Κεφάλαιο 3: Μέγιστη Πιθανοφάνεια και Bayesian Εκτίμηση Παραμέτρων

3.1 Εισαγωγή

Στο κεφάλαιο 2 είδαμε πως μπορούμε να σχεδιάσουμε έναν βέλτιστο ταξινομητή, αν όμως έχουμε γνωστές τις εκ των προτέρων πιθανότητες $P(\omega_i)$ και τις υπό συνθήκη πυκνότητες πιθανότητας $p(x|\omega_i)$. Δυστυχώς όμως στα πρακτικά προβλήματα η δομή αυτών των πιθανοτήτων είναι άγνωστη. Στις περισσότερες των περιπτώσεων έχουμε κάποια αόριστη, γενική γνώση σχετικά με την κατάσταση αυτών των πιθανοτικών δομών μαζί με έναν αριθμό επίσης από δείγματα εκπαίδευσης, που είναι αντιπροσωπευτικά των προτύπων που θέλουμε να ταξινομήσουμε. Το πρόβλημα τώρα είναι να βρούμε κάποιον τρόπο να χρησιμοποιήσουμε αυτήν την πληροφορία στο σχεδιασμό ή την εκπαίδευση του ταξινομητή.

Μια προσέγγιση στο πρόβλημα είναι να χρησιμοποιήσουμε δείγματα για να βρούμε τις άγνωστες πιθανότητες και τις πυκνότητές τους και μετά να τις χρησιμοποιήσουμε σαν να ήταν οι πραγματικές. Σε τυπικά προβλήματα αναγνώρισης προτύπων η εκτίμηση των εκ των προτέρων πιθανοτήτων δεν παρουσιάζει σοβαρές δυσκολίες. Όμως η εκτίμηση των υπό συνθήκη κατηγορίας (class-conditional) πυκνοτήτων είναι εντελώς άλλο θέμα. Ο αριθμός των διαθέσιμων δειγμάτων σχεδόν πάντα δείχνει να είναι μικρός και δημιουργούνται προβλήματα όταν το διάνυσμα χαρακτηριστικών x είναι μεγάλο. Αν γνωρίζουμε από πριν τον αριθμό των παραμέτρων και αν η γενική γνώση μας για το πρόβλημα μας επιτρέπει να παραμετροποιήσουμε τις υπο συνθήκη πυκνότητες, τότε οι δυσκολίες αυτών των προβλημάτων μπορούν να μειωθούν σημαντικά. Για παράδειγμα υποθέστε ότι $p(x|\omega_i)$ είναι μια κανονική πυκνότητα με κάποια μέση τιμή μ_i και πίνακα συνδιασποράς Σ_i., αν και δεν γνωρίζουμε τις ακριβείς τιμές,αυτών των παραμέτρων. Η γνώση αυτή απλοποιεί το πρόβλημα, αντί να ψάχνουμε να εκτιμήσουμε την άγνωστη συνάρτηση $p(x|\omega_i)$, ψάχνουμε τις παραμέτρους μ_i και Σ_i.

Το πρόβλημα της εκτίμησης παραμέτρων είναι κλασσικό στην Στατιστική και μπορεί να προσεγγιστεί με πολλούς τρόπους. Μπορούμε να θεωρήσουμε δύο γνωστές διαδικασίες, την εκτίμηση μέγιστης πιθανοφάνειας και την Bayesian εκτίμηση. Αν και τα αποτελέσματα που παίρνουμε από αυτές τις διαδικασίες μοιάζουν, εντούτοις διαφέρουν ως προς τη σύλληψη. Η εκτίμηση μέγιστης πιθανοφάνειας και άλλες μέθοδοι λαμβάνουν ως παραμέτρους κάποιες ποσότητες, οι οποίες έχουν σταθερές αλλά άγνωστες τιμές. Η βέλτιστη εκτίμηση των τιμών αυτών είναι εκείνη που μεγιστοποιεί την πιθανότητα να πάρουμε τα δείγματα που έχουμε αρχικά παρατηρήσει. Σε αντίθεση, οι μέθοδοι Bayesian εκτίμησης θεωρούν τις παραμέτρους ως τυχαίες μεταβλητές, που έχουν όμως γνωστή εκ των προτέρων κατανομή. Η παρατήρηση των δειγμάτων μετατρέπει αυτές τις κατανομές σε εκ των υστέρων πυκνότητες επιθεωρώντας έτσι τη γνώμη μας για τις πραγματικές τιμές αυτών των παραμέτρων. Στις Bayesian περιπτώσεις, θα δούμε, ότι μια τυπική επίπτωση της παρατήρησης των δειγμάτων είναι να βελτιώνουμε την εκ των υστέρων συνάρτηση πυκνότητας προκαλώντας απότομες κορυφές (peaks) κοντά στις πραγματικές τιμές των παραμέτρων. Αυτό το φαινόμενο είναι γνωστό ως μάθηση κατά Bayes. Σε οποιαδήποτε από τις δύο προαναφερθείσες περιπτώσεις χρησιμοποιούμε τις εκ των υστέρων πυκνότητες ως κανόνα ταξινόμησης όπως είδαμε προηγουμένως.

Είναι σημαντικό να θέσουμε το διαχωρισμό μεταξύ της επιβλεπόμενης μάθησης και της μη επιβλεπόμενης. Τα δείγματα x και στις 2 περιπτώσεις, υποτίθεται ότι τα

παίρνουμε επιλέγοντας μια κατάσταση από το ω_i με πιθανότητα $P(\omega_i)$ και στη συνέχεια επιλέγοντας ανεξάρτητα x σύμφωνα με το νόμο πιθανοτήτων $p(x|\omega_i)$. Η διαφορά είναι ότι με την επιβλεπόμενη μάθηση, γνωρίζουμε την κατάσταση της φύσης (ετικέτα κατηγορίας) για κάθε δείγμα, ενώ στη μη επιβλεπόμενη μάθηση η κατάσταση δεν είναι γνωστή. Όπως θα περίμενε κάποιος το πρόβλημα της μη επιβλεπόμενης μάθησης είναι δυσκολότερο. Στο κεφάλαιο αυτό θα θεωρήσουμε μόνο την περίπτωση της επιβλεπόμενης μάθησης.

3.2 Εκτίμηση Μέγιστης Πιθανοφάνειας

Η εκτίμηση μεγίστης πιθανοφάνειας έχει ένα μεγάλο αριθμό ελκυστικών ιδιοτήτων. Αρχικά έχουν σχεδόν πάντα καλές ιδιότητες σύγκλισης, όσο βέβαια ο αριθμός των δειγμάτων αυξάνει. Επίσης η εκτίμηση μεγίστης πιθανοφάνειας είναι συχνά ευκολότερη μέθοδος, από τις εναλλακτικές τεχνικές, όπως είναι η Bayesian ή άλλες μεθόδοι που παρουσιάζονται στα επόμενα υποκεφάλαια.

3.2.1 Η Γενική Αρχή

Υποθέστε ότι διαχωρίζουμε μια συλλογή δειγμάτων ανάλογα με την τάξη τους, έτσι ώστε να έχουμε c σύνολα δεδομένων, $D_1,\ldots D_c$ με τα δείγματα στο D_j να είναι επιλεγμένα ανεξάρτητα από την πιθανότητα $p(x|\omega_j)$. Τότε λέμε ότι αυτά τα δείγματα είναι α.ι.κ –ανεξάρτητες και ιδανικά κατανεμημένες τυχαίες μεταβλητές. Υποθέτουμε ότι η $p(x|\omega_j)$ έχει γνωστή παραμετρική μορφή και έτσι είναι καθορισμένη μοναδικά από την τιμή μιας παραμέτρου διανύσματος θ_j. Για παράδειγμα, μπορεί να έχουμε $p(x|\omega_j)$-$N(\mu_j,\Sigma_j)$, όπου το θ_j αποτελείται από τα μ_j και τα Σ_j. Για να δείξουμε την εξάρτηση του $p(x|\omega_j)$ και θ_j, γράφουμε το $p(x|\omega_j)$ ως $p(x|\omega_j, \theta_j)$. Το πρόβλημά μας είναι να χρησιμοποιήσουμε την πληροφορία που μας παρέχουν τα δείγματα εκπαίδευσης για να πάρουμε καλή προσέγγιση για τις άγνωστες παραμέτρους $\theta_1,\ldots\theta_c$, που σχετίζονται με κάθε κατηγορία.

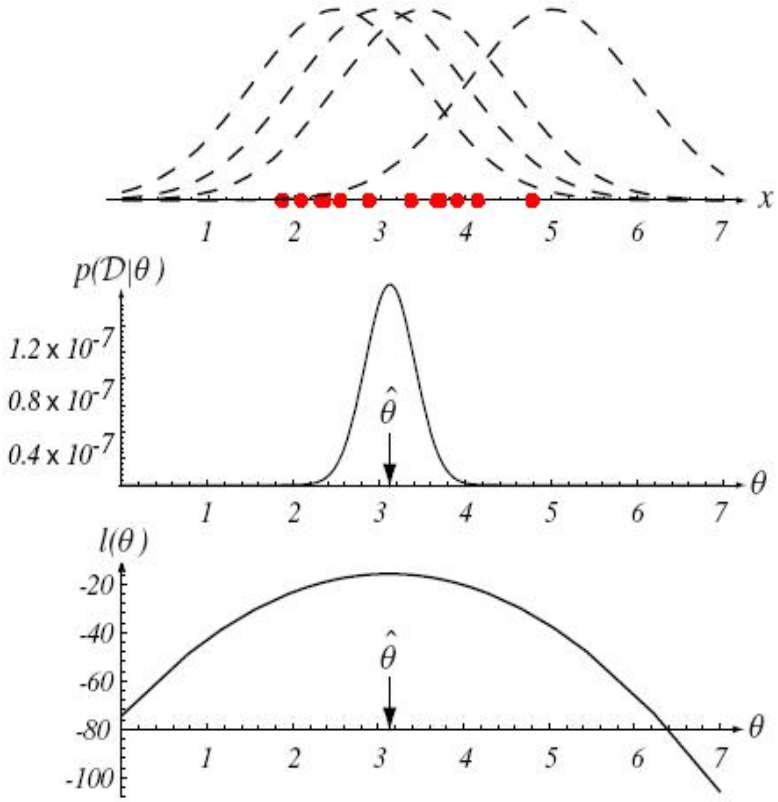
Για να απλοποιήσουμε το πρόβλημα, ας υποθέσουμε ότι τα δείγματα D_i, δε δίνουν πληροφορία για τα θ_j αν $i\neq j$, αυτό σημαίνει ότι θα υποθέσουμε ότι οι παράμετροι για τις διαφορετικές τάξεις είναι λειτουργικά ανεξάρτητες. Αυτό μας επιτρέπει να δουλέψουμε με κάθε τάξη ξεχωριστά και να απλοποιήσουμε το συμβολισμό διαγράφοντας τις ενδείξεις διαχωρισμού των κατηγοριών. Με αυτήν την υπόθεση, έχουμε c διαφορετικά προβλήματα της ακόλουθης μορφής:

Χρησιμοποιούμε ένα σύνολο D δειγμάτων εκπαίδευσης σχεδιασμένων ανεξάρτητα από την πυκνότητα πιθανότητας $p(x|\theta)$ για να εκτιμήσουμε την άγνωστη διανυσματική παράμετρο θ.

Υποθέστε ότι το D περιέχει n δείγματα, $x_1, \ldots, x_n$. Τότε επειδή τα δείγματα είναι ανεξάρτητα έχουμε:

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta) \tag{3.1}$$

Θυμηθείτε από το 2^o κεφάλαιο, ότι ο αριστερός όρος της παραπάνω εξίσωσης, αν θεωρηθεί συνάρτηση ως προς το θ καλείται πιθανοφάνεια του θ ως προς το σύνολο των δειγμάτων. Η εκτίμηση της μέγιστης πιθανοφάνειας του θ είναι εξ'ορισμού, η τιμή $\hat{\theta}$ που μεγιστοποιεί το $p(D|\theta)$. Διαισθητικά, αυτή η εκτίμηση αντιστοιχεί με μια τέτοια τιμή θ που, κατά κάποια έννοια να συμφωνεί ή να συμβαδίζει με τα πραγματικά, παρατηρούμενα, δείγματα εκπαίδευσης (Εικόνα 3.1).



Εικόνα 3.1

Για αναλυτικούς σκοπούς είναι συνήθως ευκολότερο να δουλεύει κανείς με το λογάριθμο της πιθανοφάνειας. Επειδή, ο λογάριθμος είναι αύξουσα μονότονη συνάρτηση, το $\hat{\theta}$ που μεγιστοποιεί την log-πιθανοφάνεια, μεγιστοποιεί επίσης την πιθανοφάνεια. Αν το $p(D \mid \boldsymbol{\theta})$ συμπεριφέρεται καλά, η διαφορίσιμη συνάρτηση του θ, το $\hat{\theta}$ μπορεί να βρεθεί με μεθόδους διαφορικής λογικής. Αν επίσης, ο αριθμός των παραμέτρων που πρέπει να εκτιμηθούν είναι p, τότε με το θ δηλώνουμε το p-συνιστωσών διάνυσμα $\theta = (\theta_1,...,\theta_p)^t$ και το ανάδελτα του θ ορίζεται ως:

$$\nabla\theta = \begin{bmatrix} \dfrac{\partial}{\partial\theta_1} \\ \vdots \\ \dfrac{\partial}{\partial\theta_p} \end{bmatrix}. \tag{3.2}$$

Ορίζουμε το l(θ) ως την log-συνάρτηση πιθανοφάνειας:

$$l(\boldsymbol{\theta}) \equiv \ln p(D \mid \boldsymbol{\theta}) \tag{3.3}$$

Μπορούμε να γράψουμε τη δική μας λύση τυπικά, ως το θ που μεγιστοποιεί την log-likelihood:

$$\hat{\theta} = \arg\max_{\theta} l(\theta), \tag{3.4}$$

όπου η εξάρτηση στο σύνολο D είναι έμμεση. Έτσι έχουμε από την εξίσωση 3.1:

$$l(\theta) = \sum_{\kappa=1}^{n} \ln p(x_\kappa | \theta) \tag{3.5}$$

και

$$\nabla_\theta l = \sum_{\kappa=1}^{n} \nabla_\theta \ln p(x_\kappa | \theta). \tag{3.6}$$

Έτσι ένα σύνολο από απαιτούμενες συνθήκες για τη μέγιστη εκτίμηση πιθανοφάνειας για το θ μπορεί να εξαχθεί από το σύνολο των p εξισώσεων:

$$\nabla_\theta l = 0 \tag{3.7}$$

Μία λύση έστω $\hat{\theta}$ της εξίσωσης 3.7, μπορεί να αντιπροσωπεύει ένα αληθινό, ολικό μέγιστο, ένα τοπικό μέγιστο ή ελάχιστο ή σπανιότερα ένα σημείο του l(θ). Πρέπει στο σημείο αυτό να είμαστε προσεκτικοί ώστε να ελέγξουμε εάν το ακραίο σημείο βρίσκεται στο όριο του χώρου των παραμέτρων, πράγμα το οποίο μπορεί να μην γίνει προφανές από την λύση της εξίσωσης 3.7. Αν βρεθούν όλες οι λύσεις εγγυόμαστε ότι η μια αναπαριστά το αληθινό μέγιστο, όμως μπορεί να θέλουμε να ελέγξουμε κάθε λύση ατομικά (η να υπολογίσουμε τις δεύτερες παραγώγους) για να αναγνωρίσουμε πιο είναι το ολικό βέλτιστο. Πάντως πρέπει να μην ξεχνάτε ότι το $\hat{\theta}$ είναι απλά μια εκτίμηση και ότι είναι μονάχα στο όριο ενός απειροστού αριθμού σημείων εκπαίδευσης, που μπορεί να περιμένουμε ότι η εκτίμηση μας, θα ισοδυναμεί με την αληθινή τιμή της συνάρτησης που παράγεται.

Παρατηρήσατε μέχρι εδώ ότι η σχετική τάξη των εκτιμητών –μέγιστη εκ των υστέρων ή MAP (Maximum A Posteriori)- βρίσκει την τιμή του θ που μεγιστοποιεί το l(θ)p(θ), όπου το p(θ) περιγράφει την εκ των προτέρων πιθανότητα των διαφορετικών τιμών παραμέτρων. Ο MAP εκτιμητής βρίσκει την κορυφή ή την κατάσταση της εκ των υστέρων πυκνότητας. Το μειονέκτημα των MAP εκτιμητών είναι ότι αν επιλέξουμε κάποιον αφηρημένο μη γραμμικό μετασχηματισμό του χώρου των παραμέτρων, (για παράδειγμα μια συνολική περιστροφή), η πυκνότητα θα αλλάξει και η MAP λύση, δε θα είναι πλέον η κατάλληλη.

3.2.2 Η περίπτωση Gauss: Το άγνωστο μ

Για να δείξουμε πως οι μέθοδοι της μέγιστης πιθανοφάνειας εφαρμόζονται σε μια συγκεκριμένη περίπτωση, υποθέτουμε ότι τα δείγματα προέρχονται από έναν κανονικό πληθυσμό, που όμως εμφανίζει μεγάλες αποκλίσεις με μέση τιμή μ και πίνακα συνδιασποράς Σ. Για λόγους απλότητας, θεωρήστε πρώτα την περίπτωση όπου μόνο η μέση τιμή είναι άγνωστη. Κάτω από αυτές τις συνθήκες, θεωρούμε ένα σημείο δείγμα x_κ και βρίσκουμε:

$$\ln p(x_\kappa | \mu) = -\frac{1}{2}\ln[(2\pi)^d |\Sigma|] - \frac{1}{2}(x_k - \mu)^t \Sigma^{-1}(x_k - \mu) \tag{3.8}$$

και

$$\nabla_\mu \ln p(x_\kappa | \mu) = \Sigma^{-1}(x_k - \mu) \tag{3.9}$$

Αναγνωρίζοντας το θ από το μ, βλέπουμε από τις εξισώσεις 3.6, 3.7 και 3.9 ότι η εκτίμηση μέγιστης πιθανοφάνειας για το μ πρέπει να ικανοποιεί:

$$\sum_{\kappa=1}^{n} \Sigma^{-1}(x_k - \hat{\mu}) = 0 \tag{3.10}$$

Πολλαπλασιάζοντας με Σ και αναδιαμορφώνοντας έχουμε:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.11}$$

Αυτό είναι ένα πολύ ικανοποιητικό αποτέλεσμα. Λέει ότι η εκτίμηση μέγιστης πιθανοφάνειας για τον άγνωστο μέσο όρο του πληθυσμού είναι απλά ο αριθμητικός μέσος όρος των δειγμάτων εκπαίδευσης –η μέση τιμή του δείγματος- που κάποιες φορές γράφεται $\hat{\mu}_n$, για να διευκρινίσει την εξάρτησή του από τον αριθμό των δειγμάτων. Γεωμετρικά αν σκεφτούμε τα n δείγματα ως σύννεφο σημείων, ο μέσος όρος του δείγματος είναι στο κέντρο του σύννεφου. Ο μέσος όρος του δείγματος έχει ένα αριθμό επιθυμητών στατιστικών ιδιοτήτων όπως επίσης και κάποιος θα μπορούσε να χρησιμοποιήσει αυτήν την εκτίμηση χωρίς ακόμη να γνωρίζει ότι πρόκειται για την επίλυση μέγιστης πιθανοφάνειας.

3.2.3 Η περίπτωση Gauss: Το άγνωστο μ και Σ

Σε μια πιο γενική (και πιο τυπική) περίπτωση, με δείγματα που προέρχονται από έναν κανονικό πληθυσμό, που όμως εμφανίζει μεγάλες αποκλίσεις, ούτε η μέση τιμή μ ούτε ο πίνακας συνδιασποράς Σ είναι γνωστός. Έτσι αυτές οι άγνωστες παράμετροι αποτελούν τις συνιστώσες του διανύσματος παραμέτρων θ. Θεωρήστε αρχικά την περίπτωση όπου $\theta_1=\mu$ και $\theta_2=\sigma^2$. Εδώ το log-likelihood σημείο είναι ένα απλό σημείο:

$$\ln p(x_\kappa|\theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_\kappa - \theta_1)^2 \tag{3.12}$$

και η παράγωγός του:

$$\nabla_\theta l = \nabla_\theta \ln p(x_\kappa \mid \theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_\kappa - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_\kappa - \theta_1)^2}{2\theta_2^{\,2}} \end{bmatrix} \tag{3.13}$$

Εφαρμόζοντας την εξίσωση 3.7 στην πλήρη log-likelihood οδηγούμαστε στις συνθήκες:

$$\sum_{\kappa=1}^{n} \frac{1}{\hat{\theta}_2}(x_\kappa - \hat{\theta}_1) = 0 \tag{3.14}$$

και

$$-\sum_{\kappa=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{\kappa=1}^{n} \frac{(x_\kappa - \hat{\theta}_1)^2}{\hat{\theta}_2^{\,2}} = 0 \tag{3.15}$$

όπου το $\hat{\theta}_1$ και το $\hat{\theta}_2$ είναι οι εκτιμήσεις για τη μέγιστη πιθανοφάνεια για τα θ_1 και θ_2, αντίστοιχα. Με την αντικατάσταση του $\hat{\mu} = \hat{\theta}_1$ και $\hat{\sigma}^2 = \hat{\theta}_2$ παίρνουμε:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.16}$$

και

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \quad \text{(Μονοδιάστατη Gauss)} \tag{3.17}$$

Ενώ η ανάλυση της περίπτωσης κατά την οποία υπάρχουν αποκλίσεις στον πληθυσμό είναι παρόμοια, εμπλέκονται συγκριτικά πολύ περισσότεροι χειρισμοί. Όπως θα είχαμε προβλέψει πάντως, το αποτέλεσμα είναι ότι η εκτίμηση της μέγιστης πιθανοφάνειας για το μ και το Σ δίνονται από τους τύπους:

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.18}$$

και

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t \quad \text{(Γενική Περίπτωση – Πολυδιάστατη Gauss)} \tag{3.19}$$

Έτσι, ακόμη μια φορά βρήκαμε ότι η μέγιστη πιθανοφάνεια, για το διάνυσμα της μέσης τιμής είναι απλά η μέση τιμή του δείγματος. Η εκτίμηση της μέγιστης πιθανοφάνειας για τον πίνακα συνδιασποράς είναι ο αριθμητικός μέσος όρος των n μητρείων $(x_\kappa - \hat{\mu})(x_\kappa - \hat{\mu})^t$. Επειδή ο αληθινός πίνακας συνδιασποράς είναι η προσδοκώμενη τιμή του πίνακα $(x_\kappa - \hat{\mu})(x_\kappa - \hat{\mu})^t$ αυτό είναι ένα πολύ σημαντικό αποτέλεσμα.

3.2.4 Πόλωση (Bias)

Η εκτίμηση της μέγιστης πιθανοφάνειας για τη διασπορά σ^2 είναι πολωμένη δηλαδή, η προσδοκώμενη τιμή όλων των συνόλων δεδομένων πλήθους n του δείγματος διασποράς δεν είναι ίση με την αληθινή διασπορά.

$$E\left[\frac{1}{n}\sum_{k=1}^{n}(x_i - \bar{x})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{3.20}$$

Μπορούμε να επαληθεύσουμε την εξίσωση 3.20 για την υποκείμενη κατανομή με μη μηδενική διασπορά σ^2 και στην ακραία περίπτωση του n=1, στην οποία προσδοκώμενη τιμή δίνεται από $E[\cdot] = 0 \neq \sigma^2$. Η εκτίμηση μέγιστης πιθανοφάνειας του πίνακα συνδιασποράς είναι παρομοίως πολωμένη.
Ένας στοιχειώδης μη πολωμένος εκτιμητής για το Σ δίνεται από:

$$C = \frac{1}{n-1}\sum_{k=1}^{n}(x_\kappa - \hat{\mu})(x_\kappa - \hat{\mu})^\tau \tag{3.21}$$

όπου το C είναι το λεγόμενο δείγμα του πίνακα συνδιασποράς. Αν ένας εκτιμητής είναι μη πολωμένος για όλες τις κατανομές, όπως για παράδειγμα ο εκτιμητής διασποράς στην εξίσωση 3.21, τότε λέγεται απόλυτα μη πολωμένος. Αν ο εκτιμητής τείνει να γίνει μη πολωμένος όσο ο αριθμός των δειγμάτων γίνεται πολύ μεγάλος, όπως για παράδειγμα στην εξίσωση 3.20, τότε ο εκτιμητής είναι ασυμπτωτικά μη

πολωμένος. Σε πολλά προβλήματα αναγνώρισης προτύπων με μεγάλο σύνολο δεδομένων εκπαίδευσης, οι ασυμπτωτικά μη πολωμένοι εκτιμητές είναι αποδεκτοί.

Είναι φανερό ό,τι $\hat{\Sigma} = [(n-1)/n]C$ και το $\hat{\Sigma}$ είναι ασυμπτωτικά μη πολωμένα. Αυτοί οι δύο εκτιμητές είναι ιδανικά όμοιοι για μεγάλα n. Πάντως, η ύπαρξη δύο παρόμοιων, μα και συνάμα διακριτών εκτιμητών για τον πίνακα συνδιασποράς μπορεί να είναι ανησυχητική και είναι φυσιολογικό να τίθεται το ερώτημα ποιο από τα δύο είναι ‘‘σωστό’’. Βέβαια για n > 1 η απάντηση είναι ότι αυτοί οι εκτιμητές δεν είναι ούτε σωστοί ούτε λάθος – είναι απλώς διαφορετικοί. Αυτό που όμως πραγματικά δείχνει η ύπαρξη και των δυο είναι ότι καμιά δεν κατέχει όλες τις επιθυμητές ιδιότητες. Για τους σκοπούς μας, η πιο επιθυμητή ιδιότητα είναι πολύ περίπλοκη – θέλουμε η εκτίμηση να οδηγεί στη βέλτιστη απόδοση ταξινόμησης. Ενώ είναι συνήθως συνάμα λογικό και ορθό να σχεδιάζουμε τον ταξινομητή αντικαθιστώντας την εκτίμηση της μέγιστης πιθανοφάνειας για άγνωστες παραμέτρους, μπορεί επίσης να αναρωτηθούμε αν άλλοι εκτιμητές μπορεί να μην οδηγούν σε καλύτερη απόδοση. Στη συνέχεια αντιμετωπίζουμε αυτό το ερώτημα από την Bayesian πλευρά.

Αν έχουμε ένα αξιόπιστο μοντέλο για τις υποκείμενες κατανομές και για τις εξαρτήσεις τους από το παραμετρικό διάνυσμα θ, τότε ένας ταξινομητής μεγίστης πιθανοφάνειας μπορεί να έχει βέλτιστα αποτελέσματα. Αλλά αν το μοντέλο μας είναι λάθος; Παίρνουμε έτσι κι αλλιώς τον βέλτιστο ταξινομητή από τα υποψήφια μοντέλα μας; Για παράδειγμα, τι γίνεται αν υποθέσουμε ότι μια κατανομή προέρχεται από N(μ,1) αλλά στην πραγματικότητα προέρχεται από N(μ,10); Η τιμή που βρούμε για θ=μ από τη μέγιστη πιθανοφάνεια θα παράγει τον καλύτερο ταξινομητή της μορφής N(μ,1); Δυστυχώς, η απάντηση είναι “όχι”. Συνεπώς αυτό που χρειάζεται (θυμηθείτε και από το $1^{ο}$ κεφάλαιο) είναι η αξιόπιστη πληροφορία σχετικά με το είδος του συνόλου των μοντέλων, ώστε να υποθέτουμε όχι πολύ ‘φτωχά’ μοντέλα.

3.3 Εκτίμηση κατά Bayes

Τώρα θεωρούμε έναν εκτιμητή κατά Bayes ή αλλιώς μάθηση κατά Bayes σε προβλήματα αναγνώρισης προτύπων. Αν και θα πάρουμε παρόμοιες απαντήσεις με τις μεθόδους μέγιστης πιθανοφάνειας, εντούτοις υπάρχει μια βαθύτερη νοηματική διαφορά: Ενώ στις μεθόδους μέγιστης πιθανοφάνειας θεωρήσαμε ως σταθερό το διάνυσμα παραμέτρων θ, στην μάθηση κατά Bayes θεωρούμε το θ να είναι τυχαία μεταβλητή και τα δεδομένα εκπαίδευσης μας επιτρέπουν να μετατρέψουμε την κατανομή αυτής της μεταβλητής σε εκ των υστέρων πυκνότητα πιθανότητας.

3.3.1 Οι υπό συνθήκη πυκνότητες

Ο υπολογισμός των εκ των υστέρων πιθανοτήτων $P(\omega | x_1)$ βρίσκεται στην καρδιά της ταξινόμησης κατά Bayes. Ο τύπος του Bayes, μας επιτρέπει να υπολογίζουμε αυτές τις πιθανότητες, από τις εκ των προτέρων πιθανότητες $P(\omega_i)$ και τις υπό συνθήκη κατηγορίας πυκνότητες $P(x | \omega_i)$, αλλά πως γίνεται κάτι τέτοιο αν αυτές οι τιμές είναι άγνωστες; Η γενική απάντηση είναι, ότι το καλύτερο που έχουμε να κάνουμε είναι να υπολογίσουμε το $P(\omega | x_1)$, χρησιμοποιώντας όλη την πληροφορία, που έχουμε στην κατοχή μας. Μέρος αυτής της πληροφορίας μπορεί να είναι εκ των προτέρων γνώση, όπως γνώση των συναρτησιακών τύπων για άγνωστες πυκνότητες και περιοχές για τις τιμές των άγνωστων παραμέτρων. Μέρος της πληροφορίας μπορεί να βρίσκεται στο σύνολο των δειγμάτων εκπαίδευσης. Αν πάλι δηλώσουμε ως D το σύνολο των δειγμάτων, τότε μπορούμε να δώσουμε έμφαση στο ρόλο των

δειγμάτων με το ότι ο στόχος μας είναι να υπολογίσουμε τις εκ των υστέρων πιθανότητες $P(\omega | x_l, D)$. Από αυτές τις πιθανότητες μπορούμε να πάρουμε τον ταξινομητή κατά Bayes.
Δοθέντος του δείγματος D, ο τύπος του Bayes γίνεται:

$$P(\omega_i | x, D) = \frac{P(x | \omega_i, D) P(\omega_i | D)}{\sum_{j=1}^{c} p(x | \omega_j, D) P(\omega_j | D)} \tag{3.22}$$

Όπως αυτή η εξίσωση προτείνει, μπορούμε να χρησιμοποιήσουμε την πληροφορία που μας παρέχουν τα δείγματα εκπαίδευσης, για να βοηθηθούμε στο να καθορίσουμε και τις υπό συνθήκη κατηγορία πυκνότητες και τις εκ των προτέρων πιθανότητες.
Αν και θα μπορούσαμε να διατηρήσουμε αυτή τη γενικότητα, από δω και στο εξής θα υποθέσουμε ότι οι πραγματικές τιμές των εκ των προτέρων πιθανοτήτων είναι γνωστές ή λαμβάνονται μέσα από πολύ εύκολους υπολογισμούς. Έτσι αντικαθιστούμε : $P(\omega_i) = P(\omega_i | D)$. Επιπλέον, επειδή εξετάζουμε την περίπτωση με επίβλεψη (supervised), μπορούμε να διαχωρίσουμε τα δείγματα εκπαίδευσης ανά κατηγορία σε c υποσύνολα $D_1, \ldots D_c$, με τα δείγματα στο D_i να ανήκουν στο ω_i. Όπως αναφέρθηκε όταν ασχοληθήκαμε με τις μεθόδους μέγιστης πιθανοφάνειας. Στις περισσότερες ενδιαφέρουσες περιπτώσεις (και σε όλες όσες θα θεωρήσουμε στη συνέχεια) τα δείγματα στο D_i δεν έχουν καμιά επίδραση στο $P(x | \omega_j, D)$ if $i \neq j$. Αυτό έχει δύο κυρίως αποτελέσματα. Αρχικά, μας επιτρέπει να δουλέψουμε με κάθε κατηγορία ξεχωριστά, χρησιμοποιώντας μόνο τα δείγματα του D_i για να καθορίσουμε το $P(x | \omega_i, D)$. Αν θεωρήσουμε ότι οι εκ των προτέρων πιθανότητες είναι γνωστές τότε η εξίσωση 3.22 γίνεται:

$$P(\omega_i | x, D) = \frac{p(x | \omega_i, D_i) P(\omega_i)}{\sum_{j=1}^{c} p(x | \omega_j, D_j) P(\omega_j)} \tag{3.23}$$

Έπειτα, επειδή κάθε κλάση μπορεί να αντιμετωπιστεί ανεξάρτητα, μπορούμε να προχωρήσουμε σε απλοποίηση των συμβολισμών, με το να μην λάβουμε υπόψη τους αχρείαστους περιορισμούς τάξεων. Ουσιαστικά έχουμε c διαφορετικά προβλήματα της ακόλουθης μορφής: Χρησιμοποιήστε ένα σύνολο D δειγμάτων σχηματισμένων ανεξάρτητα, αλλά σχετικά με την σταθερή αλλά άγνωστη πιθανοτική κατανομή p(x) για να καθορίσουμε το p(x|D). Αυτό είναι το κυρίως πρόβλημα στην μάθηση κατά Bayes.

3.3.2 Κατανομή παραμέτρων

Παρόλο που η επιθυμητή πυκνότητα πιθανότητας p(x) είναι άγνωστη, θα υποθέσουμε ότι έχει γνωστή παραμετρική μορφή. Το μόνο λοιπόν πράγμα που είναι άγνωστο είναι η τιμή του διανύσματος των παραμέτρων θ. Θα εκφράσουμε το γεγονός ότι το p(x) είναι άγνωστο αλλά έχει γνωστή παραμετρική μορφή με το να λέμε ότι η συνάρτηση p(x|θ) είναι γνωστή. Κάθε πληροφορία που μπορεί να έχουμε σχετικά με το θ πριν να παρατηρήσουμε τα δείγματα υποτίθεται ότι περιέχεται σε μια εκ των προτέρων πιθανότητα p(θ). Η παρατήρηση των δειγμάτων μετατρέπει αυτό σε εκ των υστέρων πυκνότητα πιθανότητας p(θ|D), για την οποία ελπίζουμε να εμφανίζει κορυφή (peak) στο πραγματικό σημείο θ.
Παρατηρείστε ότι καταφέραμε να μετατρέψουμε το πρόβλημα του να μάθουμε την συνάρτηση της πυκνότητας πιθανότητας σε εκτίμηση του παραμετρικού διανύσματος.

Ο βασικός μας σκοπός είναι να υπολογίσουμε το p(x|D), το οποίο είναι κοντά στο άγνωστο p(x). Αυτό το κάνουμε ολοκληρώνοντας την συνδυασμένη πυκνότητα πιθανότητας p(x, θ|D) ως προς θ. Δηλαδή:

$$p(x|D) = \int p(x,\theta|D)d\theta \tag{3.24}$$

όπου η ολοκλήρωση εκτείνεται σε όλο το χώρο των παραμέτρων. Τώρα μπορούμε πάντα να γράφουμε
$p(x,\theta|D)$ ως το γινόμενο $p(x|\theta,D)p(\theta|D)$. Επειδή η επιλογή του x και αυτή των δειγμάτων εκπαίδευσης D γίνονται ανεξάρτητα, ο πρώτος παράγοντας είναι αποκλειστικά και μόνο p(x|θ). Με άλλα λόγια η κατανομή του x είναι πλήρως γνωστή από τη στιγμή που γνωρίζουμε την τιμή του διανύσματος παραμέτρων. Έτσι η εξίσωση 3.24 μπορεί να ξαναγραφεί ως:

$$p(x|D) = \int p(x|\theta)p(\theta|D)d\theta \tag{3.25}$$

Αυτή η εξίσωση κλειδί συνδέει την επιθυμητή υπό συνθήκη κατηγορίας πυκνότητα p(x|D) με την εκ των υστέρων πυκνότητα p(θ|D) για το διάνυσμα παραμέτρων. Αν το p(θ|D) εμφανίζει αιχμηρή κορυφή (peak) γύρω από κάποια τιμή $\hat{\theta}$, παίρνουμε $p(x|D) \cong p(x|\hat{\theta})$, για παράδειγμα το αποτέλεσμα που θα παίρναμε αντικαθιστώντας την εκτίμηση $\hat{\theta}$ με την πραγματική τιμή του διανύσματος παραμέτρων. Αυτό το αποτέλεσμα έγκειται στην υπόθεση ότι το p(x|θ) είναι ομαλό και ότι τα όρια του ολοκληρώματος δεν έχουν μεγάλη σημασία. Αυτές οι περιπτώσεις είναι τυπικές, αλλά όχι η γενική περίπτωση. Γενικότερα αν δεν είμαστε σίγουροι για την ακριβή τιμή του θ, η παραπάνω εξίσωση μας οδηγεί απευθείας στο μέσο όρο p(x|θ) από όλες τις πιθανές τιμές του θ. Έτσι, όταν οι άγνωστες πυκνότητες έχουν γνωστό παραμετρικό τύπο, τα δείγματα ασκούν την επιρροή τους στο p(x|D) διαμέσου της εκ των υστέρων πυκνότητας p(θ|D). Θα πρέπει επίσης να υπογραμμίσουμε ότι στην πράξη η ολοκλήρωση της εξίσωσης 3.25 μπορεί να γίνει αριθμητικά για παράδειγμα με μέθοδο Monte-Carlo simulation.

3.4 Bayesian Εκτίμηση Παραμέτρων: Η Gaussian Περίπτωση

Σε αυτήν την ενότητα χρησιμοποιούμε τις Bayesian μεθόδους για να υπολογίσουμε την εκ των υστέρων πυκνότητα p(θ|D) καθώς και την επιθυμητή πυκνότητα πιθανότητας p(x|D) για την περίπτωση όπου $p(x|\mu) \sim N(\mu,\Sigma)$.

3.4.1 Η περίπτωση μιας μεταβλητής (univariate)

Θεωρήστε την περίπτωση όπου έχουμε το μ ως τη μοναδική άγνωστη παράμετρο. Για απλότητα θεωρούμε αυτήν την περίπτωση ως:

$$p(x|\mu) \sim N(\mu,\sigma^2) \tag{3.26}$$

όπου η μοναδική άγνωστη ποσότητα είναι η μέση τιμή μ. Υποθέτουμε ότι οποιαδήποτε εκ των προτέρων γνώση μπορεί να έχουμε για το μ, μπορεί να εκφραστεί με την γνωστή εκ των προτέρων πυκνότητα p(μ). Αργότερα θα κάνουμε μια πιο εκτεταμένη υπόθεση ότι:

$$p(\mu) \sim N(\mu_o,\sigma_o^2) \tag{3.27}$$

όπου συνάμα τα μ_o και σ_o^2, είναι γνωστά. Μιλώντας πρόχειρα, το μ_o αναπαριστά την καλύτερη από πριν πρόβλεψή μας για το μ και το σ_o^2 μετράει την αβεβαιότητά μας για αυτήν την πρόβλεψη. Η υπόθεση ότι, η από πριν κατανομή για το μ είναι κανονική θα απλοποιήσει τα ακολουθούμενα μαθηματικά. Πάντως, το σημαντικό εδώ δεν είναι η από πριν υπόθεση ότι η κατανομή του μ είναι κανονική αλλά ότι είναι γνωστή η κατανομή του!

Έχοντας επιλέξει την εκ των προτέρων πυκνότητα για το μ, μπορούμε να δούμε την κατάσταση ως ακολούθως. Φανταστείτε ότι μια τιμή επιλέγεται για το μ, από ένα πληθυσμό, που διέπεται από το νόμο πιθανοτήτων p(μ). Μόλις αυτή η τιμή σχηματιστεί γίνεται η πραγματική τιμή του μ και ολοκληρωτικά καθορίζει την πυκνότητα για το x. Υποθέστε τώρα ότι n δείγματα $x_1,\ldots x_n$ επιλέγονται ανεξάρτητα από τον εναπομείναντα πληθυσμό. Αν D= $[x_1,..x_n]$, χρησιμοποιούμε τον τύπο του Bayes για να πάρουμε:

$$p(\mu \mid D) = \frac{p(D \mid \mu)\,p(\mu)}{\int p(D \mid \mu)\,p(\mu)\,d\mu} = \alpha \prod_{k=1}^{n} p(x_k \mid \mu)\,p(\mu) \qquad (3.28)$$

όπου το α είναι ένας παράγοντας εξομάλυνσης που εξαρτάται από το D αλλά είναι ανεξάρτητο του μ. Αυτή η εξίσωση δείχνει το πώς η παρατήρηση ενός συνόλου δειγμάτων εκπαίδευσης επηρεάζει τις ιδέες μας σχετικά με την πραγματική τιμή του μ. Συνδέει την εκ των προτέρων πυκνότητα p(μ) με την εκ των υστέρων πυκνότητα p(μ|D). Επειδή $p(x_k \mid \mu) \sim N(\mu, \sigma^2)$ και $p(\mu) \sim N(\mu_0, \sigma_0^2)$ έχουμε:

$$\begin{aligned} p(\mu|\mathcal{D}) &= \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)} \\ &= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right] \\ &= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right], \end{aligned} \qquad (29)$$

όπου οι παράγοντες που δεν εξαρτώνται από το μ έχουν δώσει τη θέση τους στα α, α' και α''. Έτσι το p(μ|D) είναι μια εκθετική συνάρτηση μιας τετραγωνικής συνάρτησης του μ. Για παράδειγμα είναι ξανά κανονική πυκνότητα. Επειδή αυτό ισχύει για οποιοδήποτε αριθμό δειγμάτων εκπαίδευσης, το p(μ|D) παραμένει κανονικό όσο ο αριθμός των n δειγμάτων μειώνεται, το p(μ|D) λέγεται ότι είναι αναπαράγουσα πυκνότητα (reproducing density) και το p(μ) λέγεται προηγούμενος συζυγής (conjugate prior). Αν γράψουμε $p(\mu \mid D) \sim N(\mu_n, \sigma_n^2)$ τότε τα μ_n και σ_n^2 μπορούν να βρεθούν εξισώνοντας τους συντελεστές στην εξίσωση 3.29 με τους αντίστοιχους συντελεστές στη γενική Gaussian μορφή:

$$p(\mu \mid D) = \frac{1}{\sqrt{2\pi}\sigma_n} = \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] \qquad (3.30)$$

Αναγνωρίζοντας τους συντελεστές με αυτόν τον τρόπο παράγεται:

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_o^2} \tag{3.31}$$

και

$$\frac{\mu_n}{\sigma_n^2} = \frac{n}{\sigma^2}\hat{\mu}_n + \frac{\mu_o}{\sigma_o^2} \tag{3.32}$$

όπου το $\hat{\mu}_n$ είναι ο μέσος των δειγμάτων:

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.33}$$

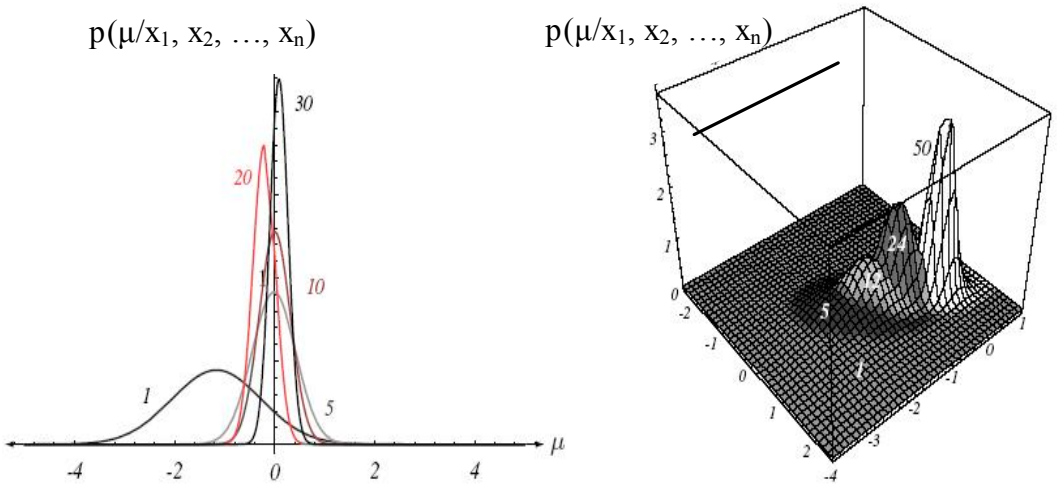
Λύνοντας ως προς μ_n και σ_n^2 παίρνουμε:

$$\mu_n = \left(\frac{n\sigma_o^2}{n\sigma_o^2 + \sigma^2}\right)\hat{\mu}_n + \frac{\sigma^2}{n\sigma_o^2 + \sigma^2}\mu_o \tag{3.34}$$

και

$$\sigma_n^2 = \frac{\sigma_o^2\sigma^2}{n\sigma_o^2 + \sigma^2} \tag{3.35}$$

Αυτές οι εξισώσεις δείχνουν πως η εκ των προτέρων πληροφορία συνδυάζεται με την εμπειρική πληροφορία που μας δίνουν τα δείγματα για να πάρουμε την εκ των υστέρων πυκνότητα p(μ|D). Μιλώντας πρόχειρα, το μ_n αναπαριστά την καλύτερη πρόβλεψή μας για το μ, αν προηγουμένως έχουμε παρατηρήσει τα n δείγματα, ενώ το σ_n^2 μετράει την αβεβαιότητά μας για αυτήν την μαντεψιά. Επειδή το σ_n^2 μειώνεται μονότονα με το n -να φτάνει το σ_n^2/n, όσο το n τείνει στο άπειρο- κάθε πρόσθετη παρατήρηση μειώνει την αβεβαιότητά μας για την αληθινή τιμή του μ. Όσο το n αυξάνει, γίνεται όλο και περισσότερη αιχμηρή η κορυφή (peak) του p(μ|D), πλησιάζοντας (θυμηθείτε από τα σήματα I) τη συνάρτηση δέλτα Dirac όσο το n τείνει στο άπειρο. Αυτή η συμπεριφορά είναι κοινώς γνωστή ως Bayesian μάθηση (σχήμα 3.2).



Εικόνα 3.2

Γενικά, το μ_n είναι γραμμικός συνδυασμός του $\hat{\mu}_n$ και του μ_o, με μη μηδενικούς συντελεστές και άθροισμα ίσο με 1. Έτσι το μ_o πάντα βρίσκεται κάπου μεταξύ του $\hat{\mu}_n$ και του μ_o. Αν $\sigma_o \neq 0$, το μ_n πλησιάζει τη μέση τιμή του δείγματος όσο το n αυξάνεται στο άπειρο. Αν $\sigma_o = 0$ έχουμε μια εκφυλισμένη περίπτωση όπου, με την εκ των προτέρων βεβαιότητά μας ότι το $\mu_o = \mu$, είναι τόσο ισχυρή ώστε κανένας αριθμός παρατηρήσεων δεν αλλάζει την εκτίμησή μας. Από την άλλη ακραία περίπτωση όταν $\sigma_o >> \sigma$ είμαστε τόσο αβέβαιοι για την μαντεψιά μας ώστε παίρνουμε $\mu_n = \hat{\mu}_n$ χρησιμοποιώντας μόνο τα δείγματα για να εκτιμήσουμε το μ. Γενικά η σχετική ισορροπία μεταξύ εκ των προτέρων γνώσης και εμπειρικών δεδομένων ορίζεται από την αναλογία του σ^2 ως προς το σ_o^2, που κάποιες φορές μπορεί να το συναντήσετε και ως δογματισμό. Αν ο δογματισμός δεν είναι άπειρος μετά από αρκετά δείγματα οι ακριβείς τιμές για τα μ_n και σ_o^2 θα είναι ελάσσονος σημασίας και επιπλέον το μ_o θα συγκλίνει στη μέση τιμή.

3.4.2. Η Περίπτωση μιας μεταβλητής (univariate): p(x|D)

Έχοντας την a posteriori πυκνότητα για τη μέση τιμή, p(μ|D), ότι περισσεύει είναι να πάρουμε και την ‘‘υπό συνθήκης κατηγορίας’’ πυκνότητα για το p(x|D), (ουσιαστικά αυτό είναι το ίδιο με το $P(x|\omega_i, D_i)$. Από τις εξισώσεις 3.25, 3.26 και 3.30 έχουμε:

$$\begin{aligned} p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\,d\mu \\ &= \int \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu \\ &= \frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma,\sigma_n), \end{aligned} \qquad 3.36$$

όπου

$$f(\sigma,\sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu.$$

Αυτό είναι μια συνάρτηση του x, με το p(x|D) να είναι ανάλογο με το $\exp[-(1/2)(x-\mu_n)^2/(\sigma^2+\sigma_n^2)]$ και επίσης το p(x|D) να είναι κανονικά κατανεμημένο με μέση τιμή μ_n και διασπορά $\sigma^2+\sigma_n^2$:

$$p(x|D) \sim N(\mu_n, \sigma^2+\sigma_n^2) \qquad (3.37)$$

Με άλλα λόγια για να πάρουμε το p(x|D), του οποίου η παραμετρική μορφή είναι γνωστό ότι είναι: $p(x|\mu) \sim N(\mu, \sigma^2)$, το μόνο που κάνουμε είναι να αντικαταστήσουμε το μ με το μ_n και το σ^2 με το $\sigma^2+\sigma_n^2$. Ως επίδραση, η υποθετική

μέση τιμή μ_n συμπεριφέρεται σαν να ήταν η πραγματική και η γνωστή διασπορά αυξάνει λόγω της πρόσθετης αβεβαιότητας του x, που είναι αποτέλεσμα της έλλειψης γνώσης για τη μέση τιμή μ. Αυτό τώρα είναι το τελικό μας αποτέλεσμα: Η πυκνότητα p(x|D) είναι η επιθυμητή υπό συνθήκη κατηγορία πυκνότητα $P(x|\omega_j, D_j)$ και μαζί με τις εκ των προτέρων πιθανότητες $P(\omega_j)$ μας δίνουν την πιθανοτική απαραίτητη πληροφορία για το σχεδιασμό του ταξινομητή. Αυτό έρχεται σε αντίθεση με τις μεθόδους της μέγιστης πιθανοφάνειας, οι οποίες μας δίνουν μόνο εκτιμήσεις για τα $\hat{\mu}$ και $\hat{\sigma}^2$ παρά εκτιμήσεις για την κατανομή του p(x|D)

3.4.3 Η Περίπτωση πολλών μεταβλητών (multivariate)

Η αντιμετώπιση της Multivariate περίπτωσης στην οποία το Σ είναι γνωστό αλλά το μ δεν είναι, αποτελεί την άμεση γενίκευση της. Για το λόγο αυτό θα χρειαστεί να πάρουμε μόνο την παράγωγο. Όπως και πριν υποθέστε ότι:

$$p(x|\mu) \sim N(\mu, \Sigma) \text{ και } p(\mu) \sim N(\mu_o, \Sigma_o) \tag{3.38}$$

όπου το Σ, Σ_o και το μ_o υποτίθεται ότι είναι γνωστά. Μετά την παρατήρηση ενός συνόλου D, n ανεξάρτητων δειγμάτων $x_1,\ldots x_n$ χρησιμοποιούμε τον κανόνα του Bayes για να πάρουμε:

$$p(\mu|D) = \alpha \prod_{k=1}^{n} p(x_k|\mu) p(\mu) =$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\mu^t (n\Sigma^{-1} + \Sigma_O^{-1})\mu - 2\mu^t \left(\Sigma^{-1} \sum_{k=1}^{n} x_k + \Sigma_o^{-1}\mu_o \right) \right) \right], \tag{3.39}$$

το οποίο έχει τη μορφή:

$$p(\mu|D) = \alpha'' \exp\left[-\frac{1}{2}\left((\mu - \mu_n)^t \Sigma_n^{-1} (\mu - \mu_n) \right) \right] \tag{3.40}$$

Έτσι έχουμε $p(\mu|D) \sim N(\mu_n, \Sigma_n)$ και έχουμε καταφέρει να αναπαράγουμε την πυκνότητα. Εξισώνοντας του συντελεστές παίρνουμε τις αντίστοιχες με τις 3.34 και 3.35 εξισώσεις :

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_o^{-1} \tag{3.41}$$

και

$$\Sigma_n^{-1}\mu_n = n\Sigma^{-1}\hat{\mu}_n + \Sigma_o^{-1}\mu_o \tag{3.42}$$

όπου το $\hat{\mu}_n$ είναι η μέση τιμή του δείγματος :

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{3.43}$$

Η λύση αυτών των εξισώσεων για τα μ_n και Σ_n απλοποιείται αν γνωρίζουμε το μητρείο πυκνοτήτων:

$$(A^{-1} + B^{-1})^{-1} = A(A+B)^{-1}B = B(A+B)^{-1}A. \tag{3.44}$$

το οποίο ισχύει για κάθε ζευγάρι από μη ιδιόμορφα d by d μητρεία A και B. Μετά από πράξεις έχουμε τα τελικά αποτελέσματα :

$$\mu_{\mathrm{n}} = \Sigma_o(\Sigma_o + \frac{1}{\mathrm{n}}\Sigma)^{-1}\hat{\mu}_{\mathrm{n}} + \frac{1}{\mathrm{n}}\Sigma(\Sigma_o + \frac{1}{\mathrm{n}}\Sigma)^{-1}\mu_o \tag{3.45}$$

και

$$\Sigma_{\mathrm{n}} = \Sigma_o(\Sigma_o + \frac{1}{\mathrm{n}}\Sigma)^{-1}\frac{1}{\mathrm{n}}\Sigma. \tag{3.46}$$

Η απόδειξη ότι $p(\mathrm{x}|\mathrm{D}) \sim \mathrm{N}(\mu_{\mathrm{n}}, \Sigma + \Sigma_{\mathrm{n}})$ γίνεται με το να κάνουμε την ολοκλήρωση:

$$p(\mathrm{x}|\mathrm{D}) = \int p(\mathrm{x}|\mu)p(\mu|\mathrm{D})\mathrm{d}\mu \tag{3.47}$$

Πάντως αυτό το αποτέλεσμα μπορεί να ληφθεί με λιγότερη προσπάθεια εάν παρατηρήσουμε ότι το x μπορεί να δωθεί ως το άθροισμα δύο αμοιβαία ανεξάρτητων τυχαίων μεταβλητών, ενός τυχαίου διανύσματος μ με $p(\mu|\mathrm{D}) \sim \mathrm{N}(\mu_{\mathrm{n}}, \Sigma_{\mathrm{n}})$ και ενός ανεξάρτητου τυχαίου διανύσματος y με $p(\mathrm{y}) \sim \mathrm{N}(0, \Sigma)$. Επειδή το άθροισμα δυο ανεξάρτητων τυχαίων, κανονικά κατανεμημένων διανυσμάτων είναι ένα κανονικά κατανεμημένο διάνυσμα επίσης, του οποίου η μέση τιμή είναι το άθροισμα των μέσων τιμών και ο πίνακας συνδιασποράς είναι το άθροισμα των πινάκων συνδιασποράς έχουμε:

$$p(\mathrm{x}|\mathrm{D}) \sim \mathrm{N}(\mu_{\mathrm{n}}, \Sigma + \Sigma_{\mathrm{n}}) \tag{3.48}$$

και η γενίκευση ολοκληρώθηκε.

3.5 Bayesian Εκτίμηση Παραμέτρων: Γενική Θεωρία

Μόλις είδαμε πως η Bayesian προσέγγιση μπορεί να χρησιμοποιηθεί για να πάρουμε τις επιθυμητές πυκνότητες p(x|D) στην ειδική περίπτωση των πολλών μεταβλητών που ακολουθούν Gaussian κατανομή. Αυτή η προσέγγιση μπορεί να γενικευθεί ώστε να μπορεί να εφαρμοστεί σε κάθε περίπτωση, στην οποία η άγνωστη πυκνότητα μπορεί να παραμετροποιηθεί. Οι βασικές υποθέσεις συνοψίζονται στα ακόλουθα:

- Η μορφή της πυκνότητας p(x|θ) υποτίθεται ότι είναι γνωστή, αλλά η τιμή του διανύσματος παραμέτρων θ δεν είναι επακριβώς γνωστή.
- Η αρχική μας γνώση σχετικά με το θ υποτίθεται ότι περιέχεται σε μια γνωστή εκ των προτέρων πυκνότητα p(θ).
- Η υπόλοιπη γνώση μας σχετικά με το θ περιέχεται στο σύνολο D των n δειγμάτων $x_1, \ldots x_n$ που προέρχονται ανεξάρτητα σχετικά με την άγνωστη πυκνότητα p(x).

Το βασικό πρόβλημα είναι να υπολογίσουμε την εκ των υστέρων πυκνότητα p(θ|D), επειδή από αυτήν θα βρούμε μέσω της εξίσωσης 3.25 το p(x|D):

$$p(\mathrm{x}|\mathrm{D}) = \int p(\mathrm{x}|\theta)p(\theta|\mathrm{D})\mathrm{d}\theta. \tag{3.49}$$

Από τον τύπο του Bayes παίρνουμε:

$$p(\theta|\mathrm{D}) = \frac{p(\mathrm{D}|\theta)p(\theta)}{\int p(\mathrm{D}|\theta)p(\theta)\mathrm{d}(\theta)} \tag{3.50}$$

και από αυτήν την ανεξάρτητη υπόθεση

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta). \quad (3.51)$$

Αυτό αποτελεί την τυπική λύση στο πρόβλημα και οι εξισώσεις 3.50 και 3.51 φωτίζουν τη σχέση που υπάρχει με τη λύση της μέγιστης πιθανοφάνειας. Υποθέστε ότι το p(D|θ) ‘‘φτάνει’’ σε μια κορυφή (peak) στο $\theta = \hat{\theta}$. Αν η εκ των προτέρων πυκνότητα p(θ) δεν είναι μηδέν στο $\theta = \hat{\theta}$ και δεν αλλάζει πολύ στο γειτονικό διάστημα, τότε το p(θ|D) επίσης εμφανίζει κορυφή (peak) στο ίδιο σημείο. Έτσι, η εξίσωση 3.49 δείχνει ότι το p(x|D) θα γίνει περίπου $p(x|\hat{\theta})$, το αποτέλεσμα δηλαδή που θα έπαιρνε κάποιος χρησιμοποιώντας την μέγιστη πιθανοφάνεια σαν να ήταν η πραγματική τιμή. Αν η κορυφή του p(D|θ) είναι πολύ αιχμηρή, τότε η επίδραση της εκ των προτέρων πληροφορίας στην αβεβαιότητα της πραγματικής τιμής θ μπορεί κάλλιστα να αγνοηθεί. Σε αυτήν αλλά και σε πιο γενική περίπτωση, πάντως, η λύση κατά Bayes μας λέει πώς να χρησιμοποιούμε όλη τη διαθέσιμη πληροφορία για να υπολογίσουμε την επιθυμητή πυκνότητα p(x|D).

Ενώ έχουμε πάρει την τυπική επίλυση κατά Bayes εντούτοις, ένας αριθμός από ενδιαφέρουσες ερωτήσεις παραμένει. Μια από αυτές αφορά το φόρτο αλλά και τον τρόπο όλων αυτών των υπολογισμών. Κάποια άλλη αναφέρεται στη σύγκλιση του p(x|D) στο p(x). Θα συζητήσουμε το θέμα της σύγκλισης σύντομα και αργότερα θα επιστρέψουμε σε θέματα υπολογιστικού φόρτου.

Για να δείχνουμε τον αριθμό των δειγμάτων ενός συνόλου σε μια κατηγορία, θα γράφουμε $D^n = [x_1,\ldots x_n]$. Από την εξίσωση 3.51 αν n>1 παίρνουμε:

$$p(D^n|\theta) = p(x_n|\theta)\,p(D^{n-1}|\theta). \quad (3.52)$$

Αντικαθιστώντας αυτήν στην εξίσωση 3.50 και χρησιμοποιώντας τον τύπο του Bayes βλέπουμε ότι η εκ των υστέρων πυκνότητα ικανοποιεί την αναδρομική σχέση:

$$p(\theta|D^n) = \frac{p(x_n|\theta)\,p(\theta|D^{n-1})}{\int p(x_n|\theta)\,p(\theta|D^{n-1})\,d\theta} \quad (3.53)$$

Εννοώντας ότι $p(\theta|D^0) = p(\theta)$η επαναλαμβανόμενη χρήση της εξίσωσης 3.53 παράγει την ακολουθία των πυκνοτήτων $p(\theta)$, $p(\theta|x_1)$, $p(\theta|x_1,x_2)$ κ.ο.κ. Θα πρέπει να γίνεται προφανές από την εξίσωση 3.53 ότι το $p(\theta|D^n)$ εξαρτάται μόνο από τα σημεία του D^nκαι όχι από την ακολουθία από την οποία επιλέχθηκαν. Αυτό λέγεται αναδρομική κατά Bayes μέθοδος για εκτίμηση παραμέτρων.

3.6 Τα Προβλήματα των Διαστάσεων

Σε πρακτικές εφαρμογές πολλών κατηγοριών δεν είναι διόλου απίθανο να αντιμετωπίζονται προβλήματα σχετικά με πενήντα ή εκατό χαρακτηριστικά, ειδικότερα όταν τα χαρακτηριστικά αυτά παίρνουν δυαδικές τιμές. Μπορεί βέβαια τυπικά να πιστεύουμε ότι κάθε χαρακτηριστικό είναι χρήσιμο για τουλάχιστον κάποιες από τις διαφοροποιήσεις, ενώ μπορεί να αμφιβάλλουμε για το αν κάθε χαρακτηριστικό παρέχει ανεξάρτητη πληροφορία,(με δική μας πρόθεση τα περίσσεια χαρακτηριστικά δεν έχουν συμπεριληφθεί). Υπάρχουν δύο θέματα που πρέπει να αντιμετωπιστούν. Το πιο σημαντικό είναι το πώς (ή πόσο) η απόδοση της ταξινόμησης εξαρτάται από τις πολλές διαστάσεις (και το ποσό των δεδομένων

εκπαίδευσης), και το δεύτερο είναι η υπολογιστική πολυπλοκότητα του σχεδιασμού του ταξινομητή.

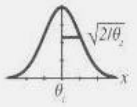
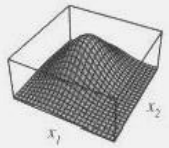
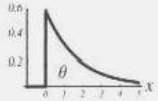
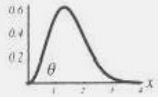
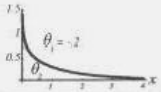
3.6.1 Ακρίβεια, Διάσταση και Εκπαίδευση μεγέθους του δείγματος

Αν τα χαρακτηριστικά είναι στατιστικά ανεξάρτητα, υπάρχουν κάποια θεωρητικά αποτελέσματα που υποστηρίζουν την πιθανότητα της τέλειας απόδοσης. Για παράδειγμα, θεωρήστε την δύο κατηγοριών Multivariate περίπτωση με την ίδια συνδιασπορά π.χ όπου $p(x|\omega_j) \sim N(\mu_j, \Sigma)$, για j=1,2. Αν οι εκ των προτέρων πιθανότητες είναι ίσες, τότε δεν είναι δύσκολο να δειχθεί ότι ο ρυθμός λάθους του Bayes δίνεται από τον τύπο:

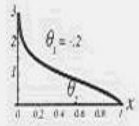
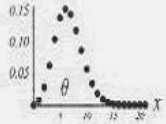
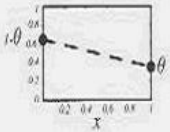
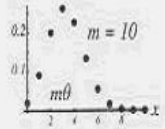
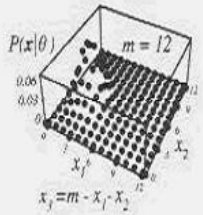
$$P(e) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du \tag{3.54}$$

όπου το r^2 είναι η τετραγωνική απόσταση Mahalanobis (Κεφάλαιο 2, Ενότητα 2.5):

$$r^2 = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2). \tag{3.55}$$

Name	Distribution	Domain		s	$[g(\mathbf{s}, \boldsymbol{\theta})]^{1/n}$
Normal	$p(x\|\boldsymbol{\theta}) = \sqrt{\frac{\theta_2}{2\pi}} e^{-(1/2)\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} x_k \\ \frac{1}{n}\sum_{k=1}^{n} x_k^2 \end{bmatrix}$	$\sqrt{\theta_2} e^{-\frac{1}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multivariate Normal	$p(\mathbf{x}\|\boldsymbol{\theta}) = \frac{\|\boldsymbol{\Theta}_2\|^{1/2}}{(2\pi)^{d/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\theta}_1)^t \boldsymbol{\Theta}_2 (\mathbf{x}-\boldsymbol{\theta}_1)}$	$\boldsymbol{\Theta}_2$ positive definite		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \\ \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^t \end{bmatrix}$	$\|\boldsymbol{\Theta}_2\|^{1/2} e^{-\frac{1}{2}[\mathrm{tr}\boldsymbol{\Theta}_2 s_2 - 2\boldsymbol{\theta}_1^t \boldsymbol{\Theta}_2 s_1 + \boldsymbol{\theta}_1^t \boldsymbol{\Theta}_2 \boldsymbol{\theta}_1]}$
Exponential	$p(x\|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta e^{-\theta s}$
Rayleigh	$p(x\|\theta) = \begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$	$\theta e^{-\theta s}$
Maxwell	$p(x\|\theta) = \begin{cases} \frac{4}{\sqrt{\pi}} \theta^{3/2} x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$	$\theta^{3/2} e^{-\theta s}$
Gamma	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \frac{1}{n}\sum_{k=1}^{n} x_k \end{bmatrix}$	$\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} s_1^{\theta_1} e^{-\theta_2 s_2}$

Πίνακας 3.1α Οι πιο γνωστές εκθετικές κατανομές και τα απαραίτητα στατιστικά στοιχεία τους.

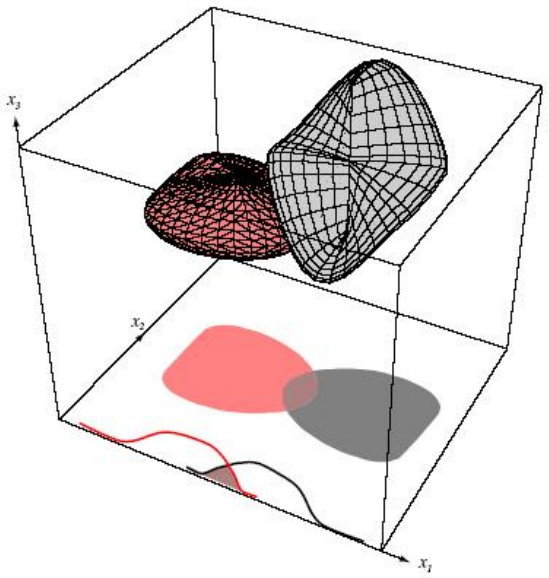
Beta	$p(x\|\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1}(1-x)^{\theta_2} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$		$\theta_1 > -1$ $\theta_2 > -1$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \left(\prod_{k=1}^{n}(1-x_k)\right)^{1/n} \end{bmatrix}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$
Poisson	$P(x\|\theta) = \frac{\theta^x}{x!}e^{-\theta}$	$x = 0, 1, 2, \ldots$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta^s e^{-\theta}$
Bernoulli	$P(x\|\theta) = \theta^x(1-\theta)^{1-x}$	$x = 0, 1$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta^s(1-\theta)^{1-s}$
Binomial	$P(x\|\theta) = \frac{m!}{x!(m-x)!}\theta^x(1-\theta)^{m-x}$	$x = 0, 1, \ldots, m$	$0 < \theta < 1$		$\frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta^s(1-\theta)^{m-s}$
Multinomial	$P(\mathbf{x}\|\boldsymbol{\theta}) = \frac{m!\prod_{i=1}^{d}\theta_i^{x_i}}{\prod_{i=1}^{d} x_i!}$	$x_i = 0, 1, \ldots, m$ $\sum_{i=1}^{d} x_i = m$	$0 < \theta_i < 1$ $\sum_{i=1}^{d}\theta_i = 1$		$\frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$	$\prod_{i=1}^{d}\theta_i^{s_i}$

Πίνακας 3.1β Οι πιο γνωστές εκθετικές κατανομές και τα απαραίτητα στατιστικά στοιχεία τους.

Έτσι η πιθανότητα του λάθους μειώνεται όσο αυξάνεται το r, πλησιάζοντας στο 0, όσο το r τείνει στο άπειρο. Στην υποθετική ανεξάρτητη περίπτωση, $\Sigma = \mathrm{diag}(\sigma_1^2,...,\sigma_d^2)$:

$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2 . \tag{3.56}$$

Αυτό δείχνει το πως κάθε χαρακτηριστικό συνεισφέρει στη μείωση της πιθανότητας του λάθους. Φυσικά, τα πιο χρήσιμα χαρακτηριστικά είναι αυτά για τα οποία η διαφορά μεταξύ των μέσων τιμών είναι μεγάλη σχετικά με τις τυπικές αποκλίσεις. Πάντως, κανένα χαρακτηριστικό δεν είναι άχρηστο αν οι μέσες τιμές για τις δύο κατηγορίες διαφέρουν. Ένας προφανής τρόπος για να μειώσουμε το ρυθμό λάθους περισσότερο είναι να προσθέσουμε νέα, ανεξάρτητα χαρακτηριστικά. Κάθε νέο χαρακτηριστικό δεν χρειάζεται να προσθέτει πολύ σε αυτήν την ελάττωση, αλλά αν το r μπορεί να αυξηθεί χωρίς όριο, η πιθανότητα λάθους μπορεί να γίνει μικρή.
Γενικά, αν η απόδοση που λαμβάνεται από ένα δοθέν σύνολο χαρακτηριστικών είναι ανεπαρκής, τότε είναι λογικό να θεωρήσουμε ό,τι πρέπει να προσθέσουμε νέα χαρακτηριστικά. Ειδικά αυτά που θα βοηθήσουν στο να ξεχωρίσουμε τα ζευγάρια των κατηγοριών που είναι συχνά δύσκολο να ταξινομηθούν χωρίς μπέρδεμα. Αν και η αύξηση του αριθμού των χαρακτηριστικών αυξάνει το κόστος και την πολυπλοκότητα και του εξαγωγέα χαρακτηριστικών αλλά και του ταξινομητή, είναι συχνά λογικό να νομίζουμε ότι η απόδοση θα βελτιωθεί. Πάντως, αν η πιθανοτική δομή του προβλήματος ήταν γνωστή, το ρίσκο του Bayes πιθανώς δε θα αυξανόταν με την πρόσθεση των νέων χαρακτηριστικών. Στη χειρότερη περίπτωση θα αγνοούσε τα νέα χαρακτηριστικά αλλά αν τα νέα χαρακτηριστικά παρέχουν οποιαδήποτε πρόσθετη πληροφορία η απόδοση πρέπει να βελτιωθεί. (Εικόνα 3.3)



Εικόνα 3.3

Δυστυχώς έχει συχνά παρατηρηθεί στην πράξη ότι η πρόσθεση περισσότερων του ενός χαρακτηριστικών στοιχείων (πέρα από ένα σημείο) αντί να βελτιώνει την

απόδοση, την μειώνει. Αυτό το φαινομενικά παράδοξο παρουσιάζεται ως σοβαρό πρόβλημα για το σχεδιασμό ταξινομητών. Η βασική πηγή δυσκολίας μπορεί αν ανιχνευθεί στο γεγονός ότι μπορεί να έχουμε επιλέξει λάθος μοντέλο (π.χ η Gaussian υπόθεση ή κάποια άλλη να είναι λανθασμένες) ή ο αριθμός των δειγμάτων εκπαίδευσης να είναι περιορισμένος και έτσι οι κατανομές να μην υπολογίζονται ακριβώς. Πάντως η ανάλυση του προβλήματος είναι δελεαστική και απαιτεί δεξιοτεχνία.

3.7 Hidden Markov Μοντέλα

Μέχρι εδώ περιορίσαμε την προσοχή μας σε προβλήματα που έχουν να κάνουν με εκτίμηση παραμέτρων, σε υπό συνθήκη κατηγορίας πυκνότητες με σκοπό να πάρουμε μια απλή απόφαση. Τώρα μεταφερόμαστε σε προβλήματα που έχουν να κάνουν με το να πάρουμε μια ακολουθία από αποφάσεις. Σε προβλήματα που έχουν έμφυτη την προσωρινότητα –αυτό σημαίνει διαδικασίες που «ξεδιπλώνονται» στο χρόνο-μπορεί να έχουμε δηλαδή μια κατάσταση στο χρόνο t που επηρεάζεται άμεσα από την κατάσταση t-1. Τα κρυμμένα μοντέλα Markov (HMM) χρησιμοποιούνται συχνά σε τέτοια προβλήματα όπως για παράδειγμα στην αναγνώριση προτύπων για ομιλία και χειρονομίες. Τα κρυμμένα μοντέλα Markov έχουν ένα αριθμό από παραμέτρους των οποίων οι τιμές εξαρτώνται από το αν είναι βέλτιστες στο να περιγράφουν τα δείγματα εκπαίδευσης για κάθε γνωστή κατηγορία. Αργότερα, ένα δοκιμαστικό πρότυπο (pattern) ταξινομείται από το μοντέλο που έχει την μεγαλύτερη εκ των υστέρων πιθανότητα δηλαδή αυτό που ‘‘εξηγεί’’ βέλτιστα το δοκιμαστικό πρότυπο (pattern).

3.8 Βιβλιογραφία

[1] Pierre Baldi. Soren Brunak. Yves Chauvin. Jacob Engel-brecht. and Anders Krogh. Hidden Markov models tor human genes. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors. Advances in Neural Information Processing Systems, volume 6, pages 761-768, Morgan Kaufmann, San Mateo, CA, 1994.

[2] Leonard E. Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state Markov chains. Annals of Mathematical Statistics, 37:1554-1563, 1966.

[3] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics. 41 (1); l64— 171, 1970.

[4] Jose M. Bernardo and Adrian F. M. Smith. Bayesian Theory. Wiley, New York, 1996. i| Christopher M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, UK, 1995.

[5] David Braverman, Learning filters for optimum pattern recognition. IRE Transactions on Information Theory, rT-8;280-285, 1962.

[6] Eugene Charniak. Statistical Language Learning. MIT Press, Cambridge, MA, 1993. i] Herman Chernoff and Lincoln E. Moses. Elementary Decision Theory. Wiley, New York, 1959.

[7] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. Introduction to Algorithms. MIT Press. Cambridge. MA. 1990.

[8] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39:1-38, 1977.

[9] Pierre A. Devijver and Josef Kittler. Patient Recognition: A Statistical Approach. Prentice-Hall, London, 1982.

[10] Ronald A. Fisher, The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7 Part II: 179-188, 1936.

[11] G. David Forney, Jr. The Viterbi algorithm. Proceedings of the IEEE, 61:268-278, 1973.
[12] Keinosuke Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press, New York, second edition, 1990.
[13] David Haussler, Michael Kearns, and Robert Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. Machine Learning, 14:84-114, 1994.
[14] Harold Jeffreys. Theory of Probability. Oxford University Press, Oxford. UK, 1961 reprint edition, 1939.
[15] Frederick Jelinek. Statistical Methods for Speech Recognition. MIT Press, Cambridge. MA, 1997.
[16] Ian T. Jolliffe. Principal Component Analysis. Springer-Verlag, New York, 1986.
[17] Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. Neural Computation, 6(2):181-214, 1994.
[18] Donald E. Knuth. The An of Computer Programming, volume 1. Addison-Wesley, Reading, MA, first edition, 1973.
[19] Gary E. Kopec and Phil A. Chou. Document image decoding using Markov source models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 16(6):602-617. 1994.
[20] Anders Krogh, Michael Brown, 1. Saira Mian, Kimmen Sjolander, and David Haussler. Hidden Markov models in computational biology: Applications to protein modelling. Journal of Molecular Biology. 235:1501-1531, 1994.
[21] Dennis Victor Lindley. The use of prior probability distributions in statistical inference and decision. In Jerzy Neyman and Elizabeth L. Scott, editors. Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability, pages 453-468. University of Califor nia Press, Berkeley, CA, 1961.
[22] Andrei Andreivich Markov, Issledovanie za mechatelnogo sluchaya zavisimykh ispytanii {in vestigation of a remarkable case of dependant trials) Izvestiya Petersburgskoi akademii nauk, 6th ser., l(3):61-80, 1907.
[23] Geoffrey J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. Wiley, New York, 1992.
[24] Geoffrey J. McLachlan and Thiriyambakam Krishnan. The EM Algorithm and Extensions. Wiley. New York, 1996.
[25] Manfred Opper and David Haussler. Generalization performance of Bayes optimal prediction algorithm for learning a perceptron. Physical Review Letters 66(20):2677-2681, 1991
[26] Lawrence Rabiner and Biing-Hwang Juang. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, 1993.
[27] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of IEEE, 77(2):257-286, 1989.
[28] Donald B. Rubin and Roderick J. A. Little. Statistical Analysis with Missing Data. Wiley, New York, 1987.
[29] Jurgen Schurmann. Pattern Classification: A Unified View of Statistical and Neural Approaches. Wiley. New York, 1996.
[30] Ross D. Shachter. Evaluating influence diagrams. Operations Research, 34(6);871-882, 1986.
[31] Padhraic Smyth, David Heckerman, and Michael Jordan. Probabilistic independence networks for hidden Markov probability models. Neural Computation, 9(2):227-269. 1997.
[32] Charles W. Therrien. Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics. Wiley, New York, 1989.
[33] D. Michael Titterington. Recursive parameter estimation using incomplete data. Journal of the Royal Statistical Society, Series B, 46(2); 257-267, 1984.

[34] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. IEEE Transactions on Information Theory, lT-13(2):260-269, 1967.