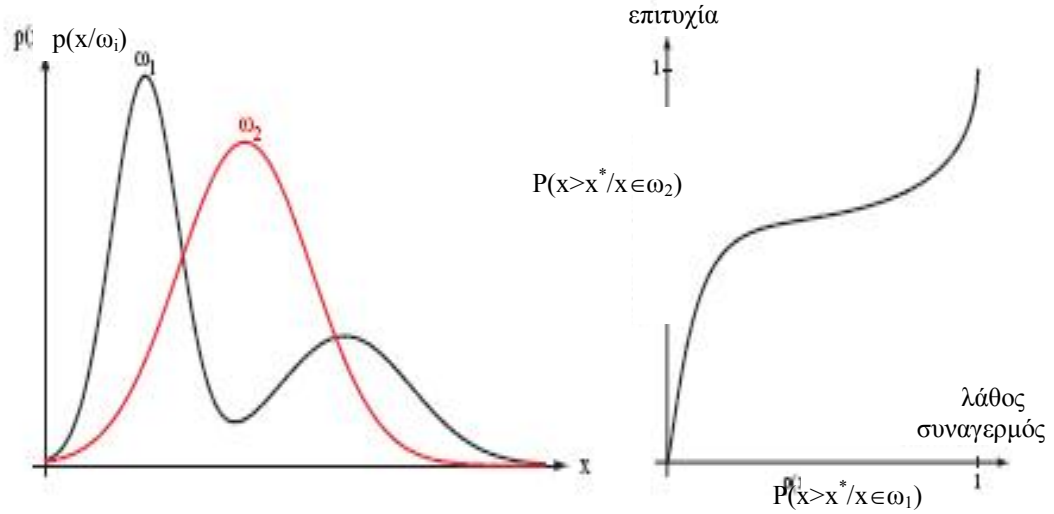


Εικόνα 2.20: Σε μία χαρακτηριστική καμπύλη λειτουργίας δέκτη (ROC), ο οριζόντιος άξονας είναι η πιθανότητα ενός λάθος συναγερμού, $P(x > x^* / x \in \omega_1)$, ενώ ο κατακόρυφος είναι η πιθανότητα μιας επιτυχίας, $P(x > x^* / x \in \omega_2)$. Από τις μετρήσεις που έγιναν για το συγκεκριμένο παράδειγμα, που αντιστοιχούν στο σημείο x^* του σχήματος 2.19, προκύπτει ότι d΄=3.

Το κύριο πλεονέκτημα αυτού του πλαισίου ανίχνευσης σημάτων είναι ότι μπορεί να διαχωριστεί λειτουργικά η διαχωρισιμότητα από την τάση απόφασης: ενώ η πρώτη αποτελεί μία εσωτερική ιδιότητα του συστήματος ανίχνευσης, η δεύτερη επηρεάζεται από τον αντίστοιχο μεταβαλλόμενο πίνακα απωλειών. Από κάθε ζεύγος πιθανοτήτων επιτυχίας και λάθος συναγερμού περνάει μία και μόνο μία ROC καμπύλη. Έτσι, για όσο διάστημα καμία από τις πιθανότητες δεν είναι ίση με 0 ή 1, η διαχωρισιμότητα μπορεί να προσδιοριστεί από αυτές τις πιθανότητες. Επιπλέον, εάν οι κατανομές είναι Gaussian, ο καθορισμός της διαχωρισιμότητας (από ένα τυχαίο x^*) επιτρέπει τον υπολογισμό του ρυθμού λάθους του Bayes, που ως γνωστόν αποτελεί το πιο σημαντικό χαρακτηριστικό κάθε ταξινομητή. Εάν ο πραγματικός ρυθμός λάθους είναι διαφορετικός από το ρυθμό Bayes που υπολογίστηκε με τον παραπάνω τρόπο, η τιμή του κατωφλίου x^* πρέπει να τροποποιηθεί ανάλογα.

Το παραπάνω παράδειγμα μπορεί εύκολα να γενικευθεί και να εφαρμοστεί στην περίπτωση δύο κατηγοριών με τυχαίες πολυδιάστατες κατανομές, Gaussian ή όχι. Έστω δύο κατανομές $p(x/\omega_1)$ και $p(x/\omega_2)$ οι οποίες υπερκαλύπτονται και επομένως έχουν μη μηδενικό λάθος ταξινόμησης κατά Bayes. Όπως και προηγουμένως, κάθε πρότυπο που ανήκει πραγματικά στην ω_2 μπορεί είτε να ταξινομηθεί σωστά στην ω_2 («επιτυχία») είτε να ταξινομηθεί λανθασμένα στην ω_1 («λάθος συναγερμός»). Στην περίπτωση αυτή όμως, αντίθετα με την περίπτωση της μιας κατηγορίας, μπορούν να υπάρχουν πολλά όρια απόφασης που αντιστοιχούν σε ένα συγκεκριμένο ρυθμό επιτυχίας, καθένα από τα οποία έχει ένα διαφορετικό ρυθμό λάθους συναγερμού. Προφανώς, στην περίπτωση αυτή δεν μπορεί να καθοριστεί ένα μέτρο για τη διαχωρισιμότητα με βάση μόνο τους ρυθμούς επιτυχίας και λάθους και χωρίς καμία επιπλέον γνώση για τον κανόνα απόφασης.

Στην ιδεατή ιδανική περίπτωση, οι ρυθμοί επιτυχίας και λάθος συναγερμού που έχουν υπολογιστεί είναι οι βέλτιστοι. Για παράδειγμα, από όλους τους κανόνες απόφασης που δίνουν το μετρημένο ρυθμό επιτυχίας χρησιμοποιείται εκείνος που έχει τον ελάχιστο ρυθμό λάθος συναγερμού. Εάν κατασκευαστεί ένας ταξινομητής πολλών διαστάσεων – ανεξάρτητα από τις κατανομές που χρησιμοποιούνται – το πρόβλημα μπορεί να αντιμετωπιστεί με αυτόν τον τρόπο, εάν και θα απαιτούσε πιθανότατα πολλούς υπολογιστικούς πόρους η αναζήτηση των βέλτιστων ρυθμών επιτυχίας και λάθους συναγερμού.



Εικόνα 2.21: Σε μία γενικής μορφής χαρακτηριστική καμπύλη λειτουργίας, ο οριζόντιος άξονας είναι η πιθανότητα ενός λάθος συναγερμού, $P(x \in R_2 / x \in \omega_1)$, ενώ ο κατακόρυφος είναι η πιθανότητα μιας επιτυχίας, $P(x \in R_2 / x \in \omega_2)$. Όπως φαίνεται και στο παραπάνω σχήμα (δεξιά), οι χαρακτηριστικές καμπύλες λειτουργίας δεν είναι γενικά συμμετρικές.

Πρακτικά, χρησιμοποιείται μια μεταβαλλόμενη παράμετρος ελέγχου για τον κανόνα απόφασης και απεικονίζονται γραφικά οι ρυθμοί επιτυχίας και λάθος συναγερμού – η καμπύλη καλείται χαρακτηριστική λειτουργίας. Είναι σύνηθες να χρησιμοποιείται μία παράμετρος ελέγχου η οποία παίρνοντας ακραίες τιμές μπορεί να δώσει είτε ένα πολύ μεγάλο ρυθμό λάθους συναγερμού είτε ένα πολύ μεγάλο ρυθμό επιτυχίας, όπως ακριβώς συμβαίνει με μια πολύ μεγάλη ή μια πολύ μικρή τιμή του x^* στην ROC καμπύλη. Πρέπει να τονιστεί ότι αφού οι κατανομές μπορούν να είναι τυχαίες, η χαρακτηριστική λειτουργίας δεν είναι απαραίτητα συμμετρική (εικόνα 2.21). Σε σπάνιες περιπτώσεις μπορεί ακόμα και να μην είναι καν κοίλα συνεχείς σε όλα τα σημεία.

Οι λειτουργικές καμπύλες ταξινόμησης έχουν αξία σε προβλήματα όπου ο πίνακας κόστους λ_{ij} μπορεί να αλλάζει. Εάν η χαρακτηριστική λειτουργίας έχει οριστεί ως συνάρτηση της παραμέτρου ελέγχου ως προς το χρόνο, είναι εύκολο, όταν η συνάρτηση κόστους αλλάζει, να μεταβληθεί η τιμή της παραμέτρου ελέγχου με τέτοιο τρόπο ώστε να ελαχιστοποιηθεί το αναμενόμενο ρίσκο.

2.9 Θεωρία Απόφασης του Bayes – Διακριτά Χαρακτηριστικά

Στις περιπτώσεις που εξετάστηκαν μέχρι τώρα το διάνυσμα των χαρακτηριστικών x θεωρείτο ότι μπορούσε να είναι οποιοδήποτε σημείο στο d-διάστατο Ευκλείδειο χώρο, R^d. Όμως, σε πολλές πρακτικές εφαρμογές τα στοιχεία του x είναι δυαδικά, ή ανήκουν σε ένα συγκεκριμένο σύνολο τιμών, έτσι ώστε το x να μπορεί να πάρει μόνο κάποια από m διακριτές τιμές $v_1, \ldots, v_m$. Σε τέτοιες περιπτώσεις, η συνάρτηση πυκνότητας πιθανότητας $p(x/\omega_j)$ γίνεται μονοδιάστατη. Ολοκληρώματα της μορφής

$$\int p(x/\omega_j)dx \tag{2.78}$$

πρέπει τότε να αντικατασταθούν από τα αντίστοιχα αθροίσματα

$$\sum_x P(x/\omega_j) \tag{2.79}$$

όπου προφανώς η άθροιση γίνεται πάνω σε όλες τις τιμές του x στην διακριτή κατανομή. Ο τύπος του Bayes στην περίπτωση αυτή περιλαμβάνει πιθανότητες, αντί για συναρτήσεις πυκνότητας πιθανότητας:

$$P(\omega_j/x) = \frac{P(x/\omega_j)P(\omega_j)}{P(x)} \tag{2.80}$$

όπου

$$P(x) = \sum_{j=1}^{c} P(x/\omega_j)P(\omega_j) \tag{2.81}$$

Ο ορισμός του υπό συνθήκη ρίσκου $R(\alpha/x)$ παραμένει ίδιος, όπως ακριβώς γίνεται και με τον κανόνα απόφασης του Bayes:
Για να ελαχιστοποιηθεί το συνολικό ρίσκο, πρέπει να επιλεχθεί η ενέργεια α_i για την οποία το $R(\alpha_i/x)$ είναι ελάχιστο, δηλαδή

$$\alpha^* = \arg\min_i R(\alpha_i/x) \tag{2.82}$$

Ο κύριος κανόνας για την ελαχιστοποίηση του ρυθμού λάθους με μεγιστοποίηση της εκ των υστέρων πιθανότητας παραμένει ίδιος όπως επίσης και οι αντίστοιχες διακρίνουσες συναρτήσεις των εξισώσεων 2.26 – 2.28 με την προφανή αντικατάσταση των πυκνοτήτων $p(\cdot)$ από τις πιθανότητες $P(\cdot)$.

2.9.1 Ανεξάρτητα Δυαδικά Χαρακτηριστικά

Ως ένα παράδειγμα ταξινόμησης που περιλαμβάνει διακριτά χαρακτηριστικά, θεωρείται το πρόβλημα δύο κατηγοριών στο οποίο τα στοιχεία του διανύσματος χαρακτηριστικών παίρνουν δυαδικές τιμές και είναι ανεξάρτητα μεταξύ τους. Πιο συγκεκριμένα, έστω $x = (x_1, \ldots, x_d)^t$, όπου τα στοιχεία x_i είναι είτε 0 είτε 1, με πιθανότητες

$$p_i = Pr[x_i = 1/\omega_1] \tag{2.83}$$

και

$$q_i = Pr[x_i = 1/\omega_2] \tag{2.84}$$

Αυτό είναι ένα μοντέλο για ένα πρόβλημα ταξινόμησης στο οποίο κάθε χαρακτηριστικό γνώρισμα παρέχει μια δυαδική απάντηση (ναι/όχι) για κάθε πρότυπο παρατήρησης. Εάν $p_i > q_i$, αναμένεται ότι το i-οστό χαρακτηριστικό θα δίνει απάντηση «ναι» συχνότερα όταν η κατάσταση της φύσης είναι η ω_1 από όταν η κατάσταση της φύσης είναι η ω_2. Ως παράδειγμα θα παρουσιαστεί το εξής: έστω δύο βιομηχανίες οι οποίες κατασκευάζουν το ίδιο μοντέλο αυτοκίνητου και καθένα από τα d ανταλλακτικά από τα οποία αποτελείται το μοντέλο μπορεί να είναι είτε

κανονικό είτε ελαττωματικό. Εάν ήταν γνωστή η αξιοπιστία των επιχειρήσεων στην κατασκευή κάθε ανταλλακτικού, τότε το παραπάνω μοντέλο θα μπορούσε να χρησιμοποιηθεί για να κρίνει ποια βιομηχανία κατασκεύασε ένα δεδομένο αυτοκίνητο με βάση τη γνώση για το ποια ανταλλακτικά είναι κανονικά και ποια ελαττωματικά. Υποθέτοντας ανεξαρτησία η $P(x/\omega_i)$ μπορεί να γραφεί ως το γινόμενο των πιθανοτήτων για τα στοιχεία του x. Με δεδομένη αυτή την υπόθεση, ένας ιδιαίτερα εύχρηστος τρόπος για να γραφούν οι υπό συνθήκη πιθανότητες των κλάσεων είναι ο ακόλουθος:

$$P(x/\omega_1) = \prod_{i=1}^{d} p_i^{x_i} (1 - p_i)^{1-x_i} \tag{2.85}$$

και

$$P(x/\omega_2) = \prod_{i=1}^{d} q_i^{x_i} (1 - q_i)^{1-x_i} \tag{2.86}$$

Επομένως, ο λόγος πιθανοφάνειας είναι ίσος με

$$\frac{P(x/\omega_1)}{P(x/\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i} \tag{2.87}$$

και η διακρίνουσα συνάρτηση που προκύπτει από την εξίσωση 2.31 είναι η εξής:

$$g(x) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1-p_i}{1-q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)} \tag{2.88}$$

Σημειώνεται ότι η παραπάνω διακρίνουσα συνάρτηση είναι γραμμική ως προς τα x_i και έτσι μπορεί να γραφεί και ως

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0 \tag{2.89}$$

όπου

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \quad i = 1, \ldots, d \tag{2.90}$$

και

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)} \tag{2.91}$$

Στη συνέχεια θα γίνει μια προσπάθεια να ερμηνευτούν τα παραπάνω αποτελέσματα. Ας σημειωθεί πρώτα ότι σύμφωνα με τον κανόνα απόφασης αποφασίζουμε ω_1 εάν $g(x) > 0$ και ω_2 εάν $g(x) \leq 0$. Επίσης, ότι η $g(x)$ αποτελεί έναν με βάρη συνδυασμό των στοιχείων του x. Η τιμή του βάρους w_i υποδεικνύει τη βαρύτητα μιας απάντησης «ναι» για το x_i στον καθορισμό της ταξινόμησής του.

Εάν $p_i = q_i$, το x_i δεν παρέχει καμία πληροφορία για την κατάσταση της φύσης και το w_i είναι ίσο με 0, όπως είναι και το αναμενόμενο.

Εάν $p_i > q_i$, τότε $1 - p_i < 1 - q_i$ και η τιμή του w_i είναι θετική. Έτσι, στην περίπτωση αυτή, μία απάντηση «ναι» για το x_i, συνεισφέρει w_i ψήφους υπέρ της κατάστασης ω_1. Επιπλέον, για κάθε καθορισμένο $q_i < 1$, η τιμή του w_i αυξάνεται με την αύξηση της τιμής του p_i. Από την άλλη πλευρά, εάν $p_i < q_i$, η τιμή του w_i είναι αρνητική και μία απάντηση «ναι» συνεισφέρει $|w_i|$ ψήφους υπέρ της κατάστασης ω_2.

Η ανεξαρτησία των χαρακτηριστικών οδηγεί σε ένα πολύ απλό (γραμμικό) ταξινομητή. Φυσικά, εάν τα χαρακτηριστικά δεν ήταν ανεξάρτητα, θα απαιτούνταν ένας πιο πολύπλοκος ταξινομητής. Γενικότερα, όσο πιο ανεξάρτητα είναι τα

χαρακτηριστικά γνωρίσματα των προς ταξινόμηση προτύπων τόσο πιο απλός μπορεί να είναι ο ταξινομητής που θα χρησιμοποιηθεί.
Η εκ των προτέρων πιθανότητες $P(\omega_i)$ συμμετέχουν στη διακρίνουσα συνάρτηση μόνο μέσα από την τιμή του βάρους κατωφλίου w_0. Αύξηση στην τιμή της $P(\omega_1)$ αυξάνει το w_0 και ωθεί την απόφαση υπέρ της κατάστασης ω_1, ενώ μείωση της $P(\omega_1)$ έχει τα ακριβώς αντίθετα αποτελέσματα. Γεωμετρικά, οι πιθανές τιμές για το x εμφανίζονται ως κόμβοι ενός d-διάστατου υπερκύβου. Η επιφάνεια απόφασης που καθορίζεται από την $g(x) = 0$ είναι ένα υπερεπίπεδο που διαχωρίζει τους κόμβους που ανήκουν στη ω_1 από τους κόμβους που ανήκουν στην ω_2.

Παράδειγμα 3 Όρια απόφασης του Bayes για δυαδικά δεδομένα 3 διαστάσεων

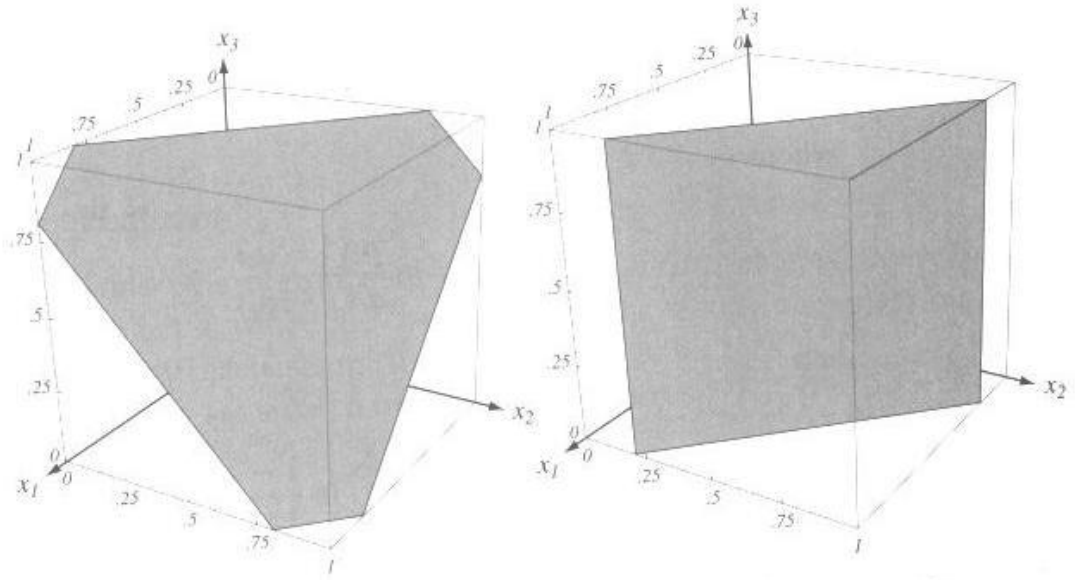
Έστω ένα πρόβλημα δύο κλάσεων, το οποίο έχει τρία ανεξάρτητα δυαδικά χαρακτηριστικά με γνωστές τις πιθανότητες των χαρακτηριστικών αυτών. Θα υπολογιστεί το όριο απόφασης του Bayes για $P(\omega_1)= P(\omega_2)=0.5$, υποθέτοντας ότι για τα ανεξάρτητα στοιχεία ισχύει $p_i=0.8$ και $q_i=0.5$, για $i=1,2,3$. Από τις εξισώσεις (2.90) και (2.91) προκύπτουν οι παρακάτω τιμές για τα βάρη:

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

ενώ η τιμή του κατωφλίου ισούται με:

$$w_o = \sum_{i=1}^{3} \ln\frac{1-0.8}{1-0.5} + \ln\frac{0.5}{0.5} = -1.75.$$

Η επιφάνεια $g(x)=0$ από την εξίσωση (2.89) φαίνεται στην αριστερή γραφική παράσταση της παρακάτω εικόνας. Όπως πιθανώς να είχε κανείς φανταστεί, το όριο τοποθετεί σημεία με δύο ή περισσότερες απαντήσεις "ναι" στην κατηγορία ω_1, διότι αυτή η κατηγορία έχει μεγαλύτερη πιθανότητα να έχει κάποιο χαρακτηριστικό με τιμή 1.



Έστω τώρα ότι, ενώ οι a priori πιθανότητες παραμένουν ίδιες, για τα ανεξάρτητα στοιχεία ισχύει $p_1=p_2=0.8$ και $p_3=0.5$, και $q_1=q_2=q_3=0.5$.
Σε αυτήν την περίπτωση το χαρακτηριστικό x_3 δεν μας δίνει καμία πρόβλεψη για τις κατηγορίες και έτσι το όριο απόφασης είναι παράλληλο με τον άξονα x_3. Σε αυτή την

διακριτή περίπτωση υπάρχει ένα μεγάλο σύνολο θέσεων για το όριο απόφασης, οι οποίες αφήνουν την κατηγοριοποίηση ανεπηρέαστη, όπως φαίνεται και από τη γραφική παράσταση στα δεξιά της παραπάνω εικόνας.

2.10 Βιβλιογραφία

[1] Subutai Ahmad and Volker Tresp. Some solutions to the missing feature problem in vision. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems, volume 5, pages 393-400, Morgan Kaufmann San Mateo, CA, 1993.

[2] Hadar Avi-Itzhak and Thanh Diep. Arbitrarily light upper and lower bounds on the Bayesian probability of error. IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-18(1):89-91, 1996.

[3] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society (London), 53:370-418. 1763.

[4] James O. Berger. Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. Journal of Multivariate Analysis, 6(2);256-264, 1976.

[5] James O. Berger. Selecting a minimax estimator of a multivariate normal mean. Annals of Statistics, 10(1):81-92, 1982.

[6] James O. Berger. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York, second

[7] Jose M. Bernardo and Adrian F. M. Smith. Bayesian Theory. Wiley, New York. 1996.

[8] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society, 35:93-110,1943.

[9] Wray L. Buntine. Operations for learning with graphical models. Journal of Artificial Intelligence Research, 2:159-225, 1994.

[10] Wray L. Buntine. A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 8(2):l95-2\0, 1996.

[11] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics, 23:493-507, 1952.

[12] Chao K. Chow. An optimum character recognition system using decision functions. IRE Transactions, pages 247-254, 1957.

[13] Chao K. Chow, On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, IT-16:41-46, 1970.

[14] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991.

[15] Morris H. DeGroot. Optimal Statistical Decisions. McGraw-Hill. New York, 1970.

[16] Bradley Efron and Carl Morris. Families of minimax estimators of the mean of a multivariate normal distribution. Annals of Statistics, 4:11-21, 1976.

[17] Thomas S. Ferguson. Mathematical Statistics: A Decision Theoretic Approach. Academic Press, New York, 1967.

[18] Simon French. Decision Theory: An Introduction to the Mathematics of Rationality. Halsted Press, New York. 1986.

[19] Keinosuke Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, second edition, 1990.

[20] Keinosuke Fukunaga and Thomas F. Krile. Calculation of Bayes recognition error for two multivariate Gaussian distributions. IEEE Transactions on Computers, C-18:220-229, 1969.

[21] Izrail M. Gelfand and Sergei Vasilevich Fomin. Calculus of Variations, Prentice-Hall, Englewood Cliffs, NJ, translated from the Russian by Richard A. Silverman, 1963.

[22] David M. Green and John A. Swets. Signal Detection Theory and Psychophysics. Wiley, New York. 1974.

[23] David J. Hand. Construction and Assessment of Classifi-cation Rules. Wiley, New York, 1997.

[24] Peter E. Hart and Jamey Graham. Query-free information retrieval. IEEE Expert: Intelligent Systems and Their Application, 12(5):32-37,1997.
[25] David Heckerman. Probabilistic Similarity Networks. ACM Doctoral Dissertation Award Series. MIT Press, Cambridge, MA, 1991.
[26] Anil K. Jain. On an estimate of the Bhattacharyya distance. IEEE Transactions on Systems, Man and Cybernetics, SMC-16(11):763-766, 1976.
[27] Michael I. Jordan, editor. Learning in Graphical Models. MIT Press, Cambridge, MA, 1999.
[28] Bernard Kolman. Elementary Linear Algebra. Macmil-lan. New York, fifth edition, 1991.
[29] Pierre Simon Laplace. Theorie Analytique des Proba-billies. Courcier, Paris, France. 1812.
[30] Peter M Lee. Bayesian Statistics: An Introduction. Edward Arnold, London, 1989.
[31] Dennis V. Lindley. Making Decisions. Wiley, New York, 1991.
[32] Jerzy Neyman and Egon S. Pearson, On the problem of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society, London, 231:289-337, 1928.
[33] Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo. CA, 1988.
[34] Sheldon M. Ross. Introduction to Probability and Statistics for Engineers. Wiley, New York, 1987.
[35] Donald B. Rubin and Roderick J. A. Little. Statistical Analysis with Missing Data. Wiley, New York, 1987.
[36] Claude E. Shannon. A mathematical theory of communication. Bell Systems Technical Journal, 27:379-423. 623-656,1948.
[37] George B. Thomas, Jr. and Ross L. Finney. Calculus and Analytic Geometry. Addison-Wesley, New York, ninth edition, 1996.