Ταξινόμηση

Ορισμός του προβλήματος

✓ Ο στόχος της ταξινόμησης είναι να καταχωρήσει ένα αντικείμενο (πρότυπο) σε μία από ένα σύνολο δυνατών κλάσεων των οποίων ο αριθμός είναι γνωστός.

Παραδείγματα:

- Μαστογραφία με ακτίνες-Χ: ταξινόμηση των περιοχών της εικόνας σε αυτές που προέρχονται από καλοήθη ιστό (κλάση 1) και σε αυτές που αντιστοιχούν σε κακοήθεια (κλάση 2)
- Οπτική αναγνώριση χαρακτήρων: η εικόνα που αντιστοιχεί σε κάθε γράμμα του αλφάβητου πρέπει να αναγνωριστεί και να αντιστοιχιστεί σε μια από 24 κλάσεις (για το ελληνικό αλφάβητο).
- Αναγνώριση του συγγραφέα ενός κειμένου: σύστημα που επεξεργάζεται ένα κείμενο και αναγνωρίζει τον συγγραφέα του από ένα σύνολο συγγραφέων (κλάσεις)
- Ισοστάθμιση τηλεπικοινωνιακού καναλιού: ταξινόμηση του σήματος στο δέκτη σε μια από τις κλάσεις που αντιστοιχούν σε όλα τα πιθανά σύμβολα που μπορούν να εκπεμφθούν από τον πομπό.

Τα στάδια της ταξινόμησης

1. Αναπαράσταση δεδομένων

- Το πρώτο βήμα σε μια διαδικασία ταξινόμησης είναι να αποφασίσουμε πως θα αναπαραστήσουμε κάθε πρότυπο στον υπολογιστή. Η πληροφορία που υπάρχει στα ακατέργαστα δεδομένα (raw data) θα πρέπει να κωδικοποιηθεί αποτελεσματικά.
- Η αναπαράσταση των δεδομένων συνήθως πραγματοποιείται μετασχηματίζοντας τα ακατέργαστα δεδομένα σε κάποιο άλλο χώρο, όπου κάθε πρότυπο αναπαρίσταται από ένα διάνυσμα, $\boldsymbol{x} \in \mathbb{R}^l$. Αυτό είναι γνωστό ως διάνυσμα χαρακτηριστικών και τα l στοιχεία του $\boldsymbol{x}$ ονομάζονται χαρακτηριστικά (features). Έτσι κάθε πρότυπο γίνεται ένα σημείο σε έναν l-διάστατο χώρο, τον χώρο των χαρακτηριστικών (feature space).
- Αυτή η προ-επεξεργασία ονομάζεται στάδιο γέννησης χαρακτηριστικών (feature generation). Συνήθως ξεκινάμε με ένα μεγάλο αριθμό K χαρακτηριστικών και επιλέγουμε τα l χαρακτηριστικά που είναι “πιο πλούσια” σε πληροφορία (επιλογή χαρακτηριστικών-feature selection).

Τα στάδια της ταξινόμησης

2. Σχεδιασμός του ταξινομητή μέσω εκπαίδευσης

- Έχοντας αποφασίσει για το χώρο των χαρακτηριστικών, το επόμενο βήμα είναι να εκπαιδεύσουμε έναν ταξινομητή (classifier) δηλαδή ένα προβλέπτη (predictor). Αυτό επιτυγχάνεται επιλέγοντας ένα σύνολο δεδομένων των οποίων οι κλάσεις είναι γνωστές. Αυτά είναι τα δεδομένα εκπαίδευσης (training set).

- Τα δεδομένα εκπαίδευσης είναι ένα σύνολο από ζεύγη $(\boldsymbol{x}_n, y_n), n = 1,2, \dots, N$, όπου η μεταβλητή (εξόδου) y_n δείχνει την κλάση στην οποία ανήκει το $\boldsymbol{x}_n$ (είσοδος) και είναι γνωστή ως ετικέτα της κλάσης (class label). Οι ετικέτες, y, παίρνουν διακριτές τιμές, π.χ. $y \in \{1,2, \dots, M\}$ για ένα πρόβλημα ταξινόμησης με M κλάσεις.

- Με βάση το σύνολο δεδομένων εκπαίδευσης, μπορεί κανείς να σχεδιάσει μια συνάρτηση $f(\boldsymbol{x})$ η οποία μπορεί να προβλέψει την έξοδο όταν της δοθεί η είσοδος. Η συνάρτηση αυτή ονομάζεται ταξινομητής.

Τα στάδια της ταξινόμησης

3. Χρήση του ταξινομητή για προβλέψεις

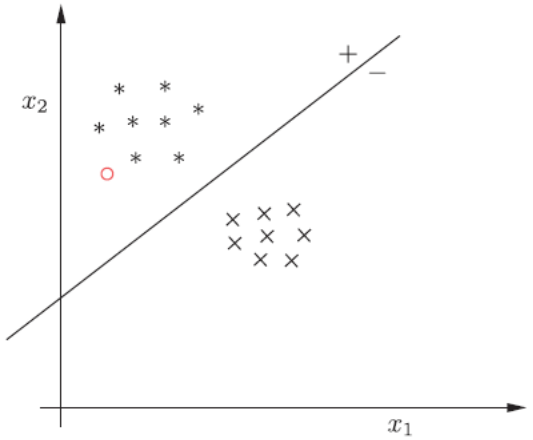
- Όταν ο σχεδιασμός του ταξινομητή έχει ολοκληρωθεί, το σύστημα είναι έτοιμο να πραγματοποιήσει προβλέψεις. Αν λοιπόν δοθεί ένα άγνωστο πρότυπο από τα ακατέργαστα δεδομένα, σχηματίζουμε το αντίστοιχο διάνυσμα χαρακτηριστικών, $\boldsymbol{x}$, και ανάλογα με την τιμή της $\hat{y} = f(\boldsymbol{x})$, το πρότυπο ταξινομείται σε μία από τις δυνατές κλάσεις.
- Για παράδειγμα σε ένα δυαδικό πρόβλημα ταξινόμησης (2 κλάσεις), με $y \in \{1, -1\}$, η ετικέτα του $\boldsymbol{x}$ μπορεί να προβλεφθεί από το $\hat{y} = \mathrm{sgn}\{f(\boldsymbol{x})\}$.

Παράδειγμα ταξινόμησης

- Στο παρακάτω σχήμα φαίνεται η διαδικασία της ταξινόμησης. Αρχικά μας δίνεται ένα σύνολο από σημεία εκπαίδευσης στον δισδιάστατο χώρο (χρησιμοποιούνται δύο χαρακτηριστικά, x_1, x_2). Τα σημεία-αστέρια ανήκουν στην μία κλάση (ω_1) και τα σημεία-σταυροί στην άλλη κλάση (ω_2), σε ένα δυαδικό πρόβλημα ταξινόμησης. Με βάση αυτά τα σημεία "μαθαίνουμε" τον ταξινομητή. Στην απλή αυτή περίπτωση, ο ταξινομητής είναι μια γραμμική συνάρτηση,

$$f(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2, \quad \boldsymbol{x} = [x_1, x_2]^T$$

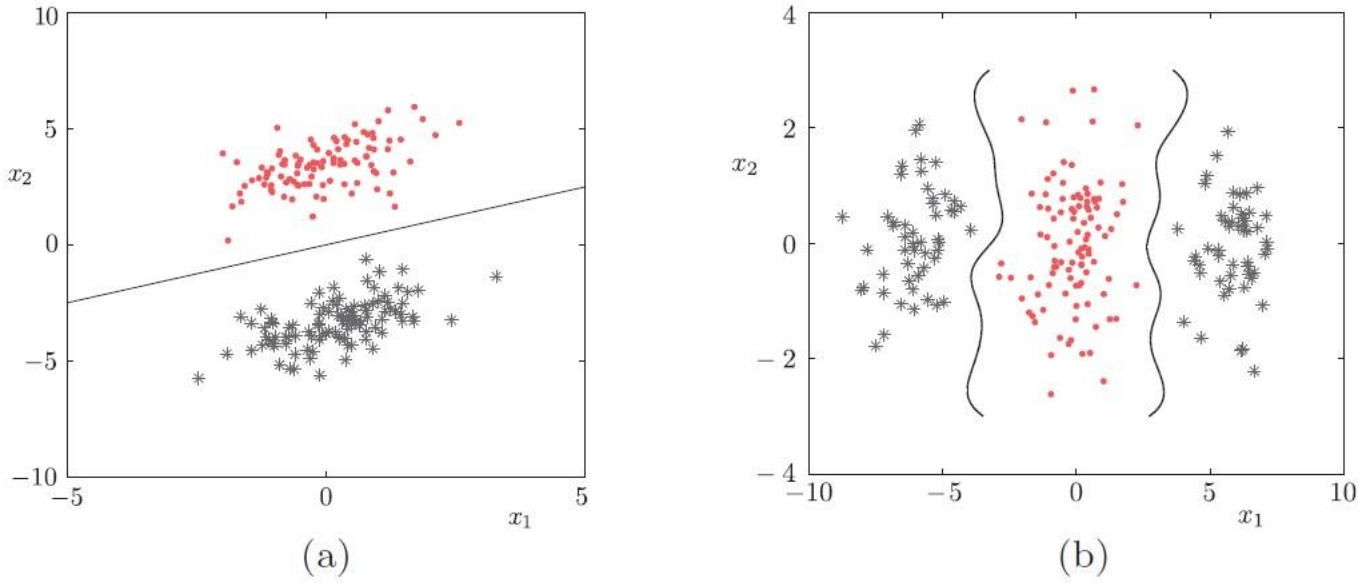
Η γραφική παράσταση της συνάρτησης $f(\boldsymbol{x}) = 0$ είναι η ευθεία που φαίνεται στο σχήμα.



Ο ταξινομητής (γραμμικός σε αυτό το παράδειγμα) σχεδιάστηκε έτσι ώστε να ξεχωρίζει τα σημεία εκπαίδευσης σε δύο κλάσεις, έχοντας στη θετική του πλευρά τα σημεία που ανήκουν στη μία κλάση και στην αρνητική του πλευρά τα σημεία της δεύτερης κλάσης.
Το "κόκκινο" σημείο, του οποίου η κλάση είναι άγνωστη, ταξινομείται στην ίδια κλάση με τα σημεία-αστέρια, καθώς βρίσκεται στη **θετική πλευρά** του ταξινομητή

Ταξινομητής ελαχίστων τετραγώνων

- Όπως έχουμε αναφέρει, στην ταξινόμηση στόχος μας είναι ο σχεδιασμός ενός ταξινομητή, δηλαδή μιας συνάρτησης $f(\boldsymbol{x})$, ή ισοδύναμα μιας επιφάνειας απόφασης (decision surface) $f(\boldsymbol{x}) = 0$ στο $\mathbb{R}^l$. Η επιφάνεια αυτή θα πρέπει να διαχωρίζει το χώρο στον οποίο βρίσκονται τα διανύσματα χαρακτηριστικών σε περιοχές και να αντιστοιχίζει κάθε περιοχή σε μία κλάση.
- Στην περίπτωση της γραμμικής ταξινόμησης η "επιφάνεια" απόφασης είναι ένα σημείο ($l = 1$), μια ευθεία ($l = 2$), ένα επίπεδο ($l = 3$) ή γενικότερα ένα υπερεπίπεδο ($l > 3$).



Παραδείγματα ταξινόμησης σε δύο κλάσεις. a) Περίπτωση γραμμικά διαχωρίσιμων κλάσεων.
b) Μη-γραμμικά διαχωρίσιμες κλάσεις.

Ταξινομητής ελαχίστων τετραγώνων

Έστω ότι μας δίνεται ένα σύνολο προτύπων εκπαίδευσης $\boldsymbol{x}_n \in \mathbb{R}^l, n = 1,2, \dots, N$ που ανήκουν σε κάποια από δύο κλάσεις, έστω ω_1 και ω_2. Στόχος μας είναι να σχεδιάσουμε ένα υπερεπίπεδο

$$f(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_l x_l = \boldsymbol{\theta}^T \boldsymbol{x} = 0 \quad \text{όπου}$$

$$\boldsymbol{\theta} = [\theta_0, \theta_1, \dots, \theta_l]^T \quad \text{και} \quad \boldsymbol{x} = [1, x_1, \dots, x_l]^T$$

Δηλαδή θέλουμε να βρούμε το διάνυσμα $\boldsymbol{\theta}$, ώστε το υπερεπίπεδο να τοποθετηθεί κατάλληλα ανάμεσα στις δύο κλάσεις. Για τα σημεία πάνω στο υπερεπίπεδο προφανώς ισχύει $f(\boldsymbol{x}) = 0$, ενώ για τα σημεία που βρίσκονται στις δύο πλευρές του θα είναι είτε $f(\boldsymbol{x}) > 0$, είτε $f(\boldsymbol{x}) < 0$. Θα πρέπει λοιπόν να εκπαιδεύσουμε με τέτοιο τρόπο τον ταξινομητή μας ώστε για τα σημεία που ανήκουν στην κλάση ω_1 η f να είναι θετική και για τα σημεία που ανήκουν στην κλάση ω_2 η f να είναι αρνητική.

Ταξινομητής ελαχίστων τετραγώνων

Λύση. Δίνουμε ετικέτα 1 σε όλα τα σημεία που ανήκουν στην κλάση ω_1 ($y_n = 1, \forall n: \boldsymbol{x}_n \in \omega_1$) και $y_n = -1, \forall n: \boldsymbol{x}_n \in \omega_2$. Στη συνέχεια ελαχιστοποιούμε το παρακάτω κόστος ελαχίστων τετραγώνων:

$$J(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2$$

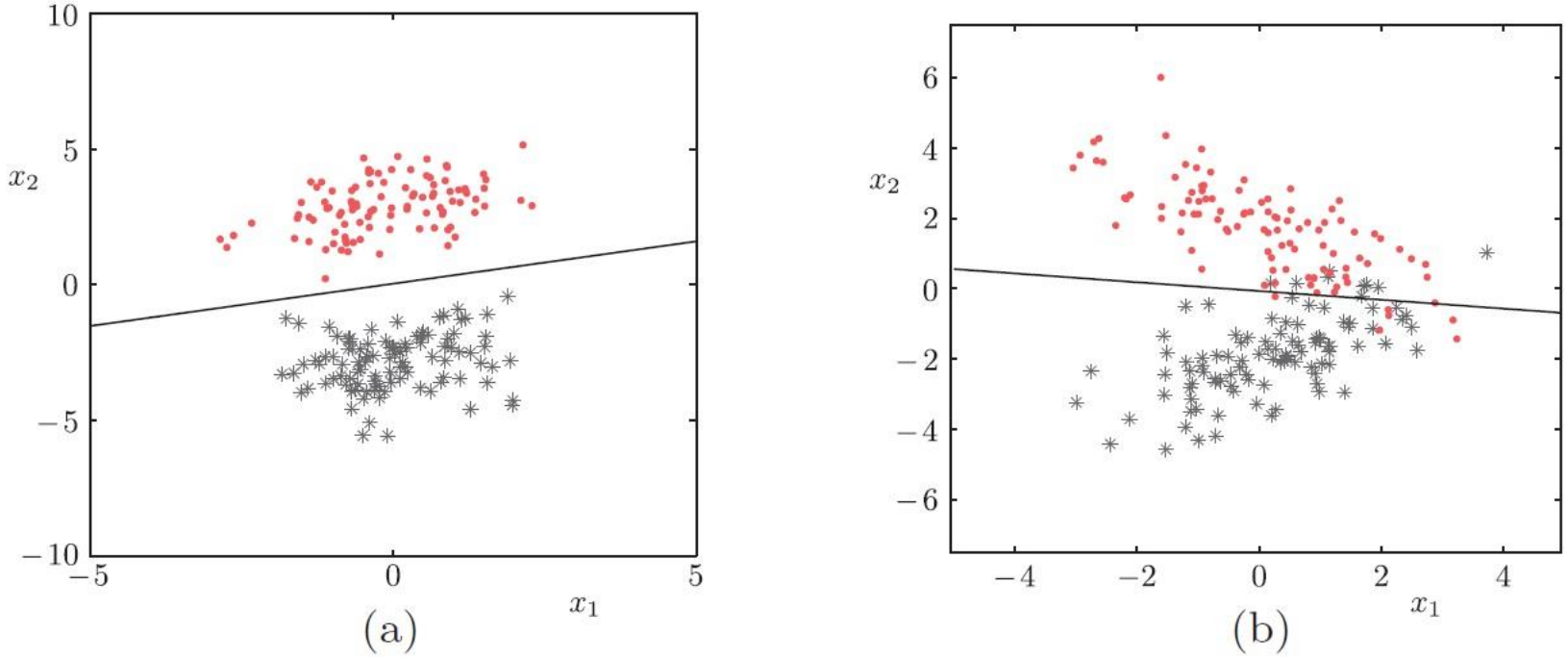
Η ελαχιστοποίηση της $J(\boldsymbol{\theta})$ δίνει:

$$\widehat{\boldsymbol{\theta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

$$X = \begin{bmatrix} 1 & \boldsymbol{x}_1^T \\ 1 & \boldsymbol{x}_2^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_N^T \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Ταξινομητής ελαχίστων τετραγώνων

Η μέθοδος μπορεί να είναι αποτελεσματική αν έχουμε γραμμικά διαχωρίσιμες κλάσεις. Διαφορετικά η επίδοση της μεθόδου χειροτερεύει.



Σχεδιασμός ενός γραμμικού ταξινομητή $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$ με βάση το κριτήριο ελαχίστων τετραγώνων. (a) Γραμμικά διαχωρίσμες κλάσεις, (b) Μη-γραμμικά διαχωρίσιμες κλάσεις. Στη δεύτερη περίπτωση ο ταξινομητής δεν μπορεί να διαχωρίσει πλήρως τις δύο κλάσεις. Το καλύτερο που μπορεί να κάνει είναι να τοποθετήσει την επιφάνεια απόφασης ώστε να ελαχιστοποιήσει τη διακύμανση μεταξύ των πραγματικών ετικετών και των προβλεπόμενων τιμών εξόδου $\hat{y} = \boldsymbol{x}^T \widehat{\boldsymbol{\theta}}$ με την έννοια των ελαχίστων τετραγώνων.

Ο ταξινομητής Bayes

- Στη συνέχεια θα δούμε το πρόβλημα της ταξινόμησης με βάση τη θεωρία αποφάσεων κατά Bayes (Bayesian decision theory). Στην καρδιά της προσέγγισης αυτής βρίσκεται η εκτίμηση της από κοινού ΣΠΠ $p(\boldsymbol{x}, y), y \in D, x \in \mathbb{R}^l$, όπου D είναι ένα διακριτό σύνολο ετικετών κλάσεων.

- **Bayesian κανόνας ταξινόμησης**: Έστω ότι μας δίνεται ένα σύνολο από M κλάσεις $\omega_i, i = 1,2, \dots, M$ καθώς και οι αντίστοιχες εκ των υστέρων (posterior) πιθανότητες, $P(\omega_i|\boldsymbol{x})$. Ταξινόμησε το διάνυσμα χαρακτηριστικών, $\boldsymbol{x}$, σύμφωνα με τον κανόνα,

$$\text{Ταξινόμησε το } \boldsymbol{x} \text{ στην κλάση } \omega_i = \arg\max_{\omega_j} P(\omega_j|\boldsymbol{x}), \quad j = 1,2, \dots, M$$

- Δηλαδή, το άγνωστο πρότυπο, που αναπαρίσταται με $\boldsymbol{x}$, καταχωρείται στην κλάση για την οποία η posterior πιθανότητα γίνεται μέγιστη.

Ο ταξινομητής Bayes

- Θα πρέπει να σημειωθεί ότι πριν λάβουμε οποιαδήποτε νέα παρατήρηση, η αβεβαιότητα σε ό,τι αφορά τις κλάσεις εκφράζεται μέσω των εκ των προτέρων (prior) πιθανοτήτων που συμβολίζονται με $P(\omega_i), i = 1,2, \dots, M$.

- Από τη στιγμή που παίρνουμε μια νέα παρατήρηση $\boldsymbol{x}$, αυτή η επιπλέον πληροφορία απομειώνει την αρχική μας αβεβαιότητα, και η σχετική στατιστική πληροφορία παρέχεται τώρα μέσω των posterior πιθανοτήτων, $P(\omega_i|\boldsymbol{x}), i = 1,2, \dots, M$, οι οποίες χρησιμοποιούνται για την ταξινόμηση του $\boldsymbol{x}$.

Ο ταξινομητής Bayes

Από το θεώρημα Bayes έχουμε

$$P(\omega_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_j)P(\omega_j)}{P(\boldsymbol{x})}$$

όπου $P(\boldsymbol{x}|\omega_j)$ είναι οι σχετικές υπό συνθήκη (δεσμευμένες) ΣΠΠ. Με βάση το παραπάνω ο κανόνας ταξινόμησης κατά Bayes γίνεται:

Ταξινόμησε το $\boldsymbol{x}$ στην κλάση $\omega_i = \arg\max_{\omega_j} P(\boldsymbol{x}|\omega_j)P(\omega_j)\,, \quad j = 1,2,\dots,M$

Δηλαδή, ο ταξινομητής εξαρτάται από τις a-priori πιθανότητες των κλάσεων καθώς και από τις σχετικές δεσμευμένες ΣΠΠ.

Αξίζει να σημειωθεί επίσης ότι

$$P(\boldsymbol{x}|\omega_j)P(\omega_j) = P(\omega_j, \boldsymbol{x}) := P(y, \boldsymbol{x})$$

Ο ταξινομητής Bayes

- Εκπαίδευση των Bayesian ταξινομητών: Ας υποθέσουμε ότι μας δίνεται ένα σύνολο από σημεία εκπαίδευσης, $(y_n, \boldsymbol{x}_n) \in D \times \mathbb{R}^l, n = 1,2, \dots, N$ και ας θεωρήσουμε το γενικό πρόβλημα ταξινόμησης σε M κλάσεις. Έστω ότι κάθε κλάση $\omega_i, i = 1,2, \dots, M$ αντιπροσωπεύεται με N_i σημεία στο σύνολο εκπαίδευσης, με $\sum_{i=1}^{M} N_i = N$. Τότε, οι a-priori πιθανότητες μπορούν να προσεγγιστούν ως εξής:

$$P(\omega_i) \approx \frac{N_i}{N}, \qquad i = 1,2, \dots, M$$

- Για τις υπό συνθήκη ΣΠΠ $p(\boldsymbol{x}|\omega_i), i = 1,2, \dots, M$, μπορούμε να χρησιμοποιήσουμε κάθε μέθοδο που μπορεί να εκτιμήσει ΣΠΠ με βάση τα δεδομένα εκπαίδευσης για κάθε μια από τις κλάσεις. Για παράδειγμα, η μέθοδος μεγίστης πιθανοφάνειας ή εναλλακτικά μη-παραμετρικές τεχνικές που βασίζονται στο ιστόγραμμα.

Ο ταξινομητής Bayes ελαχιστοποιεί την πιθανότητα σφάλματος ταξινομησης

- Υπενθυμίζεται ότι ο στόχος κάθε ταξινομητή είναι να διαμερίσει το χώρο, στον οποίο βρίσκονται τα χαρακτηριστικά διανύσματα, σε περιοχές και να αντιστοιχίσει κάθε μία από τις περιοχές σε μία και μόνο μία κλάση.
- Έστω ένα πρόβλημα με δύο κλάσεις και έστω $\mathcal{R}_1, \mathcal{R}_2$ δύο περιοχές στον $\mathbb{R}^l$, μέσα στις οποίες αποφασίζουμε υπέρ των κλάσεων ω_1 και ω_2, αντίστοιχα. Τότε, η πιθανότητα σφάλματος ταξινόμησης δίνεται από την ακόλουθη σχέση:

$$P_e = P(\boldsymbol{x} \in \mathcal{R}_1, \boldsymbol{x} \in \omega_2) + P(\boldsymbol{x} \in \mathcal{R}_2, \boldsymbol{x} \in \omega_1).$$

Δηλαδή, ισούται με την πιθανότητα ένα χαρακτηριστικό διάνυσμα να ανήκει στην κλάση $\omega_1(\omega_2)$ και ταυτόχρονα να βρίσκεται στην "λάθος" περιοχή $\mathcal{R}_2(\mathcal{R}_1)$ στο χώρο των χαρακτηριστικών.

Ο ταξινομητής Bayes ελαχιστοποιεί την πιθανότητα σφάλματος ταξινομησης

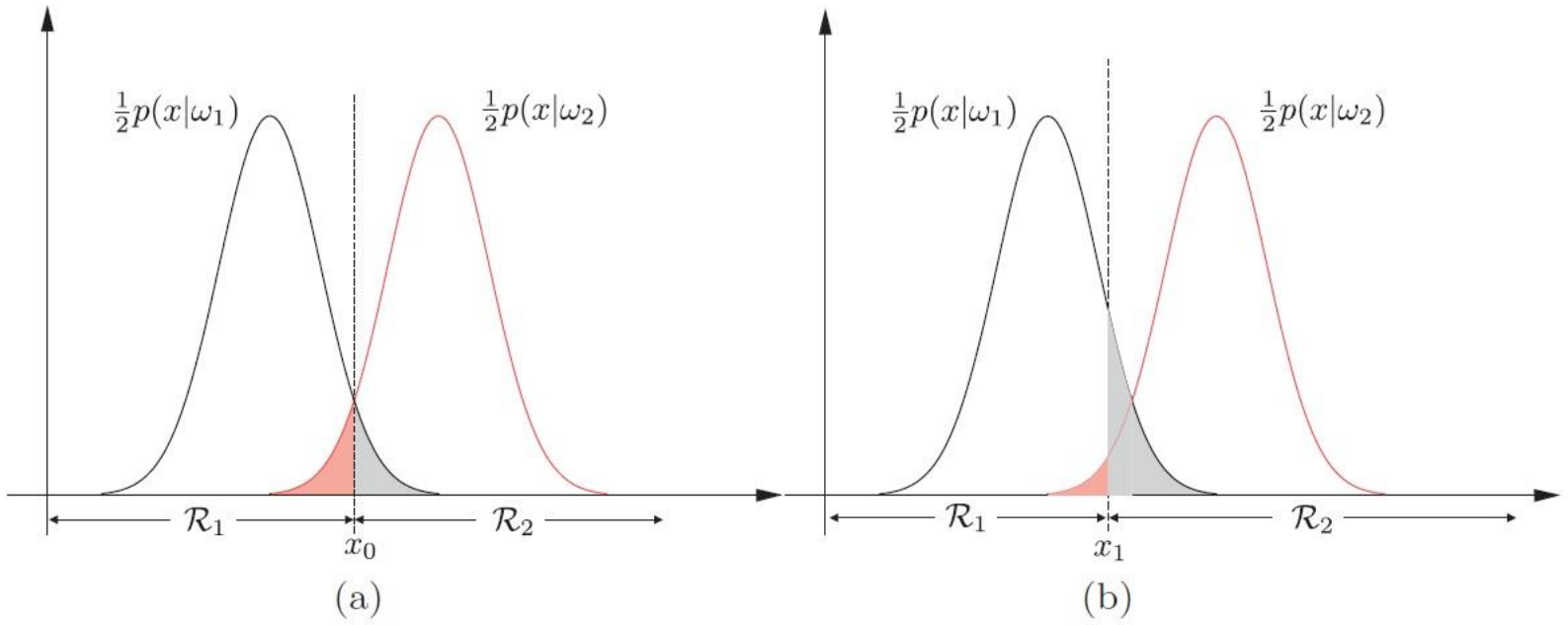
- Η προηγούμενη εξίσωση γράφεται ως εξής:

$$P_e = P(\omega_2)\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2)d\boldsymbol{x} + P(\omega_1)\int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1)d\boldsymbol{x}$$

- Μπορεί να δειχτεί ότι ο ταξινομητής Bayes ελαχιστοποιεί την πιθανότητα σφάλματος ταξινόμησης P_e ως προς τα $\mathcal{R}_1$ και $\mathcal{R}_2$. Αυτό ισχύει και στο γενικότερο πρόβλημα με M κλάσεις.

Ο ταξινομητής Bayes ελαχιστοποιεί την πιθανότητα εσφαλμένης ταξινομησης

- Η ιδιότητα του ταξινομητή να ελαχιστοποιεί το P_e γίνεται κατανοητή από το παρακάτω σχήμα.



(a) Για δύο ισοπίθανες κλάσεις, η πιθανότητα σφάλματος ταξινόνησης κατά τη διαμέριση του χώρου των χαρακτηριστικών από έναν βέλτιστο Bayesian ταξινομητή ισούται με το εμβαδόν της σκιασμένης περιοχής. (b) Αν μετακινήσουμε την τιμή κατωφλίου μακριά από την τιμή που αντιστοιχεί στον βέλτιστο κανόνα κατά Bayes, η πιθανότητα σφάλματος αυξάνεται, όπως προκύπτει από την αύξηση του εμβαδού της σκιασμένης περιοχής.

Ελαχιστοποιώντας το μέσο ρίσκο

- Η πιθανότητα σφάλματος ταξινόμησης δεν είναι πάντα το καλύτερο κριτήριο που μπορεί να χρησιμοποιηθεί για ταξινόμηση. Αυτό συμβαίνει γιατί αποδίδει την ίδια σπουδαιότητα σε όλα τα σφάλματα. Ωστόσο υπάρχουν περιπτώσεις, στις οποίες μερικές εσφαλμένες αποφάσεις μπορεί να έχουν πιο σοβαρές επιπτώσεις απ’ ότι άλλες.

- Για παράδειγμα, είναι πολύ πιο σοβαρό για ένα γιατρό να λάβει μια λάθος απόφαση που θα έχει ως αποτέλεσμα να διαγνωσθεί ένας κακοήθης όγκος ως καλοήθης, παρά το αντίστροφο. Αν ένας καλοήθης όγκος διαγνωσθεί ως κακοήθης, η εσφαλμένη απόφαση θα διορθωθεί σε επόμενες κλινικές εξετάσεις. Αντίθετα, τα αποτελέσματα από μια λανθασμένη διάγνωση για έναν κακοήθη όγκο μπορεί να είναι μοιραία.

Ελαχιστοποιώντας το μέσο ρίσκο

- Σε τέτοιες περιπτώσεις είναι ορθότερο να ορίζεται ένας παράγοντας ποινής (penalty) που θα σταθμίζει κάθε σφάλμα ανάλογα με τη σπουδαιότητά του. Η συνάρτηση που προκύπτει ονομάζεται μέσο ρίσκο (average risk).
- Για την περίπτωση δύο κλάσεων, το ρίσκο που σχετίζεται με κάθε μια από τις κλάσεις ορίζεται ως εξής:

$$r_1 = \lambda_{11} \int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x} + \lambda_{12} \underbrace{\int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_1) d\boldsymbol{x}}_{\sigma\varphi\alpha\lambda\mu\alpha}$$

$$r_2 = \lambda_{21} \underbrace{\int_{\mathcal{R}_1} p(\boldsymbol{x}|\omega_2) d\boldsymbol{x}}_{\sigma\varphi\alpha\lambda\mu\alpha} + \lambda_{22} \int_{\mathcal{R}_2} p(\boldsymbol{x}|\omega_2) d\boldsymbol{x}$$

- Συνήθως $\lambda_{11} = \lambda_{22} = 0$, καθώς αντιστοιχούν σε ορθές αποφάσεις.

Ελαχιστοποιώντας το μέσο ρίσκο

- Το μέσο ρίσκο που πρέπει να ελαχιστοποιηθεί δίνεται από τη σχέση:

$$r = P(\omega_1)r_1 + P(\omega_2)r_2$$

- Για δύο κλάσεις, είναι φανερό ότι ο κανόνας ταξινόμησης γίνεται:

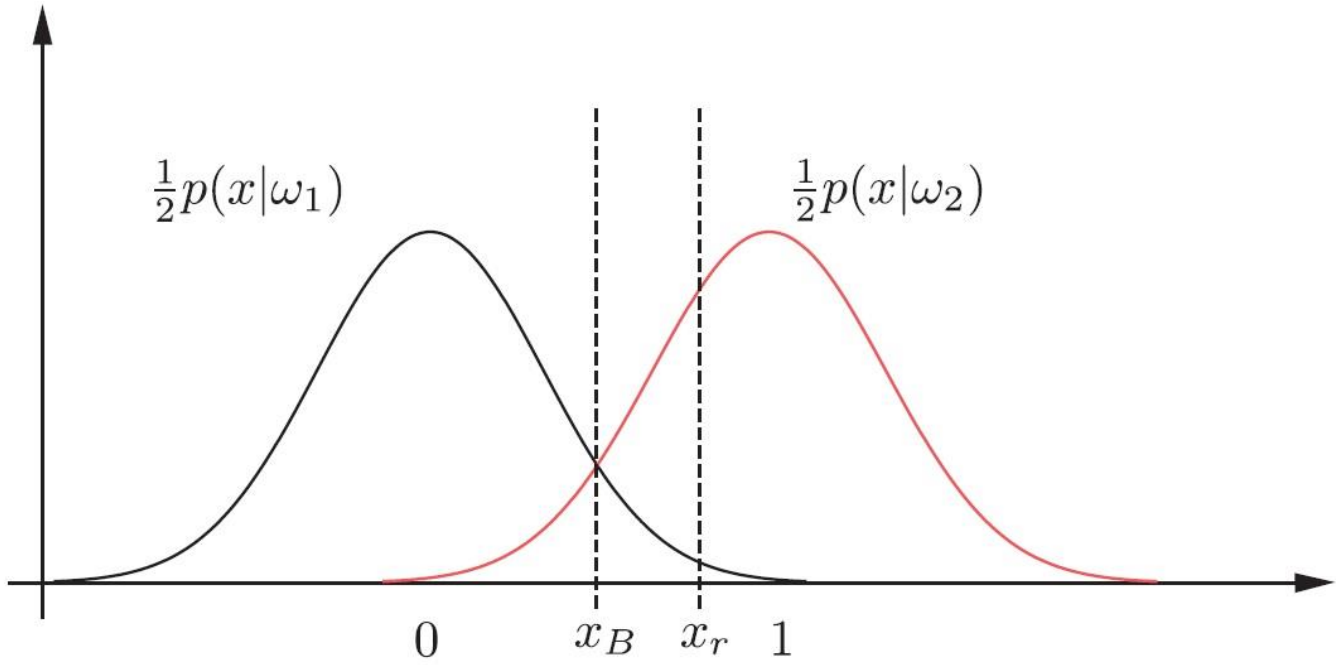
Ταξινόμησε το $\boldsymbol{x}$ στην κλάση $\omega_1(\omega_2)$ αν: $\lambda_{12}P(\omega_1|\boldsymbol{x}) > (<)\lambda_{21}P(\omega_2|\boldsymbol{x})$

ή ισοδύναμα

Ταξινόμησε το $\boldsymbol{x}$ στην κλάση $\omega_1(\omega_2)$ αν: $\underbrace{\lambda_{12}P(\omega_1)}_{P'(\omega_1)}P(\boldsymbol{x}|\omega_1) > (<)\ \underbrace{\lambda_{21}P(\omega_2)}_{P'(\omega_2)}P(\boldsymbol{x}|\omega_2)$

- Θα πρέπει να σημειωθεί ότι αν το λ_{12} είναι μεγάλο, αυτό σημαίνει ότι η κλάση ω_1 είναι πιο “σημαντική” και αυτό είναι ισοδύναμο με το να αυξήσουμε σχετικά την a-priori πιθανότητα της κλάσης ω_1 σε σχέση με την κλάση ω_2 $\left(\frac{P'(\omega_1)}{P'(\omega_2)} > \frac{P(\omega_1)}{P(\omega_2)}\right)$

Παράδειγμα Bayesian ταξινόμησης



- Η ελαχιστοποίηση του μέσου ρίσκου μεγαλώνει την περιοχή στην οποία αποφασίζουμε υπέρ της πιο σημαντικής κλάσης, ω_1.

Υπερεπιφάνειες απόφασης

- Ο στόχος κάθε ταξινομητή είναι να διαμερίσει το χώρο των χαρακτηριστικών σε περιοχές. Η διαμέριση επιτυγχάνεται με σημεία στον $\mathbb{R}$, καμπύλες στον $\mathbb{R}^2$, επιφάνειες στον $\mathbb{R}^3$και υπερεπιφάνειες στον $\mathbb{R}^l$. Κάθε υπερεπιφάνεια, S, εκφράζεται με βάση μια συνάρτηση,

$$g: \mathbb{R}^l \longmapsto \mathbb{R}$$

  και περιλαμβάνει όλα τα σημεία για τα οποία ισχύει:

$$S = \{\boldsymbol{x} \in \mathbb{R}^l : g(\boldsymbol{x}) = 0\}$$

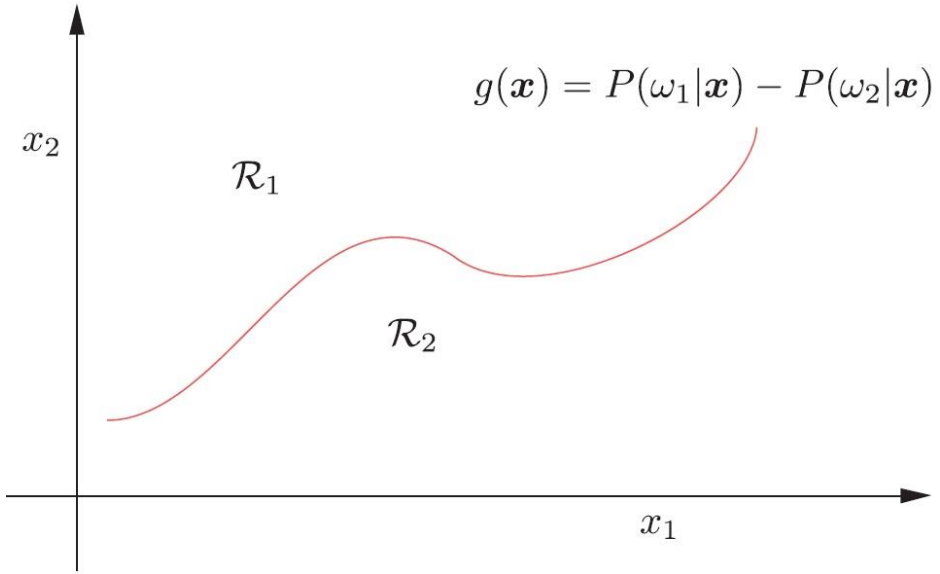
- Υπενθυμίζουμε ότι για όλα τα σημεία που βρίσκονται στην μία πλευρά της υπερεπιφάνειας ισχύει $g(\boldsymbol{x}) > 0$ και για αυτά που βρίσκονται στην άλλη πλευρά $g(\boldsymbol{x}) < 0$. Η υπερεπιφάνεια που προκύπτει ονομάζεται υπερεπιφάνεια απόφασης.

Υπερεπιφάνειες απόφασης

- Για παράδειγμα, η υπερεπιφάνεια απόφασης που σχηματίζεται από τον Bayesian ταξινομητή για ένα πρόβλημα δύο κλάσεων είναι:

$$g(\boldsymbol{x}) \coloneqq P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x}) = 0$$

- Πράγματι, αποφασίζουμε υπερ της κλάσης ω_1, αν το $\boldsymbol{x}$ βρίσκεται στη θετική πλευρά της παραπάνω υπερεπιφάνειας και υπερ της κλάσης ω_2, για σημεία που απαντώνται στην αρνητική πλευρά της (περιοχές $\mathcal{R}_1$ και $\mathcal{R}_2$ στο παρακάτω σχήμα).



Η περίπτωση της κανονικής κατανομής

- Ας υποθέσουμε ότι τα δεδομένα σε κάθε κλάση ακολουθούν την κανονική κατανομή,

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right), \quad i = 1, 2, \ldots, M.$$

- Ορίζουμε για κάθε κλάση τη συνάρτηση

$$g_i(\boldsymbol{x}) := \ln P(\omega_i|\boldsymbol{x}) = \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i), \quad i = 1, 2, \ldots, M,$$

συναρτήσεις διάκρισης

- Για δύο κλάσεις η υπερεπιφάνεια απόφασης θα είναι:

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) = 0,$$

$$g(\boldsymbol{x}) = \underbrace{\frac{1}{2}\left(\boldsymbol{x}^{\mathrm{T}}\Sigma_2^{-1}\boldsymbol{x} - \boldsymbol{x}^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{x}\right)}_{\text{quadratic terms}}$$

$$\underbrace{+\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{x} - \boldsymbol{\mu}_2^{\mathrm{T}}\Sigma_2^{-1}\boldsymbol{x}}_{\text{linear terms}}$$

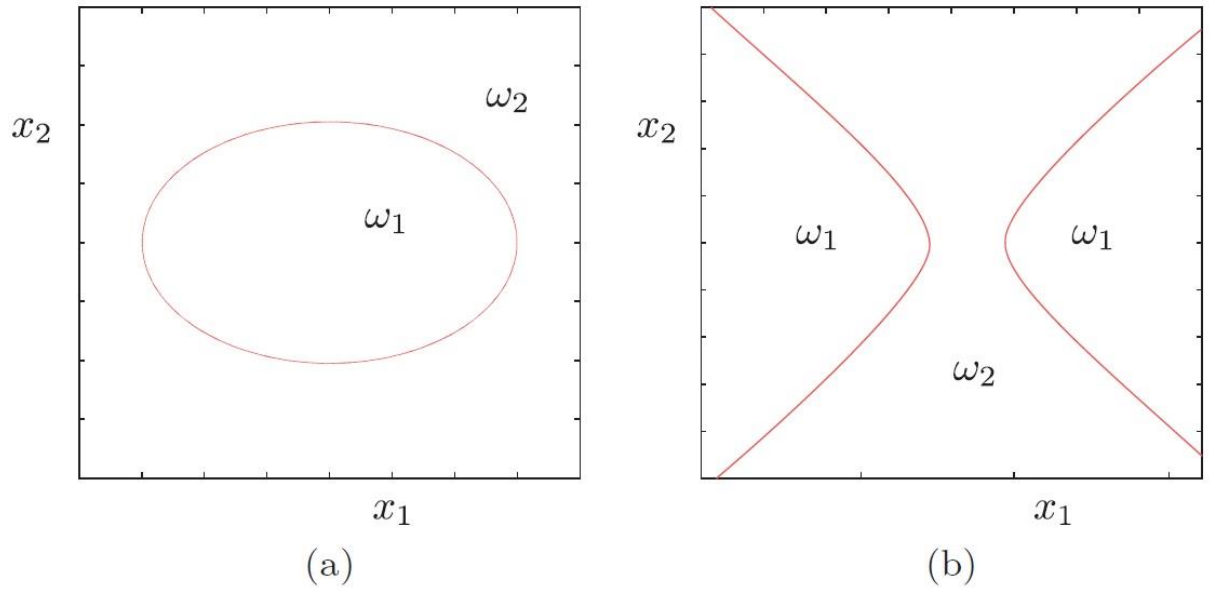
$$\underbrace{-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\Sigma_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\Sigma_2^{-1}\boldsymbol{\mu}_2 + \ln\frac{P(\omega_1)}{P(\omega_2)} + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|}}_{\text{constant terms}} = 0.$$

Η περίπτωση της κανονικής κατανομής

- Το παρακάτω σχήμα δείχνει δύο παραδείγματα στο χώρο των δύο διαστάσεων με $P(\omega_1) = P(\omega_2)$ και

$$\text{(a)} \quad \boldsymbol{\mu}_1 = [0,0]^{\mathrm{T}},\ \boldsymbol{\mu}_2 = [4,0]^{\mathrm{T}},\ \Sigma_1 = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.35 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 1.2 & 0.0 \\ 0.0 & 1.85 \end{bmatrix},$$

$$\text{(b)} \quad \boldsymbol{\mu}_1 = [0,0]^{\mathrm{T}},\ \boldsymbol{\mu}_2 = [3.2,0]^{\mathrm{T}},\ \Sigma_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.75 \end{bmatrix},\ \Sigma_2 = \begin{bmatrix} 0.75 & 0.0 \\ 0.0 & 0,1 \end{bmatrix},$$



Ο απλοϊκός Bayesian ταξινομητής

- Για διανυσματικούς χώρους μεγάλης διάστασης (l μεγάλο), η εκτίμηση του πίνακα συμμεταβλητότητας απαιτεί ένα μεγάλο αριθμό σημείων (data points), ώστε να πάρουμε μια στατιστικά καλή εκτίμηση.
- Σε μια τέτοια περίπτωση, αρκούμαστε σε υποβέλτιστες (suboptimal) λύσεις. Πράγματι, έναν υιοθετήσουμε μια βέλτιστη μέθοδο, η οποία χρησιμοποιεί κακές εκτιμήσεις των απαιτούμενων παραμέτρων, η επίδοση της μεθόδου δε θα είναι καλή.
- Ο απλοϊκός Bayesian ταξινομητής είναι ένα τυπικό και δημοφιλές παράδειγμα ενός υποβέλτιστου ταξινομητή. Η βασική υπόθεση που γίνεται εδώ είναι ότι τα στοιχεία (χαρακτηριστικά) του διανύσματος χαρακτηριστικών είναι στατιστικά ανεξάρτητα. Έτσι η από κοινού ΣΠΠ μπορεί να εκφραστεί σαν το γινόμενο l περιθώριων κατανομών, δηλ.,

$$p(\boldsymbol{x}|\omega_i) = \prod_{k=1}^{l} p(x_k|\omega_i), \qquad i = 1,2,\dots,M$$

Θεώρημα Bayes: Παράδειγμα

Έστω ότι έχουμε κατάστημα ηλεκτρολογικού υλικού και προμηθευόμαστε λαμπτήρες από *τρεις* κατασκευαστές: τον **A**, τον **B** και τον **C**. Πιο συγκεκριμένα ο **A** μας προμηθεύει το *80%* των λαμπτήρων που πουλάμε, ο **B** το *15%* και ο **C** το υπόλοιπο *5%*. Επίσης, οι κατασκευαστές μας έχουν ενημερώσει ο μεν **A** ότι το *4%* των λαμπτήρων του είναι ελλαττωματικό, ο **B** το *6%* και ο **C** το *9%*. Δεδομένου ότι ένας πελάτης μας επιστρέφει **έναν λαμπτήρα** πίσω ως **ελλαττωματικό**, **ποια** είναι η **πιθανότητα** να έχει κατασκευαστεί από τον **A**;

Λύση

$P(A) = 0{,}8, P(B) = 0{,}15, P(C) = 0{,}05$

$P(E|A) = 0{,}04, P(E|B) = 0{,}06, P(E|C) = 0{,}09$

Εφαρμογή θεωρήματος Bayes

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)} = \frac{0{,}04 \cdot 0{,}8}{0{,}04 \cdot 0{,}8 + 0{,}06 \cdot 0{,}15 + 0{,}09 \cdot 0{,}05} \approx 0{,}7033$$

Άρα η πιθανότητα ο ελλαττωματικός λαμπτήρας να έχει κατασκευαστεί από τον Α είναι **περίπου 70,33%**

Αφελής Μπεϋζιανός Ταξινομητής: Παράδειγμα

Παρατηρήσεις βροχόπτωσης

Θερμοκρασία	**Υγρασία**	**Βροχή**
Κρύο	Υψηλή	Ναι
Κρύο	Χαμηλή	Όχι
Μέση	Χαμηλή	Ναι
Μέση	Μέτρια	Όχι
Ζέστη	Μέτρια	Όχι
Ζέστη	Υψηλή	Όχι

Αν σήμερα η *Θερμοκρασία* είναι **Μέση** και η *Υγρασία* **Υψηλή**, θα βρέξει ή όχι;

Λύση

- $P(B) = \frac{2}{6}$, $P(OB) = \frac{4}{6}$
- $P(M|B) = \frac{1}{2}$, $P(M|OB) = \frac{1}{4}$
- $P(Y|B) = \frac{1}{2}$, $P(Y|OB) = \frac{1}{4}$
- $P(B|M,Y) \propto P(B)\ P(M|B)P(Y|B) = \frac{2}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{12}$
- $P(OB|M,Y) \propto P(OB)\ P(M|OB)P(Y|OB) = \frac{4}{6} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{24}$
- Επειδή $P(B|M,Y) > P(OB|M,Y)$ ο αφελής μπεϋζιανός ταξινομητής προβλέπει ότι <u>θα βρέξει</u> σήμερα

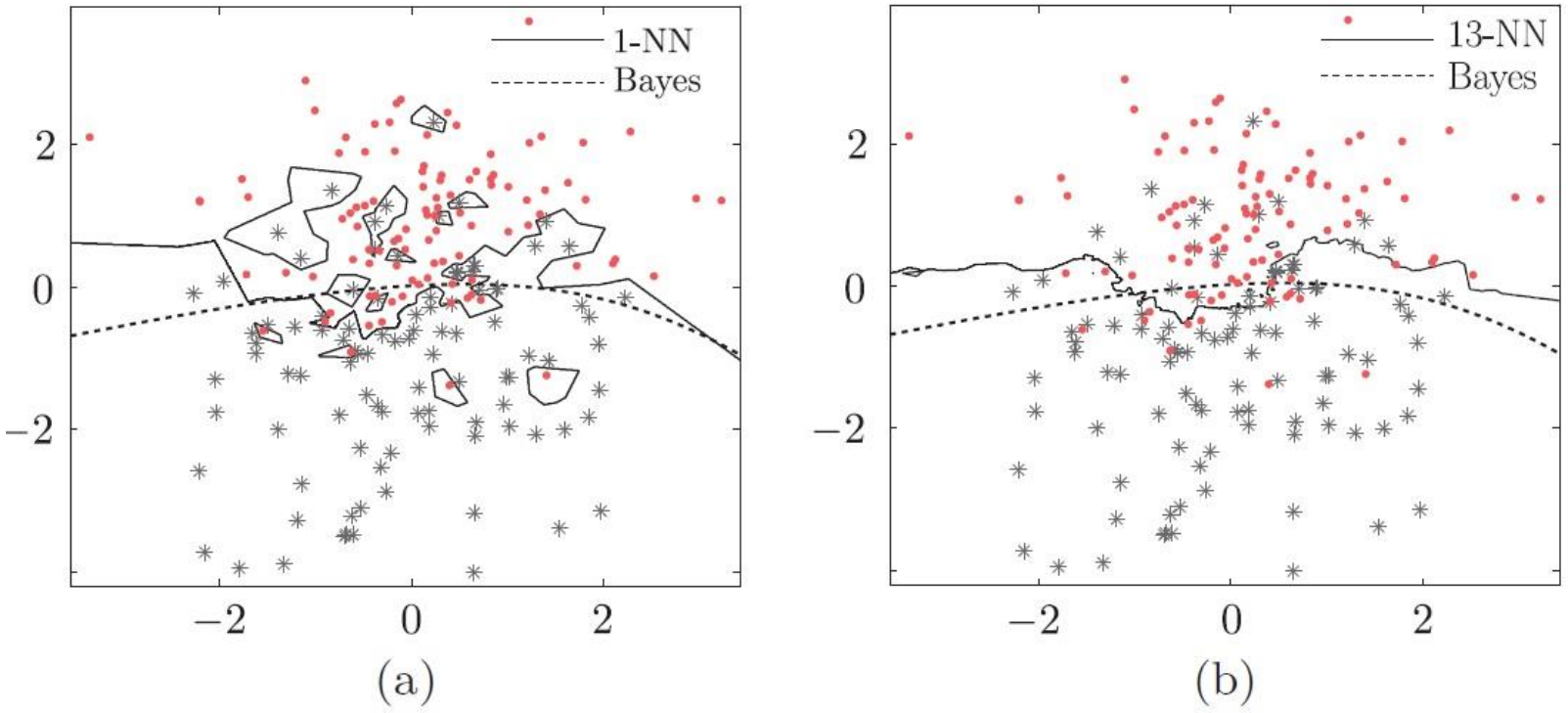
Ο ταξινομητής των k-πλησιέστερων γειτόνων

- Αν και ο ταξινομητής Bayes οδηγεί στη βέλτιστη λύση με βάση την πιθανότητα του σφάλματος ταξινόμησης, για να τον χρησιμοποιήσουμε θα πρέπει να εκτιμήσουμε τις σ.π.π. των κλάσεων. Αυτό μπορεί να μην είναι εύκολο, ιδιαίτερα αν η διάσταση του χώρου χαρακτηριστικών είναι μεγάλη. Αυτό μας οδηγεί στο να αναζητήσουμε εναλλακτικούς κανόνες ταξινόμησης.
- Ο κανόνας των k-πλησιέστερων γειτόνων (k-nearest neighbor rule, k-NN) είναι ένας τυπικός μη-παραμετρικός ταξινομητής και είναι μεταξύ των πιο γνωστών και δημοφιλών ταξινομητών. Παρά την απλότητά του χρησιμοποιείται συχνά στην πράξη και συναγωνίζεται πιο σύνθετους ταξινομητές.
- Ένα χαρακτηριστικό γνώρισμα του k-NN είναι ότι δεν χρειάζεται το στάδιο της εκπαίδευσης.

Ο ταξινομητής των k-πλησιέστερων γειτόνων

- Ας θεωρήσουμε N σημεία εκπαίδευσης $(\boldsymbol{x}_n, y_n), n = 1,2, \dots, N$ σε ένα πρόβλημα ταξινόμησης με M κλάσεις. Αν δοθεί ένα άγνωστο διάνυσμα χαρακτηριστικών $\boldsymbol{x}$, ταξινομείται σύμφωνα με τον κανόνα k-NN ως εξής:
  - Από τα N διανύσματα εκπαίδευσης, προσδιόρισε τα k πλησιέστερα (στο $\boldsymbol{x}$), ανεξαρτήτως κλάσης. Το k επιλέγεται να είναι περιττό για πρόβλημα δύο κλάσεων, και γενικά όχι πολλαπλάσιο του πλήθους των κλάσεων M.
  - Από αυτά τα k δείγματα, προσδιόρισε των αριθμό διανυσμάτων, k_i, που ανήκουν στην κλάση $\omega_i, i = 1,2, \dots, M$. Προφανώς, $\sum_{i=1}^{M} k_i = k$.
  - Καταχώρησε το $\boldsymbol{x}$ στην κλάση ω_i με τον μέγιστο αριθμό δειγμάτων k_i.
- Μπορούν να χρησιμοποιηθούν διάφορα μέτρα απόστασης, συμπεριλαμβανομένης της Ευκλείδειας και της Mahalanobis απόστασης.
- Η απλούστερη εκδοχή του κανόνα k-NN είναι για $k = 1$, γνωστή και ως ο κανόνας του πλησιέστερου γείτονα (nearest neighbor rule – NN). Με άλλα λόγια, ένα άγνωστο διάνυσμα χαρακτηριστικών $\boldsymbol{x}$ ταξινομείται στην κλάση του πλησιέστερου γείτονά του.

Ο ταξινομητής των k-πλησιέστερων γειτόνων



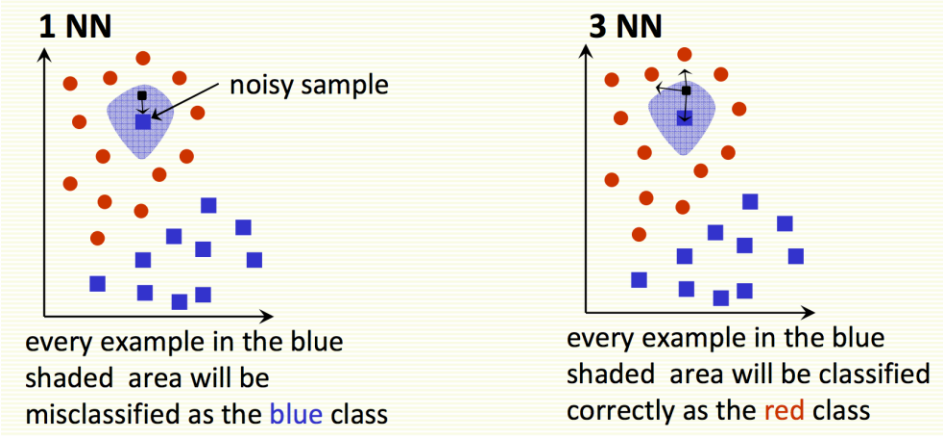
Για ένα πρόβλημα ταξινόνησης 2 κλάσεων στο δισδιάστατο χώρο, τα σχήματα δείχνουν τις καμπύλες απόφασης του Bayes και των 1-NN και 13-NN ταξινομητών. Παράγουμε 100 σημεία από κάθε κλάση από την κανονική κατανομή. Η καμπύλη απόφασης του Bayes ταξινομητή είναι παραβολή, ενώ του 1-NN είναι εξαιρετικά μη γραμμική. Η καμπύλη απόφασης τους 13-NN προσεγγίζει αυτή του Bayes ταξινομητή.

k-πλησιέστεροι γείτονες

- Ο ταξινομητής πλησιέστερου γείτονα είναι «*ευαίσθητος*» στην ύπαρξη *θορύβου* στα δεδομένα

- **Λύση**
  - «*Ομαλοποίηση*» της ταξινόμησης μέσω «*ψηφοφορίας*» των k πλησιέστερων γειτόνων

- **Αλγόριθμος**
  1. Βρες τα k πλησιέστερα παράδειγμα $(\boldsymbol{x}^{(i)};\ t^{(i)})$ από τα δεδομένα εκπαίδευσης τα οποία είναι «εγγύτερα» στο νέο στιγμιότυπο $\boldsymbol{x}$
  2. Ανάθεσε στο $\boldsymbol{x}$ την ετικέτα y:
     - $\mathrm{y} = \underset{t^{(z)}}{argmax} \sum_{r=1}^{k} \delta(t^{(z)}, t^{(r)})$



k-πλησιέστεροι γείτονες: Συμπεράσματα

- Σχηματίζουν περίπλοκα όρια απόφασης, τα οποία προσαρμόζονται στην πυκνότητα των δεδομένων εκπαίδευσης
- Σε περιπτώσεις που τα δεδομένα εκπαίδευσης είναι πολλά, η μέθοδος των k-πλησιέστερων γειτόνων λειτουργεί ικανοποιητικά
- **Ζητήματα**
  1. Ευαισθησία στο θόρυβο
  2. Ευαισθησία στο εύρος τιμών των χαρακτηριστικών των δεδομένων
  3. Η έννοια της απόστασης μεταξύ στιγμιοτύπων δεδομένων χάνει τη σημασία της όσο οι διαστάσεις μεγαλώνουν
  4. Γραμμική υπολογιστική πολυπλοκότητα συναρτήσει του πλήθους των δεδομένων εκπαίδευσης (ψευδο-πολυωνυμικός αλγόριθμος)

Λογιστική παλινδρόμηση

- Στην Bayesian ταξινόμηση οι posterior σ.π.π. υπολογίζονται από τις αντίστοιχες δεσμευμένες σ.π.π. $p(\omega_i|\boldsymbol{x})$, κάτι το οποίο δεν είναι συχνά εύκολο.

- Με τη μέθοδο της λογιστικής παλινδρόμησης, οι posterior πιθανότητες μοντελοποιούνται άμεσα, αντί να εκτιμηθούν από τα δεδομένα.

- Πρόκειται για μια μέθοδο ταξινόμησης, παρ' όλο που το όνομά της παραπέμπει σε παλινδρόμηση. Είναι ένα τυπικό παράδειγμα τεχνικής διακριτικής μοντελοποίησης (discriminative modeling), όπου δε μας ενδιαφέρει η κατανομή των δεδομένων. Αντίθετα, ο ταξινομητής Bayes είναι ένα χαρακτηριστικό παράδειγμα τεχνικής αναπαραγωγικής μοντελοποίησης (generative modeling), όπου η κατανομή των δεδομένων είναι απαραίτητη.

Εκπαίδευση του μοντέλου της λογιστικής παλινδρόμησης

- Η παράμετρος $\boldsymbol{\theta}$ εκτιμάται με τη μέθοδο της μεγίστης πιθανοφάνειας που εφαρμόζεται πάνω στα δεδομένα εκπαίδευσης $(\boldsymbol{x}_n, y_n), n = 1,2, \dots, N, y_n \in \{0,1\}$. Η συνάρτηση πιθανοφάνειας γράφεται ως εξής:

$$P(y_1, \dots, y_N; \boldsymbol{\theta}) = \prod_{n=1}^{N} (\sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n))^{y_n} (1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n))^{1-y_n}$$

- Συνήθως ελαχιστοποιούμε την αρνητική log-likelihood:

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} (y_n \ln s_n + (1 - y_n) \ln(1 - s_n)), \qquad s_n \coloneqq \sigma(\boldsymbol{\theta}^T \boldsymbol{x}_n)$$

- Η $L(\boldsymbol{\theta})$ είναι γνωστή ως cross-entropy σφάλμα.

Λογιστική παλινδρόμηση

- Επικεντρώνουμε στο πρόβλημα δύο κλάσεων. Αρχικά μοντελοποιείται ο λόγος των posterior σ.π.π. ως εξής:

$$\ln \frac{P(\omega_1|\boldsymbol{x})}{P(\omega_2|\boldsymbol{x})} = \boldsymbol{\theta}^T \boldsymbol{x}$$

όπου τα $\boldsymbol{x}, \boldsymbol{\theta}$, ορίζονται όπως πριν.

- Από τη σχέση $P(\omega_1|\boldsymbol{x}) + P(\omega_2|\boldsymbol{x}) = 1$ και ορίζοντας:

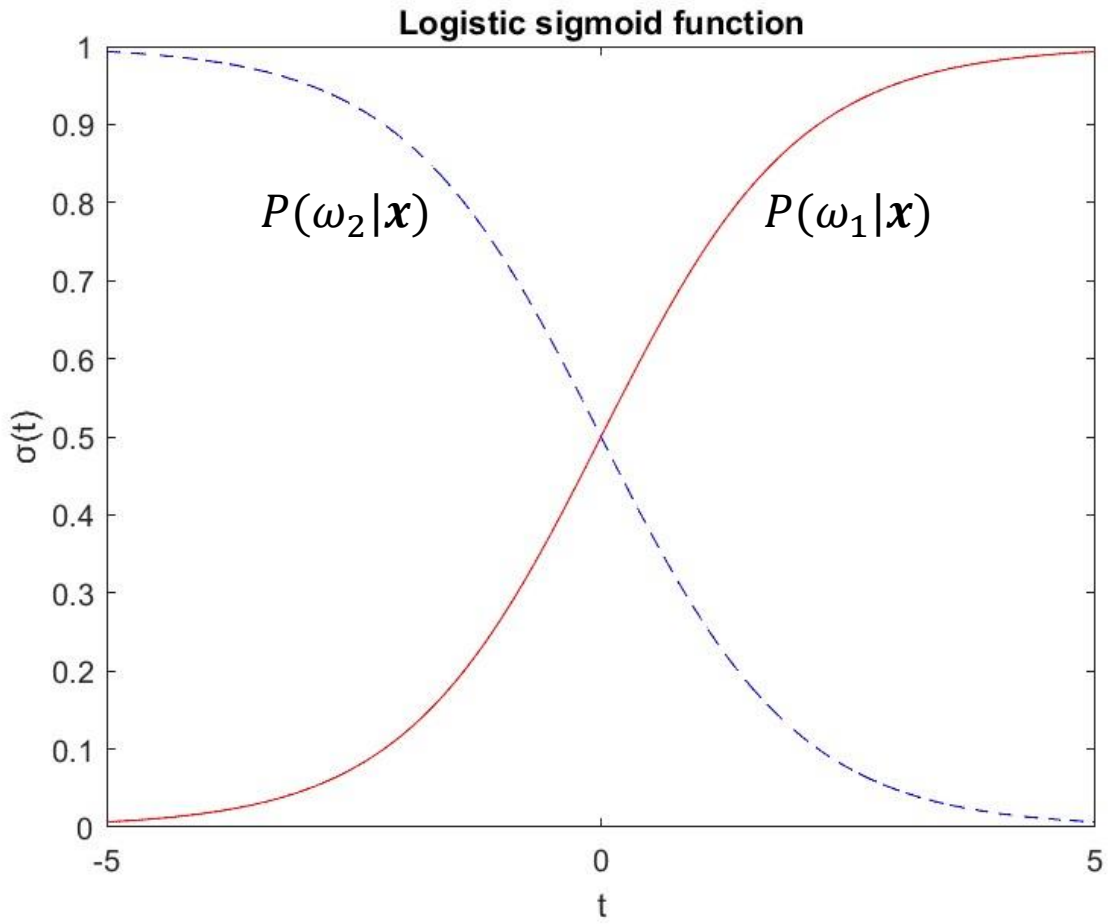
$$t := \boldsymbol{\theta}^T \boldsymbol{x}$$

προκύπτει εύκολα ότι

$$P(\omega_1|\boldsymbol{x}) = \sigma(t) := \frac{1}{1 + e^{-t}}$$

$$P(\omega_2|\boldsymbol{x}) = 1 - \sigma(t) := \frac{e^{-t}}{1 + e^{-t}}$$

Λογιστική παλινδρόμηση



Βιβλιογραφία

- S. Theodoridis, Machine Learning: A Bayesian and Optimization Perspective, 2nd Edition, Academic Press, 2020.
- Σ. Θεοδωρίδης, Διαφάνειες του παραπάνω συγγράμματος (στα αγγλικά).
- Σ. Θεοδωρίδης, Κ. Κουτρούμπας, Αναγνώριση Προτύπων, Εκδ. Π.Χ. Πασχαλίδης, Αθήνα, 2011.
- Θ. Ροντογιάννης, Διαφάνειες μαθήματος ΠΠΣ
- Γ. Αλεξανδρίδης, Διαφάνειες «Εισαγωγή στην ταξινόμηση»