Μηχανική Μάθηση

Μιχάλης Τίτσιας

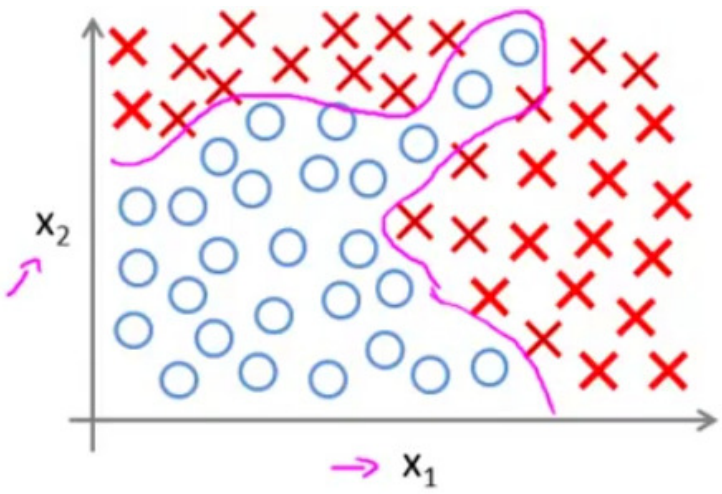
Διάλεξη 5ή

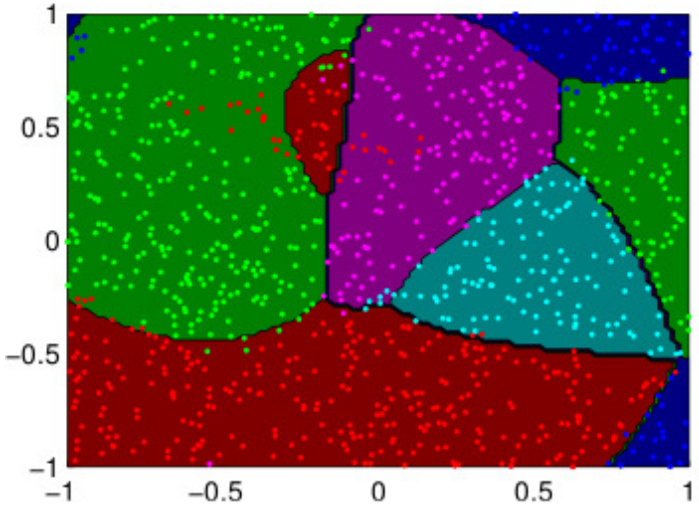
Κ κοντινότεροι γείτονες, περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης, *naive Bayes*

Περιεχόμενα

- Λογιστική παλινδρόμηση αποτελεί διαχωριστική μέθοδος κατηγοριοποίησης
- K κοντινότεροι γείτονες
- Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης
- Naive Bayes
- Σύγκριση περιγραφικών και διαχωριστικών μοντέλων

Λογιστική παλινδρόμηση αποτελεί διαχωριστική μέθοδος κατηγοριοποίησης

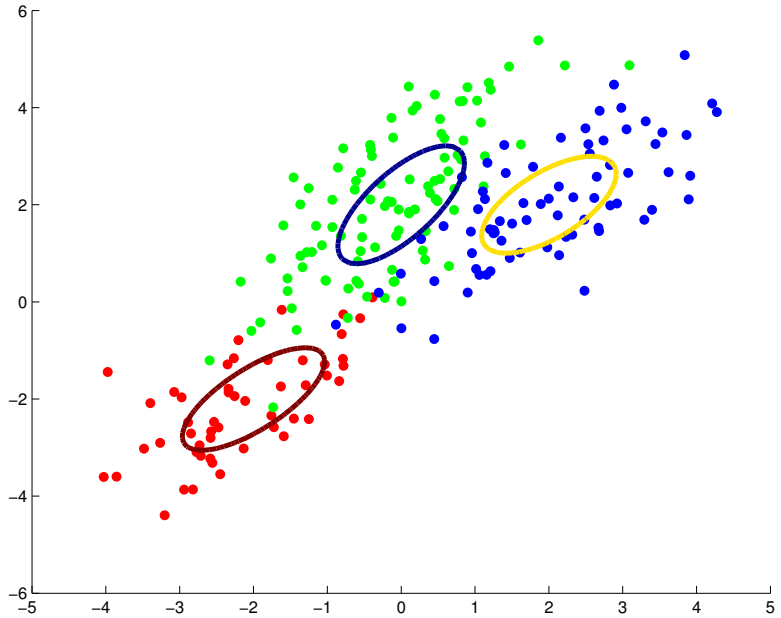




Η λογιστική παλινδρόμηση για 2 ή και περισσότερες κατηγορίες αυτό που κάνει είναι να μοντελοποιεί τη γραμμή/επιφάνεια διαχωρισμού των καρηγοριών

- αποτελεί αυτό που αναφέρεται ως διαχωριστική μέθοδος (discriminative method)

Λογιστική παλινδρόμηση αποτελεί διαχωριστική μέθοδος κατηγοριοποίησης



Ένας άλλος τρόπος κατασκευής συστημάτων κατηγοριοποίησης είναι να μοντελοποιήσουμε την ομοιότητα των δεδομένων εντός της ίδιας κατηγορίας

- αποτελεί την περιγραφική μέθοδο (descriptive/generative method)
- υπό την έννοια ότι προσπαθεί να περιγράψει την κατανομή των δεδομένων εντός της κάθε κατηγορίας

K κοντινότεροι γείτονες

- Κάθε δεδομένο $\mathbf{x}$ ανήκει σε μια από K κατηγορίες $\mathcal{C}_k, k = 1, \ldots, K$.
- Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης $\{\mathbf{x}_n, \mathbf{t}_n\}_{n=1}^{N}$, θα θέλαμε να κατασκευάσουμε ένα σύστημα που να κατηγοριοποιεί κάθε $\mathbf{x}_*$ άγνωστης κατηγορίας
- Μέθοδοι κοντινότερων γειτόνων (nearest neighbors) είναι ίσως οι απλούστερες περιγραφικές τεχνικές
  - Ιδέα: Για άγνωστο $\mathbf{x}_*$ βρες το κοντινότερο δεδομένο εισόδου από το σύνολο εκπαίδευσης και ανέθεσε στο $\mathbf{x}_*$ την κατηγορία του κοντινότερου αυτού δεδομένου

K κοντινότεροι γείτονες

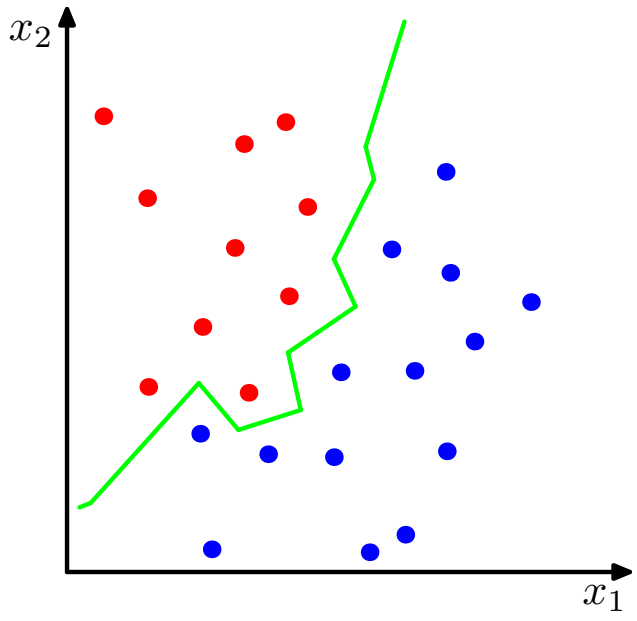
Κοντινότερος γείτονας (K=1)

- Για διανύσματα $\mathbf{x}$ και $\mathbf{x}'$ που αναπαριστούν δύο διαφορετικά δεδομένα, μέτρουμε 'κοντινότητα' χρησιμοποιώντας μια συνάρτηση $d(\mathbf{x}, \mathbf{x}')$. Μια συνηθισμένη επιλογή είναι η τετραγωνισμένη Ευκλείδεια (L_2) απόσταση

$$d(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T(\mathbf{x} - \mathbf{x}') = ||\mathbf{x} - \mathbf{x}'||^2$$

- Βάσει αυτής της απόστασης, το σύνορο απόφασης (δηλ. διαχωρισμού των κατηγοριών) καθορίζεται από τα perdendicular bisectors των κοντινότερων δεδομένων εισόδου εκπαίδευσης που ανήκουν σε διαφορετικές κατηγορίες. Αυτό διαχωρίζει το χώρο σε υποπεριοχές

Κοντινότερος γείτονας (K=1)



(b)

Το σύνορο απόφασης (για δεδομένα δύο κατηγοριών) είναι piecewise linear όπου το κάθε τμήμα αντιστοιχεί στο perpendicular bisector μεταξύ δύο δεδομένων διαφορετικών κατηγοριών

Κοντινότερος γείτονας (K=1): Επιλογή απόστασης

- Η Ευκλείδεια απόσταση δεν λαμβάνει υπόψη το μήκος της κλίμακας (length scale) των διαφορετικών συνιστωσών του $\mathbf{x} = (x_1, x_2, \ldots, x)_d)$, π.χ. $x_1 \in [-1, 1]$ και $x_2 \in [10^5, 10^9]$.
- Οπότε ενδέχεται η συνιστώσα με το μεγαλύτερο μήκος κλίμακας να έχει κυρίαρχη συμβολή στη τιμή της απόστασης, ενώ άλλες συνιστώσες (με ενδεχομένως πολύ χρήσιμη πληροφορία για την κατηγοριοποίηση) να μη λαμβάνονται υπόψη
- Τετραγωνισμένη Ευκλείδεια απόσταση με βάρη

$$d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\sigma_i^2}$$

- Mahalanobis απόσταση

$$d(\mathbf{x}, \mathbf{x}') = (\mathbf{x}_i - \mathbf{x}_i')^T \Sigma^{-1} (\mathbf{x}_i - x_i')$$

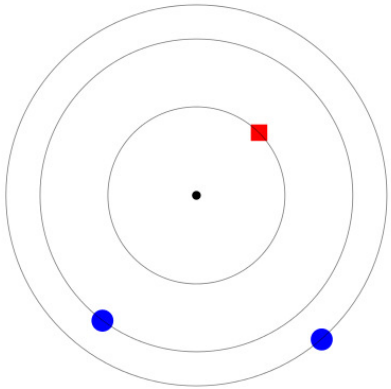
όπου Σ θετικά ορισμένος πίνακας (συνήθως ο covariance πίνακας όλων των δεδομένων από όλες τις κατηγορίες)

K **Κοντινότεροι γείτονες** ($K > 1$)

- Όταν ο κοντινότερος γείτονας έχει (λόγω θορύβου) λάθος label ή δεν είναι ένα αντιπροσωπευτικό δεδομένο της κατηγορίας του, ενδέχεται να συμβαίνουν λάθη στην κατηγοριοποίηση
- Συμπεριλαμβάνοντας περισσότερους $K > 1$ κοντινότερους γείτονες, ελπίζουμε να πάρουμε καλύτερο σύστημα κατηγοριοποίησης με πιο ομαλό (smooth) σύνορο απόφασης
- Αν χρησιμοποιήσουμε την Ευκλείδεια απόσταση, η μέθοδος κατά κάποιο τρόπο σχηματίζει μια υπερσφαίρα με κέντρο το δεδομένο $\mathbf{x}_*$
- Η ακτίνα της υπερσφαίρας μεγαλώνει ώσπου να συμπεριλάβει K ακριβώς δεδομένα από το σύνολο εκπαίδευσης

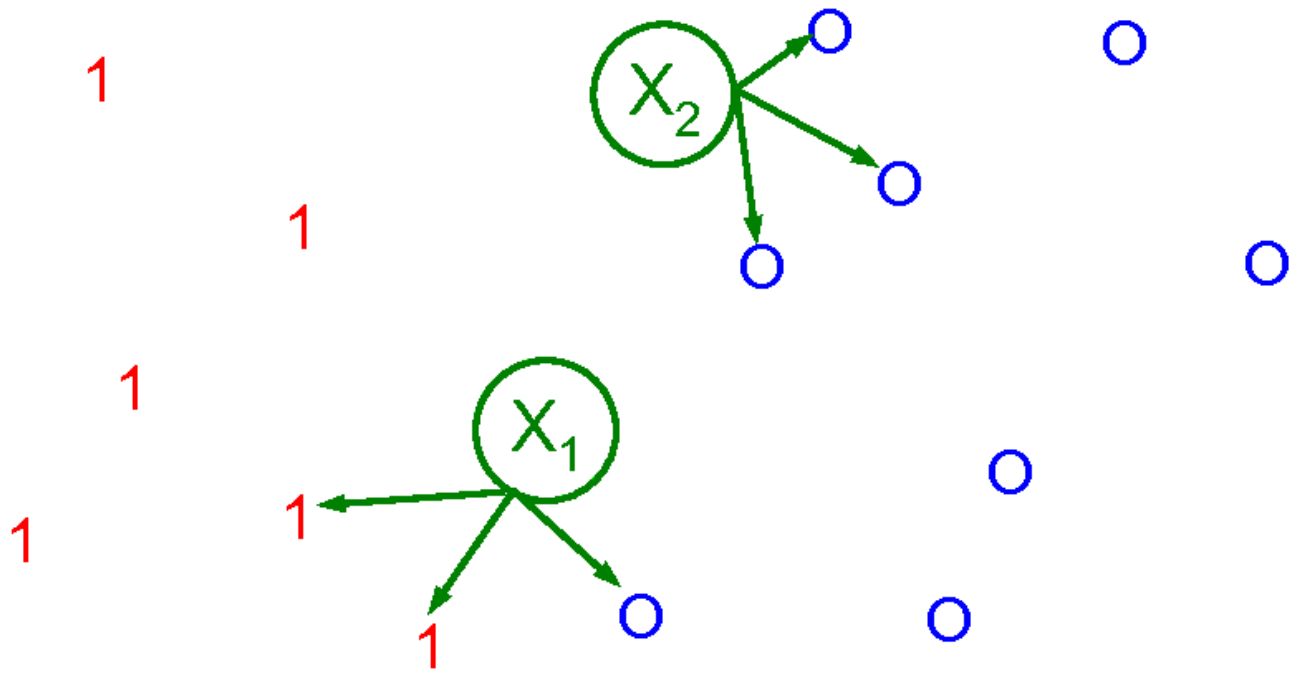
K κοντινότεροι γείτονες

K **Κοντινότεροι γείτονες** ($K > 1$)



- Θέλουμε να κατηγοριοποιήσουμε την κεντρική μαύρη κουκκίδα. Ο εσωτερικός κύκλος περικλύει τον κοντινότερο γείτονα $\Rightarrow$ το δεδομένο θα κατηγοριοποιηθεί ως κόκκινο τετράγωνο χρησιμοποιώντας ένα κοντινότερο γείτονα
- Ωστόσο χρησιμοποιώντας $K = 3$ κοντινότερους γείτονες θα ανατεθεί στην κατηγορία των μπλε κουκκίδων

K **Κοντινότεροι γείτονες**

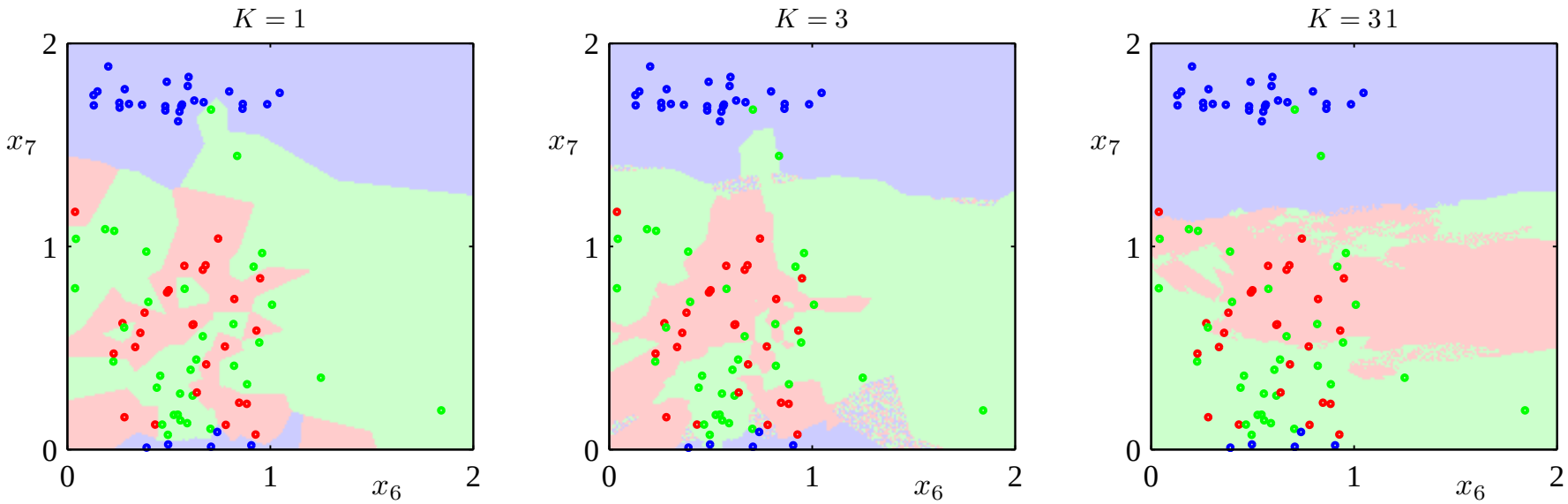


Εφαρμογή της μεθόδου των κοντινότερων γειτόνων στο χώρο $\mathbb{R}^2$ για δύο (άγνωστα) δεδομένα x_1 και x_2

K **Κοντινότεροι γείτονες: Επιλογή του** K

- Για $K = 1$, το σφάλμα (ο αριθμός των δεδομένων στο σύνολο εκπαίδευσης που κατηγοριοποιούνται λανθασμένα) είναι μηδέν
  - Overfitting είναι πολύ πιθανόν να συμβεί
- Για $K = N$, προβλέπουμε πάντα την ίδια κατηγορία (την πολυπληθέστερη) ανεξαρτήτου της τιμής του $\mathbf{x}_*$!
- Οπότε η μέθοδος που γενικεύει καλύτερα θα αντιστοιχεί συνήθως σε κάποια ενδιάμεση τιμή του K
- Πώς μπορούμε να καθορίσουμε το K;
  - $\Rightarrow$ cross validation

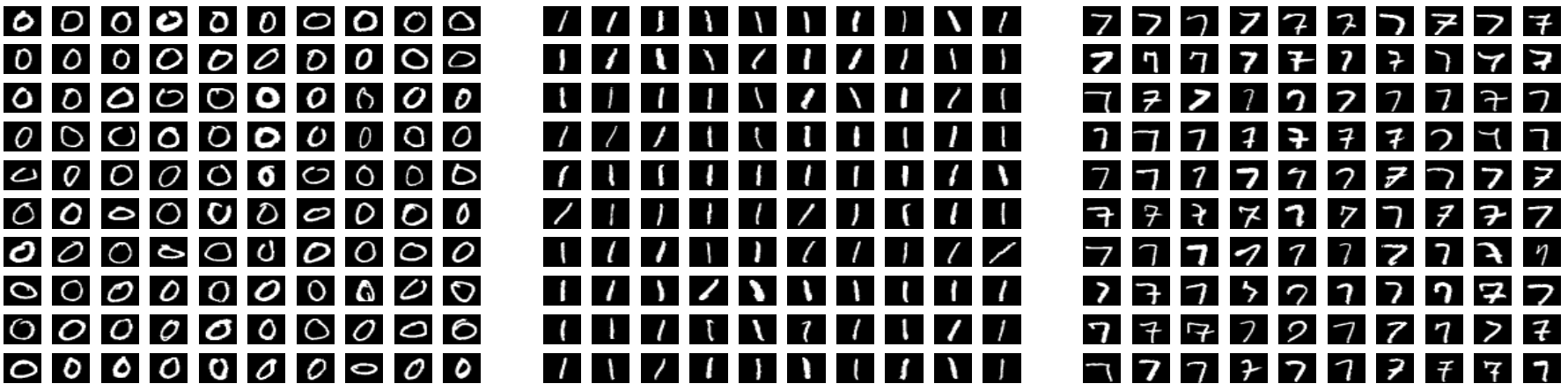
K **Κοντινότεροι γείτονες: Επιλογή του** K



Καθώς το K μεγαλώνει τα σύνορα απόφασης γίνονται πιο ομαλά (smooth) με λιγότερες μικρές περιοχές

K κοντινότεροι γείτονες

Παράδειγμα αναγνώρισης χειρόγραφων χαρακτήρων



Κάποια από τα δεδομένα του συνόλου εκπαίδευσης για τα ψηφία 0, 1, και 7. Υπάρχουν 300 παραδείγματα στο σύνολο εκπαίδευσης για καθένα από τα τρία ψηφία

K κοντινότεροι γείτονες

Η κατηγορία του ψηφίου 1 ενάντια στη κατηγορία του 0

- Έστω ότι μας ενδιαφέρει να διαχωρίσουμε χειρόγραφους χαρακτήρες δύο κατηγοριών που αντιστοιχούν στα ψηφία 1 και 0
- Κάθε δεδομένο αποτελεί μια εικόνα $28 \times 28 = 784$ pixels. Το σύνολο εκπαίδευσης περιλαμβάνει 300 παραδείγματα του 0 και 300 του 1
- Χρησιμοποιήσαμε τη μέθοδο του κοντινότερου γείτονα ($K = 1$) βασισμένη στην Ευκλείδεια απόσταση
- Η ικανότητα γενικεύσης μετρήθηκε σε ένα ανεξάρτητο σύνολο ελέγχου (test set) με 600 παραδείγματα
- Η μέθοδος δεν είχε κανένα σφάλμα. Γιατι;

K κοντινότεροι γείτονες

Η κατηγορία του ψηφίου 1 ενάντια στη κατηγορία του 7

- Επαναλάβαμε το πείραμα με τη διαφορά ότι τώρα θα θέλαμε να διαχωρίσουμε τις κατηγορίες των ψηφίων 1 και 7
- Είχαμε 18 σφάλματα, δηλ. 3%

Σημειωτέον ότι οι καλύτερες μέθοδοι μηχανικής μάθησης επιτυγχάνουν σφάλμα (όταν κατηγοριοποιούν συγχρόνως δεδομένα απ΄ όλα τα 10 ψηφία) λιγότερο του $1\% \Rightarrow$ **καλύτερη από την επίδοση ενός ανθρώπου**

K κοντινότεροι γείτονες

Παράδειγμα αναγνώρισης χειρόγραφων χαρακτήρων

- Πάνω είναι τα 18 ψηφία τα οποία ταξινομήθηκαν εσφαλμένα
- Κάτω είναι οι αντίστοιχοι κοντινότεροι γείτονες από το σύνολο εκπαίδευσης

K κοντινότεροι γείτονες

Πλεονεκτήματα:

- Ευκολία υλοποίησης
- Μπορεί να δουλέψει καλά αν χρησιμοποιηθεί η κατάλληλη απόσταση

Μειονεκτήματα:

- Η επιλογή της απόστασης μπορεί να είναι δύσκολη
- Απαιτεί την αποθήκευση όλων των δεδομένων εκπαίδευσης
- Η εύρεση των K κοντινότερων γειτόνων μπορεί να είναι πολύ δαπανηρή $\Rightarrow$ η πολυπλοκότητα είναι ανάλογη του αριθμού των δεδομένων εκπαίδευσης και της διάστασης
- Δεν μαθαίνει κάποια βαθύτερη δομή που διέπει τα δεδομένα $\Rightarrow$ απλά τα απομνημονεύει

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Πώς θα μπορούσαμε να κατασκευάσουμε πιθανοτικά περιγραφικά μοντέλα κατηγοριοποίησης;

- Θα βασιστούμε στο θεώρημα του Bayes

Εκ των προτέρων πιθανότητες

- Έστω ότι έχουμε δεδομένα εισόδου **x** που μπορούν να ανήκουν σε K κατηγορίες: $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$.
- Έκ των προτέρων πιθανότητα $p(\mathcal{C}_k)$: Η πιθανότητα ένα οποιοδήποτε δεδομένο (δηλ. προτού παρατηρηθεί κάποιο συγκεκριμένο δεδομένο) να ανήκει στη κατηγορία $\mathcal{C}_k$
- Προφανώς ισχύει

$$\sum_{k=1}^{K} p(\mathcal{C}_k) = 1$$

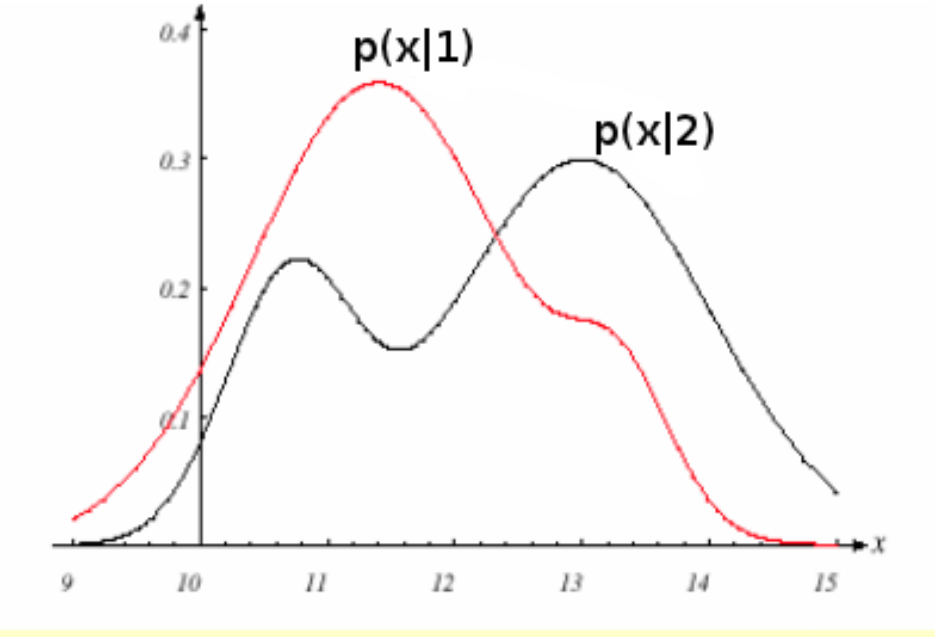
Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Εκ των προτέρων πιθανότητες

- Οι πιθανότητες $p(\mathcal{C}_k), k = 1, \ldots, K$ αποτελούν την εκ των προτέρων πληροφορία
- Βέλτιστη πρόβλεψη χρησιμοποιώντας μόνο την εκ των προτέρων πληροφορία βασίζεται στον κανόνα
  - $\mathcal{C}_k^* = \arg\max_{\mathcal{C}_k} \{p(\mathcal{C}_k), k = 1, \ldots, K\}$

Η υπό συνθήκη κατανομή της κατηγορίας

- $p(\mathbf{x}|\mathcal{C}_k)$: Περιγράφει το πως κατανέμεται το δεδομένο $\mathbf{x}$ δοθέντος της κατηγορίας $\mathcal{C}_k$.
  Π.χ. η παρακάτω εικόνα δείχνει δύο υπό συνθήκη κατανομές για μονοδιάστατο δεδομένο εισόδου



Κανόνας του Bayes

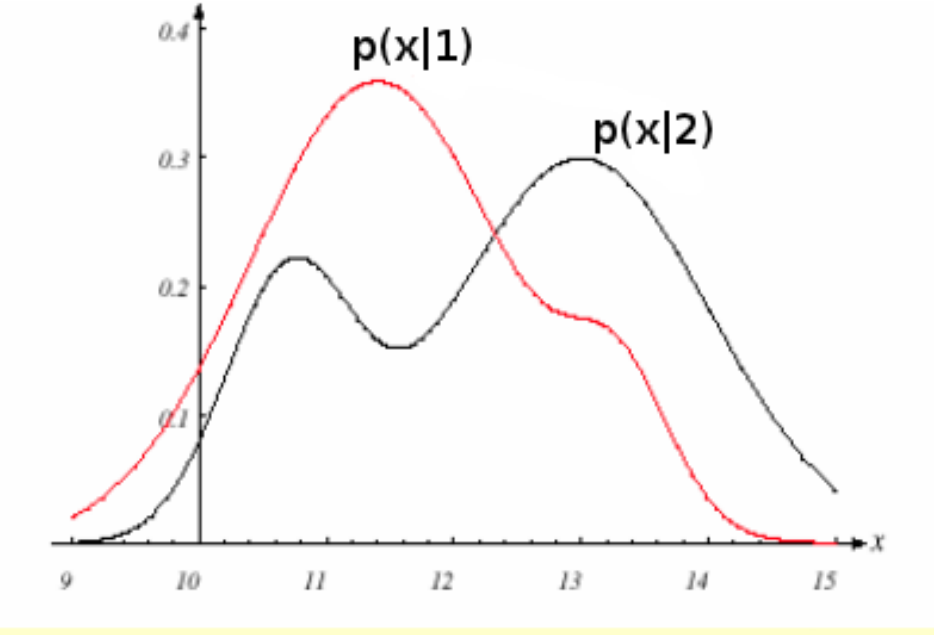
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

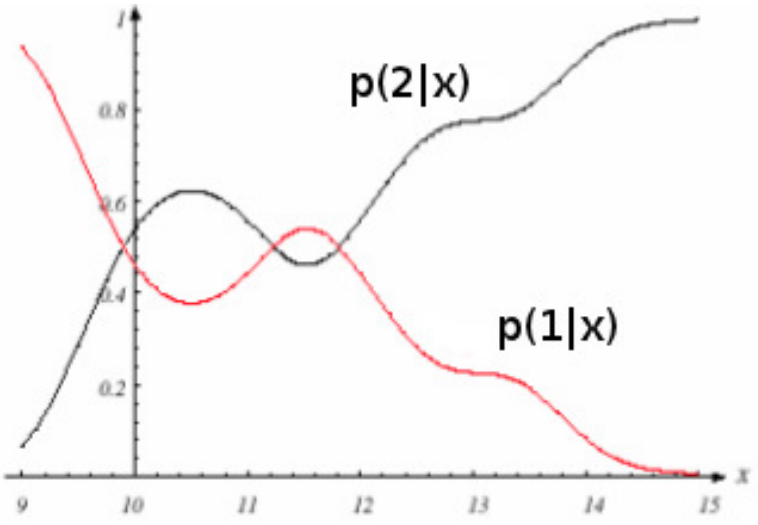
όπου

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

Η $p(\mathcal{C}_k|\mathbf{x})$ εκφράζει την εκ των υστέρων πιθανότητα (δηλ. αφού παρατηρηθεί ένα συγκεκριμένο $\mathbf{x}$) να ισχύει η κατηγορία $\mathcal{C}_k$

Κανόνας του Bayes





- $p(\mathcal{C}_1) = 2/3$, $p(\mathcal{C}_2) = 1/3$
- Για $x = 14$, $p(\mathcal{C}_1|x) = 0.08$ και $p(\mathcal{C}_2|x) = 0.92$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Βέλτιστη πρόβλεψη χρησιμοποιώντας την εκ των υστέρων πληροφορία γίνεται με βάση τον κανόνα

- $\mathcal{C}_k^* = \arg\max_{\mathcal{C}_k}\{p(\mathcal{C}_k|\mathbf{x}), k = 1, \ldots, K\}$

Δηλ. για το $\mathbf{x}$ η κατηγορία την οποία αποφασίζουμε ότι ανήκει είναι αυτή με τη μέγιστη εκ των υστέρων πιθανότητα

Αν οι εκ των προτέρων πιθανότητες $p(\mathcal{C}_k), k = 1, \ldots, K$ και υπό συνθήκη κατανομές των κατηγοριών $p(\mathbf{x}|\mathcal{C}_k), k = 1, \ldots, K$ είναι οι (άγνωστες) πραγματικές ποσότητες, τότε ο παραπάνω κανόνας είναι βέλτιστος

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Στην πράξη όμως δεν γνωρίζουμε τις πιθανότητες $p(\mathcal{C}_k)$ και τις υπό συνθήκη κατανομές των κατηγοριών $p(\mathbf{x}|\mathcal{C}_k)$

Πρέπει να κάνουμε υποθέσεις για τη μορφή των υπό συνθήκη κατανομών

Μοντέλο/υπόθεση: υποθέτουμε ένα μοντέλο/κατανομή $p(\mathbf{x}|\boldsymbol{\theta}_k)$ που στοχεύει να προσεγγίσει την άγνωστη $p(\mathbf{x}|\mathcal{C}_k)$ και εξαρτάται από παραμέτρους $\boldsymbol{\theta}_k$

Για τις πιθανότητες $p(\mathcal{C}_k), k = 1, \ldots, K$, δεν χρειάζεται να κάνουμε κάποια υπόθεση. Απλώς τις θεωρούμε ως άγνωστες παραμέτρους

Έπειτα ο σκοπός μας είναι να εκπαιδεύσουμε το σύστημα, δηλ. να βρούμε τιμές για τις παραμέτρους $(p(\mathcal{C}_k), \boldsymbol{\theta}_k)_{k=1}^K$ χρησιμοποιώντας ένα σύνολο δεδομένων εκπαίδευσης και αλγορίθμους μάθησης

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Έστω δεδομένα $(X, T) = (\mathbf{x}_n, \mathbf{t}_n)_{n=1}^N$ τα οποία υποθέτουμε ότι έχουν παραχθεί ανεξάρτητα μεταξύ τους

Δηλ. το κάθε ζεύγος $(\mathbf{x}_n, \mathbf{t}_n)$ ακολουθεί

$$p(\mathbf{x}_n, \mathbf{t}_n) = p(\mathbf{x}_n|\mathbf{t}_n)p(\mathbf{t}_n)$$

όπου

$$p(\mathbf{t}_n) = \prod_{k=1}^{K} [p(\mathcal{C}_k)]^{t_{nk}}$$

$$p(\mathbf{x}_n|\mathbf{t}_n) = \prod_{k=1}^{K} [p(\mathbf{x}|\boldsymbol{\theta}_k)]^{t_{nk}}$$

Πιο συνοπτικά μπορούμε να γράψουμε

$$p(\mathbf{x}_n, \mathbf{t}_n) = \prod_{k=1}^{K} [p(\mathbf{x}|\boldsymbol{\theta}_k)p(\mathcal{C}_k)]^{t_{nk}}$$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Πιθανοφάνεια

$$p(X, T) = \prod_{n=1}^{N} \prod_{k=1}^{K} [p(\mathbf{x}|\boldsymbol{\theta}_k)p(\mathcal{C}_k)]^{t_{nk}}$$

Μπορούμε να χωρίσουμε τα δεδομένα σε K υποσύνολα, δηλ. ανά κατηγορία έτσι ώστε το $\mathcal{N}_k = \{n|t_{nk} = 1\}$ σύνολο υποδεικνύει όλα τα δεδομενα για τα οποία η κατηγορία είναι η k

Ισοδύναμα η παραπάνω πιθανοφάνεια γράφεται ως

$$p(X, T) = \prod_{k=1}^{K} \left(\prod_{n \in \mathcal{N}_k} p(\mathbf{x}_n|\boldsymbol{\theta}_k) \right) p(\mathcal{C}_k)^{N_k}$$

όπου $N_k = |\mathcal{N}_k|$ είναι το πλήθος δεδομένων της κατηγορίας k

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Λογαριθμική πιθανοφάνεια

$$\begin{aligned}
\mathcal{L}(\{\boldsymbol{\theta}_k, p(\mathcal{C}_k)\}_{k=1}^K) &= \log \prod_{k=1}^K \left(\prod_{n \in \mathcal{N}_k} p(\mathbf{x}_n|\boldsymbol{\theta}_k) \right) p(\mathcal{C}_k)^{N_k} \\
&= \sum_{k=1}^K \log \left(\prod_{n \in \mathcal{N}_k} p(\mathbf{x}_n|\boldsymbol{\theta}_k) \right) + \sum_{k=1}^K N_k \log p(\mathcal{C}_k) \\
&= \sum_{k=1}^K \mathcal{L}(\boldsymbol{\theta}_k) + \sum_{k=1}^K N_k \log p(\mathcal{C}_k)
\end{aligned}$$

όπου

$$\mathcal{L}(\boldsymbol{\theta}_k) = \sum_{n \in \mathcal{N}_k} \log p(\mathbf{x}_n|\boldsymbol{\theta}_k)$$

είναι η λογαριθμική πιθανοφάνεια που προκύπτει από τα δεδομένα της κατηγορίας k

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

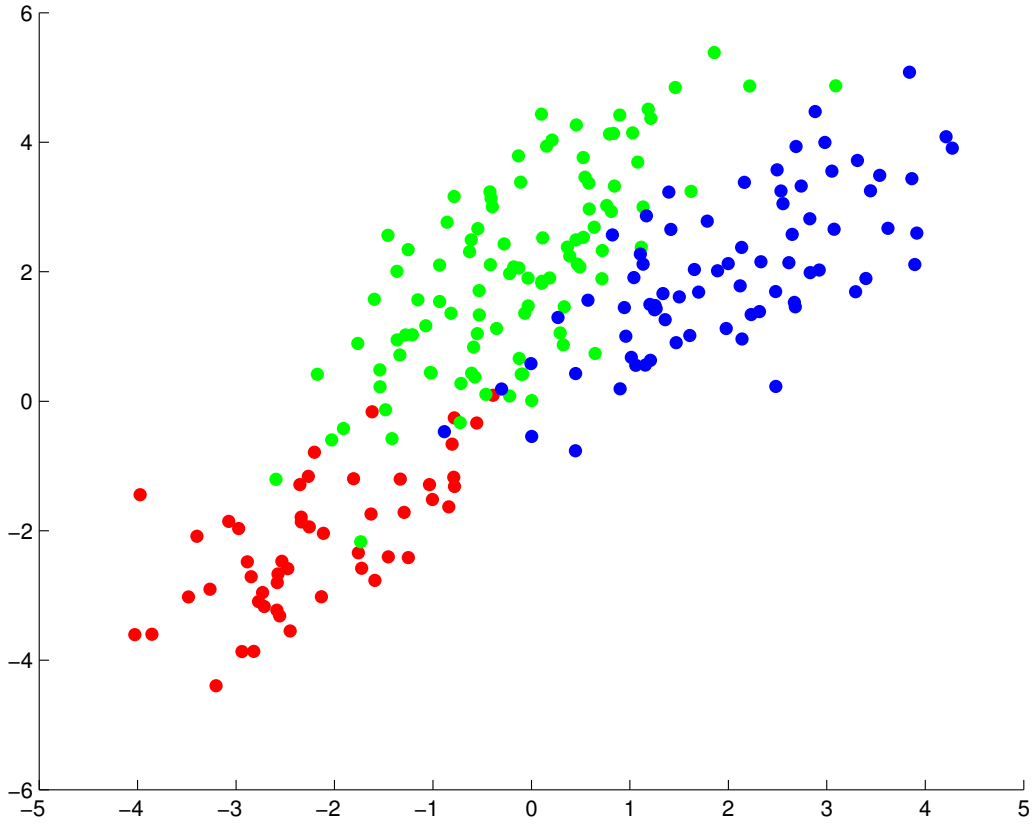
$$\mathcal{L}(\{\boldsymbol{\theta}_k, p(\mathcal{C}_k)\}_{k=1}^K) = \sum_{k=1}^K \mathcal{L}(\boldsymbol{\theta}_k) + \sum_{k=1}^K N_k \log p(\mathcal{C}_k)$$

Μεγιστοποίηση της πιθανοφάνειας δίνει για τις εκ των προτέρων πιθανότητες την ακόλουθη σχέση

$$p(\mathcal{C}_k) = \frac{N_k}{\sum_{j=1}^K N_j} = \frac{N_k}{N}$$

Ο κάθε όρος $\mathcal{L}(\boldsymbol{\theta}_k)$ μεγιστοποιείται ανεξάρτητα (από τους υπόλοιπους $\mathcal{L}(\boldsymbol{\theta}'_k)$, $k' \neq k$) εφόσον δεν υπάρχουν κοινές παράμετροι, δηλ. $\boldsymbol{\theta}_k \cap \boldsymbol{\theta}_{k'} = \emptyset$, για κάθε $k \neq k'$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης



Έστω σύνολο δεδομένων εκπαίδευσης τριών κατηγοριών όπου $N_1 = 50$, $N_2 = 100$ και $N_3 = 75$ είναι το πλήθος δεδομένων για κάθε κατηγορία αντίστοιχα

Θέλουμε μέσω αυτών των δεδομένων να «εκπαιδεύσουμε» ένα σύστημα κατηγοριοποίησης

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Θα πρέπει να κάνουμε μια υπόθεση για την μορφή που θα έχει η υπό συνθήκη κατανομή της κάθε κατηγορίας

Έχουμε 3 κατηγορίες και επομένως 3 υπό συνθήκη κατανομές

$$p(\mathbf{x}|\mathcal{C}_1), p(\mathbf{x}|\mathcal{C}_2) \quad \text{και} \quad p(\mathbf{x}|\mathcal{C}_3)$$

όπου το δεδομένο εισόδου είναι δισδιάστατο $\mathbf{x} = [x_1 \ x_2]^T$

Υποθέτουμε ότι οι κατανομές αυτές είναι Gaussian στο δισδιάστατο χώρο

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

(Παρένθεση για πολυδιάστατες Gaussian κατανομές)

Η μονοδιάστατη Gaussian κατανομή (την οποία έχουμε συναντήσει) ορίζεται ως

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

όπου (μ, σ^2) είναι οι παράμετροι

Η κατανομή αυτή γενικεύεται και στις πολλές διαστάσεις (δηλ. όπου το x γίνεται διάνυσμα $\mathbf{x}$)

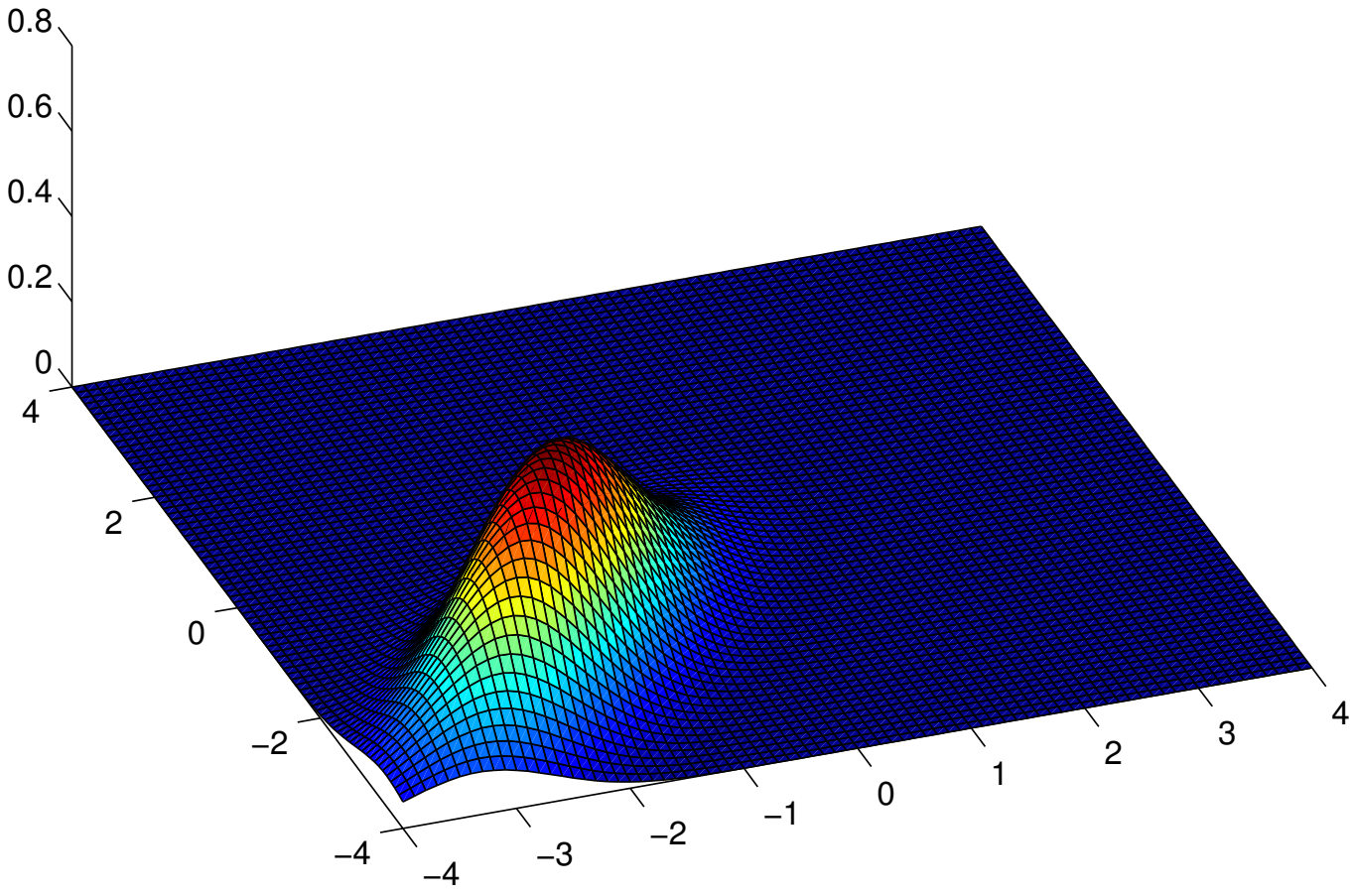
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

όπου το διάνυσμα $\boldsymbol{\mu}$ είναι η μέση τιμή και Σ ο πίνακας συμμεταβλητότητας (covariance matrix)

- Ο Σ είναι συμμετρικός και θετικά ορισμένος

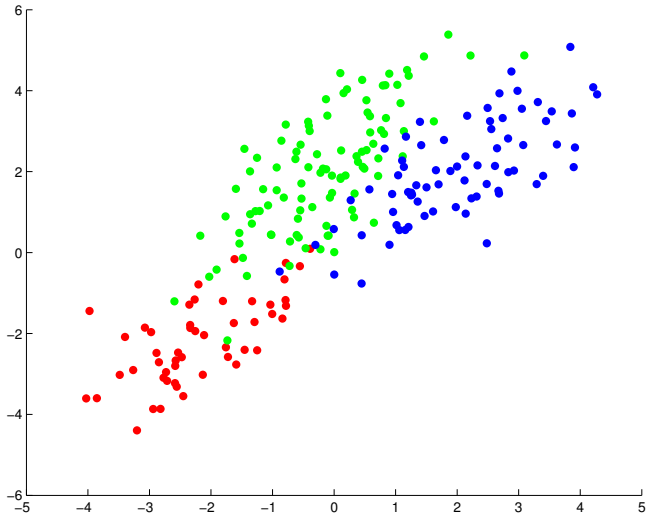
Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

(Παρένθεση για πολυδιάστατες Gaussian κατανομές)



$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \text{ όπου } D = 2$$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης



Υποθέτουμε ότι οι υπό συνθήκη κατανομές των κατηγοριών είναι

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{2}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}, \quad k = 1, 2, 3$$

όπου $(\boldsymbol{\mu}_k, \Sigma_k)$ οι άγνωστοι παράμετροι της κατανομής k, όπου $\boldsymbol{\mu}_k$ είναι δισδιάστατο διάνυσμα και Σ_k είναι ένας 2×2 συμμετρικός και θετικά ορισμένος πίνακας

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

θα πρέπει να εκτιμήσουμε τις άγνωστες παραμέτρους του στατιστικού μοντέλου που είναι

$$(p(\mathcal{C}_1), p(\mathcal{C}_2), p(\mathcal{C}_3), \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \Sigma_1, \Sigma_2, \Sigma_3)$$

Χρησιμοποιούμε τον αλγόριθμο εκπαίδευσης βάσει της μέγιστης πιθανοφάνειας

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Λογαριθμική πιθανοφάνεια

$$\mathcal{L}(\{\boldsymbol{\mu}_k, \Sigma_k, p(\mathcal{C}_k)\}_{k=1}^3) = \sum_{k=1}^{3} \mathcal{L}(\boldsymbol{\mu}_k, \Sigma_k) + \sum_{k=1}^{3} N_k \log p(\mathcal{C}_k)$$

όπου

$$\mathcal{L}(\boldsymbol{\mu}_k, \Sigma_k) = -\frac{N_k}{2}\log(2\pi) - \frac{N_k}{2}\log|\Sigma_k| - \frac{1}{2}\sum_{n\in\mathcal{N}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

Μεγιστοποίηση της πιθανοφάνειας δίνει

$$p(\mathcal{C}_k) = \frac{N_k}{N} \;\Rightarrow\; p(\mathcal{C}_1) = \frac{50}{225},\; p(\mathcal{C}_2) = \frac{100}{225},\; p(\mathcal{C}_3) = \frac{75}{225}$$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

Η μεγιστοποίηση του κάθε όρου

$$\mathcal{L}(\boldsymbol{\mu}_k, \Sigma_k) = -\frac{N_k}{2}\log(2\pi) - \frac{N_k}{2}\log|\Sigma_k| - \frac{1}{2}\sum_{n\in\mathcal{N}_k}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

δίνει

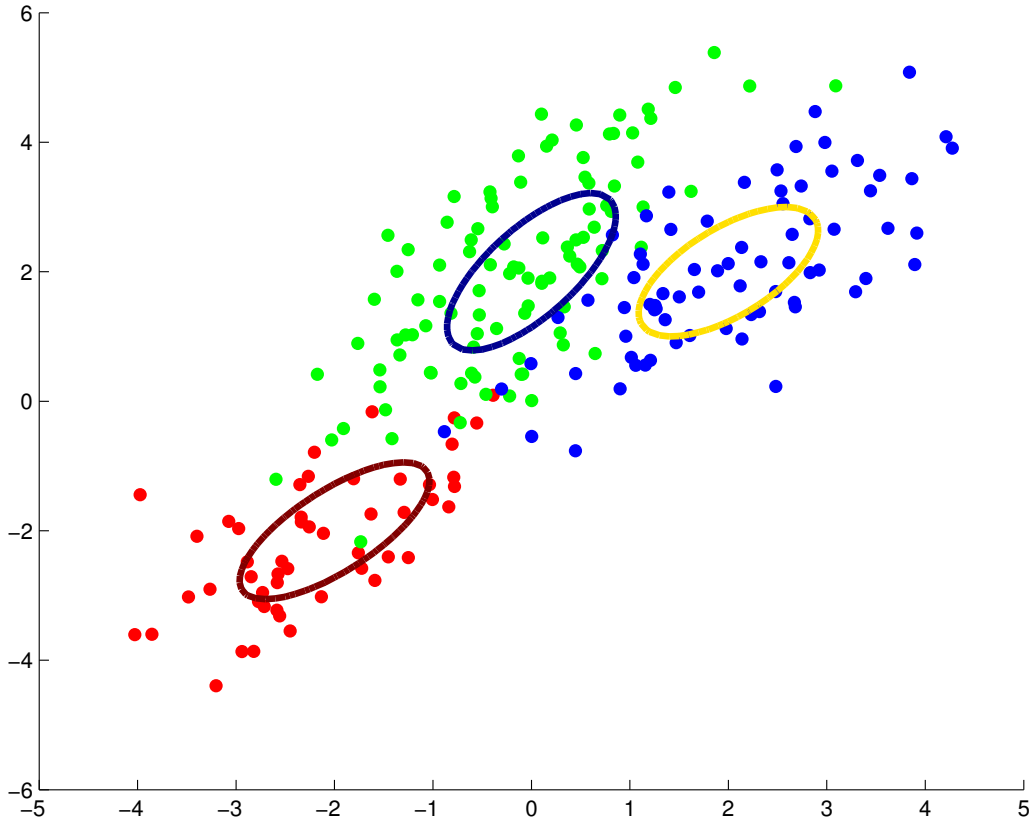
$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k}\sum_{n\in\mathcal{N}_k}\mathbf{x}_n, \quad k = 1, 2, 3$$

$$\hat{\Sigma}_k = \frac{1}{N_k}\sum_{n\in\mathcal{N}_k}(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k)^T, \quad k = 1, 2, 3$$

Παρατήρησε ότι όταν το x είναι μονοδιάστατο η εξίσωση για το Σ_k απλοποιείται σε

$$\hat{\sigma}_k^2 = \frac{1}{N_k}\sum_{n\in\mathcal{N}_k}(x_n - \hat{\mu}_k)^2$$

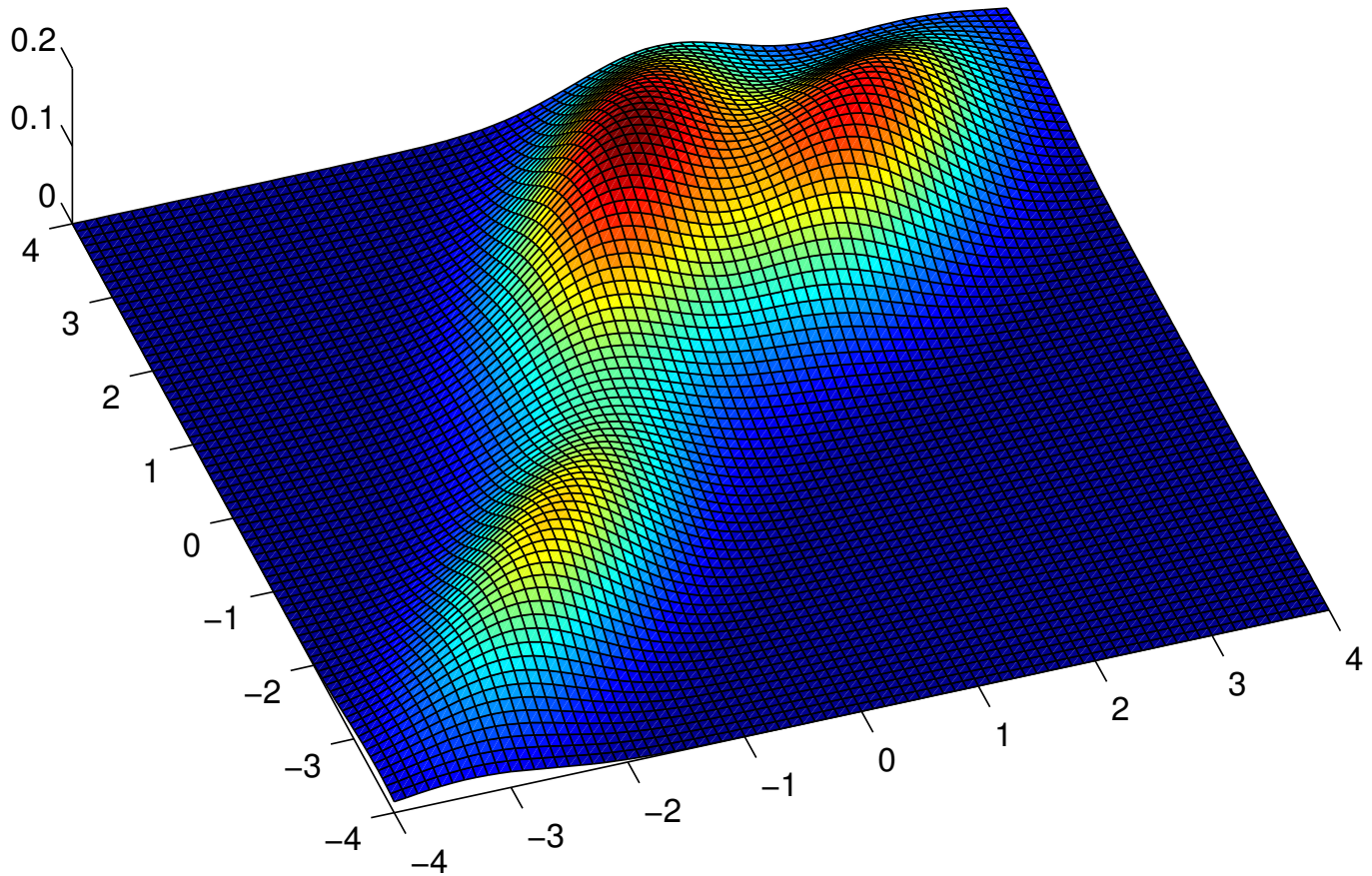
Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης



Οι ελλείψεις (η κάθε μια αποτέλει ισοπίθανη γραμμή ως προς $\mathbf{x}$, δηλ. που προκύπτει από την εξίσωση $c = p(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)$ για κάποια σταθερά c) απεικονίσουν τις δισδιάστατες κανονικές κατανομές οι οποίες πλέον έχουν ταιριάξει στα δεδομένα

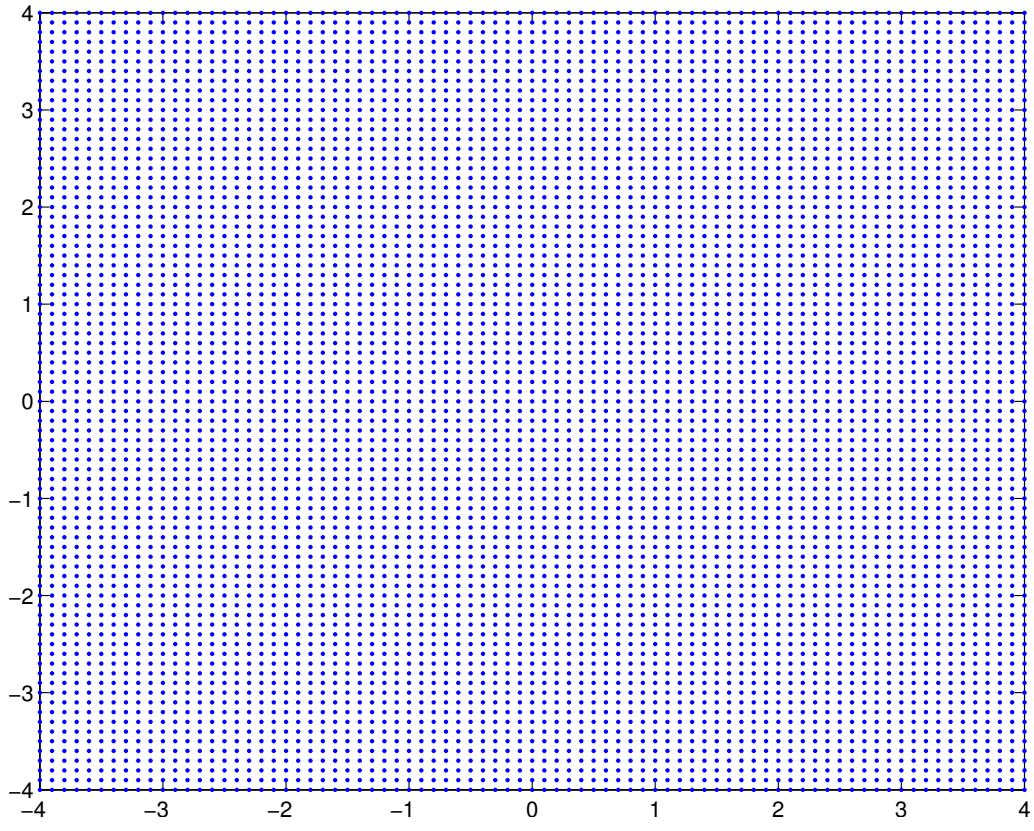
Το κέντρο της κάθε έλλειψης είναι το αντίστοιχο $\boldsymbol{\mu}_k$, ενώ το σχήμα και η κατεύθυνση καθορίζεται από την τιμή του πίνακα Σ_k

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης



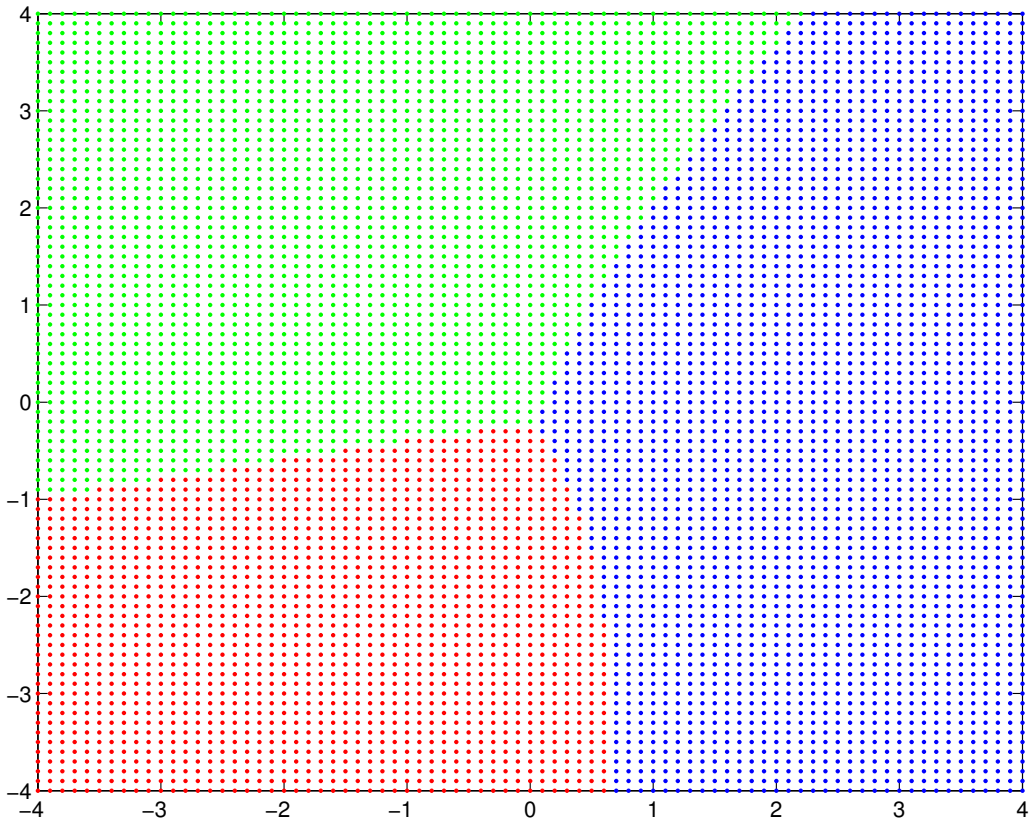
Τρισδιάστατη απεικόνιση των κατανομών (συγκεκριμένα της ολικής κατανομής
$p(\mathbf{x}) = p(\mathbf{x}|\boldsymbol{\mu}_1, \Sigma_1)p(\mathcal{C}_1) + p(\mathbf{x}|\boldsymbol{\mu}_2, \Sigma_2)p(\mathcal{C}_2) + p(\mathbf{x}|\boldsymbol{\mu}_3, \Sigma_3)p(\mathcal{C}_3))$

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης



Έστω δεδομένα ελέγχου που αντιστοιχούν σε όλες τις κουκκίδες του σχήματος

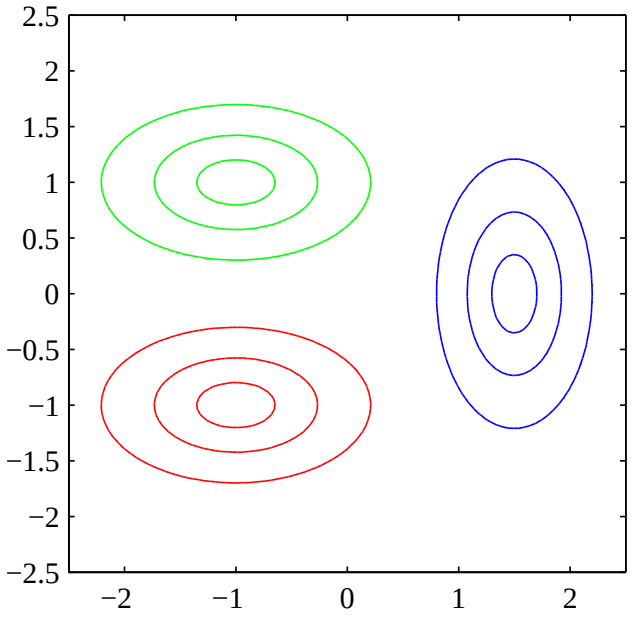
Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης

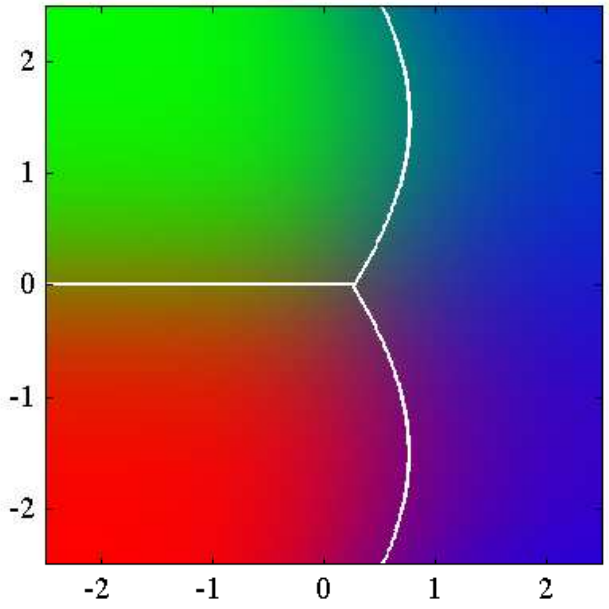


Τα δεδομένα κατηγοριοποιημένα

Μια σημαντική παρατήρηση είναι ότι τα σύνορα απόφασης έχουν ελλειψοειδή μορφή (αυτό έχει να κάνει με τη μορφή της Gaussian κατανομής)

Περιγραφικά πιθανοτικά μοντέλα κατηγοριοποίησης





Αριστερά φαίνονται η υπό συνθήκη κατανομές τριών κατηγοριών όπου οι δύο από αυτές (κόκκινη και πράσινη) έχουν κοινό covariance πίνακα

Δεξιά φαίνονται τα σύνορα απόφασης: μεταξύ κόκκινης και πρασινής κατηγορίας είναι γραμμικό, ενώ μεταξύ της μπλε και των άλλων δύο είναι ελλειψοειδή

Naive Bayes

Όταν το δεδομένου εισόδου $\mathbf{x} = (x_1, \ldots, x_D)$ έχει πολύ μεγάλη διάσταση, δηλ. $D >> 1$, οι παράμετροι που θα πρέπει να εκτιμηθούν για τις υπό συνθήκη κατανομές των κατηγοριών $p(\mathbf{x}|\boldsymbol{\theta}_k)$ αυξάνονται δραματικά και για τις Gaussian κατανομές είναι $O(D^2)$

- Ο αριθμός των δεδομένων μπορεί να είναι πολύ μικρός για να επιτρέψει ικανοποιητική εκτίμηση τόσων πολλών παραμέτρων

Σε περιπτώσεις όπου το $\mathbf{x}$ παίρνει διακριτές τιμές είναι δύσκολο να ορίσουμε κατανομές που να λαμβάνουν υπόψη την εξάρτηση των συνιστωσών του $\mathbf{x}$

Ιδέα του naive Bayes: Υπέθεσε απλουστευμένες κατανομές $p(\mathbf{x}|\boldsymbol{\theta}_k)$ για τις διαφορετικές κατηγορίες όπου οι συνιστώσες του $\mathbf{x}$ είναι ανεξάρτητες δοθέντος της κατηγορίας

- $\Rightarrow O(D)$ αριθμός παραμέτρων

Naive Bayes

Έστω το κάθε δεδομένο εισόδου $\mathbf{x} = (x_1, \ldots, x_D)$ είναι μεγάλης διάστασης

Ως μοντέλα για την κάθε άγνωστη κατανομή $p(\mathbf{x}|\mathcal{C}_k)$ της κατηγορίας k υποθέτουμε μοντέλα όπου οι διαστάσεις είναι υπό συνθήκη ανεξάρτητες, δηλ. ανεξάρτητες δοθείσης της κατηγορίας

$$p(\mathbf{x}|\boldsymbol{\theta}_k) = \prod_{d=1}^{D} p(x_d|\theta_{k,d})$$

όπου $\theta_{k,d}$ είναι οι παράμετροι που σχετίζονται με την διάσταση d

Naive Bayes

(Παρένθεση: Παράδειγμα υπό συνθήκη ανεξαρτησίας)

- Έστω ότι η κατηγορία είναι η ηλικία ενός ατόμου, δηλ. $\mathcal{C}_k =$ ηλικία, και το $\mathbf{x} =$ (ύψος, γκρίζο-μαλλί)
- Διαισθητικά ισχύει

$$p(\text{ύψος}, \text{γκρίζο-μαλλί}|\text{ηλικία}) = p(\text{ύψος}|\text{ηλικία})p(\text{γκρίζο-μαλλί}|\text{ηλικία})$$

Π.χ. αν η ηλικία ενός ατόμου είναι 10 ετών η πεποίθηση σου για το αν θα έχει γκρίζα μαλλιά δεν θα αλλάξει όταν πληροφορηθείς το ύψος του

- Ωστόσο δεν ισχύει

$$p(\text{ύψος}, \text{γκρίζο-μαλλί}) = p(\text{ύψος})p(\text{γκρίζο-μαλλί})$$

- Δηλαδή «ύψος» και «γκρίζο μαλλί» δεν είναι ανεξάρτητα. Π.χ. Αν μάθεις ότι το ύψος ενός ατόμου είναι 1 και 30 εκατοστά, αυτό θα αυξήσει την πεποίθηση σου ότι το άτομο αυτό δεν έχει γκρίζα μαλλιά (λόγω ότι είναι μικρής ηλικίας)

Naive Bayes

Naive Bayes χρησιμοποιείται εκτενώς για διακριτά, π.χ. δυαδικά, δεδομένα

Έστω $\mathbf{x} = (x_1, \ldots, x_D)$ είναι ένα διάνυσμα δυαδικών τιμών, δηλ. κάθε $x_d \in \{0, 1\}$

Για δυαδικά δεδομένα

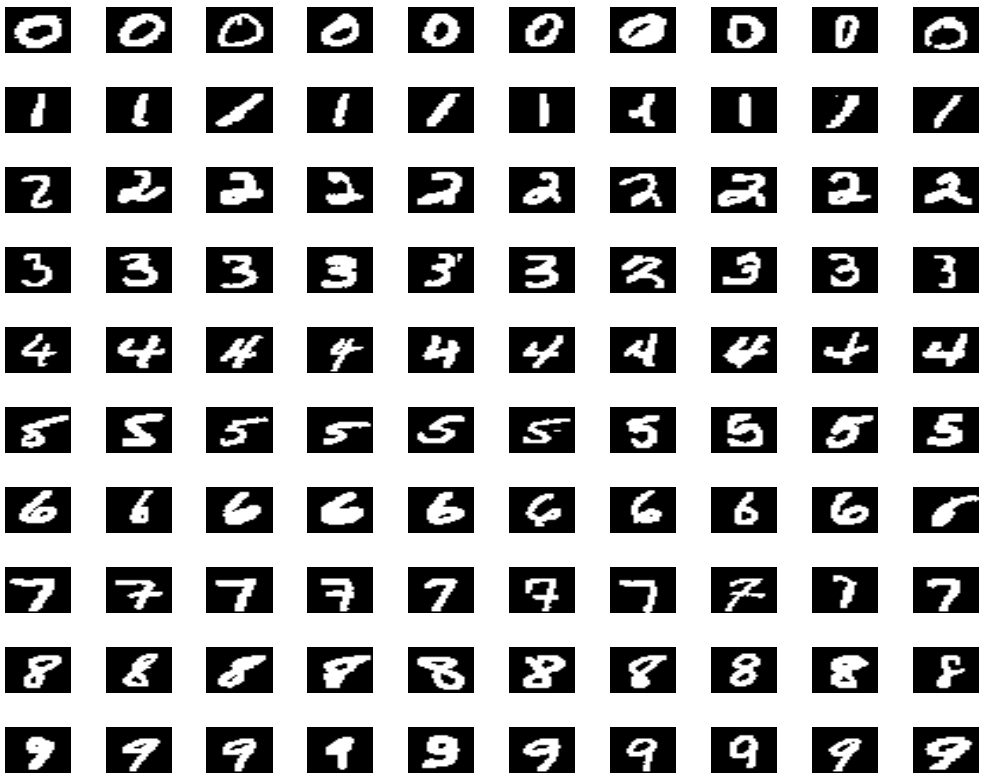
$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{d=1}^{D} \mu_{k,d}^{x_d}(1 - \mu_{k,d})^{1-x_d}$$

όπου κάθε $\mu_{k,d}^{x_d}(1 - \mu_{k,d})^{1-x_d}$ είναι μια ξεχωριστή Bernoulli κατανομή

Ένα τέτοιο μοντέλο χρησιμοποιείται συχνά για ταξινόμηση κειμένου όπου το δεδομένο εισόδου $\mathbf{x}$ είναι δυαδικό και περιγραφεί την παρουσία/απουσία κάποιων λέξεων κλειδιών

Naive Bayes

Παράδειγμα εφαρμογής naive Bayes στους χειρόγραφους χαρακτήρες



Έχουμε ένα σύνολο 60000 δεδομένα χειρόγραφων ψηφίων (10 κατηγορίες)

Κάθε δεδομένο εισόδου αποτελεί μια δυαδική εικόνα διάστασης 28×28, οπότε κάθε δεδομένο εισόδου έχει διάσταση 784

Naive Bayes

Παράδειγμα εφαρμογής naive Bayes στους χειρόγραφους χαρακτήρες

Στο σχήμα απεικονίζονται οι τιμές των παραμέτρων για τις 10 κατηγορίες

$$\mu_{k,d}, d = 1, \ldots, 784, \quad k = 1, \ldots, K$$

που προκύπτουν από την εκπαίδευση με μέγιστη πιθανοφάνεια

Naive Bayes

Παράδειγμα εφαρμογής naive Bayes στους χειρίογραφους χαρακτήρες

Αυτό που κάνει ο naive Bayes στην προκειμένη περίπτωση είναι να περιγράψει την κάθε κατηγορία με την μέση τιμή όλων των δεδομένων εκπαίδευσης της κατηγορίας αυτής

Παράδειγμα εφαρμογής naive Bayes στους χειρόγραφους χαρακτήρες

Κατά το έλεγχο του συστήματος κατηγοριοποίησης χρησιμοποιούμε 10000 δεδομένα

Το ολικό σφάλμα της μεθόδου ήταν

$$error = 15.83\%$$

Ως σύγκριση οι διαχωριστικές μέθοδοι που εξετάσαμε (δες προηγούμενο μάθημα) είχαν πολύ καλύτερη επίδοση. Συγκεκριμένα

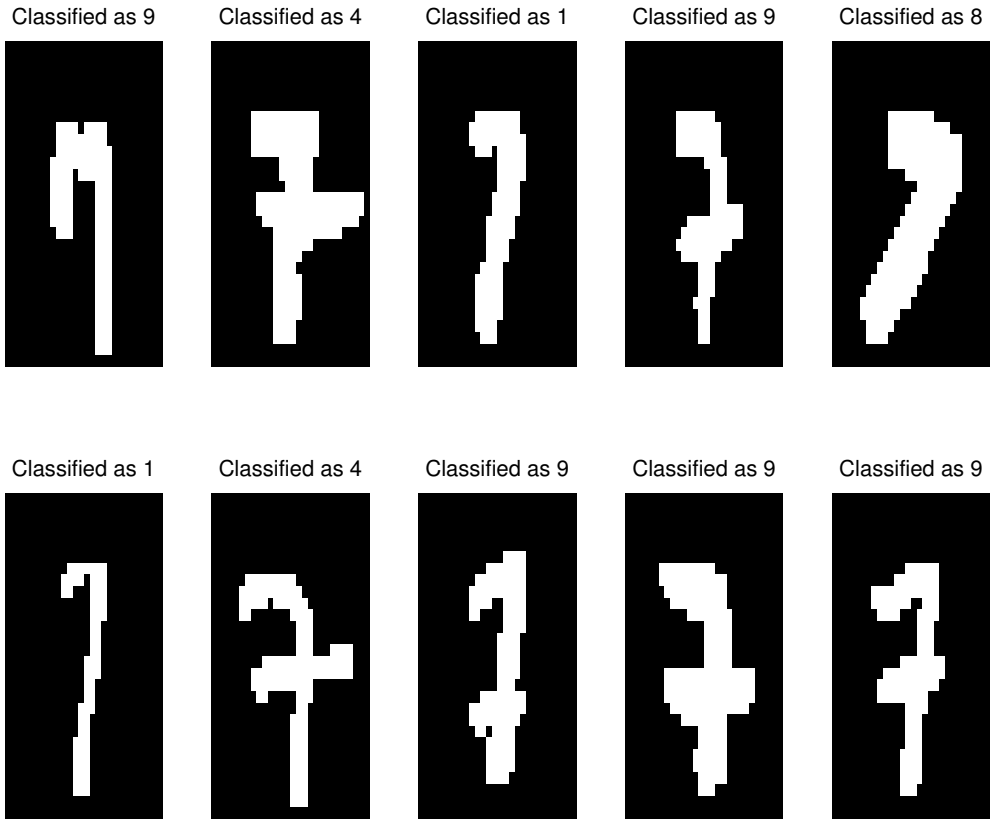
- Λογιστική παλινδρόμηση για πολλές κατηγοριές

$$error = 8.18\%$$

- Νεωρωνικό δίκτυο

$$error = 3.31\%$$

Παράδειγμα εφαρμογής naive Bayes στους χειρόγραφους χαρακτήρες



Στο σχήμα απεικονίζονται περιπτώσεις εσφαλμένα κατηγοριοποιημένων παραδειγμάτων του χαρακτήρα 7

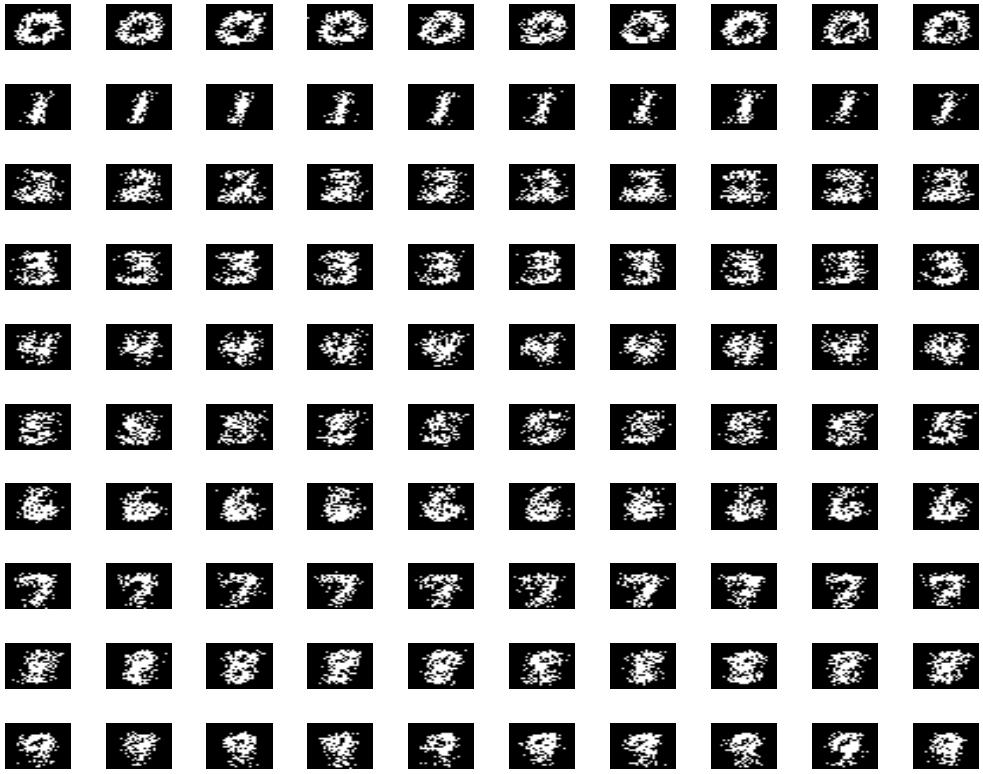
Γιατί δεν έχει τόσο καλή επίδοση ο naive Bayes;

Αν υποθέσουμε ότι δεν έχει συμβαίνει υπερεκπαίδευση, τότε η αιτία πρέπει να αφορά τις υποθέσεις μας και συνίσταται στο γεγονός ότι τα μοντέλα $p(\mathbf{x}|\boldsymbol{\theta}_k)$ δεν περιγράφουν ικανοποιητικά τις άγνωστες κατανομές των κατηγοριών $p(\mathbf{x}|\mathcal{C}_k)$

Ο παραπάνω λόγος αφορά γενικότερα όλα τα περιγραφικά μοντέλα κατηγοριοποίησης

Ένας διαισθητικός τρόπος προκειμένου να ελέγξουμε αν η $p(\mathbf{x}|\boldsymbol{\theta}_k)$ περιγράφει ικανοποιητικά την $p(\mathbf{x}|\mathcal{C}_k)$ είναι να γεννήσουμε φανταστικά δεδομένα από την $p(\mathbf{x}|\boldsymbol{\theta}_k)$ και να δούμε αν αυτά μοιάζουν με τα δεδομένα εκπαίδευσης (τα οποία έχουν γεννηθεί από την άγνωστη $p(\mathbf{x}|\mathcal{C}_k)$)

Naive Bayes



Το σχήμα δείχνει φανταστικά δεδομένα που προέκυψαν δειγματοληπτώντας τυχαία από τα εκπαιδευμένα μοντέλα

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{d=1}^{D} \mu_{k,d}^{x_d}(1-\mu_{k,d})^{1-x_d}, \quad k = 1, \ldots, K$$

Τα δεδομένα αυτά απέχουν πολύ από αυτά του συνόλου εκπαίδευσης!

Σύγκριση περιγραφικών και διαχωριστικών μοντέλων

Το σημαντικό μειονέκτημα των περιγραφικών μοντέλων: Είναι δύσκολο να κατασκευάσουμε πολύ ευέλικτα περιγραφικά μοντέλα

- Αυτό έχει ως συνέπεια τα σύνορα απόφασης που προκύπτουν μέσω των μοντέλων αυτών να μην είναι ιδιαίτερα περίπλοκα

Το σημαντικό πλεονέκτημα των περιγραφικών μοντέλων: Περιγραφικές μέθοδοι μπορούν εύκολα να ανταπεξέλθουν σε ελλειπή (μη παρατηρήσιμα) δεδομένα εισόδου

Το σημαντικό πλεονέκτημα των διαχωριστικών μοντέλων: Οι διαχωριστικές μέθοδοι είναι πολύ πιο ευέλικτες και μπορούν σχετικά εύκολα να μοντελοποιήσουν περίπλοκα σύνορα απόφασης, π.χ. μπορούμε να αντικαταστήσουμε το δεδομένο εισόδου $\mathbf{x}$ με $\phi(\mathbf{x})$ και οι αλγόριθμοι εφαρμόζονται όπως αρχικά

Το σημαντικό μειονέκτημα των διαχωριστικών μοντέλων: Δεν μπορούν να ανταπεξέλθουν σε ελλειπή δεδομένα

Επίλογος

(Η μέθοδος των κοντινότερων γειτόνων μπορεί επίσης να ερμηνευτεί ως περιγραφικό πιθανοτικό μοντέλο κατηγοριοποίησης (δες Barber, chapter 14))

- Διάβασμα για το σπίτι: Barber: chapters 14 και 10. Bishop: sections 4.2, 2.3 (ως σελίδα 85) και 2.5.

- Επόμενο μάθημα: Support Vector Machines