ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΙΔΙΚΕΥΣΗΣ

ΤΑΞΙΝΟΜΗΣΗ ΜΕ ΜΙΚΤΕΣ ΚΑΝΟΝΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

ΜΙΧΑΗΛ Κ. ΤΙΤΣΙΑΣ

Ιωάννινα, Ιούνιος 2001

Η εργασία εκπονήθηκε στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων κατά το τέταρτο εξάμηνο σπουδών. Ωστόσο αποτελεί μια επίμονη προσπάθεια δύο χρόνων. Ουσιαστική και συνεχής ήταν η καθοδήγηση του επιβλέποντα καθηγητή κ. Α. Λύκα, τον οποίο ευχαριστώ όχι μόνο για τις γνώσεις που μου προσέφερε αλλά και διότι με το παράδειγμά του με δίδαξε πώς να εργάζομαι μεθοδικά. Θα ήθελα επίσης να ευχαριστήσω τους καθηγητές του τμήματος Πληροφορικής κ.κ. Ι. Λαγαρή και Δ. Φωτιάδη για την βοήθεια και τις συμβουλές τους, καθώς και τον συμφοιτητή μου Κ. Κωνσταντινόπουλο για τις εποικοδομητικές συζητήσεις μας γύρω από διάφορα θέματα της στατιστικής αναγνώρισης προτύπων.

Τέλος ευχαριστώ του γονείς μου για την ειλικρινή τους συμπαράσταση.

Μιχαήλ Κ. Τίτσιας

Ταξινόμηση με Μικτές Κανονικές Κατανομές

Μιχαήλ Κ. Τίτσιας

Πανεπιστήμιο Ιωαννίνων
Σχολή Θετικών Επιστημών
Τμήμα Πληροφορικής

Επιβλέπων: Α. Λύκας

Περίληψη

Στη παρούσα εργασία εξετάζονται μέθοδοι ταξινόμησης που βασίζονται στη χρήση μικτών κατανομών για την εκτίμηση της δεσμευμένης κατανομής της κάθε κατηγορίας. Η ευρέως χρησιμοποιούμενη προσέγγιση στο πρόβλημα ταξινόμησης με χρήση μικτών κατανομών βασίζεται στη υπόθεση ότι η δεσμευμένη κατανομή της κατηγορίας μοντελοποιείται από μια ξεχωριστή μικτή κατανομή. Αυτό σημαίνει ότι κάθε πυρήνας μιας μικτής κατανομής μπορεί να αναπαριστά δεδομένα μόνο της αντίστοιχης κατηγορίας. Μια εναλλακτική προσέγγιση είναι η χρήση μικτών κατανομών με κοινούς πυρήνες. Προτείνουμε μια γενίκευση των δύο προγενέστερων τεχνικών εισάγοντας ένα νέο μοντέλο που επιτρέπει κάθε πυρήνα να είναι κοινός σε ένα υποσύνολο των κατηγοριών. Παρουσιάζουμε μια ανάλυση η οποία υποδηλώνει ότι για σταθερό συνολικό αριθμό πυρήνων, το μοντέλο με την καλύτερη επίδοση ταξινόμησης αποτελεί μια ειδική περίπτωση του γενικού μοντέλου. Προκειμένου να ανακαλύψουμε μια αποτελεσματική μοντελοποίηση των δεσμευμένων κατανομών χρησιμοποιούμε τον αλγόριθμο EM προσαρμόζοντας κατά την εκπαίδευση όχι μόνο τις παραμέτρους του μοντέλου αλλά και τον βαθμό που ένας πυρήνας συνεισφέρει στη αναπαράσταση δεδομένων των κατηγοριών. Η μέθοδος εκπαίδευσης είναι αρκετά γενική και επιτρέπει την εισαγωγή διαφόρων αλγορίθμων εκπαίδευσης μικτών δεσμευμένων κατανομών σε προβλήματα ταξινόμησης.

Περιεχόμενα

Σχήματα

Κεφάλαιο 1

Εισαγωγή

1.1 Στατιστική αναγνώριση προτύπων

Ο όρος αναγνώριση προτύπων αναφέρεται σε ένα πλήθος προβλημάτων επεξεργασίας πληροφορίας όπως είναι η αναγνώριση φωνής, η αναγνώριση χειρόγραφων χαρακτήρων, κτλ. Τέτοιου είδους προβλήματα είναι συνήθως απλά για την ανθρώπινη νοημοσύνη, π.χ. ένας άνθρωπος έχει την ικανότητα να αναγνωρίζει χειρόγραφους χαρακτήρες ακόμη και σε περιπτώσεις που αυτοί είναι γραμμένοι με ιδιόμορφο τρόπο. Ωστόσο η επίλυση τέτοιων προβλημάτων χρησιμοποιώντας υπολογιστικές μηχανές έχει αποδειχθεί ιδιαίτερα δύσκολη και έχει αποτελέσει το επίκεντρο σημαντικής ερευνητικής προσπάθειας. Προκειμένου να κατασκευαστούν αποδοτικά συστήματα για την επίλυση προβλημάτων αναγνώρισης προτύπων πρέπει να υιοθετηθεί μια γενική προσέγγιση η οποία θα προσφέρει ένα πλαίσιο αρχών και εννοιών πάνω στο οποίο θα βασιστεί στη συνέχεια η ερευνητική προσπάθεια.

Η στατιστική προσέγγιση προσπαθεί να αναδείξει την πιθανοτική φύση του προβλήματος. Ο τομέας της στατιστικής αναγνώρισης προτύπων είναι ο παλαιότερος και καλύτερα θεμελιωμένος και βασίζεται σε έννοιες της θεωρίας στατιστικής και πιθανοτήτων.

Σε αυτό το εισαγωγικό κεφάλαιο περιγράφουμε έννοιες και μεθόδους της στατιστικής αναγνώρισης προτύπων που θα χρησιμοποιήσουμε στα επόμενα κεφάλαια.

1.1.1 Στατιστική θεωρία απόφασης

Το πρόβλημα ταξινόμησης ορίζεται ως εξής: Έστω ότι έχουμε ένα πρότυπο x το οποίο αποτελεί ένα διάνυσμα χαρακτηριστικών που στη γενική περίπτωση παίρνει τιμές στο συνεχή d-διάστατο χώρο (R^d). Το δεδομένο x ανήκει σε μια κατηγορία

C_k, όπου το k παίρνει τιμές από το πεπερασμένο σύνολο $\{1, \ldots, K\}$ με $(K \geq 2)$. Η κατηγορία του x θεωρείται άγνωστη και το πρόβλημα ταξινόμησης αφορά την εύρεση της άγνωστης κατηγορίας του δεδομένου x παρατηρώντας τις τιμές των χαρακτηριστικών του.

Η στατιστική προσέγγιση στο πρόβλημα ταξινόμησης θεωρεί το δεδομένο x και την κατηγορία C_k ως τυχαίες μεταβλητές. Το x αποτελεί συνεχή τυχαία μεταβλητή[1] αφού ανήκει στο R^d, ενώ η κατηγορία C_k αποτελεί διακριτή τυχαία μεταβλητή. Για την κατηγορία C_k ορίζεται μια τιμή πιθανότητας $P(C_k)$, που θα καλείται εκ των προτέρων πιθανότητα της κατηγορίας, έτσι ώστε $\sum_{k=1}^{K} P(C_k) = 1$. Ομοίως για το x ορίζεται μια συνάρτηση πυκνότητας πιθανότητας $p(x)$ η οποία μπορεί να γραφεί σαν ολική συνάρτησης πυκνότητας πιθανότητας ως εξής:

$$p(x) = \sum_{k=1}^{K} p(x|C_k)P(C_k) \tag{1.1}$$

όπου $p(x|C_k)$ είναι η δεσμευμένη ως προς την κατηγορία C_k συνάρτηση πυκνότητας πιθανότητας του δεδομένου x (εν συντομία θα την αναφέρουμε ως δεσμευμένη κατανομή της κατηγορίας C_k).

Επιδιώκουμε με βάση τις κατανομές και τις πιθανότητες (προς το παρόν θεωρούνται γνωστές ποσότητες) που έχουμε ορίσει παραπάνω να κατασκευάσουμε ένα σύστημα ταξινόμησης (ή απόφασης) που θα επιλύει το πρόβλημα ταξινόμησης. Προκειμένου να δούμε για το πώς μπορούμε τα ταξινομούμε με βάση κατανομές και πιθανότητες ας θεωρήσουμε το εξής απλό παράδειγμα. Υποθέτουμε ότι είμαστε αναγκασμένοι να κατασκευάσουμε το σύστημα ταξινόμησης με το περιορισμό ότι δεν επιτρέπεται η χρήση της πληροφορίας που δίνει το ίδιο το δεδομένο, δηλαδή οι τιμές των χαρακτηριστικών του. Προφανώς το μόνο που μπορούμε χρησιμοποιήσουμε είναι η πιθανότητα $P(C_k)$. Φαίνεται λογικό πως ο καλύτερος τρόπος απόφασης θα ήταν η επιλογή της κατηγορίας με την μεγαλύτερη πιθανότητα $P(C_k)$ και οπότε το σύστημα θα αποφάσιζε με βάση τον κανόνα: *επέλεξε την* C_k *εάν* $P(C_k) > P(C_\ell)$ *για κάθε* $\ell \neq k$. Είναι προφανές ότι με τον κανόνα αυτό ελαχιστοποιείται η πιθανότητα λανθασμένης ταξινόμησης. Επίσης από το προηγούμενο παράδειγμα είναι φανερό ότι η πιθανότητα $P(C_k)$ εκφράζει την εκ των προτέρων γνώση ή πεποίθησή μας για το ποια είναι η κατηγορία του δεδομένου, προτού αυτό εμφανιστεί (και παρατηρηθούν οι τιμές των χαρακτηριστικών του) στο σύστημα ταξινόμησης. Έτσι εξηγείται και το γεγονός ότι η

[1] Θα μπορούσε να ήταν διακριτή αν κάθε συνιστώσα του διανύσματος χαρακτηριστικών λάμβανε διακριτές τιμές

πιθανότητα αυτή καλείται εκ των προτέρων πιθανότητα της κατηγορίας.

Ωστόσο ο παραπάνω τρόπος είναι υπερβολικά 'απλοϊκός' αφού για οποιοδήποτε νέο πρότυπο επιλέγουμε πάντα την ίδια κατηγορία. Προχειμένου να βελτιώσουμε την μέθοδο απόφασης μας πρέπει να χρησιμοποιήσουμε την τιμή x του άγνωστου δεδομένου. Προφανώς για κάθε κατηγορία υπάρχουν κάποιες περιοχές του χώρου δεδομένων από όπου είναι πιθανότερο να προέρθουν δεδομένα που ανήκουν σ' αυτήν και επιπλέον κάποιες άλλες περιοχές όπου η πιθανότητα αυτή είναι μικρότερη. Μια τέτοια πληροφορία εκφράζεται πλήρως από την δεσμευμένη κατανομή $p(x|C_k)$. Εάν οι δεσμευμένες κατανομές και οι εκ των προτέρων πιθανότητες είναι γνωστές, τότε η πιθανότητα της κατηγορίας C_k δοθέντος του x (τώρα παρατηρούμε τις συγκεκριμένες τιμές των χαρακτηριστικών του) δίνεται από τον κανόνα του Bayes:

$$P(C_k|x) = \frac{p(x|C_k)P(C_k)}{p(x)}. \tag{1.2}$$

Η πιθανότητα $P(C_k|x)$ αντιπροσωπεύει την εκ των υστέρων (a posterior) 'πίστη' μας σχετικά με το ποια είναι η κατηγορία του δεδομένου x. Επομένως είναι λογικό να αποφασίσουμε ότι το x ανήκει στην κατηγορία με την μεγαλύτερη εκ των υστέρων πιθανότητα, δηλαδή να εισάγουμε τον εξής κανόνα:

επέλεξε την C_k εάν $P(C_k|x) > P(C_\ell|x)$ για κάθε $\ell \neq k$.

Ο κανόνας ονομάζεται *κανόνας απόφασης του Bayes*. Λόγω του ότι η ολική συνάρτηση πυκνότητας πιθανότητας $p(x)$ είναι ανεξάρτητη της κάθε κατηγορίας μπορεί αυτή να παραληφθεί στη σχέση ορισμού του κανόνα απόφασης, οπότε παίρνουμε τον ισοδύναμο κανόνα:

επέλεξε την C_k εάν $p(x|C_k)P(C_k) > p(x|C_\ell)P(C_\ell)$ για κάθε $\ell \neq k$.

Στη συνέχεια θα δούμε ότι η προηγούμενη μέθοδος απόφασης ορίζει ένα βέλτιστο σύστημα ταξινόμησης όπου το κριτήριο που βελτιστοποιείται είναι αυτό της πιθανότητας λανθασμένης ταξινόμησης.

1.1.2 Ο κανόνας απόφασης του Bayes και ελαχιστοποίηση της πιθανότητας λανθασμένης ταξινόμησης

Επιδιώκουμε να υπολογίσουμε την πιθανότητα λανθασμένης ταξινόμησης για τον κανόνα απόφασης του Bayes που παρουσιάστηκε στο τέλος της προηγούμενης ενότητας. 'Εστω ότι για το άγνωστο δεδομένο x που εμφανίζεται στο σύστημα ο κανόνας απόφασης του Bayes ισχυρίζεται ότι αυτό ανήκει στη κατηγορία C_k. Θα έχουμε κάνει λάθος αν η σωστή κατηγορία είναι μια από τις υπόλοιπες κατηγορίες, δηλαδή

$$P(error|x) = 1 - P(C_k|x). \tag{1.3}$$

όπου με $P(error|x)$ συμβολίζουμε την πιθανότητα λανθασμένης ταξινόμησης του δεδομένου x. Η ελάχιστη τιμή της $P(error|x)$ με βάση την σχέση (1.3) συμβαίνει εάν επιλέξουμε την κατηγορία C_ℓ για την οποία ισχύει:

$$P(C_\ell|x) > P(C_i|x), \quad \forall i \neq \ell, \tag{1.4}$$

που προφανώς αποτελεί τον κανόνα με τον οποίο αποφασίσαμε, δηλαδή $\ell = k$. Επομένως ο κανόνας του Bayes ελαχιστοποιεί την πιθανότητα λανθασμένης ταξινόμησης για το νέο δεδομένο που εμφανίζεται στο σύστημα.

'Ενας κανόνας απόφασης για κάθε σημείο του χώρου δεδομένων ορίζει μια κατηγορία, όποτε μπορεί να θεωρηθεί ότι ο κανόνας ορίζει ένα διαχωρισμό του χώρου σε K υποπεριοχές. Με βάση το γεγονός αυτό προκύπτει ότι ο χώρος δεδομένων είναι η ένωση K ξένων περιοχών $R_1, \ldots, R_K$ (για οποιεσδήποτε δύο περιοχές ισχύει $R_k \cap R_\ell = \emptyset$), έτσι ώστε όταν ένα δεδομένο x βρίσκεται στην περιοχή R_k ο κανόνας αποφασίζει ότι αυτό ανήκει στην κατηγορία C_k. Για προφανή λόγο οι περιοχές αυτές ονομάζονται περιοχές απόφασης και η κάθε μια δεν είναι ανάγκη να είναι συνεχής αλλά μπορεί να αποτελείται από μη γειτονικές υποπεριοχές που όμως όλες σχετίζονται με την ίδια κατηγορία. Τα όρια αυτών των περιοχών ονομάζονται όρια ή επιφάνειες απόφασης.

Αποδεικνύεται ότι οι περιοχές απόφασης που ορίζονται από τον κανόνα απόφασης του Bayes είναι βέλτιστα οριοθετημένες. Συγκεκριμένα επιτυγχάνεται η βέλτιστη τοποθέτηση των ορίων απόφασης ως προς το κριτήριο της συνολικής πιθανότητας λανθασμένης ταξινόμησης ($P(error)$). Θα το δείξουμε αυτό για την απλή περίπτωση δύο κατηγοριών που τα πρότυπα τους ανήκουν σε ένα υποσύνολο $R_0 \subset R$. Οι δύο περιοχές απόφασης είναι η R_1 και η R_2 για τις οποίες προφανώς ισχύει $R_1 \cup R_2 = R_0$. Αναζητούμε εκείνα τα όρια απόφασης για τα οποία ελαχιστοποιείται η ακόλουθη πιθανότητα

$$P(error) = \int_{R_0} P(error, x)dx, \tag{1.5}$$

ή σε μορφή αναμένόμενης τιμής της ποσότητας $P(error|x)$

$$P(error) = \int_{R_0} P(error|x)p(x)dx. \tag{1.6}$$

Όταν το x βρίσκεται στην περιοχή R_1 ο χανόνας απόφασης αποφαίνεται ότι αυτό έχει προέρθει από την κατηγορία C_1, οπότε η πιθανότητα λάθους για οποιοδήποτε $x \in R_1$ είναι $P(error|x) = P(C_2|x)$. Ομοίως όταν το $x \in R_2$ ο χανόνας απόφασης αποφαίνεται ότι ανήκει στην C_2 και η πιθανότητα λάθους είναι ίση με $P(C_1|x)$. Επομένως η σχέση (1.6) γράφεται στην μορφή:

$$P(error) = \int_{R_1} P(C_2|x)p(x)dx + \int_{R_2} P(C_1|x)p(x)dx \qquad (1.7)$$

η οποία με βάση το γεγονός ότι $P(C_k|x)p(x) = P(x|C_k)p(C_k)$ γράφεται και ως

$$P(error) = \int_{R_1} P(x|C_2)p(C_2)dx + \int_{R_2} P(x|C_1)p(C_1)dx \qquad (1.8)$$

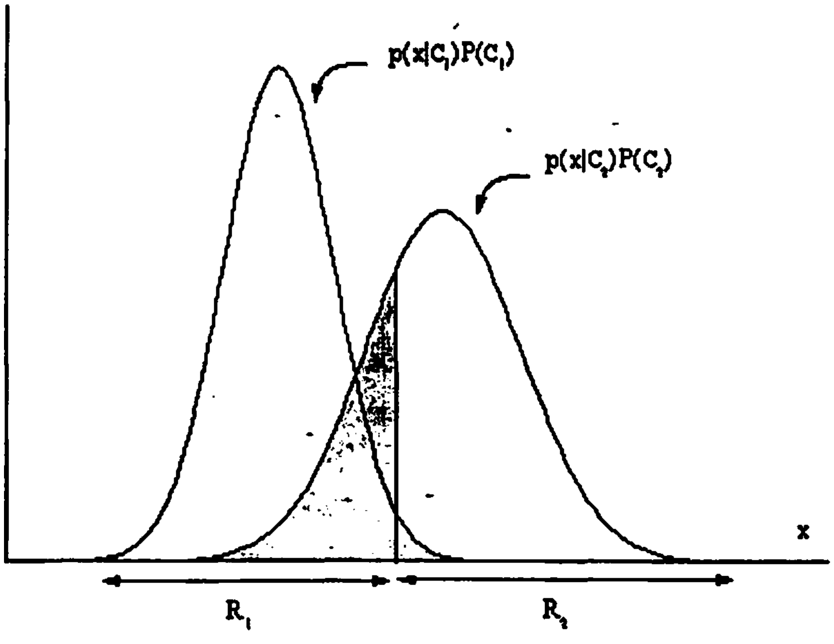
Αν οι δεσμευμένες κατανομές είναι αυτές που φαίνονται στο Σχήμα 1.1, η παραπάνω συνολική πιθανότητα λάθους είναι ίση με το γραμμοσχιασμένο εμβαδόν. Σε αυτό το σχήμα το όριο απόφασης είναι εκείνο το x από το οποίο περνάει η κάθετη γραμμή, ενώ εκατέρωθεν της γραμμής αυτής υπάρχουν οι περιοχές R_1 και R_2. Εάν επιλέξουμε ως χανόνα απόφασης αυτόν του Bayes δηλαδή για κάθε δεδομένο επιλέγουμε ότι ανήκει στην C_1 εφόσον ισχύει $P(x|C_1)p(C_1) > P(x|C_2)p(C_2)$ και αντίστοιχα για την C_2, το όριο απόφασης όπως φαίνεται στο Σχήμα 1.1 θα μετακινηθεί αριστερά και θα τοποθετηθεί στην τομή των δύο γραμμών, δηλαδή το όριο απόφασης θα είναι εκείνο το x για το οποίο ισχύει $P(x|C_1)p(C_1) = P(x|C_2)p(C_2)$. Για τις νέες περιοχές απόφασης το γραμμοσχιασμένο εμβαδόν είναι το ελάχιστο δυνατό και οπότε και η συνολική πιθανότητα λάθους είναι η ελάχιστη.

Με παρόμοιο τρόπο τα παραπάνω μπορούν να γενικευθούν στην περίπτωση K κατηγοριών και d-διάστατου χώρου δεδομένων [3, 9].

1.2 Εκτίμηση των $P(C_k)$ και $p(x|C_k)$

Προηγουμένως είδαμε πώς μπορούμε να σχεδιάσουμε ένα βέλτιστο σύστημα ταξινόμησης με κριτήριο την ελαχιστοποίηση της πιθανότητας λάθους. Η προσέγγισή μας προϋπέθετε ότι οι εκ των προτέρων πιθανότητες καθώς και οι δεσμευμένες κατανομές των κατηγοριών ήταν γνωστές. Ωστόσο, στη πράξη αυτές οι ποσότητες είναι άγνωστες και ο υπολογισμός τους είναι το πραγματικό πρόβλημα που οφείλουμε να αντιμετωπίσουμε κατά τη διαδικασία εκπαίδευσης.

Στην στατιστική αναγνώριση προτύπων η εκτίμηση κατανομών και πιθανοτήτων βασίζεται στη πληροφορία που εμπεριέχεται σε ένα σύνολο γνωστών παρατη-



Σχήμα 1.1: Απεικόνιση των από κοινού συναρτήσεων πυκνότητας $P(x, C_k) = P(x|C_k)p(C_k)$ $(k = 1, 2)$ των δύο κατηγοριών. Η κάθετη γραμμή βρίσκεται πάνω στο όριο απόφασης. Για αυτό το όριο απόφασης το γραμμοσκιασμένο εμβαδόν αντιστοιχεί στην συνολική πιθανότητα λανθασμένης απόφασης. Ελαχιστοποιώντας αυτό το εμβαδόν ελαχιστοποιούμε την πιθανότητα αυτή. Κάτι τέτοιο επιτυγχάνεται μετακινώντας το όριο αριστερά έτσι ώστε να τοποθετηθεί στην τομή των δύο γραμμών.

ρήσεων ή δεδομένων[2]. Έστω ότι διαθέτουμε ένα τέτοιο σύνολο X από δεδομένα γνωστής κατηγορίας. Για το λόγο ότι κάθε $x \in X$ ανήκει σε μια κατηγορία το αρχικό σύνολο μπορεί να διαχωριστεί σε K ξένα μεταξύ τους υποσύνολα X_k, $k = 1, ..., K$ που το κάθε ένα περιέχει τα δεδομένα της κατηγορίας C_k. Το ζητούμενο είναι με αυτά τα δεδομένα να εκτιμήσουμε τις εκ των προτέρων πιθανότητες των κατηγοριών $P(C_k)$ καθώς και τις δεσμευμένες κατανομές $p(x|C_k)$. Ο υπολογισμός των εκ των προτέρων πιθανοτήτων είναι εύκολος. Συγκεκριμένα για να υπολογίζουμε την $P(C_k)$ βρίσκουμε τον αριθμό των στοιχείων του συνόλου X_k καθώς και τον συνολικό αριθμό δεδομένων. Έπειτα εκτιμούμε τη $P(C_k)$ με βάση το ακόλουθο κλάσμα:

$$P(C_k) = \frac{|X_k|}{|X|} \quad k = 1, \dots, K, \tag{1.9}$$

όπου με $|Y|$ συμβολίζουμε το πλήθος των στοιχείων του συνόλου Y. Από την άλλη, ο υπολογισμός των δεσμευμένων κατανομών είναι πολύ πιο περίπλοκος. Οι συναρτήσεις αυτές επιδιώκουμε να προσεγγίσουν την άγνωστη κατανομή των δεδομένων της κάθε κατηγορίας. Ωστόσο η συναρτησιακή μορφή της κατανομής αυτής μπορεί να είναι οσοδήποτε περίπλοκη, πράγμα που δυσχεραίνει την εκτίμησή της. Επιπλέον ο αριθμός των διαθέσιμων δεδομένων μπορεί να μην είναι αρκετός για μια τέτοια εκτίμηση, ιδιαίτερα αν η διάσταση του χώρου δεδομένων είναι μεγάλη.

Το σύνολο X_k έχει προέρθει με βάση την κατανομή $p(x|C_k)$, επομένως μπορεί να θεωρήσουμε ότι στο X_k δεν εμπεριέχεται καμιά πληροφορία σχετικά με τις υπόλοιπες δεσμευμένες κατανομές. Αυτό σημαίνει ότι το πρόβλημα εκτίμησης των δεσμευμένων κατανομών μπορεί να διαχωριστεί σε K ανεξάρτητα προβλήματα που το καθένα είναι της μορφής: Δοθέντος ενός συνόλου δεδομένων X το ζητούμενο είναι η εκτίμηση της άγνωστης κατανομής $p(x)$ από την οποία έχει προέρθει το X. Το πρόβλημα αυτό αναφέρεται ως *εκτίμηση συνάρτησης πυκνότητας πιθανότητας* και περιγράφεται αρκετά αναλυτικά στη συνέχεια.

1.3 Εκτίμηση συνάρτησης πυκνότητας πιθανότητας

Παραπάνω είδαμε πώς προκύπτει το πρόβλημα εκτίμησης συνάρτησης πυκνότητας πιθανότητας σε ένα πρόβλημα ταξινόμησης. Η μέθοδος ανήκει στην κατηγορία των τεχνικών μάθησης χωρίς επίβλεψη [9, 10].

[2]Πολλές τεχνικές μοντελοποίησης προβλημάτων βασίζονται σε μάθηση μέσω γνωστών δεδομένων, όπως νευρωνικά δίκτυα κλπ.

Κατά την μάθηση χωρίς επίβλεψη έχουμε να αντιμετωπίσουμε το εξής πρόβλημα. Δοθέντος ενός συνόλου δεδομένων X επιδιώχουμε να κατασκευάσουμε ένα μοντέλο περιγραφής αυτών. Οι πληροφορίες που αναζητούμε μέσω μιας τέτοιας περιγραφής αφορούν την μορφή ή την δομή των δεδομένων στο αντίστοιχο χώρο. Η κατανομή $p(x)$ από την οποία έχουν προέρθει τα δεδομένα X δίνει πλήρη περιγραφή του X για αυτό και η εκτίμηση συνάρτησης πυκνότητας πιθανότητας θεωρείται ως η πιο γενική τεχνική μάθησης χωρίς επίβλεψη.

Οι γνωστές μέθοδοι εκτίμησης συνάρτησης πυκνότητας πιθανότητας διακρίνονται σε δύο μεγάλες κατηγορίες: στις παραμετρικές και στις μη παραμετρικές [9, 10]. Η βασική διαφορά μεταξύ των δύο είναι ότι οι πρώτες υποθέτουν ένα παραμετρικό μοντέλο για την άγνωστη κατανομή ενώ οι δεύτερες δεν υποθέτουν κάτι τέτοιο αλλά προσπαθούν να εκφράσουν την άγνωστη κατανομή απευθείας από τα δεδομένα. Θα αναφερθούμε μόνο στις παραμετρικές μεθόδους αφού μόνο αυτές θα μας απασχολήσουν στην συνέχεια της εργασίας.

1.3.1 Παραμετρικά μοντέλα

Ένας τρόπος προσέγγισης του προβλήματος εκτίμησης άγνωστης κατανομής είναι να υποθέσουμε ότι η άγνωστη κατανομή είναι μια συγκεκριμένη συνάρτηση εξαρτώμενη από ένα διάνυσμα παραμέτρων. Η εκτίμηση σ' αυτή την περίπτωση δεν είναι τίποτε άλλο παρά η εύρεση εκείνων των παραμέτρων[3] έτσι ώστε η συνάρτηση να 'ταιριάζει' όσο το δυνατό καλύτερα στην κατανομή των δεδομένων. Επομένως υποθέτουμε ότι η $p(x)$ εξαρτάται από ένα διάνυσμα παραμέτρων Θ για το λόγο αυτό και θα την γράφουμε ως $p(x|\Theta)$.

Υπόθεση απλής κατανομής. Για την μορφή της $p(x|\Theta)$ μπορούμε μια υποθέσουμε μια από τις γνωστές συναρτήσεις πυκνότητας πιθανότητας. Η παραμετρική συνάρτηση στην οποία έχει δοθεί η μεγαλύτερη προσοχή από κάθε άλλη είναι η κανονική (ή Gaussian) κατανομή. Η δημοτικότητα της οφείλεται κυρίως στις καλές αναλυτικές και στατιστικές ιδιότητες που διαθέτει. Η κανονική κατανομή στην γενική μορφή έχει την ακόλουθη μορφή:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{1.10}$$

όπου μ είναι ένα d-διάστατο διάνυσμα που αναπαριστά το μέσο της κατανομής και Σ είναι ο $d \times d$ πίνακας συμμεταβλητότητας. Ο παράγοντας μπροστά στο

[3] Η Μπεϋσιανή μάθηση δεν βρίσκει απλά τιμές παραμέτρων, άλλα μια κατανομή ως προς τις παραμέτρους που εκφράζει το πόσο καλά αναπαριστά τα δεδομένα η κάθε τιμή παραμέτρων.

εκθετικό μέρος της συνάρτησης εγγυάται ότι ισχύει $\int p(x|\mu,\Sigma)dx = 1$. Το μέσο μ και ο πίνακας Σ ορίζονται από τις σχέσεις:

$$\mu = E[x] = \int x p(x|\mu_j,\Sigma)dx, \tag{1.11}$$

$$\Sigma = E[(x-\mu)(x-\mu)^T] = \int (x-\mu)(x-\mu)^T p(x|\mu_j,\Sigma)dx. \tag{1.12}$$

Κάθε συνιστώσα μ_i του μέσου καθώς και κάθε στοιχείο σ_{ij} του Σ ορίζονται από τις σχέσεις:

$$\mu_i = E[x_i] = \int x_i p(x|\mu_j,\Sigma)dx, \tag{1.13}$$

$$\sigma_{ij} = E[(x_i-\mu_i)(x_j-\mu_j)^T] = \int (x_i-\mu_i)(x_j-\mu_j)^T p(x|\mu_j,\Sigma)dx. \tag{1.14}$$

Αν στο στοιχείο σ_{ij} το $i = j$, τότε η παράμετρος αναπαριστά διακύμανση της συνιστώσας i, ενώ διαφορετικά αναπαριστά συμμεταβλητότητα της συνιστώσας i με την j. Ένας συνηθισμένος συμβολισμός της κανονικής κατανομής είναι $N(\mu,\Sigma)$, ενώ προκειμένου να δηλώσουμε ότι η μεταβλητή x ακολουθεί την προηγούμενη κανονική κατανομή γράφουμε $x \sim N(\mu,\Sigma)$ και λέμε ότι η x ακολουθεί κανονική κατανομή με μέση τιμή μ και πίνακα συμμεταβλητότητας Σ.

Από τη σχέση (1.12) προκύπτει ότι ο Σ είναι πάντα συμμετρικός και θετικά ημιορισμένος πίνακας. Λόγω της συμμετρίας μπορεί να περιγραφεί με $d(d+1)/2$ ανεξάρτητους παραμέτρους, οπότε συμπεριλαμβανομένων και των d παραμέτρων για το μέσο η συνάρτηση καθορίζεται πλήρως από $d + d(d+1)/2$ παραμέτρους. Η παρακάτω ποσότητα

$$\Delta^2 = (x-\mu)^T\Sigma^{-1}(x-\mu), \tag{1.15}$$

η οποία εμφανίζεται στο εκθετικό μέρος της (1.10) ονομάζεται απόσταση Mahalanobis μεταξύ x και μ. Προκειμένου να κατανοήσουμε πώς κατανέμονται τα δείγματα μιας κανονικής κατανομής ας σκεφτούμε το εξής: Για σταθερή απόσταση Δ^2, όλα τα x ανήκουν σε μια υπερελλειψοειδή επιφάνεια που έχει ως κέντρο το μ ενώ το σχήμα της καθορίζεται από τον πίνακα Σ. Προφανώς για όλα τα x μιας τέτοιας επιφάνειας η τιμή της συνάρτησης είναι σταθερή, πράγμα που σημαίνει ότι τα πρότυπα που παράγονται με βάση την κατανομή (1.10) ομαδοποιούνται έτσι ώστε να σχηματίζουν υπερελλειψοειδείς πυρήνες.

Μερικές φορές είναι βολικότερο να χρησιμοποιήσουμε μια απλούστερη μορφή της πολυδιάστατης κανονικής κατανομής. Αν για παράδειγμα υποθέσουμε ότι δεν υπάρχει συμμεταβλητότητα μεταξύ των συνιστωσών του x, δηλαδή $\sigma_{ij} = 0$ για κάθε $i \neq j$, τότε ο πίνακας συμμεταβλητότητας μετατρέπεται σε έναν διαγώνιο $\Sigma = diag(\sigma_1^2 \dots \sigma_d^2)$. Με βάση την υπόθεση αυτή η κατανομή παίρνει τη μορφή

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sigma_1 \dots \sigma_d} \exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} \dots - \frac{(x_d-\mu_d)^2}{2\sigma_d^2}\right\}, \quad (1.16)$$

ή

$$p(x|\mu, \Sigma) = \prod_{i=1}^{d} \frac{1}{(2\pi)^{1/2}\sigma_i} \exp\left\{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right\}. \quad (1.17)$$

Η παραπάνω μορφή καθορίζεται από $2d$ ανεξάρτητους παραμέτρους. Υποθέτοντας ότι οι διακυμάνσεις της κάθε συνιστώσας είναι ίσες μεταξύ τους, δηλαδή ισχύει $\sigma_i^2 = \sigma^2$ για όλα i, καταλήγουμε σε μια επιπλέον απλούστευση της σχέσης (1.17). Σε αυτή την απλουστευμένη μορφή της κατανομής ο αριθμός των παραμέτρων έχει μειωθεί στις $d+1$ και η κανονική κατανομή γράφεται στην μορφή:

$$p(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left\{-\frac{||x-\mu||^2}{2\sigma^2}\right\}, \quad (1.18)$$

όπου με την νόρμα $||x-\mu||$ συμβολίζουμε την ευκλείδεια απόσταση των διανυσμάτων x και μ. Σε αυτή την περίπτωση για σταθερή απόσταση Mahalanobis τα διανύσματα x με ίσες τιμές πιθανότητας $p(x)$ ορίζουν μια υπερσφαίρα στο d-διάστατο χώρο, επομένως πρότυπα κατανεμημένα με βάση την σχέση (1.18) ομαδοποιούνται έτσι ώστε να σχηματίζουν υπερσφαίρες. Η απλουστευμένη αυτή μορφή της κανονικής κατανομής έχει τις λιγότερες παραμέτρους αλλά υστερεί προφανώς σε γενικότητα.

Μικτές κατανομές. Μια μικτή κατανομή [32, 19] ορίζεται ως μια ειδική περίπτωση γραμμικού συνδυασμού ενός πεπερασμένου αριθμού συναρτήσεων πυκνότητας πιθανότητας. Δηλαδή η πιθανότητα μιας τυχαίας μεταβλητής x που ακολουθεί μικτή κατανομή γράφεται ως άθροισμα συναρτήσεων πυκνότητας πιθανότητας με βάρη και στην γενική περίπτωση των M τέτοιων συναρτήσεων δίνεται από την ακόλουθη σχέση:

$$p(x|\Theta) = \sum_{j=1}^{M} \pi_j p(x|j, \theta_j). \quad (1.19)$$

Τον πυρήνα j τον ονομάζουμε συστατικό πυρήνα ή απλώς πυρήνα, ενώ την αντίστοιχη κατανομή $p(x|j, \theta_j)$ (που εξαρτάται από ένα διάνυσμα παραμέτρων θ_j) του μικτού μοντέλου την ονομάζουμε συστατική συνάρτηση πυκνότητας πιθανότητας της ολικής κατανομής. Το βάρος π_j αποτελεί παράμετρο που εκφράζει την εκ των προτέρων πιθανότητα σύμφωνα με την οποία η παραγωγή ενός δεδομένου οφείλεται στον συστατικό πυρήνα j. Το σύνολο των παραμέτρων της μικτής κατανομής είναι προφανώς $\Theta = \{(\pi_j, \theta_j), j = 1, \ldots, M\}$. Οι παράμετροι π_j δεν μπορούν να λάβουν αρνητικές τιμές και υπόκεινται στον εξής περιορισμό:

$$\sum_{j=1}^{M} \pi_j = 1. \tag{1.20}$$

Η συνάρτηση $p(x|j, \theta_j)$ εκφράζει την δεσμευμένη κατανομή βάσει της οποίας ο πυρήνας j παράγει το δεδομένο x. Προκειμένου να παράγουμε ένα πρότυπο που ακολουθεί μικτή κατανομή της μορφής (1.19) επιλέγουμε, καταρχήν, έναν πυρήνα j από το σύνολο των M πυρήνων με πιθανότητα π_j και στην συνέχεια παράγουμε το πρότυπο με βάση την συστατική κατανομή $p(x|j, \theta_j)$.

Είναι δυνατόν να υποθέσουμε μια μικτή κατανομή για την άγνωστη συνάρτησης πυκνότητας πιθανότητας ενός συνόλου δεδομένων. Όπως είδαμε στην περίπτωση των παραμετρικών μεθόδων η υπόθεση ήταν ότι το σύνολο των δεδομένων έχει παραχθεί από μια εκ των γνωστών συναρτήσεων πυκνότητας (π.χ. τη κανονική κατανομή). Στην περίπτωση ενός μικτού μοντέλου η υπόθεση είναι πιο γενική λόγω του ότι το σύνολο δεδομένων λαμβάνεται ως ένα μίγμα συστατικών πληθυσμών ο καθένας εκ των οποίων σχετίζεται με μια συστατική κατανομή και την αντίστοιχη εκ των προτέρων πιθανότητα.

Είναι αξιοσημείωτο ότι στο μικτό μοντέλο η έννοια της εκ των προτέρων πιθανότητας και της συστατικής κατανομής ενός πυρήνα χρησιμοποιείται ακριβώς ανάλογα με την έννοια της εκ των προτέρων πιθανότητας και της δεσμευμένης κατανομής της κατηγορίας στο πρόβλημα ταξινόμησης. Ωστόσο υπάρχει μια σημαντική διαφορά που αφορά την φύση του προβλήματος[4]. Στο πρόβλημα ταξινόμησης τα πρότυπα είναι 'χαρακτηρισμένα' ως προς την κατηγορία που ανήκουν πράγμα που αποτελεί σημαντικό πλεονέκτημα κατά την διαδικασία μάθησης. Όπως αναφέρθηκε προηγουμένως, μπορούμε να διαχωρίσουμε το σύνολο δεδομένων εκπαίδευσης σε τόσα υποσύνολα όσες είναι και οι κατηγορίες και στη συνέχεια να εκτιμήσουμε την υπό συνθήκη κατανομή της κάθε κατηγορίας χρησιμοποιώντας μόνο τα δεδομένα της. Αντιθέτως κατά την εκτίμηση πυκνότητας πιθανότητας με μικτό μοντέλο δεν γνωρίζουμε σε ποιον πυρήνα ανήκει κάθε δεδομένο και επομένως έχουμε ένα επιπρόσθετο πρόβλημα σχετικά με την αντιστοίχηση δεδομένων και πυρήνων.

[4]Επιπλέον της προφανούς διαφοράς, δηλαδή ότι η εκτίμηση πυκνότητας και η ταξινόμηση-ως τεχνικές μάθησης- έχουν διαφορετικούς στόχους.

Μια σημαντική ιδιότητα των μικτών μοντέλων είναι ότι με κατάλληλες επιλογές συστατικών συναρτήσεων κατανομής μπορούν να προσεγγίσουν οποιαδήποτε συνεχή κατανομή με οσοδήποτε ακρίβεια εφόσον χρησιμοποιηθεί επαρκής αριθμός πυρήνων [32].

Είναι ενδιαφέρον να δούμε τις πληροφορίες ομαδοποίησης που μπορεί να μας εξασφαλίσει η εκτίμηση συνάρτησης πυκνότητας πιθανότητας με μικτές κατανομές. Ας υποθέσουμε ότι με κάποια διαδικασία μάθησης έχουν καθοριστεί όλοι οι παράμετροι της μικτής κατανομής. Καταρχήν, η εκ των προτέρων πιθανότητα ενός πυρήνα εκφράζει την αναλογία του αριθμού των δεδομένων που παράγονται από τον πυρήνα αυτό σε σχέση με το σύνολο των δεδομένων. Επιπλέον μέσω των συστατικών κατανομών παίρνουμε πληροφορίες σχετικά με τα χαρακτηριστικά της κάθε ομάδας (π.χ. κέντρο, διακύμανση). Και τέλος για ένα οποιοδήποτε δεδομένο x μπορούμε να υπολογίσουμε την εκ των υστέρων πιθανότητα να ανήκει σε ένα πυρήνα j κάνοντας χρήση του κανόνα του Bayes:

$$P(j|x,\Theta) = \frac{\pi_j p(x|j,\theta_j)}{\sum_{i=1}^{M} \pi_i p(x|i,\theta_i)}. \tag{1.21}$$

Οι εκ των υστέρων πιθανότητες ικανοποιούν την σχέση

$$\sum_{j=1}^{M} P(j|x;\Theta) = 1. \tag{1.22}$$

1.4 Μέθοδοι εκτίμησης παραμέτρων

1.4.1 Μέγιστη πιθανοφάνεια

Στην παρούσα ενότητα θα παρουσιάσουμε μια μέθοδο εύρεσης κατάλληλων τιμών παραμέτρων για τα παραμετρικά μοντέλα και θα δούμε πώς εφαρμόζεται στην περίπτωση της κανονικής κατανομής.

Έστω ότι έχουμε αποφασίσει για το ποια θα είναι η παραμετρική συνάρτηση που θα χρησιμοποιήσουμε αυτό που απομένει είναι να ορίσουμε τρόπους με τους οποίους θα βρίσκουμε κατάλληλες τιμές για τις παραμέτρους. Μια από τις πιο ευρέως χρησιμοποιούμενες μεθόδους είναι αυτή της μέγιστης πιθανοφάνειας. Με βάση την μέθοδο της μέγιστης πιθανοφάνειας αναζητούμε εκείνες τις τιμές των παραμέτρων οι οποίες μεγιστοποιούν μια συγκεκριμένη συνάρτηση, την οποία ονομάζουμε συνάρτηση πιθανοφάνειας.

Υποθέτουμε ότι έχουμε στην διάθεσή μας ένα σύνολο δειγμάτων X, όπου κάθε στοιχείο $x \in X$ ανήκει στον d-διάστατο χώρο. Επιπλέον υποθέτουμε ότι τα στοιχεία του X έχουν παραχθεί ανεξάρτητα το ένα από το άλλο με βάση την κατανομή $p(x|\Theta)$. Η από κοινού συνάρτηση πυκνότητας πιθανότητας των δεδομένων δίνεται από την σχέση:

$$P(X|\Theta) = \prod_{x \in X} p(x|\Theta). \tag{1.23}$$

Η $P(X|\Theta)$ όταν λαμβάνεται ως συνάρτηση των παραμέτρων Θ ονομάζεται πιθανοφάνεια του συνόλου δεδομένων X. Ο εκτιμητής μέγιστης πιθανοφάνειας είναι εξ ορισμού εκείνο το διάνυσμα παραμέτρων $\hat{\Theta}$ για το οποίο μεγιστοποιείται η πιθανοφάνεια. Μεγιστοποιώντας την ποσότητα (1.23) φαίνεται λογικό ότι η $p(x|\Theta)$ θα ταιριάζει όσο το δυνατό καλύτερα στην άγνωστη κατανομή των δεδομένων (ακριβέστερα στα δεδομένα X). Για αναλυτικούς κυρίως λόγους προκειμένου να βρούμε τον εκτιμητή μέγιστης πιθανοφάνειας είναι βολικότερο να εργαστούμε με το λογάριθμο της σχέσης (1.23). Λόγω του ότι ο λογάριθμος είναι γνησίως μονότονη (αύξουσα) συνάρτηση το μέγιστο της λογαριθμικής πιθανοφάνειας είναι συγχρόνως και το μέγιστο της πιθανόφανειας. Η λογαριθμική πιθανοφάνεια έχει την παρακάτω μορφή:

$$L(\Theta) = \log P(X|\Theta) = \sum_{x \in X} \log p(x|\Theta). \tag{1.24}$$

Εφόσον η $L(\Theta)$ είναι παραγωγίσιμη συνάρτηση ως προς το διάνυσμα Θ και, τότε ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\Theta}$ πρέπει να είναι στάσιμο σημείο της (1.24), δηλαδή να αποτελεί λύση της εξίσωσης

$$\nabla_{\Theta} L(\hat{\Theta}) = 0, \tag{1.25}$$

όπου με ∇_{Θ} συμβολίζουμε τον τελεστή παραγώγισης ως προς το διάνυσμα Θ. Στις περισσότερες των περιπτώσεων την παραπάνω εξίσωση δεν μπορούμε να την λύσουμε αναλυτικά πράγμα που σημαίνει ότι απαιτείται να καταφύγουμε σε κάποια μέθοδο βελτιστοποίησης προκειμένου να προσεγγίσουμε τον εκτιμητή μέγιστης πιθανοφάνειας. Ωστόσο επιλέγοντας ένα πλήθος κατανομών για την $p(x|\Theta)$ όπως, π.χ. αυτές που ανήκουν στην οικογένεια των εκθετικών κατανομών [9] μπορούμε να βρούμε τον εκτιμητή μέγιστης πιθανοφάνειας με ένα άμεσο τρόπο. Παρακάτω δείχνουμε μια τέτοια περίπτωση όπου η $p(x|\Theta)$ είναι η κανονική κατανομή.

Εφαρμογή σε κανονική κατανομή. Η σχέση (1.24) χρησιμοποιώντας ως $p(x|\Theta)$ την κατανομή από την σχέση (1.10) παίρνει την μορφή:

$$L(\Theta) = \sum_{x \in X} \left\{ -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}. \quad (1.26)$$

Παραγωγίζοντας την παραπάνω σχέση ως προς μ και Σ και θέτοντας τις αντίστοιχες μερικές παραγώγους ίσες με το μηδέν παίρνουμε τελικά:

$$\hat{\mu} = \frac{1}{|X|} \sum_{x \in X} x, \quad (1.27)$$

$$\hat{\Sigma} = \frac{1}{|X|} \sum_{x \in X} (x-\hat{\mu})(x-\hat{\mu})^T. \quad (1.28)$$

Παρατηρούμε ότι στον εκτιμητή μέγιστης πιθανοφάνειας η μέση τιμή $\hat{\mu}$ ορίζεται ως ο μέσος όρος των τιμών όλων των προτύπων. Κάτι τέτοιο φαίνεται λογικό επειδή η εκτιμούμενη τιμή του μέσου θα είναι πλησιέστερα στις τιμές των δεδομένων που συναντώνται συχνότερα στο σύνολο δεδομένων. Ακριβώς ανάλογα ο πίνακας συμμεταβλητότητας είναι ο μέσος όρος όλων των ποσοτήτων $(x-\hat{\mu})(x-\hat{\mu})^T$.

Εφαρμογή σε μικτές κατανομές. Προηγουμένως παρουσιάσαμε τη μέθοδο της μέγιστης πιθανοφάνειας στην περίπτωση της κανονικής κατανομής. Η μέθοδος χρησιμοποιείται και στην περίπτωση των μικτών μοντέλων, ωστόσο δεν είναι δυνατόν να επιτευχθεί μια άμεση αναλυτική λύση όπως στην περίπτωση των διάφορων απλών κατανομών.

Υποθέτουμε ότι έχουμε μια μικτή κανονική κατανομή $p(x|\Theta)$ η οποία ορίζεται από την σχέση (1.19) ενώ οι συστατικοί της πυρήνες προς το παρόν υποθέτουμε ότι μπορούν να έχουν οποιαδήποτε μορφή. Η λογαριθμική πιθανοφάνεια έχει την παρακάτω μορφή:

$$L(\Theta) = \log \prod_{x \in X} \sum_{j=1}^{M} \pi_j p(x|j, \theta_j) \sum_{x \in X} \log \sum_{j=1}^{M} \pi_j p(x|j, \theta_j). \quad (1.29)$$

Η $L(\Theta)$ για το συγκεκριμένο σύνολο X αποτελεί μια συνάρτηση του διανύσματος Θ. Η μεγιστοποίηση της (1.29) δεν είναι μια απλή διαδικασία όπως είναι στην περίπτωση των παραμετρικών μεθόδων. Η βασική δυσκολία συνίσταται στο ότι η συνάρτηση έχει υψηλή μη γραμμικότητα (λόγω του αθροίσματος μέσα στο λογάριθμο), και διαθέτει πολλά τοπικά μέγιστα πράγμα που σημαίνει ότι, αναζητώντας τον εκτιμητή μέγιστης πιθανοφάνειας μέσω ενός αλγορίθμου βελτιστοποίησης, είναι εύκολο να εγκλωβιστούμε σε ένα τοπικό μέγιστο. Εκτός των παραπάνω για την λογαριθμική πιθανοφάνεια μικτών κατανομών υπάρχουν διάφορα θεωρητικά ζητήματα σχετικά με την μοναδικότητα του εκτιμητή μέγιστης πιθανοφάνειας. Ειδικότερα λόγω των πολλών τοπικών μεγίστων ενδεχομένως το ολικό μέγιστο να προκύπτει για πολλά διαφορετικά διανύσματα παραμέτρων (που ορίζουν διαφορετικά μοντέλα), οπότε το βέλτιστο διάνυσμα δεν ορίζεται μοναδικά. Επίσης το πρόβλημα ύπαρξης μοναδικής λύσης προέρχεται και από την ίδια την κατανομή για το λόγο ότι μπορεί να μην είναι ταυτοποιήσιμη συνάρτηση[5]. Για τέτοιου είδους θεωρητικά ζητήματα ο αναγνώστης μπορεί να ανατρέξει στο [25, 19].

Ο εκτιμητής μέγιστης πιθανοφάνειας αντιστοιχεί σε κάποιο από τα στάσιμα σημεία της συνάρτησης πιθανοφάνειας. Επομένως ως μια πρώτη προσέγγιση στο πρόβλημα καθορισμού του εκτιμητή μέγιστης πιθανοφάνειας μπορούμε να βρούμε το σύστημα εξισώσεων που ικανοποιεί. Όπως θα δούμε η μορφή των εξισώσεων αυτών δεν επιτρέπει μια άμεση λύση. Αν $\hat{\Theta}$ είναι στάσιμο σημείο της (1.29), τότε ικανοποιεί τις εξισώσεις:

$$\sum_{x \in X} P(j|x, \hat{\Theta}) \nabla_{\theta_j} \log p(x|j, \hat{\theta}_j) = 0, \tag{1.30}$$

για κάθε διάνυσμα θ_j και

$$\hat{\pi}_{jk} = \frac{1}{|X|} \sum_{x \in X} P(j|x, \hat{\Theta}). \tag{1.31}$$

για κάθε εκ των προτέρων πιθανότητα π_j. Από την παραπάνω μορφή των εξισώσεων είναι φανερό ότι δεν μπορεί να βρεθεί αναλυτική λύση για το διάνυσμα παραμέτρων. Η μορφή της εξίσωσης (1.30) εξαρτάται κάθε φορά από την επιλογή της συνάρτησης πυρήνα $p(x|j, \theta_j)$. Παρακάτω εμφανίζονται οι εξισώσεις που παίρνουμε από την (1.30) στη περίπτωση που η $p(x|j, \theta_j)$ είναι η κανονική κατανομή:

$$\hat{\mu}_j = \frac{\sum_{x \in X} P(j|x, \hat{\Theta}) x}{\sum_{x \in X} P(j|x, \hat{\Theta})}, \tag{1.32}$$

$$\hat{\Sigma}_j = \frac{\sum_{x \in X} P(j|x, \hat{\Theta})(x - \hat{\mu}_j)(x - \hat{\mu}_j)^T}{\sum_{x \in X} P(j|x, \hat{\Theta})}. \tag{1.33}$$

Σημειωτέον ότι οι δεύτερες παράγωγοι της λογαριθμικής πιθανοφάνειας ως προς τις εκ των προτέρων πιθανότητες π_j δεν μπορούν να είναι θετικές:

$$\nabla_{\pi_j, \pi_i} L(\Theta) = -\frac{1}{\pi_j \pi_i} \sum_{x \in X} P(j|x, \Theta) P(i|x, \Theta) \leq 0. \tag{1.34}$$

[5]Μια παραμετρική κατανομή $p(x|\Theta)$ (ή και γενικότερα ένα παραμετρικό μοντέλο) είναι ταυτοποιήσιμη συνάρτηση αν για κάθε $\Theta_1 \neq \Theta_2$ υπάρχει τουλάχιστον ένα x τέτοιο ώστε $p(x|\Theta_1) \neq p(x|\Theta_2)$.

Για τον λόγο αυτό ο Εισιανός πίναχας έχει αρνητιχούς αριθμούς στη χύρια διαγώνιο και κατά συνέπεια δεν μπορεί να είναι θετικά ορισμένος. Αυτό έχει ως αποτέλεσμα να μη υπάρχει κανένα στάσιμο σημείο της λογαριθμικής πιθανοφάνειας που να είναι ελάχιστο, πράγμα που αποτελεί μια γενική ιδιότητα των μικτών κατανομών [22].

1.4.2 Μπεϋζιανή Μάθηση

Με την μέθοδο της μέγιστης πιθανοφάνειας αναζητούμε μια μοναδική λύση για το διάνυσμα παραμέτρων. Ωστόσο είναι δυνατόν τα δεδομένα X να αναπαριστώνται εξίσου καλά από διάφορες τιμές παραμέτρων και οι διαφορετικές τιμές παραμέτρων να δίνουν εναλλακτικές πιθανές ερμηνείες για την προέλευση των δεδομένων του X. Επομένως μια γενικότερη προσέγγιση εκτίμησης παραμέτρων είναι να βρούμε μια κατανομή εξαρτώμενη από τα δεδομένα που να εκφράζει την καταλληλότητα της κάθε δυνατής τιμής των παραμέτρων. Κάτι τέτοιο επιτυγχάνεται με την Μπεϋζιανή μάθηση.

Στη Μπεϋζιανή μάθηση υποθέτουμε ότι η άγνωστη συνάρτηση πυκνότητας πιθανότητας έχει μια γνωστή παραμετρική μορφή $p(x|\Theta)$ όπου το διάνυσμα παραμέτρων Θ θεωρείται άγνωστο και συγχρόνως αποτελεί τυχαία μεταβλητή. Μέρος της πληροφορίας μας για τις τιμές παραμέτρων Θ εκφράζεται μέσω μιας εκ των προτέρων κατανομής $P(\Theta)$, ενώ το υπόλοιπο μέρος της πληροφορίας προέρχεται από το σύνολο X (στη μέγιστη πιθανοφάνεια η πληροφορία προέρχονταν μόνο από το X) που υποτίθεται ότι έχει παραχθεί από την $p(x|\Theta)$. Αν τα δεδομένα X έχουν παραχθεί ανεξάρτητα μεταξύ τους, τότε η εκ των υστέρων κατανομή $p(\Theta|X)$ δίνεται από τον κανόνα του Bayes:

$$p(\Theta|X) = \frac{P(X|\Theta)p(\Theta)}{\int P(X|\Theta)p(\Theta)d\Theta} \tag{1.35}$$

Εφόσον έχει υπολογιστεί η εκ των υστέρων κατανομή, η εκτιμούμενη κατανομή θα δίνεται με βάση την σχέση

$$p(x|X) = \int p(x|\Theta)p(\Theta|X)d\Theta. \tag{1.36}$$

Η παραπάνω εκτίμηση εξαρτάται από την μορφή της κατανομής $p(\Theta|X)$. Όσο ο αριθμός των δεδομένων αυξάνει ο όρος της πιθανοφάνειας στη σχέση (1.35) γίνεται ισχυρότερος, ενώ καθώς ο αριθμός των δεδομένων τείνει στο άπειρο η $p(\Theta|X)$ είναι ανάλογη της $P(X|\Theta)$. Αν διαθέτουμε λίγα δεδομένα ο όρος $p(\Theta)$ επηρεάζει σημαντικά την λύση και τότε η Μπεϋζιανή μέθοδος ενδεχομένως να δώσει τελείως διαφορετική λύση από την μέγιστη πιθανοφάνεια. Λόγω του ότι η εφαρμογή της Μπεϋζιανής μεθόδου είναι δύσκολη υπολογιστικά (αφού απαιτεί την ολοκλήρωση ως προς Θ), πολλές φορές χρησιμοποιείται η προσέγγιση $p(x|X) \approx p(x|\Theta_{MAP})$, όπου Θ_{MAP} μεγιστοποιεί την $p(\Theta|X)$. Μια διαφορετική προσέγγιση υπολογισμού του ολοκληρώματος (1.36) είναι μέσω της μεθόδου ολοκλήρωσης Monte Carlo η οποία προυποθέτει την δειγματοληψία με βάση την κατανομή $p(\Theta|X)$ [13].

1.5 Βελτιστοποίηση μέσω του αλγορίθμου ΕΜ

1.5.1 Ορισμός του αλγορίθμου ΕΜ

Ο αλγόριθμος ΕΜ (Expectation-Maximization) [8] ορίζεται ως μια γενική διαδικασία μεγιστοποίησης λογαριθμικών πιθανοφανειών σε προβλήματα όπου κάποιες μεταβλητές δεν έχουν παρατηρηθεί (μη παρατηρήσιμες ή κρυμμένες μεταβλητές). Θα δώσουμε, καταρχήν, ένα γενικό ορισμό του αλγορίθμου και εν συνεχεία θα δούμε την μορφή που παίρνει στο πρόβλημα εκτίμησης πυκνότητας πιθανότητας υποθέτοντας μικτή κατανομή.

Η λειτουργία του ΕΜ βασίζεται στην σχέση μεταξύ δύο συνόλων. Το πρώτο σύνολο το ονομάζουμε ελλιπές σύνολο (incomplete set) και το δεύτερο πλήρες σύνολο (complete set). Ελλιπή σύνολα δεδομένων είναι συνήθως δείγματα δεδομένων που παίρνουμε από πειράματα ή στατιστικές μετρήσεις, για αυτό το λόγο και τέτοιου είδους σύνολα αποτελούν πραγματικά δεδομένα. Αντιθέτως πλήρη σύνολα δεδομένων είναι συνήθως υποθετικά σύνολα και εκφράζουν την μορφή που θα θέλαμε να έχουν τα δεδομένα μας σε ένα πείραμα. Ωστόσο στην πράξη μια τέτοια μορφή δεν είναι διαθέσιμη, δηλαδή τα σύνολα αυτά είναι μη παρατηρήσιμα.

Υποθέτουμε ότι έχουμε ένα ελλιπές σύνολο προτύπων X για το οποίο ορίζεται η από κοινού κατανομή $P(X|\Theta)$ η οποία εξαρτάται από το άγνωστο διάνυσμα παραμέτρων Θ. Υποθέτουμε επίσης ένα πλήρες σύνολο $Y = (X, Z)$ όπου Z είναι ένα σύνολο μη παρατηρήσιμων μεταβλητών. Η κατανομή $P(Y|\Theta)$ εξαρτάται από το ίδιο διάνυσμα παραμέτρων Θ. Οι δύο κατανομές, δηλαδή του ελλιπούς και του πλήρους συνόλου δεδομένων συνδέονται με την σχέση:

$$P(X|\Theta) = \int_Z P(X, Z|\Theta)dZ \tag{1.37}$$

Επίσης οι λογαριθμικές πιθανοφάνειες των δύο συνόλων είναι $L(\Theta) = \log P(X|\Theta)$ και $L_C(\Theta) = \log P(Y|\Theta)$, αντίστοιχα.

Το πρόβλημα μας είναι να βρούμε εκείνο το διάνυσμα παραμέτρων για το οποίο μεγιστοποιείται η λογαριθμική πιθανοφάνεια του ελλιπούς συνόλου. Ο αλγόριθμος EM προσπαθεί να μεγιστοποιήσει την ποσότητα αυτή (την $L(\Theta)$) αναδεικνύοντας την σχέση μεταξύ των δύο συνόλων. Συγκεκριμένα ο EM προσεγγίζει το πρόβλημα μεγιστοποίησης έμμεσα εφαρμόζοντας μια επαναληπτική διαδικασία για την λογαριθμική πιθανοφάνεια $L_C(\Theta)$ του πλήρους συνόλου. Επειδή όμως το σύνολο Y (συγκεκριμένα το Z) είναι μη παρατηρήσιμο και επομένως η λογαριθμική πιθανοφάνεια $L_C(\Theta)$ είναι ακαθόριστη, ο EM την λαμβάνει ως τυχαία μεταβλητή και υπολογίζει την αναμενόμενη τιμή της ως προς την κατανομή $P(Z|X,\Theta)$, όπου Θ λαμβάνει την τρέχουσα τιμή των παραμέτρων. Ειδικότερα εάν βρισκόμαστε στην $t+1$ επανάληψη του αλγορίθμου και το τρέχον διάνυσμα είναι το $\Theta^{(t)}$ η προηγούμενη ποσότητα ορίζεται ως εξής:

$$Q(\Theta;\Theta^{(t)}) = E[L_C(\Theta)|X,\Theta^{(t)}] = \int_Z L_C(\Theta)P(Z|X,\Theta^{(t)}), \tag{1.38}$$

όπου

$$P(Z|X,\Theta) = \frac{P(X,Z|\Theta)}{P(X|\Theta)}. \tag{1.39}$$

Κάθε επανάληψη του αλγόριθμου αποτελείται από δύο βήματα: το E-βήμα (Expectation-step) στο οποίο καθορίζεται η $Q(\Theta;\Theta^{(t)})$ και το M-βήμα (Maximization-step) στο οποίο μεγιστοποιείται η ποσότητα αυτή ως προς το διάνυσμα παραμέτρων. Πιο συγκεκριμένα τα βήματα στην $t+1$ επανάληψη ορίζονται ως εξής:

E-βήμα.: Υπολογισμός της ποσότητας $Q(\Theta;\Theta^{(t)})$.

M-βήμα: $\Theta^{(t+1)} = \arg\max_\Theta Q(\Theta;\Theta^{(t)})$.

Σύμφωνα με τις ιδιότητες του αλγορίθμου η λογαριθμική πιθανοφάνεια του ελλιπούς συνόλου δεν μειώνεται μετά από μια επανάληψη του αλγορίθμου, δηλαδή ισχύει:

$$L(\Theta^{(t+1)}) \geq L(\Theta^{(t)}). \tag{1.40}$$

Από τον τρόπο που ορίζεται ο αλγόριθμος δεν είναι ξεκάθαρο για το πώς ορίζεται το σύνολο των μη παρατηρήσιμων μεταβλητών Z και γιατί η μεγιστοποίηση της ποσότητας $Q(\Theta;\Theta^{(t)})$ σε κάθε επανάληψη έχει ως αποτέλεσμα την αύξηση της $L(\Theta)$. Για αυτά τα ζητήματα ο αναγνώστης μπορεί να ανατρέξει στον άρθρο εισαγωγής του αλγορίθμου [8] ή στο βιβλίο των McLachlan και Krishnan [21] το οποίο αναφέρεται αποκλειστικά στον αλγόριθμο αυτόν. Στην επόμενη ενότητα θα εφαρμόσουμε τον αλγόριθμο στο πρόβλημα μεγιστοποίησης της πιθανοφάνειας στην περίπτωση της μικτής κατανομής.

1.5.2 Εφαρμογή του EM σε μικτές κατανομές

'Οπως είδαμε στην ενότητα (1.4.1) ο εκτιμητής μέγιστης πιθανοφάνειας δεν μπορεί να βρεθεί αναλυτικά. Προχειμένου να λύσουμε το πρόβλημα με τον αλγόριθμο EM πρέπει να ορίσουμε το σύνολο των μη παρατηρήσιμων μεταβλητών Z και ακολούθως την συνάρτησης $Q(\Theta;\Theta^{(t)})$.

Επιστρέφοντας σε όσα είχαμε πει συγκρίνοντας το μικτό μοντέλο και το στατιστικό μοντέλο ταξινόμησης τα αντίστοιχα σύνολα προτύπων εκπαίδευσης έχουν την εξής διαφορά. Κάθε πρότυπο στο πρόβλημα ταξινόμησης αποτελεί ένας ζεύγος της μορφής (x, C_k) όπου C_k υποδεικνύει την κατηγορία του x, δηλαδή το σύνολο εκπαίδευσης αποτελείται από ανεξάρτητα σύνολα δεδομένων ένα για κάθε κατηγορία. Αντιθέτως αν και το μικτό μοντέλο θεωρεί τα δεδομένα ως ένα μίγμα ανεξάρτητων υποσυνόλων, ένα για κάθε συστατικό πυρήνα, τα υποσύνολα αυτά στην πράξη είναι αχαθόριστα δεδομένου ότι δεν υπάρχει καμιά πληροφορία υπόδειξης σχετικά με τον πυρήνα προέλευσης του κάθε δεδομένου x. Υπό αυτή την έννοια το σύνολο εκπαίδευσης θεωρείται πως είναι ελλιπές[6]. Προφανώς θα θέλαμε τα πρότυπα εκπαίδευσης να ορίζονται όπως στο πρόβλημα ταξινόμησης, δηλαδή να είναι της μορφής (x, z) όπου z είναι ένας ακέραιος, $z \in \{1, \dots, M\}$ που υποδεικνύει τον πυρήνα από τον οποίο έχει παραχθεί το x. Εάν τα πρότυπα μας είχαν αυτή την μορφή η εύρεση των παραμέτρων της μικτής κατανομής θα ήταν εύκολη. Για παράδειγμα στην περίπτωση των κανονικών πυρήνων θα υπολογίζαμε τις παραμέτρους της κάθε κανονικής κατανομής χρησιμοποιώντας το αντίστοιχο υποσύνολο, που όπως δείξαμε γίνεται με αναλυτικό τρόπο.

Με βάση τα προηγούμενα αν το σύνολο εκπαίδευσης είναι το X και η αντίστοιχη λογαριθμική πιθανοφάνεια δίνεται από (1.29), για κάθε $x \in X$ εισάγουμε την μεταβλητή $z(x)$ ως ένα ακέραιο που παίρνει τιμές από το σύνολο $\{1, \dots, M\}$ και υποδεικνύει τον πυρήνα που παρήγαγε το x[7] [?]. Το πλήρες σύνολο είναι το εξής:

$$Y = \{(x, z(x)), x \in X\},$$

ενώ η λογαριθμική πιθανοφάνεια του παραπάνω συνόλου είναι η ακόλουθη:

$$L_C(\Theta) = \log \prod_{x \in X} \pi_{z(x)} p(x|z(x), \theta_{z(x)}) = \sum_{x \in X} \log \pi_{z(x)} p(x|z(x), \theta_{z(x)}). \quad (1.41)$$

[6]Ο χαρακτηρισμός του συνόλου X προφανώς σχετίζεται με την υπόθεση της μικτής κατανομής που κάναμε για την προέλευση του X.

[7]Ισοδύναμα κάθε μη παρατηρήσιμη μεταβλητή θα μπορούσε να είχε οριστεί ως ένα M-διάστατο διάνυσμα υπόδειξης που παίρνει τιμές μηδέν ή ένα ως εξής: $z_j(x) = 1$, αν ο πυρήνας j παρήγαγε το x και $z_j(x) = 1$ διαφορετικά.

Η παραπάνω σχέση προκύπτει ως εξής: Το πρότυπο x γνωρίζουμε ότι ανήκει στο πυρήνα $z(x)$ πράγμα που σημαίνει ότι παράγεται με βάση την πιθανότητα $p(x, z(x)|\Theta) = \pi_{z(x)}p(x|z(x), \theta_{z(x)})$. Επομένως η από κοινού πιθανότητα των στοιχείων του Y είναι $P(Y|\Theta) = \prod_{x\in X} \pi_{z(x)}p(x|z(x), \theta_{z(x)})$ από την οποία προκύπτει η σχέση (1.41).

Στην πραγματικότητα οι μεταβλητές $z(x)$ είναι άγνωστες, κάτι που σημαίνει ότι το πλήρες σύνολο (X, Z) είναι ακαθόριστο (όπως προαναφέραμε κάθε τέτοιο σύνολο είναι υποθετικό). Επομένως η ποσότητα $L_C(\Theta)$ είναι επίσης ακαθόριστη. Υπάρχουν πολλές επιλογές για την μορφή του πλήρους συνόλου που προκύπτουν αν σκεφτούμε ότι για κάθε x η μεταβλητή z μπορεί να πάρει τιμές από 1 έως M. Συγκεκριμένα υπάρχουν M^N διαφορετικές επιλογές του συνόλου και ισάριθμες εκδοχές της λογαριθμικής πιθανοφάνειας του πλήρους συνόλου. Από τον ορισμό της συνάρτησης Q και δεδομένου ότι οι μεταβλητές του συνόλου Z παίρνουν διακριτές τιμές έχουμε:

$$Q(\Theta; \Theta^{(t)}) = \sum_Z L_C(\Theta)P(Z|X, \Theta^{(t)}), \tag{1.42}$$

ή με βάση την (1.29), (1.39) και (1.41)

$$Q(\Theta; \Theta^{(t)}) = \sum_Z \sum_{x\in X} \log\left\{\pi_{z(x)}p(x|z(x), \theta_{z(x)})\right\} \frac{\prod_{x\in X} \pi^{(t)}_{z(x)}p(x|z(x), \theta^{(t)}_{z(x)})}{\prod_{x\in X} p(x|\Theta^{(t)})} \tag{1.43}$$

και χρησιμοποιώντας την (1.21)

$$Q(\Theta; \Theta^{(t)}) = \sum_Z \sum_{x\in X} \log \pi_{z(x)}p(x|z(x), \theta_{z(x)}) \prod_{x\in X} P(z(x)|x, \Theta^{(t)}). \tag{1.44}$$

Από την παραπάνω σχέση με λίγες πράξεις και χρησιμοποιώντας την σχέση (1.22) προκύπτει τελικά

$$Q(\Theta; \Theta^{(t)}) = \sum_{x\in X} \sum_{j=1}^{M} P(j|x, \Theta^{(t)}) \log \pi_j p(x|j, \theta_j). \tag{1.45}$$

Η παραπάνω ποσότητα υπολογίζεται στο E-βήμα. Στο M-βήμα υπολογίζεται το διάνυσμα παραμέτρων $\Theta^{(t+1)}$ μεγιστοποιώντας την ποσότητα (1.45). Στην περίπτωση που κάθε $p(x|j, \theta_j)$ αποτελεί κανονική κατανομή, το διάνυσμα $\Theta^{(t+1)}$ προκύπτει με απλή παραγώγιση της συνάρτησης Q ως προς κάθε παράμετρο λαμβάνοντας υπόψη τον περιορισμό (1.20) για τις εκ των προτέρων πιθανότητες. Οι εξισώσεις ενημέρωσης των παραμέτρων είναι οι ακόλουθες:

$$\mu_j^{(t+1)} = \frac{\sum_{x\in X} P(j|x, \Theta^{(t)})x}{\sum_{x\in X} P(j|x, \Theta^{(t)})}, \tag{1.46}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{x\in X} P(j|x,\Theta^{(t)})(x-\mu_j^{(t+1)})(x-\mu_j^{(t+1)})^T}{\sum_{x\in X} P(j|x,\Theta^{(t)})}, \tag{1.47}$$

$$\pi_j^{(t+1)} = \frac{1}{|X|}\sum_{x\in X} P(j|x,\Theta^{(t)}), \tag{1.48}$$

για κάθε $j = 1,\ldots,M$. Η επαναληπτική διαδικασία ξεκινά με αρχικοποίηση του διανύσματος παραμέτρων και εναλλάσσεται μεταξύ δύο βημάτων. Στο E-βήμα υπολογίζονται οι εκ των υστέρων πιθανότητες με βάση την τρέχουσα τιμή των παραμέτρων και στο M-βήμα δίνονται νέες τιμές στις παραμέτρους με βάση τις παραπάνω σχέσεις. Ο αλγόριθμος αυξάνει σε κάθε επανάληψη την πιθανοφάνεια έως ότου έχουμε σύγκλιση στο τελικό διάνυσμα παραμέτρων Θ^* (η πιθανοφάνεια έχει μεγιστοποιηθεί έστω και τοπικά).

1.6 Ανασκόπηση της εργασίας

Στο δεύτερο κεφάλαιο παρουσιάζουμε την κύρια θεωρητική συνεισφορά της εργασίας που αποτελεί η εισαγωγή του Z-μοντέλου. Το μοντέλο αυτό αποτελεί γενίκευση των γνωστών μεθόδων εκτίμησης δεσμευμένων κατανομών με χρήση μικτών κατανομών που επίσης περιγράφονται στο κεφάλαιο αυτό. Η βασική ιδιότητα του Z-μοντέλου είναι ότι μοντελοποιεί τις δεσμευμένες κατανομές με μικτές κατανομές τέτοιες ώστε ο κάθε πυρήνας να χρησιμοποιείται συγχρόνως από ένα υποσύνολο των κατηγοριών. Ο τρόπος που οι πυρήνες συνεισφέρουν στις διάφορες κατηγορίες καθορίζεται από ένα πίνακα υπόδειξης Z. Η εφαρμογή του μοντέλου προϋποθέτει τον καθορισμό του πίνακα αυτού. Δοθέντος ότι ο πίνακας Z παίρνει εξ αρχής μια σταθερή τιμή, δείχνουμε πώς ο αλγόριθμος EM μπορεί να χρησιμοποιηθεί για την βελτιστοποίηση των παραμέτρων. Επίσης στο κεφάλαιο αυτό παρουσιάζουμε μια εκτενή ανάλυση (βασιζόμενοι στη μέθοδο της μέγιστης πιθανοφάνειας) των περιπτώσεων που η χρήση κοινών πυρήνων είναι ωφέλιμη από την σκοπιά της ταξινόμησης και αντιστρόφως. Η ανάλυση είναι σημαντική αφού μας προσφέρει κατευθύνσεις όσον αφορά την επιλογή του πίνακα Z.

Στο τρίτο κεφάλαιο περιγράφουμε μια μέθοδο εκπαίδευσης του Z-μοντέλου που ουσιαστικά καθορίζει τις τιμές του πίνακα Z. Ο αλγόριθμος εκπαίδευσης βασίζεται στη εισαγωγή μιας κατάλληλης αντικειμενικής συνάρτησης, ενώ η βελτιστοποίηση των παραμέτρων γίνεται μέσω του αλγορίθμου EM.

Στο τέταρτο κεφάλαιο παρουσιάζουμε συγκριτικά αποτελέσματα των μεθόδων σε πέντε γνωστά προβλήματα ταξινόμησης. Τέλος στο επίλογο δίνουμε μια μικρή περίληψη της εργασίας καθώς και συγκεκριμένες κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2

Μικτές Κατανομές για Ταξινόμηση

2.1 Γενικά

Σε αυτό το κεφάλαιο μελετώνται μέθοδοι εκτίμησης των δεσμευμένων κατανομών των κατηγοριών για το πρόβλημα ταξινόμησης. Οι μέθοδοι βασίζονται στη χρήση μικτών κατανομών για την μοντελοποίηση της δεσμευμένης κατανομής κάθε κατηγορίας. Οι μικτές κατανομές (Κεφάλαιο 1) αποτελούν παραμετρικά μοντέλα με γενικές ιδιότητες εκτίμησης πυκνότητας πιθανότητας μιας άγνωστης κατανομής.

Η εκτίμηση των δεσμευμένων κατανομών αποτελεί το υπολογιστικά δύσκολο μέρος της κατασκευής ενός συστήματος ταξινόμησης με βάση την στατιστική προσέγγιση. Όπως εξηγήσαμε στο εισαγωγικό κεφάλαιο για την παραπάνω εκτίμηση απαιτείται ο διαχωρισμός των δεδομένων σε υποσύνολα με βάση την κατηγορία στην οποία ανήκει το καθένα. Στη συνέχεια υπολογίζεται η κάθε δεσμευμένη κατανομή χρησιμοποιώντας τα δεδομένα του αντίστοιχου υποσυνόλου. Αυτό υποδηλώνει ότι τα παραμετρικά μοντέλα $p(x|C_k;\Theta_k)$ που θα υποθέσουμε μπορεί κάλλιστα να είναι λειτουργικά ανεξάρτητα[1], δηλαδή να μην περιέχουν κοινές παραμέτρους. Για παράδειγμα η υπόθεση ότι κάθε δεσμευμένη κατανομή αποτελεί μια ξεχωριστή κανονική κατανομή ανήκει στη προηγούμενη κατηγορία μεθόδων.

Η πρόθεση μας είναι να χρησιμοποιήσουμε μικτές κατανομές για την μοντελοποίηση των δεσμευμένων κατανομών και σύμφωνα με τα παραπάνω μια προφανής λύση είναι να χρησιμοποιηθούν ανεξάρτητες μικτές κατανομές. Ωστόσο σε αυτό το κεφάλαιο θα δούμε επίσης μεθόδους βασισμένες σε μικτές κατανομές όπου

[1] Ο όρος εμφανίζεται στο βιβλίο των Duda και Heart [9].

τα μοντέλα των δεσμευμένων κατανομών δεν είναι λειτουργικά ανεξάρτητα αλλά περιέχουν κοινές παραμέτρους. Οι κοινές παράμετροι θα προκύπτουν από τη χρήση κοινών πυρήνων. Παρουσιάζουμε μια εκτενή ανάλυση των πλεονεκτημάτων και μειονεκτημάτων της μεθόδου των ξεχωριστών μικτών κατανομών και αυτή των μικτών κατανομών με κοινούς πυρήνες. Η ανάλυση γίνεται με βάση τον περιορισμό ότι και οι δύο μέθοδοι πρέπει να χρησιμοποιούν συνολικά τον ίδιο αριθμό πυρήνων. Παρουσιάζουμε παραδείγματα όπου η μια μέθοδος είναι καλύτερη από την άλλη και αντιστρόφως. Τελικά εισάγουμε μια νέα πιο γενική μέθοδο εκτίμησης δεσμευμένων κατανομών με μικτές κατανομές και εξετάζουμε προβλήματα ταξινόμησης όπου η προτεινόμενη μέθοδος δίνει το αποδοτικότερο σύστημα ταξινόμησης.

2.2 Ξεχωριστές μικτές κατανομές

Για την εκτίμηση των δεσμευμένων κατανομών υποθέτουμε K μικτές κατανομές τέτοιες ώστε η καθεμιά να έχει το δικό της σύνολο πυρήνων. Αν M_k είναι ο αριθμός των πυρήνων που χρησιμοποιούνται από το μικτό μοντέλο της κατηγορίας C_k, τότε η αντίστοιχη δεσμευμένη κατανομή δίνεται από την σχέση:

$$p(x|C_k;\pi_k,\theta_k) = \sum_{j_k=1}^{M_k} \pi_{j_k k} p(x|j_k;\theta_{j_k}) \qquad k = 1,\ldots,K, \tag{2.1}$$

όπου $\pi_k = \{\pi_{j_k k}, j_k = 1,\ldots,M_k\}$ και $\theta_k = \{\theta_{j_k}, j_k = 1,\ldots,M_k\}$. Η παράμετρος π_{j_k} εκφράζει την εκ των προτέρων πιθανότητα σύμφωνα με την οποία ένα δεδομένο της κατηγορίας C_k προέρχεται από τον πυρήνα j. Για τις εκ των προτέρων πιθανότητες ισχύει ο περιορισμός:

$$\sum_{j_k=1}^{M_k} \pi_{j_k k} = 1, \tag{2.2}$$

για κάθε k. Λόγω του γεγονότος ότι οι δεσμευμένες κατανομές δεν περιέχουν κοινές παραμέτρους η παραπάνω υπόθεση για τις κατανομές θα καλείται *μοντέλο ανεξάρτητων μικτών κατανομών*. Οι ιδιότητες της μεθόδου των ανεξάρτητων μικτών κατανομών περιγράφονται στο [15]. Επίσης διάφορες εφαρμογές μοντέλου σε προβλήματα ταξινόμησης εμφανίζονται στα [2, 35].

2.3 Το μοντέλο των κοινών πυρήνων

Έστω ότι έχουμε στη διάθεση μας ένα σύνολο M πυρήνων. Αν σε κάθε πυρήνα αντιστοιχεί μια κατανομή με παραμέτρους θ_j συμβολίζουμε με θ το διάνυσμα όλων των κατανομών των πυρήνων, δηλαδή $\theta = (\theta_1, \ldots, \theta_M)$. Υποθέτουμε ότι οι δεσμευμένες κατανομές των κατηγοριών δίνονται από τις ακόλουθες μικτές κατανομές:

$$p(x|C_k; \pi_k, \theta) = \sum_{j=1}^{M} \pi_{jk} p(x|j; \theta_j) \qquad k = 1, \ldots, K, \tag{2.3}$$

όπου έχουμε χρησιμοποιήσει ανάλογους συμβολισμούς με αυτούς της προηγούμενης ενότητας. Επίσης με Θ συμβολίζουμε το διάνυσμα όλων των παραμέτρων (εκ των προτέρων πιθανοτήτων π_{jk} και παραμέτρους πυρήνων θ_j). Οι εκ των προτέρων πιθανότητες δεν λαμβάνουν αρνητικές τιμές και ικανοποιούν το περιορισμό

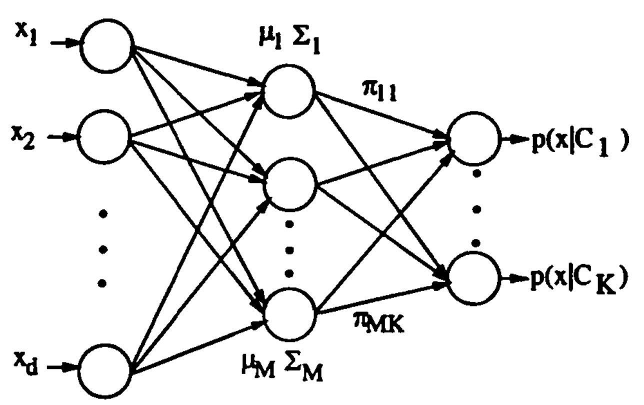
$$\sum_{j=1}^{M} \pi_{jk} = 1. \tag{2.4}$$

Παρατηρούμε ότι το μοντέλο έχει την ιδιότητα ότι οι δεσμευμένες κατανομές είναι λειτουργικά εξαρτημένες αφού οι παράμετρες των πυρήνων είναι κοινές. Η προηγούμενη υπόθεση για τις δεσμευμένες κατανομές θα αναφέρεται στη συνέχεια ως *μοντέλο κοινών πυρήνων* [29]. Η μέθοδος των ξεχωριστών μικτών κατανομών διαφέρει από την προηγούμενη στο γεγονός ότι υποθέτει λειτουργικά ανεξάρτητες μικτές κατανομές (δεν υπάρχουν κοινοί πυρήνες). Με άλλα λόγια στο μοντέλο των κοινών πυρήνων κάθε πυρήνας έχει την δυνατότητα να συνεισφέρει σε όλες τις δεσμευμένες κατανομές, ενώ αντιθέτως στο αντίστοιχο μοντέλο των ανεξάρτητων μικτών κατανομών κάθε πυρήνας συνεισφέρει μόνο σε μια δεσμευμένη κατανομή. Αν κατά την υπόθεση των ανεξάρτητων μικτών κατανομών έχουμε αποφασίσει να χρησιμοποιήσουμε M συνολικά πυρήνες, τότε μπορούμε να εκφράσουμε το μοντέλο αυτό μέσω του αντίστοιχου των κοινών πυρήνων ως εξής: Αν M_k είναι οι πυρήνες που θα χρησιμοποιηθούν από την κατηγορία C_k, θέτουμε για κάθε πυρήνα j ενός συνόλου πυρήνων με M_k στοιχεία (όλα τα σύνολα είναι ξένα μεταξύ τους) το περιορισμό $\pi_{j\ell} = 0$, για κάθε $\ell \neq k$.

Έστω ότι διαθέτουμε ένα σύνολο δεδομένων X όπως αυτό ορίστηκε στη ενότητα 1.2. Αν υποθέσουμε ότι τα δεδομένα κάθε υποσυνόλου X_k έχουν παραχθεί ανεξάρτητα με βάση την κατανομή $p(x|C_k; \pi_k, \theta)$, τότε η λογαριθμική πιθανοφάνεια του X είναι

$$L(\Theta) = \log P(X|\Theta) = \sum_{k=1}^{K} \sum_{x \in X_k} \log p(x|C_k; \pi_k, \theta) = \sum_{k=1}^{K} L_k(\pi_k, \theta), \tag{2.5}$$

όπου L_k είναι η λογαριθμική πιθανοφάνεια των δεδομένων της κατηγορίας C_k (αντιστοιχεί στο X_k). Στο [29, 30] περιγράφουμε ένα αλγόριθμο EM για την



Σχήμα 2.1: Η αρχιτεκτονική του μοντέλου των κοινών πυρήνων.

μεγιστοποίηση της λογαριθμικής πιθανοφάνειας ως προς τις παραμέτρους του μοντέλου. Αν υποθέσουμε ότι κάθε $p(x|j;\theta_j)$ αποτελεί μια κανονική κατανομή της μορφής (1.10), οι εξισώσεις ενημέρωσης των παραμέτρων είναι οι ακόλουθες:

$$\mu_j^{(t+1)} = \frac{\sum_{k=1}^{K}\sum_{x\in X_k} P(j|x,C_k;\pi_k^{(t)},\theta^{(t)})x}{\sum_{k=1}^{K}\sum_{x\in X_k} P(j|x,C_k;\pi_k^{(t)},\theta^{(t)})}, \tag{2.6}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{k=1}^{K}\sum_{x\in X_k} P(j|x,C_k;\pi_k^{(t)},\theta^{(t)})(x-\mu_j^{(t+1)})(x-\mu_j^{(t+1)})^T}{\sum_{k=1}^{K}\sum_{x\in X_k} P(j|x,C_k;\pi_k^{(t)},\theta^{(t)})}, \tag{2.7}$$

$$\pi_{jk}^{(t+1)} = \frac{1}{|X_k|}\sum_{x\in X_k} P(j|x,C_k;\pi_k^{(t)},\theta^{(t)}) \quad k=1,\ldots,K, \tag{2.8}$$

για κάθε j. Στις παραπάνω εξισώσεις με $P(j|x,C_k;\pi_k,\theta)$ συμβολίζουμε την εκ των υστέρων πιθανότητα σύμφωνα με την οποία ένα δεδομένο x, που ανήκει στη κατηγορία C_k, έχει προέρθει από τον πυρήνα j και δίνεται από τον κανόνα του Bayes

$$P(j|x,C_k;\pi_k,\theta) = \frac{\pi_{jk}p(x|j;\theta_j)}{\sum_{i=1}^{M}\pi_{ik}p(x|i;\theta_i)}. \tag{2.9}$$

Το βασικό επιχείρημα σχετικά με την χρησιμότητα του μοντέλου των κοινών πυρήνων είναι ότι μπορεί να είναι αποτελεσματικό σε περιπτώσεις προβλημάτων ταξινόμησης με σημαντικό βαθμό επικάλυψης των κατηγοριών. Αυτό συμβαίνει διότι χρησιμοποιώντας κοινούς πυρήνες γίνεται δυνατή η ταυτόχρονη αναπαράσταση δεδομένων που ανήκουν σε διαφορετικές κατηγορίες. Ωστόσο σε πραγματικά προβλήματα δεν είναι γνωστή η μορφή της επικάλυψης των δεδομένων των κατηγοριών και ενδεχομένως εφαρμόζοντας το μοντέλο των κοινών πυρήνων να οδηγηθούμε σε μια αναξιόπιστη αναπαράσταση από τη σκοπιά της ταξινόμησης.

Συγκεκριμένα αν βασιστούμε στη μέθοδο της μέγιστης πιθανοφάνειας είναι δυνατόν να βρούμε τέτοιες τιμές παραμέτρων για κάποιους πυρήνες ώστε οι πυρήνες αυτοί να αναπαριστούν δεδομένα διαφορετικών κατηγοριών ακόμα και αν δεν υπάρχει τοπική επικάλυψη μεταξύ των κατηγοριών. Στην επόμενη ενότητα μελετώνται περιπτώσεις όπου κάτι τέτοιο συμβαίνει και επηρεάζει σημαντικά την επίδοση του συστήματος.

2.4 Σύγκριση των δύο μεθόδων

Σε αυτή την ενότητα παρουσιάζουμε συγκριτικά τα πλεονεκτήματα και μειονεκτήματα του μοντέλου των κοινών πυρήνων και των ανεξάρτητων μικτών κατανομών. Η ανάλυση βασίζεται στην υπόθεση ότι και οι δύο μέθοδοι χρησιμοποιούν συνολικά ίσο αριθμό πυρήνων. Θα μας απασχολήσουν τα ακόλουθα δύο ζητήματα:

- υποθέσεις που απαιτούνται κατά την εφαρμογή της κάθε μεθόδου σχετικά με τον αριθμό πυρήνων
- καθορισμός περιπτώσεων όπου η χρήση κοινών πυρήνων οδηγεί σε καλύτερη αναπαράσταση των δεδομένων από τη σκοπιά της ταξινόμησης και αντιστρόφως.

Σε ό,τι αφορά το πρώτο ζήτημα, κατά την εφαρμογή του μοντέλου των κοινών πυρήνων απαιτείται να επιλεγεί μόνο ο συνολικός αριθμός των πυρήνων M. Από την άλλη, η εφαρμογή των ανεξάρτητων μικτών κατανομών απαιτεί τον καθορισμό ενός διαχωρισμού των M συνολικά πυρήνων σε K υποσύνολα έτσι ώστε κάθε κατηγορία C_k να χρησιμοποιεί M_k πυρήνες. Θεωρούμε την προηγούμενη διαφορά των δύο μεθόδων ως πλεονέκτημα του μοντέλου των κοινών πυρήνων λόγω του ότι δεν είναι φανερό πώς μπορεί να επιτευχθεί ένας αποτελεσματικός διαχωρισμός των M πυρήνων κατά την εφαρμογή των ανεξάρτητων μικτών κατανομών[2].

Όσον αφορά το δεύτερο ζήτημα, είναι φανερό ότι εφαρμόζοντας το μοντέλο των κοινών πυρήνων κάποιοι πυρήνες μπορούν να αναπαριστούν δεδομένα που ανήκουν σε περισσότερες από μια κατηγορίες, ή διαφορετικά να συνεισφέρουν στην εκτίμηση της δεσμευμένης κατανομής πολλών κατηγοριών. Κάτι τέτοιο

[2]Εφόσον δεν υπάρχει καμιά εκ των προτέρων πληροφορία σχετικά με την πολυπλοκότητα της κάθε δεσμευμένης κατανομής, οι πυρήνες διανέμονται ισάριθμα στις κατηγορίες.

απαγορεύεται κατά την εφαρμογή του μοντέλου των ανεξάρτητων μικτών κατανομών. Επομένως το μοντέλο των κοινών πυρήνων επιτρέπει την μείωση του συνολικού αριθμού πυρήνων που απαιτούνται για την αναπαράσταση των δεδομένων. Ωστόσο, μένει να ερευνήσουμε αν η προηγούμενη ιδιότητα των κοινών πυρήνων είναι ωφέλιμη από τη πλευρά της ταξινόμησης. Παρακάτω αναλύονται κάποιες περιπτώσεις προβλημάτων ταξινόμησης όπου η χρήση κοινών πυρήνων έχει ως συνέπεια την βελτίωση της γενικευτικής ικανότητας του συστήματος καθώς και ορισμένες άλλες όπου έχει αντίθετα αποτελέσματα.

Εξετάζουμε αρχικά τις δύο περιπτώσεις μέσω δύο απλών παραδειγμάτων ταξινόμησης. Τα δύο παραδείγματα αποτελούν προβλήματα ταξινόμησης δύο κατηγοριών και τα δεδομένα ανήκουν στο δισδιάστατο χώρο (Σχήμα 2.2 και 2.3 αντίστοιχα). Και στις δύο περιπτώσεις υποθέτουμε ότι ο συνολικός αριθμός πυρήνων είναι δύο. Στο πρώτο παράδειγμα (Σχήμα 2.2), τα δεδομένα της πρώτης κατηγορίας προέρχονται από δύο ομάδες δεδομένων, ενώ τα αντίστοιχα της δεύτερης κατηγορίας ανήκουν σε μια ομάδα δεδομένων η οποία όμως έχει σημαντικό βαθμό επικάλυψης με μια εκ των δύο ομάδων της πρώτης κατηγορίας. Το μοντέλο των κοινών πυρήνων με δύο πυρήνες αναπαριστά ικανοποιητικά τα δεδομένα και των δύο κατηγοριών[3] (Σχήμα 2.2α). Προκειμένου να πάρουμε μια ανάλογη αναπαράσταση των δεδομένων εφαρμόζοντας το μοντέλο των ανεξάρτητων μικτών κατανομών απαιτούνται συνολικά τρεις πυρήνες (δύο για την πρώτη κατηγορία και ένας για την δεύτερη). Είναι φανερό ότι η χρήση δύο μόνο πυρήνων (ένας ανά κατηγορία) οδηγεί σε μια ανεπαρκή αναπαράσταση (Σχήμα 2.2β).

Το δεύτερο παράδειγμα (Σχήμα 2.3) είναι ανάλογο του πρώτου, με τη διαφορά ότι ο βαθμός επικάλυψης των δύο διαφορετικής κατηγορίας ομάδων δεδομένων έχει μειωθεί αισθητά (ασθενής επικάλυψη). Το μοντέλο των κοινών πυρήνων δίνει τέτοια λύση για τις παραμέτρους των πυρήνων ώστε ένας από τους πυρήνες τοποθετείται πάνω στο όριο απόφασης (Σχήμα 2.3α). Για το λόγο αυτό ο ταξινομητής παρουσιάζει αυξημένο σφάλμα γενίκευσης. Αντιθέτως η μέθοδος των ανεξάρτητων μικτών κατανομών (ένας πυρήνας ανά κατηγορία), αν και η λύση που δίνει δεν είναι καθόλου ικανοποιητική από τη σκοπιά της εκτίμησης πυκνότητας πιθανότητας, προσεγγίζει ωστόσο το πραγματικό όριο απόφασης με ικανοποιητική ακρίβεια (Σχήμα 2.3β).

Το φαινόμενο της τοποθέτησης κάποιων πυρήνων πάνω στο όριο απόφασης κατά την εφαρμογή του μοντέλου των κοινών πυρήνων εξηγείται αν εξετάσουμε

[3]Και στα δύο παραδείγματα χρησιμοποιήθηκαν κανονικοί πυρήνες, ενώ οι λύσεις που απεικονίζονται στα σχήματα αποτελούν εκτιμητές μέγιστης πιθανοφάνειας.

τις συνθήκες που ισχύουν για τις παραμέτρους των πυρήνων σε στάσιμα σημεία (συνήθως τοπικά μέγιστα) της πιθανοφάνειας. Αν το $\hat{\Theta}$ είναι ένα στάσιμο σημείο της λογαριθμικής πιθανοφάνειας (2.5), για το διάνυσμα παραμέτρων θ_j κάθε πυρήνα j ισχύει

$$\sum_{k=1}^{K} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0, \tag{2.10}$$

και για κάθε εκ των προτέρων πιθανότητα π_{jk}

$$\hat{\pi}_{jk} = \frac{1}{|X_k|} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}). \tag{2.11}$$

Με βάση την εξίσωση (2.10) μπορούμε να αναγνωρίσουμε δύο περιπτώσεις σχετικά με τη μορφή της λύση των παραμέτρων του πυρήνα j. Η πρώτη περίπτωση αφορά τιμές παραμέτρων που αποτελούν συγχρόνως στάσιμα σημεία της κάθε λογαριθμικής πιθανοφάνειας L_k, δηλαδή ικανοποιούν την ακόλουθη εξίσωση για κάθε k:

$$\sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0. \tag{2.12}$$

Αυτή η κατηγορία λύσεων προκύπτει είτε σε περιπτώσεις όπου ο πυρήνας j τοποθετείται σε μια περιοχή δεδομένων όπου υπάρχει σημαντική επικάλυψη μεταξύ διαφορετικών κατηγοριών είτε σε περιπτώσεις όπου ο πυρήνας αναπαριστά δεδομένα μόνο μιας κατηγορίας. Η δεύτερη περίπτωση αναφέρεται σε τιμές παραμέτρων που ικανοποιούν την εξίσωση (2.10) χωρίς συγχρόνως να ικανοποιούν την (2.12) για κάθε k.

Βασιζόμενοι στις παραπάνω παρατηρήσεις όσον αφορά το μοντέλο των κοινών πυρήνων δίνουμε τον ακόλουθο ορισμό που ισχύει για κάθε μέθοδο εκτίμησης δεσμευμένων κατανομών των κατηγοριών χρησιμοποιώντας μικτές κατανομές.

Υποθέτουμε ένα μοντέλο μικτής κατανομής $p(x|C_k; \pi_k, \theta_k)$ για την εκτίμηση της δεσμευμένης κατανομής της κατηγορίας C_k, όπου με θ_k συμβολίζουμε τις παραμέτρους όλων των πυρήνων που χρησιμοποιούνται από τη μικτή κατανομή. Ένα στάσιμο σημείο $\hat{\Theta}$ της λογαριθμικής πιθανοφάνειας (που ορίζεται ομοίως με τη σχέση (2.5)) θα καλείται λύση *τύπου 1*, εάν για κάθε k το διάνυσμα παραμέτρων $\hat{\theta}_k$ είναι στάσιμο σημείο της λογαριθμικής πιθανοφάνειας L_k. Διαφορετικά το $\hat{\Theta}$ θα καλείται λύση *τύπου 2*.

Είναι φανερό ότι το μοντέλο των ανεξάρτητων μικτών κατανομών δίνει πάντα λύσεις τύπου 1, αφού η μεγιστοποίηση της λογαριθμικής πιθανοφάνειας L ανάγεται σε K ανεξάρτητα προβλήματα μεγιστοποίησης που το καθένα αντιστοιχεί στην λογαριθμική πιθανοφάνεια L_k.

Επιστρέφοντας στην ανάλυση των δύο παραδειγμάτων, παρατηρούμε ότι στο πρώτο παράδειγμα εφαρμόζοντας το μοντέλο των κοινών πυρήνων οδηγούμαστε σε μια λύση (Σχήμα 2.2α) που είναι προσεγγιστικά τύπου 1. Αντιθέτως στο δεύτερο παράδειγμα η λύση (Σχήμα 2.3α) δεν είναι τύπου 1 αφού ο πυρήνας στη δεξιά πλευρά του σχήματος έχει τοποθετηθεί σε μια περιοχή ασθενούς επικάλυψης μεταξύ των δεδομένων των δύο κατηγόριων. Αυτό έχει ως συνέπεια οι τιμές παραμέτρων του να μην ικανοποιούν την συνθήκη στάσιμου σημείου και για τις δύο λογαριθμικές συναρτήσεις πιθανοφάνειας L_1 και L_2. Επιπλέον, στο πρώτο παράδειγμα όπου και δυο μέθοδοι δίνουν λύσεις τύπου 1, η καλύτερη μέθοδος με κριτήριο την επίδοση γενίκευσης είναι το μοντέλο των κοινών πυρήνων και αυτό διότι δίνει μεγαλύτερη τιμή συνολικής λογαριθμικής πιθανοφάνειας. Βασιζόμενοι στις προηγούμενες παρατηρήσεις και υποθέτοντας πάντα σταθερό συνολικό αριθμό πυρήνων παραθέτουμε τα ακόλουθα διαισθητικά συμπεράσματα:

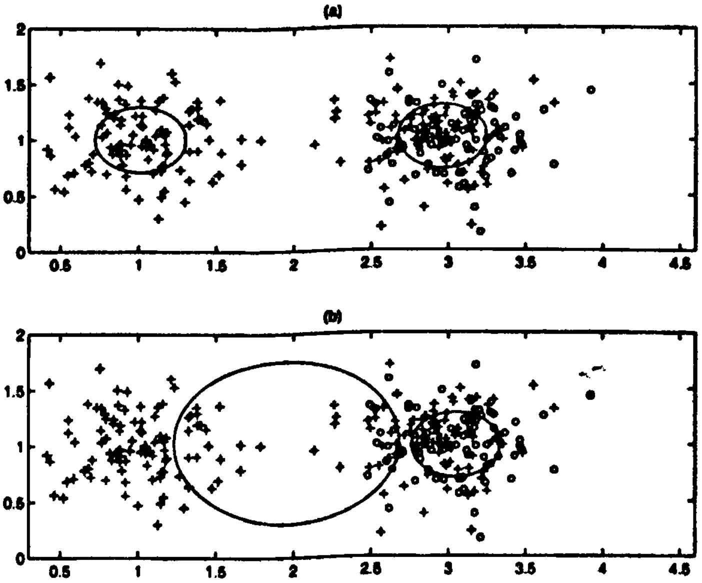
- Οι λύσεις τύπου 2 δεν συνιστούν αποτελεσματική τοποθέτηση των πυρήνων από τη σκοπιά της ταξινόμησης

- Από τις λύσεις τύπου 1 εκείνες που αναμένεται να έχουν καλύτερη επίδοση ταξινόμησης είναι αυτές με την μεγαλύτερη τιμή συνολικής πιθανοφάνειας

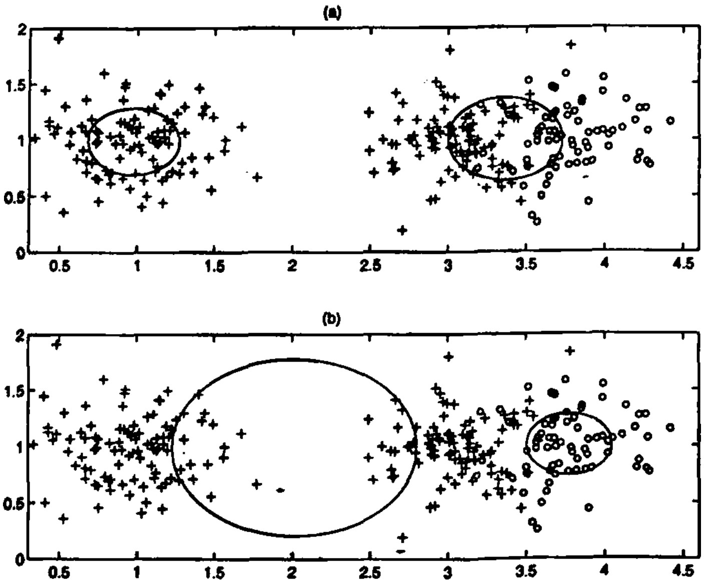
Τελικά αν ένα πρόβλημα περιέχει περιοχές δεδομένων με σημαντικό βαθμό επικάλυψης μεταξύ δεδομένων διαφορετικών κατηγοριών και συγχρόνως αντίστοιχες περιοχές με ασθενή επικάλυψη, τότε ούτε το μοντέλο των κοινών πυρήνων και επίσης ούτε οι ανεξάρτητες μιχτές κατανομές αναμένεται να δώσουν το αποτελεσματικότερο σύστημα ταξινόμησης (για σταθερό αριθμό πυρήνων). Ένα τέτοιο πρόβλημα απαιτεί μια πιο γενική μέθοδο εκτίμησης δεσμευμένων κατανομών βασισμένη σε μικτές κατανομές.

2.5 Το Z-μοντέλο

Το μοντέλο των κοινών πυρήνων (2.3) μπορεί να γενικευτεί έτσι ώστε ένα ορισμένο *υποσύνολο* του συνόλου των πυρήνων M να χρησιμοποιείται από κάθε δεσμευμένη κατανομή [31]. Προκειμένου να ορίσουμε μια τέτοια μοντελοποίηση εισάγουμε έναν $M \times K$ πίνακα Z που θα υποδεικνύει το πώς διανέμονται οι πυρήνες στις διάφορες κατηγορίες. Κάθε στοιχείο z_{jk} του πίνακα ορίζεται ως



Σχήμα 2.2: Παράδειγμα προβλήματος ταξινόμησης όπου το μοντέλο των κοινών πυρήνων έχει καλύτερη επίδοση γενίκευσης. Τα δεδομένα της κάθε κατηγορίας έχουν παραχθεί με βάση τις κατανομές $p(x|C_1) = 0.5N([1\ 1]^T, 0.08) + 0.5N([2.9\ 1]^T, 0.08)$ και $p(x|C_2) = N([3\ 1]^T, 0.08)$, ενώ οι εκ των προτέρων πιθανότητες των κατηγοριών ήταν $P(C_1) = 0.7$ και $P(C_2) = 0.3$, αντίστοιχα. Σημειωτέον ότι αφού οι πυρήνες είναι σφαιρικές κανονικές κατανομές χρησιμοποιούμε τον συμβολισμό $N(\mu, \sigma^2)$, όπου μ είναι το διάνυσμα του μέσου και σ^2 είναι η κοινή τιμή διακύμανσης όλων των συνιστωσών. Δημιουργήθηκαν δύο σύνολα δεδομένων, ένα για εκπαίδευση και ένα για έλεγχο. Το μοντέλο των κοινών πυρήνων (α) έδωσε σφάλμα γενίκευσης 27% και τιμή λογαριθμικής πιθανοφάνειας $L = -238.62$. Το αντίστοιχο σφάλμα και η τιμή της λογαριθμικής πιθανοφάνειας για τις ανεξάρτητες μικτές κατανομές (β) ήταν 32.2% και $L = -465.97$.



Σχήμα 2.3: Παράδειγμα προβλήματος ταξινόμησης όπου το μοντέλο των ανεξάρτητων μικτών κατανομών έχει καλύτερη επίδοση γενίκευσης. Τα δεδομένα της κάθε κατηγορίας έχουν παραχθεί με βάση τις κατανομές $p(x|C_1) = 0.5N([1\ 1]^T, 0.08) + 0.5N([3\ 1]^T, 0.08)$ και $p(x|C_2) = N([3.8\ 1]^T, 0.08)$, ενώ οι εκ των προτέρων πιθανότητες των κατηγοριών ήταν $P(C_1) = 0.7$ και $P(C_2) = 0.3$. Δημιουργήθηκαν δύο σύνολα δεδομένων, ένα για εκπαίδευση και ένα για έλεγχο. Το μοντέλο των κοινών πυρήνων (α) έδωσε σφάλμα γενίκευσης 26.1% και τιμή λογαριθμικής πιθανοφάνειας $L = -326.23$. Το αντίστοιχο σφάλμα και η τιμή της λογαριθμικής πιθανοφάνειας για τις ανεξάρτητες μικτές κατανομές (β) ήταν 7% και $L = -489.18$.

εξής:

$$z_{jk} = \begin{cases} 1 & \text{αν ο πυρήνας } j \text{ συνεισφέρει στην δεσμευμένη κατανομή της } C_k \\ 0 & \text{διαφορετικά} \end{cases} \tag{2.13}$$

Προκειμένου να αποκλειστούν περιπτώσεις όπου είτε κάποιοι πυρήνες δεν χρησιμοποιούνται από καμία δεσμευμένη κατανομή είτε υπάρχει ένα μοντέλο δεσμευμένης κατανομής που δεν χρησιμοποιεί κανένα πυρήνα, υποθέτουμε ότι κάθε γραμμή και στήλη ενός έγκυρου πίνακα Z περιέχει τουλάχιστον ένα στοιχείο με τιμή την μονάδα. Ένας τρόπος για να εισάγουμε τον περιορισμό z_{jk} στο μοντέλο που περιγράφεται από την (2.3) είναι να θέσουμε περιορισμούς στις εκ των προτέρων πιθανότητες π_{jk}. Συγκεκριμένα θέτουμε την τιμή της εκ των προτέρων πιθανότητα π_{jk} σταθερά ίση με μηδέν αν $z_{jk} = 0$. Σε μια τέτοια περίπτωση, μπορούμε να θεωρήσουμε ότι η δεσμευμένη κατανομή της κατηγορίας C_k εξακολουθεί να δίνεται από (2.3), αλλά με τη διαφορά ότι ο αρχικός χώρος παραμέτρων έχει περιοριστεί σε έναν υποχώρο που ορίζεται από τους περιορισμούς στις εκ των προτέρων πιθανότητες (τιμή του πίνακα Z), δηλαδή

$$p(x|C_k; z_k, \pi_k, \theta) = \sum_{j=1}^{M} \pi_{jk} p(x|j; \theta_j) = \sum_{j: z_{jk}=1} \pi_{jk} p(x|j; \theta_j), \tag{2.14}$$

όπου με z_k συμβολίζουμε την k-ιοστή στήλη του Z και $\{j : z_{jk} = 1\}$ το σύνολο όλων των πυρήνων j για τους οποίους $z_{jk} = 1$. Η σχέση (2.4) για τις εκ των προτέρων πιθανότητες είναι έγκυρη, ωστόσο αυτό που πραγματικά ισχύει είναι

$$\sum_{j: z_{jk}=1} \pi_{jk} = 1, \tag{2.15}$$

για κάθε k. Είναι προφανές ότι το μοντέλο των κοινών πυρήνων προκύπτει ως μια ειδική περίπτωση του Z-μοντέλου όπου $z_{jk} = 1$ για κάθε j, k. Επίσης οποιοδήποτε μοντέλο ανεξάρτητων μικτών κατανομών με συνολικό αριθμό πυρήνων M προκύπτει από το Z-μοντέλο αν επιλέξουμε τον πίνακα έτσι ώστε κάθε γραμμή του να περιέχει ακριβώς ένα στοιχείο με τιμή μονάδα.

Έστω τώρα ότι έχουμε ένα σύνολο δεδομένων εκπαίδευσης X. Αν υποθέσουμε ότι τα δεδομένα του υποσυνόλου X_k έχουν παραχθεί ανεξάρτητα μεταξύ τους με βάση την $p(x|C_k; z_k, \pi_k, \theta)$, η πιθανοφάνεια του συνόλου X ορίζεται από τη σχέση

$$P(X|\Theta) = \prod_{k=1}^{K} \prod_{x \in X_k} p(x|C_k; z_k, \pi_k, \theta). \tag{2.16}$$

Πρέπει να σημειωθεί ότι για την εφαρμογή του παραπάνω μοντέλου ο πίνακας Z καθορίζεται αρχικά και παραμένει σταθερός, με άλλα λόγια τα στοιχεία z_{jk} δεν αποτελούν προσαρμοζόμενες παραμέτρους. Σύμφωνα με την μέθοδο της μέγιστης πιθανοφάνειας επιθυμούμε να βρούμε εκείνες τις τιμές παραμέτρων που μεγιστοποιούν την παραπάνω συνάρτηση ή ισοδύναμα τον λογάριθμό της

$$L(\Theta) = \sum_{k=1}^{K} \sum_{x \in X_k} \log p(x|C_k; z_k, \pi_k, \theta). \tag{2.17}$$

Η εκπαίδευση του Z-μοντέλου μπορεί να εκτελεστεί χρησιμοποιώντας τον αλγόριθμο ΕΜ με τρόπο ακριβώς ανάλογο της εκπαίδευσης του μοντέλου των κοινών πυρήνων. Λεπτομέρειες της εφαρμογής του αλγορίθμου και εξισώσεις ενημέρωσης των παραμέτρων υπάρχουν στο Παράρτημα Α.

Ομοίως με την προηγούμενη ενότητα μπορεί να δειχθεί ότι σε κάποιο στάσιμο σημείο της πιθανοφάνειας, η αναγκαία συνθήκη για τις παραμέτρους θ_j ενός πυρήνα j είναι

$$\sum_{k: z_{jk}=1} \sum_{x \in X_k} P(j|x, C_k; z_k, \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0 \tag{2.18}$$

και για την εκ των προτέρων πιθανότητα π_{jk}

$$\hat{\pi}_{jk} = \frac{1}{|X_k|} \cdot \sum_{x \in X_k} P(j|x, C_k; z_k, \hat{\pi}_k, \hat{\theta}), \tag{2.19}$$

όπου (2.18) ισχύει για κάθε j, ενώ (2.19) για κάθε j και k τέτοια ώστε $z_{jk} = 1$. Επίσης η εκ των υστέρων πιθανότητα $P(j|x, C_k; z_k, \pi_k, \theta)$ δίνεται ομοίως με τη σχέση (2.9) από

$$P(j|x, C_k; z_k, \pi_k, \theta) = \frac{\pi_{jk} p(x|j; \theta_j)}{\sum_{i: z_{ik}=1} \pi_{ik} p(x|i; \theta_i)}. \tag{2.20}$$

Για τα στάσιμα σημεία της λογαριθμικής πιθανοφάνειας ενός Z-μοντέλου ισχύει η ακόλουθη πρόταση:

Πρόταση 1. *Κάθε στάσιμο σημείο $\hat{\Theta}$ της λογαριθμικής πιθανοφάνειας (2.17) του Z-μοντέλου (για αυθαίρετο Z) είναι επίσης στάσιμο σημείο της λογαριθμικής πιθανοφάνειας (2.5) που αντιστοιχεί στο μοντέλο των κοινών πυρήνων για τον ίδιο αριθμό πυρήνων.*

Απόδειξη: Μια τιμή του διανύσματος παραμέτρων αποτελεί στάσιμο σημείο της λογαριθμικής πιθανοφάνειας του μοντέλου των κοινών πυρήνων (2.5), αν η συνθήκη (2.10) ισχύει για κάθε $j = 1, \ldots, M$ καθώς επίσης η (2.11) ισχύει για κάθε $j = 1, \ldots, M$ και $k = 1, \ldots, K$.

Αφού $\hat{\Theta}$ είναι στάσιμο σημείο για το Z-μοντέλο, η εξίσωση (2.18) ισχύει για κάθε j. Τώρα, για όλα τα k : $z_{jk} = 1$ και $x \in X_k$ ισχύει $P(j|x, C_k; z_k, \hat{\pi}_k, \hat{\theta}) = P(j|x, C_k; \hat{\pi}_k, \hat{\theta})$ σύμφωνα με τις σχέσεις (2.9), (2.14) και (2.20), οπότε η (2.18) γράφεται ως

$$\sum_{k:z_{jk}=1} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0. \quad (2.21)$$

Επιπλέον, εφόσον για όλα τα k τέτοια ώστε $z_{jk} = 0$ ισχύει $\pi_{jk} = 0$, κάθε εκ των υστέρων πιθανότητα $P(j|x, C_k; \hat{\pi}_k, \hat{\theta})$ είναι μηδέν σύμφωνα με την (2.9). Επομένως ισχύει ότι $\sum_{k:z_{jk}=0} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0$, και ακολούθως η (2.21) μπορεί να γραφεί

$$\sum_{k:z_{jk}=1} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) + \quad (2.22)$$
$$\sum_{k:z_{jk}=0} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0,$$

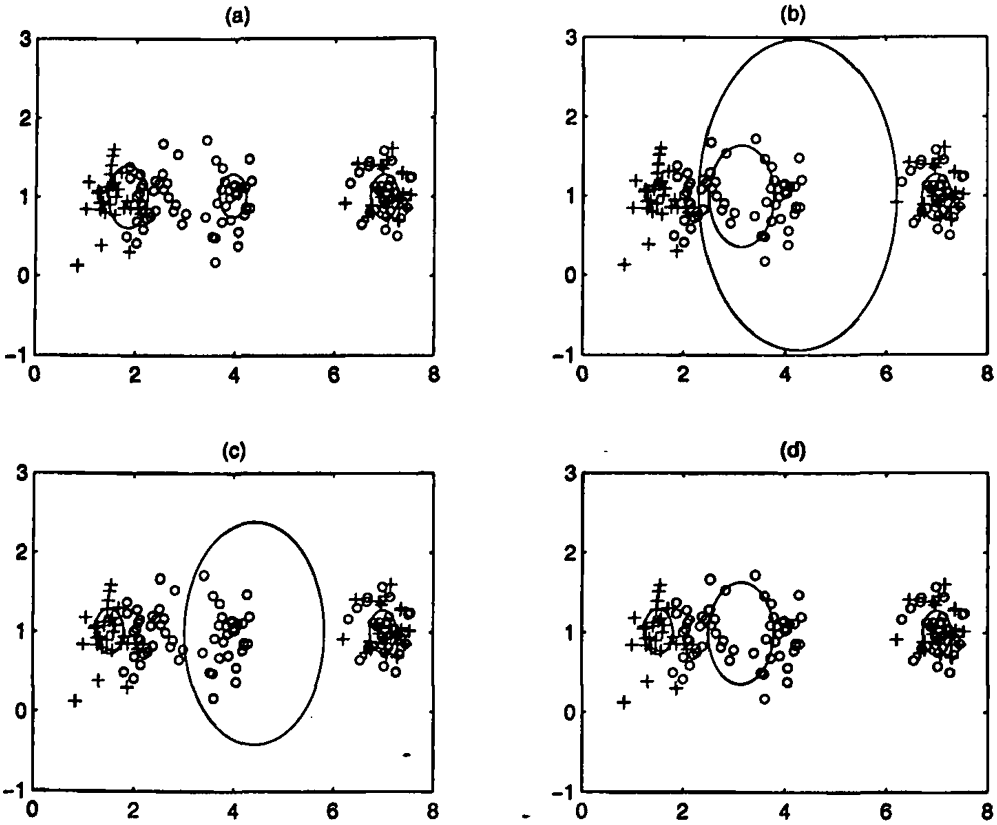
ή

$$\sum_{k=1}^{K} \sum_{x \in X_k} P(j|x, C_k; \hat{\pi}_k, \hat{\theta}) \nabla_{\theta_j} \log p(x|j; \hat{\theta}_j) = 0, \quad (2.23)$$

η οποία δεν είναι τίποτα άλλο παρά ή συνθήκη (2.10) και ισχύει για κάθε j. Σε ό,τι αφορά τις εκ των προτέρων πιθανότητες των πυρήνων, έχουμε ότι η (2.11) ισχύει για όλα τα $\hat{\pi}_{jk}$ με $z_{jk} = 1$. Προφανώς, η (2.11) ισχύει επίσης για κάθε εκ των προτέρων πιθανότητα π_{jk} για την οποία $z_{jk} = 0$ αφού σε μια τέτοια περίπτωση και οι δύο πλευρές της εξίσωσης είναι μηδέν (όπως δείχθηκε προηγουμένως όλες οι σχετιζόμενες εκ των υστέρων πιθανότητες $P(j|x, C_k; \hat{\pi}_k, \hat{\theta})$ είναι ίσες με μηδέν). ∎

Σύμφωνα με την ανάλυση που παρουσιάστηκε στην Ενότητα 2.2, οι περιπτώσεις Z-μοντέλων οι οποίες μπορούν να δώσουν συστήματα ταξινόμησης με καλές επιδόσεις ταξινόμησης είναι αυτές που προσεγγιστικά δίνουν λύσεις τύπου 1 με υψηλές τιμές λογαριθμικής πιθανοφάνειας. Παρακάτω, παρουσιάζουμε ένα παράδειγμα προβλήματος ταξινόμησης όπου ένα Z-μοντέλο έχει καλύτερη επίδοση ταξινόμησης συγκρινόμενο με το μοντέλο των κοινών πυρήνων και αυτό των ανεξάρτητων μικτών κατανομών (Σχήμα 2.4).

Το παράδειγμα προβλήματος ταξινόμησης αποτελεί συνδυασμό των παραδειγμάτων που παρουσιάστηκαν στην ενότητα 2.2. Τα δεδομένα της πρώτης κατηγορίας σχηματίζουν τρεις ομάδες δεδομένων, ενώ τα αντίστοιχα της δεύτερης κατηγορίας σχηματίζουν δύο ομάδες. Το πρόβλημα είναι κατασκευασμένο με



Σχήμα 2.4: Παράδειγμα προβλήματος ταξινόμησης όπου μια κατάλληλη επιλογή ενός Z-μοντέλου οδηγεί σε καλύτερη επίδοση γενίκευσης. Τα δεδομένα της κάθε κατηγορίας έχουν παραχθεί με βάση τις κατανομές $p(x|C_1) = 0.33N([2.3\ 1]^T, 0.08) + 0.33N([4\ 1]^T, 0.08) + 0.33N([7\ 1]^T, 0.08)$ και $p(x|C_2) = 0.5N([1.5\ 1]^T, 0.08) + 0.5N([7\ 1]^T, 0.08)$, ενώ οι εκ των προτέρων πιθανότητες των κατηγοριών ήταν $P(C_1) = P(C_2) = 0.5$. Δημιουργήθηκαν δύο σύνολα δεδομένων, ένα για εκπαίδευση και ένα για έλεγχο, ενώ σε κάθε περίπτωση βρέθηκε ο εκτιμητής μέγιστης πιθανοφάνειας. Το σφάλμα γενίκευσης e και η τελική τιμή της λογαριθμικής πιθανοφάνειας L για κάθε μέθοδο ξεχωριστά είναι: α) Μοντέλο κοινών πυρήνων: $e = 33.33\%$ και $L = -1754.51$ β) Ανεξάρτητες μικτές κατανομές (δύο πυρήνες για την C_1 και ένας για τη C_2): $e = 24.33\%$ $L = -2683.25$, γ) Ανεξάρτητες μικτές κατανομές (ένας πυρήνας για την C_1 και δύο πυρήνες για την C_2): $e = 34\%$ και $L = -3748.42$ και δ) Ένας πυρήνας κοινός και για τις δύο δεσμευμένες κατανομές, ενώ οι άλλοι δύο συνεισφέρουν ο καθένας σε μια κατηγορία: $e = 21.67\%$ και $L = -1822.53$.

τέτοιο τρόπο ώστε υπάρχει ένα ζεύγος από ομάδες διαφορετικών κατηγοριών οι οποίες έχουν σημαντική επικάλυψη, ενώ για ένα διαφορετικό ζεύγος ομάδων υπάρχει ασθενή επικάλυψη. Υποθέτουμε ότι ο συνολικός αριθμός πυρήνων είναι τρία.

Η λύση που βρήκαμε εφαρμόζοντας το μοντέλο των κοινών πυρήνων (Σχήμα 2.4α) αν και δίνει μια ικανοποιητική αναπαράσταση της περιοχής όπου συμβαίνει σημαντική επικάλυψη, στη περίπτωση της περιοχής με ασθενή επικάλυψη τοποθετεί τον πυρήνα πάνω στο όριο απόφασης πράγμα που αυξάνει δραματικά το σφάλμα γενίκευσης. Όσον αφορά το μοντέλο των ανεξάρτητων μικτών κατανομών υπάρχουν δύο επιλογές: i) χρησιμοποίηση δύο πυρήνων για την δεσμευμένη κατανομή της πρώτης κατηγορίας και ενός πυρήνα για την αντίστοιχη κατανομή της δεύτερης κατηγορίας (Σχήμα 2.4β) και ii) χρησιμοποίηση ενός πυρήνα για την πρώτη κατηγορία και δύο για την δεύτερη (Σχήμα 2.4γ). Το μοντέλο με την καλύτερη επίδοση ταξινόμησης δίνεται από ένα Z-μοντέλο με ένα μόνο κοινό πυρήνα και με τους υπόλοιπους δύο να συνεισφέρουν ο καθένας σε μόνο μια κατηγορία (Σχήμα 2.4δ).

2.6 Συζήτηση

Το Z-μοντέλο αποτελεί μια γενίκευση των προηγούμενων τεχνικών μοντελοποίησης δεσμευμένων κατανομών των κατηγοριών χρησιμοποιώντας μικτές κατανομές, δηλαδή του μοντέλου των κοινών πυρήνων και των ανεξάρτητων μικτών κατανομών. Ως γενίκευση μπορεί να συνδυάζει με ευέλικτο τρόπο τα χαρακτηριστικά των προγενέστερων μεθόδων.

Συγκεκριμένα το Z-μοντέλο μπορεί να διαθέτει κοινούς πυρήνες για ορισμένες κατηγορίες καθώς και πυρήνες που χρησιμοποιούνται αποκλειστικά μόνο από μια κατηγορία. Ένας κοινός πυρήνας έχει την δυνατότητα να αναπαριστά συγχρόνως δεδομένα διαφορετικών κατηγοριών. Κάτι τέτοιο είναι ιδιαίτερα σημαντικό διότι έτσι μειώνεται ο αριθμός των απαιτούμενων πυρήνων πράγμα που με την σειρά του αποτελεί υπολογιστικό πλεονεκτήματα δεδομένου ότι διαθέτουμε πάντα ένα πεπερασμένο σύνολο δεδομένων εκπαίδευσης. Για να γίνει πιο εμφανές το τελευταίο ας θεωρήσουμε την εξής περίπτωση. Έστω ότι σε ένα πρόβλημα ταξινόμησης υπάρχει μια περιοχή που χώρου δεδομένων όπου οι άγνωστες δεσμευμένες κατανομές των κατηγοριών είναι τοπικά πανομοιότυπες (π.χ. τα δεδομένα παράγονται από την ίδια ομάδα). Τότε αν χρησιμοποιήσουμε για την συγκεκριμένη τοπική αναπαράσταση των δεδομένων πυρήνες που ανήκουν μόνο σε μια κατηγορία, ενδεχομένως να μη έχουμε καλή τοπική εκτίμηση ιδιαίτερα αν από κάθε κατηγορία διαθέτουμε λίγα δεδομένα εκπαίδευσης. Αντιθέτως αν χρησιμοποιούσαμε ένα κοινό πυρήνα, τότε η εκτίμηση θα είναι πιο αξιόπιστη αφού χρησιμοποιείται αυξημένος αριθμός δεδομένων (όλων των κατηγοριών). Γενικότερα σε προβλήματα ταξινόμησης όπου από κάθε κατηγορία διαθέτουμε μικρό αριθμό δεδομένων εκπαίδευσης (< 100), ενώ πιθανόν έχουμε πολλές κατηγορίες (> 10) η χρήση κοινών πυρήνων είναι πολύ σημαντική αφού θα έκανε περισσότερο αξιόπιστη την εκτίμηση των δεσμευμένων κατανομών. Ωστόσο η χρήση κοινών πυρήνων (σε αντίθεση με την χρήση μη κοινών πυρήνων) δεν δίνει καλές αναπαραστάσεις σε περιοχές δεδομένων ασθενούς επικάλυψης οι οποίες αποτελούν περιοχές κοντά στα όρια απόφασης και κατά συνέπεια είναι κρίσιμες για την επίδοση του συστήματος ταξινόμησης.

Από τα παραπάνω είναι προφανές ότι η χρησιμότητα του Z-μοντέλου σχετίζεται με την επιλογή του πίνακα Z, δηλαδή των υποθέσεων σχετικά με την χρήση των πυρήνων από τις διάφορες κατηγορίες. Κάτι τέτοιο προφανώς εξαρτάται κάθε φορά από τα δεδομένα. Στο επόμενο κεφάλαιο παρουσιάζουμε μια μέθοδο επιλογής του πίνακα Z.

Κεφάλαιο 3

Μέθοδοι εκπαίδευσης του Z-μοντέλου

3.1 Το πρόβλημα καθορισμού του πίνακα Z

Στο προηγούμενο κεφάλαιο δεν έγινε ιδιαίτερη αναφορά για το πώς καθορίζονται οι τιμές του πίνακα Z κατά την εφαρμογή του Z-μοντέλου σε ένα δεδομένο πρόβλημα ταξινόμησης.

Προκειμένου να εφαρμόσουμε το Z-μοντέλο πρέπει να καθοριστεί αρχικά η τιμή του πίνακα Z. Σύμφωνα με την ανάλυση του προηγούμενου κεφαλαίου η επιλογή του πίνακα Z οφείλει να ανταποκρίνεται στη γεωμετρία του συνόλου δεδομένων, δηλαδή στο τύπο της επικάλυψης που συμβαίνει στις διάφορες περιοχές του χώρου δεδομένων. Ωστόσο μια τέτοια πληροφορία είναι εκ των προτέρων άγνωστη και πρέπει να εξαχθεί από τα δεδομένα κατά την διάρκεια της μάθησης. Μια πρώτη προσέγγιση θα ήταν να διερευνήσουμε όλο τον χώρο τιμών του πίνακα Z, έπειτα να εκπαιδεύσουμε το κάθε μοντέλο χρησιμοποιώντας τον αλγόριθμο EM και τελικά να ταξινομήσουμε τα μοντέλα με βάση μια εκτίμηση της γενικευτικής τους ικανότητας (μετρούμενη, π.χ. χρησιμοποιώντας μεθόδους διασταυρωμένης επικύρωσης (cross-validation)). Ωστόσο απαιτείται υπολογιστικά σημαντικός χρόνος για την εξερεύνηση όλων των δυνατών Z-μοντέλων, ιδιαίτερα αν τα M και K λαμβάνουν μεγάλες τιμές. Επιπλέον η προσέγγιση αυτή πάσχει και από ένα σοβαρό πρόβλημα πρακτικής φύσεως που αφορά στην αρχικοποίηση των παραμέτρων των πυρήνων: εάν ένας πυρήνας έχει καθοριστεί (από τις τιμές του Z) να είναι κοινός για κάποια μοντέλα δεσμευμένων κατανομών, τότε είναι λογικό, στην αρχικοποίηση των παραμέτρων κατά την εφαρμογή του αλγορίθμου EM ο πυρήνας να τοποθετείται σε μια περιοχή δεδομένων όπου συμβαίνει επικάλυψη μεταξύ δεδομένων των αντίστοιχων κατηγοριών. Κάτι τέτοιο, ωστόσο δεν είναι καθόλου εύκολο να επιτευχθεί, αφού δεν διαθέτουμε καμιά πληροφορία για την επικάλυψη των δεδομένων μεταξύ διαφορετικών κατηγοριών.

Προκειμένου να αντιμετωπίσουμε τα παραπάνω προβλήματα εφαρμογής του Z-μοντέλου έχουμε αναπτύξει μια υπολογιστικά αποδοτική μέθοδο [31] η οποία βασίζεται στη εισαγωγή μιας νέας αντικειμενικής συνάρτησης. Με την μέθοδο αντιμετωπίζεται συγχρόνως η επιλογή της τιμής του πίνακα Z καθώς και η αρχικοποίηση των παραμέτρων των πυρήνων.

3.2 Μια μέθοδος επιλογής ενός Z-μοντέλου

Η βασική ιδέα της μεθόδου είναι ότι προσαρμόζουμε κατά την διάρκεια της βελτιστοποίησης όχι μόνο παραμέτρους των δεσμευμένων κατανομών αλλά και παραμέτρους περιορισμού (που καθορίζουν την συνεισφορά του κάθε πυρήνα στη εκτίμηση μιας δεσμευμένης κατανομής).

Ας ξεκινήσουμε την περιγραφή της μεθόδου παρατηρώντας το εξής: Οι τιμές των στοιχείων του πίνακα Z περιορίζουν το διάνυσμα παραμέτρων Θ σε ένα υποχώρο του ευρύτερου χώρου παραμέτρων που αντιστοιχεί στο μοντέλο των κοινών πυρήνων ($z_{jk} = 1$ για κάθε j, k). Για το λόγο αυτό, αρχικά μπορούμε να υποθέσουμε ότι οι δεσμευμένες κατανομές ακολουθούν το πιο γενικό μοντέλο των κοινών πυρήνων. Στη συνέχεια κατά την διαδικασία μάθησης μπορούμε να περιορίζουμε τον χώρο παραμέτρων ανάλογα με τα χαρακτηριστικά του συνόλου δεδομένων εκπαίδευσης. Βέβαια ο αντικειμενικός σκοπός είναι να καταλήξουμε σε ένα αποδοτικό σύστημα ταξινόμησης για το λόγο αυτό και η αναζήτηση του τελικού υποχώρου παραμέτρων οφείλει να συμβαίνει προς την κατεύθυνση αυτή. Όπως εξηγήθηκε στο προηγούμενο κεφάλαιο, η βασική 'οδηγία' για την αναζήτηση ενός αποδοτικού συστήματος ταξινόμησης είναι ότι ένας πυρήνας δεν πρέπει να αναπαριστά δεδομένα διαφορετικών κατηγοριών με εξαίρεση τις περιπτώσεις που υπάρχει τοπικά σημαντική επικάλυψη. Η μέθοδος που παρουσιάζουμε έχει την ιδιότητα ότι να κάνει ανταγωνιστική την μάθηση των δεσμευμένων κατανομών ως προς την δέσμευση πυρήνων. Κάτι τέτοιο μπορεί να επιλύσει το πρόβλημα του μοντέλου των κοινών πυρήνων (και γενικότερα της χρήσης των κοινών πυρήνων) σχετικά με την περίπτωση της ασθενούς επικάλυψης μεταξύ δεδομένων διαφορετικών κατηγοριών.

Για να ορίσουμε μια μέθοδο εκπαίδευσης που προσαρμόζει ταυτόχρονα με τις παραμέτρους Θ και τους περιορισμούς χρήσης των πυρήνων από τις κατηγορίες, ορίζουμε τις παραμέτρους περιορισμού r_{jk} όπου $0 \leq r_{jk} \leq 1$ και για κάθε j

ικανοποιούν:

$$\sum_{k=1}^{K} r_{jk} = 1. \tag{3.1}$$

Ο ρόλος της κάθε παραμέτρου r_{jk} είναι ανάλογος της μεταβλητής z_{jk}: καθορίζουν τον βαθμό που επιτρέπεται ο πυρήνας j να χρησιμοποιηθεί από το μοντέλο της δεσμευμένης κατανομής της κατηγορίας C_k (να αναπαριστά δεδομένα της κατηγορίας αυτής).

Οι παράμετροι r_{jk} χρησιμοποιούνται προκειμένου να οριστούν οι ακόλουθες συναρτήσεις που είναι όμοιες με τις συναρτήσεις των δεσμευμένων κατανομών[1]:

$$\varphi(x; C_k, r_k, \pi_k, \theta) = \sum_{j=1}^{M} r_{jk}\pi_{jk}p(x|j;\theta_j) \quad k = 1, \dots, K. \tag{3.2}$$

Η εξίσωση (3.2) αποτελεί μια επέκταση της (2.3) όπου έχουν εισαχθεί ειδικοί παράμετροι περιορισμού μέσα στο γραμμικό άθροισμα. Όπως θα γίνει φανερό στη συνέχεια, οι παράμετροι r_{jk} με $k = 1, \dots, K$ εκφράζουν τον ανταγωνισμό μεταξύ των κατηγοριών ως προς την δέσμευση του πυρήνα j.

Αν για μια παράμετρο περιορισμού ισχύει $r_{jk} = 0$, τότε εξ ορισμού η αντίστοιχη εκ των προτέρων πιθανότητα είναι $\pi_{jk} = 0$. Επομένως, οι τιμές των εκ των προτέρων πιθανοτήτων ικανοποιούν

$$\sum_{j:r_{jk}=1} \pi_{jk} = 1, \tag{3.3}$$

για κάθε k. Σημειωτέον ότι για κάθε κατηγορία C_k πρέπει να υπάρχει τουλάχιστον ένα $r_{jk} > 0$. Έτσι, αποκλείουμε την περίπτωση όπου μια συνάρτηση φ, άρα και η αντίστοιχη δεσμευμένη κατανομή, είναι ίση με μηδέν.

Γενικά οι συναρτήσεις φ δεν αποτελούν συναρτήσεις πυκνότητας πιθανότητας ως προς x λόγω του γεγονότος ότι $\int \varphi(x; C_k, r_k, \pi_k, \theta)dx \leq 1$. Εξαίρεση αποτελεί η περίπτωση όπου οι περιορισμοί r_{jk} παίρνουν τιμές μηδέν ή ένα[2]. Μάλιστα τότε οι περιορισμοί r_{jk} είναι ισοδύναμοι με τους περιορισμούς z_{jk}. Ωστόσο, γενικά ισχύει ότι $\varphi(x; C_k, r_k, \pi_k, \theta) \geq 0$ και $\int \varphi(x; C_k, r_k, \pi_k, \theta)dx > 0$ που οφείλεται στην (3.3).

Προκειμένου οι συναρτήσεις φ να χρησιμοποιηθούν για την προσαρμογή όλων των παραμέτρων (Θ, r) (όπου r είναι το διάνυσμα όλων των r_{jk}) είναι απαραίτητο

[1]Οι συναρτήσεις δεν αποτελούν τα μοντέλα των δεσμευμένων κατανομών που θέλουμε να εκπαιδεύσουμε, αλλά χρησιμοποιούνται προκειμένου να βρούμε κατάλληλες τιμές παραμέτρων για τις πραγματικές δεσμευμένες κατανομές που περιγράφονται πάντα από την σχέση (2.3).

[2]Σε αυτήν την ειδική περίπτωση κάθε συνάρτηση $\varphi(x; C_k, r_k, \pi_k, \theta)$ ταυτίζεται με την αντίστοιχη $p(x|C_k; \pi_k, \theta)$.

να θεωρήσουμε τις συναρτήσεις αυτές ως δεσμευμένες κατανομές. Με βάση αυτή την λογική, εισάγουμε μια αντικειμενική συνάρτηση ανάλογη της λογαριθμικής πιθανοφάνειας ως ακολούθως:

$$L(\Theta, r) = \sum_{k=1}^{K} \sum_{x \in X_k} \log \varphi(x; C_k, r_k, \pi_k, \theta). \tag{3.4}$$

Μέσω της μεγιστοποίησης της παραπάνω ποσότητας προσαρμόζουμε τις τιμές των μεταβλητών r_{jk} (που ισοδυναμούν με τη προσαρμογή του βαθμού που διαμοιράζονται οι πυρήνες στις δεσμευμένες κατανομές) και αυτό αυτόματα επηρεάζει την λύση των παραμέτρων των μοντέλων των δεσμευμένων κατανομών Θ.

Ο αλγόριθμος [8] μπορεί να χρησιμοποιηθεί για την μεγιστοποίηση της αντικειμενικής συνάρτησης (3.4). Παρότι ο αλγόριθμος EM χρησιμοποιείται ως επί των πλείστων για τη μεγιστοποίηση της λογαριθμικής πιθανοφάνειας ή και για λογαριθμικής εκ των υστέρων κατανομής, το γεγονός ότι η συνάρτηση μας γενικά δεν αποτελεί καμιά από τις δύο αυτές περιπτώσεις δεν συνιστά πρόβλημα. Στο Παράρτημα Γ αποδεικνύεται ότι η βασική ιδιότητα του EM όσον αφορά την εγγυημένη μονότονη αύξηση της αντικειμενικής συνάρτησης ισχύει και στη περίπτωση της $L(\Theta, r)$.

Η συνάρτηση Q που υπολογίζεται στο E-βήμα της $t+1$ επανάληψης του EM είναι:

$$Q(\Theta, r; \Theta^{(t)}, r^{(t)}) = \sum_{k=1}^{K} \sum_{x \in X_k} \sum_{j=1}^{M} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)}) \log\{r_{jk}\pi_{jk} p(x|j;\theta_j)\}, \tag{3.5}$$

όπου

$$\Phi_j(x; C_k, r_k, \pi_k, \theta) = \frac{r_{jk}\pi_{jk} p(x|j;\theta_j)}{\sum_{i=1}^{M} r_{ik}\pi_{ik} p(x|i;\theta_i)}. \tag{3.6}$$

Ο αλγόριθμος στο M-βήμα μεγιστοποιεί την προηγούμενη ποσότητα ως προς το διάνυσμα παραμέτρων (Θ, r). Στο Παράρτημα Β παρουσιάζουμε πώς εξάγεται η συνάρτηση Q (3.5) καθώς επίσης τις εξισώσεις ενημέρωσης των παραμέτρων για την περίπτωση πυρήνων που ακολουθούν κανονική κατανομή.

Σε αυτό το σημείο θα ήταν χρήσιμο να γράψουμε τις εξισώσεις ενημέρωσης (δίνονται στο Παράρτημα Β) για τις εκ των προτέρων πιθανότητες π_{jk} και τις παραμέτρους περιορισμού r_{jk} προκειμένου να γίνει σαφές ο τρόπος που δουλεύει ο αλγόριθμος:

$$\pi_{jk}^{(t+1)} = \frac{1}{|X_k|} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)}), \tag{3.7}$$

για κάθε πιθανότητα π_{jk} και

$$r_{jk}^{(t+1)} = \frac{\sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)})}{\sum_{i=1}^{K} \sum_{x \in X_i} \Phi_j(x; C_i, r_i^{(t)}, \pi_i^{(t)}, \theta^{(t)})}, \tag{3.8}$$

για κάθε παράμετρο r_{jk}. Με βάση την εξίσωση (3.7), η (3.8) μπορεί να γραφεί ως

$$r_{jk}^{(t+1)} = \frac{\pi_{jk}^{(t+1)}|X_k|}{\sum_{i=1}^{K} \pi_{ji}^{(t+1)}|X_i|}. \tag{3.9}$$

Από την παραπάνω εξίσωση φαίνεται η εξάρτηση των παραμέτρων r_{jk} από τις νέες τιμές των εκ των προτέρων πιθανοτήτων π_{jk} κατά την εφαρμογή μιας επανάληψης του ΕΜ. Αν υποθέσουμε ότι οι κατηγορίες έχουν περίπου τον ίδιο αριθμό δεδομένων, τότε κατά την διάρκεια της μάθησης η χρήση του πυρήνα j από την κατηγορία C_k εξαρτάται από την τιμή της πιθανότητας π_{jk} καθώς και το άθροισμα όλων των αντίστοιχων πιθανοτήτων που σχετίζονται με τον ίδιο πυρήνα (π_{jk}, $k = 1, \ldots, K$). Με αυτόν τον τρόπο, όσο περισσότερο ο πυρήνας j συνεισφέρει στη δεσμευμένη κατανομή της κατηγορίας C_k, δηλαδή η εκ των προτέρων πιθανότητα π_{jk} αυξάνει, ανάλογα επηρεάζεται και η νέα τιμή r_{jk}. Καθώς αυξάνεται η τιμή μιας παραμέτρου r_{jk} σε μια επανάληψη, στη επόμενη επανάληψη η πιθανότητα π_{jk} γίνεται ακόμη μεγαλύτερη (λόγω της (3.7) και (3.6)) κοκ. Έτσι ερμηνεύεται ο ανταγωνισμός μεταξύ των κατηγοριών σχετικά με την δέσμευση πυρήνων ο οποίος επιτυγχάνεται μέσω των παραμέτρων περιορισμού r_{jk}. Σύμφωνα με τον ανταγωνισμό αυτό είναι λιγότερο πιθανό ένας πυρήνας να τοποθετηθεί πάνω σε κάποιο όριο απόφασης, για το λόγο ότι σε μια τέτοια περίπτωση η κατηγορία με τα περισσότερα δεδομένα σ' αυτήν την περιοχή θα τραβήξει τον πυρήνα προς το μέρος των δεδομένων που ανήκουν σ' αυτήν. Από την άλλη, η μέθοδος δεν φαίνεται να επηρεάζει σημαντικά το πλεονέκτημα της χρήσης κοινών πυρήνων στην περίπτωση αναπαράστασης δεδομένων διαφορετικών κατηγοριών με σημαντικό βαθμό επικάλυψης. Κάτι τέτοιο ερμηνεύεται από την εξίσωση (3.6). Αν ένας πυρήνας j αναπαριστά μια περιοχή σημαντικού βαθμού επικάλυψης, η κατανομή $p(x|j;\theta_j)$ θα δίνει υψηλές τιμές για δεδομένα όλων των κατηγοριών που εμπλέκονται. Επομένως, παρά το γεγονός ότι οι παράμετροι περιορισμού μπορεί να είναι μεγαλύτερες για κάποιες κατηγορίες, η τιμή της Φ_j (3.6) θα παραμένει υψηλή για τα δεδομένα όλων των κατηγοριών που εμπλέκονται.

Έστω ότι ο αλγόριθμος ΕΜ συγκλίνει σε ένα τοπικό βέλτιστο σημείο παραμέτρων (Θ^*, r^*). Τότε οι τιμές των παραμέτρων περιορισμού r_{jk}^* χρησιμοποιούνται για επιλογή Z-μοντέλου, δηλαδή τον καθορισμό των τιμών z_{jk}^*. Μια εύλογη επιλογή είναι η ακόλουθη

$$z_{jk}^* = \begin{cases} 1 & \varepsilon\alpha\nu \ r_{jk}^* > 0 \\ 0 & \varepsilon\alpha\nu \ r_{jk}^* = 0 \end{cases} \tag{3.10}$$

Ο καθορισμός του πίνακα Z^* βασίζεται στο επιχείρημα ότι αν $r_{jk}^* > 0$, ο πυρήνας j συμβάλλει στην εκτίμηση της δεσμευμένης κατανομής της κατηγορίας C_k (αφού $\pi_{jk}^* > 0$) και, επομένως, ο j πρέπει να συμπεριληφθεί στη μικτή κατανομή για την αναπαράσταση των δεδομένων της C_k. Το αντίθετο ισχύει όταν $r_{jk}^* = 0$. Εφόσον καθοριστεί το Z^*-μοντέλο, εφαρμόζεται ο αλγόριθμος EM ως την σύγκλιση εκκινώντας από την τιμή παραμέτρων Θ^*. Οι τελικές τιμές παραμέτρων Θ_f χρησιμοποιούνται για τον υπολογισμό των δεσμευμένων κατανομών από την σχέση (2.14).

Η παραπάνω μέθοδος εφαρμόστηκε στο πρόβλημα της ενότητας 2.5 (Σχήμα 2.4). Η λύση Θ_f ήταν ακριβώς αυτή που απεικονίζεται στο Σχήμα 2.4δ, όπου ο πίνακας Z είχε επιλεγεί κατάλληλα ώστε να ταιριάζει στη γεωμετρία του προβλήματος. Είναι αξιοσημείωτο, ότι η διαφορά (με βάση την ℓ_1 νόρμα) στις τιμές παραμέτρων Θ_f και Θ^* ήταν 0.03 και αφορούσε μόνο τις τιμές των εκ των προτέρων πιθανοτήτων του πυρήνα που αναπαριστά την περιοχή με σημαντική επικάλυψη[3].

3.3 Το μοντέλο λPRBF

Η μέθοδος επιλογής ενός Z-μοντέλου που περιγράφτηκε παραπάνω θα μπορούσε να χρησιμοποιηθεί εναλλακτικά θεωρώντας τις παραμέτρους περιορισμού r σταθερές και καθορισμένες πριν από την εφαρμογή του αλγορίθμου. Η εφαρμογή της μεθόδου με τέτοιο τρόπο μπορεί να έχει αρκετές ομοιότητες με την εφαρμογή του Z-μοντέλου, όπου ο πίνακας Z έχει καθοριστεί εξ αρχής και παραμένει σταθερός κατά την εκπαίδευση. Μάλιστα υπάρχουν περιπτώσεις επιλογής των μεταβλητών r όπου η μέθοδος εκπαίδευσης είναι ισοδύναμη με την εκπαίδευση ενός Z-μοντέλου. Για παράδειγμα οποιαδήποτε εκπαίδευση ανεξάρτητων μικτών κατανομών με συνολικό αριθμό πυρήνων M μπορεί ισοδύναμα να πραγματοποιηθεί μέσω της τεχνικής εκπαίδευσης του r-μοντέλου επιλέγοντας κατάλληλα τις

[3]Έχουμε παρατηρήσει ότι σε πολλά προβλήματα ταξινόμησης η τιμή Θ_f είναι πολύ κοντά στη αντίστοιχη Θ^*, πράγμα που σημαίνει ότι μεγιστοποιώντας την συνάρτηση (3.4) λαμβάνουμε μια λύση που είναι κοντά σε ένα τοπικό μέγιστο της (2.5). Μια ικανή συνθήκη προκειμένου να ισχύει $\Theta_f = \Theta^*$ είναι ότι οι παράμετροι r_{jk}^* πρέπει να λαμβάνουν τιμές μηδέν ή ένα. Σε μια τέτοια περίπτωση η λύση του Z-μοντέλου που παίρνουμε εφαρμόζοντας τη μέθοδο αντιστοιχεί σε ανεξάρτητες μικτές κατανομές όπου οι πυρήνες έχουν διανεμηθεί δυναμικά στις διάφορες κατηγορίες.

τιμές των μεταβλητών r_{jk}, που στη περίπτωση αυτή θα παίρνουν τιμές μηδέν ή ένα.

Παρακάτω περιγράφουμε μια ενδιαφέρουσα τεχνική εκπαίδευσης που προκύπτει θέτοντας σταθερές τις τιμές των παραμέτρων περιορισμού. Η τεχνική περιγράφεται στο [30] όπου καλείται με το όνομα μοντέλο λPRBF. Έστω ότι οι M διαθέσιμοι πυρήνες διαχωρίζονται σε K ανεξάρτητα σύνολα T_k, $k = 1, \ldots, K$, με το κάθε υποσύνολο πυρήνων T_k να αντιστοιχεί στην κατηγορία C_k και $|T_1| + \ldots + |T_K| = M$. Εισάγουμε την παράμετρο $\lambda \in [0, 1]$ που ο ρόλος της είναι να ρυθμίζει τον βαθμό χρήσης των ομαδοποιημένων πυρήνων στις διάφορες κατηγορίες. Ειδικότερα υποθέτουμε ότι η δεσμευμένη κατανομή της κατηγορίας C_k χρησιμοποιεί πλήρως τους πυρήνες της ομάδας T_k, ενώ οι υπόλοιποι πυρήνες χρησιμοποιούνται σε κάποιο βαθμό καθοριζόμενο από την τιμή της παραμέτρου λ. Η παράμετρος λ παίρνει μια σταθερή τιμή αρχικά και παραμένει αμετάβλητη κατά την διάρκεια της εκπαίδευσης.

Οι παραπάνω απαιτήσεις για το τρόπο περιορισμού των πυρήνων μπορεί να εκφραστούν μέσω ενός r-μοντέλου που προκύπτει αν θεωρήσουμε ότι οι τιμές των παραμέτρων r_{jk} δίνονται ως ακολούθως:

$$r_{jk} = \begin{cases} \frac{1}{1+\lambda(K-1)} & j \in T_k \\ \frac{\lambda}{1+\lambda(K-1)} & j \notin T_k \end{cases} \tag{3.11}$$

όπου η έκφραση $j \notin T_k$ αναφέρεται σε όλους τους πυρήνες του συνόλου $\bigcup_{k' \neq k} T_{k'}$. Σημειωτέον ότι με βάση την ανάθεση τιμών των παραμέτρων r_{jk} ικανοποιείται ο περιορισμός (3.1). Τώρα η αντικειμενική συνάρτηση (3.4) παίρνει την μορφή

$$L(\Theta; \lambda) = \sum_{k=1}^{K} \sum_{x \in X_k} \log \varphi(x; C_k, \lambda, \pi_k, \theta), \tag{3.12}$$

όπου

$$\varphi(x|C_k, \lambda, \pi_k, \theta) = \sum_{j \in T_k} \pi_{jk} p(x|j; \theta_j) + \lambda \sum_{j \notin T_k} \pi_{jk} p(x|j; \theta_j). \tag{3.13}$$

και όπου στη σχέση (3.12) έχουμε αφαιρέσει ένα σταθερό όρο που δεν έχει περιέχει προσαρμοζόμενους παραμέτρους. Η παραπάνω εξίσωση (3.13) περιγράφει το μοντέλο λPRBF όπως αυτό ορίστηκε στο [30], ενώ η συνάρτηση (3.12) μεγιστοποιείται χρησιμοποιώντας το αλγόριθμο EM.

Αξίζει να σημειωθεί ότι όταν $\lambda = 0$, η (3.12) είναι μια λογαριθμική πιθανοφάνεια που αντιστοιχεί σε μια περίπτωση ανεξάρτητων μικτών κατανομών (η μικτή κατανομή της κατηγορίας C_k χρησιμοποιεί μόνο τους πυρήνες του συνόλου T_k).

Εάν $\lambda = 1$, η (3.12) ισούται με την (2.5), που είναι η λογαριθμική πιθανοφάνεια του μοντέλου των κοινών πυρήνων.

Στο Παράρτημα Δ περιγράφεται αναλυτικά ο αλγόριθμος EM για το λPRBF μοντέλο.

3.3.1 Μέσος όρος ως προς λ

Ωστόσο στο μοντέλο λPRBF η τιμή της παραμέτρου λ καθορίζεται εξ αρχής και δεν είναι προφανής κάποιος τρόπος εύρεσης μιας βέλτιστης τιμής. Προκειμένου να αντιμετωπίσουμε το πρόβλημα αυτό εφαρμόζουμε την μέθοδο του μέσου όρου [27] ως προς διάφορα μοντέλα. Ειδικότερα, εκπαιδεύουμε διάφορα λPRBF μοντέλα για διαφορετικές τιμές λ. Η τελική δεσμευμένη κατανομή της κατηγορίας C_k δίνεται ως ο μέσος όρος όλων των δεσμευμένων κατανομών $p(x|C_k, \lambda)$ που υπολογίζονται για τις διαφορετικές τιμές λ.

Πιο συγκεκριμένα, επιλέγουμε ένα σύνολο τιμών της παραμέτρου λ $\{\lambda_i,\ i = 1, \ldots, L\}$ και για κάθε τιμή λ_i παίρνουμε μια εκτίμηση της δεσμευμένης κατανομής $p(x|C_k, \lambda_i)$, όπου $k = 1, \ldots, K$, εκπαιδεύοντας το αντίστοιχο μοντέλο λPRBF χρησιμοποιώντας το ίδιο σύνολο δεδομένων X. Η τιμή της δεσμευμένης κατανομής $p(x|C_k)$ για ένα νέο δεδομένο $\bar{x}$ που εμφανίζεται στο σύστημα είναι

$$p(\bar{x}|C_k) \approx \frac{1}{L}\sum_{i=1}^{L} p(\bar{x}|C_k, \lambda_i). \tag{3.14}$$

Στο κεφαλαίο 4 δείχνουμε ότι εφαρμόζοντας την προηγούμενη μέθοδο μέσου όρου βελτιώνεται σημαντικά η επίδοση ταξινόμησης.

3.4 Συζήτηση

Η ιδέα της μεθόδου εκπαίδευσης του Z-μοντέλου που παρουσιάστηκε σε αυτό το κεφάλαιο βασίστηκε στη χρήση προσαρμοζόμενων παραμέτρων περιορισμού από τους οποίους τελικά εξάγουμε τις τιμές των Z περιορισμών. Η εύρεση λύσεων για τις παραμέτρους των δεσμευμένων κατανομών γίνεται σε δύο φάσεις. Στην πρώτη φάση προσαρμόζονται όλες οι παράμετροι (περιορισμού και των δεσμευμένων κατανομών) (Θ, r) μέσω της μεγιστοποίησης μιας κατάλληλα ορισμένης αντικειμενικής συνάρτησης. Με το πέρας της πρώτης φάσης έχει καθοριστεί ο πίνακας Z καθώς επίσης και η αρχικοποίηση των παραμέτρων για την εφαρμογή του EM αλγορίθμου κατά την δεύτερη φάση.

Η αντικειμενική συνάρτηση που εισήχθηκε μπορεί να θεωρηθεί ότι αποτελεί μια μορφή κανονικοποίησης της λογαριθμικής πιθανοφάνειας του μοντέλου των κοινών πυρήνων (2.5). Ωστόσο η κανονικοποίηση (regularization) υλοποιείται μέσω παραμέτρων περιορισμού που εμφανίζονται εμφωλευμένοι στην αρχική συνάρτηση λογαριθμικής πιθανοφάνειας. Ένας διαφορετικός τρόπος κανονικοποίησης θα ήταν να εισάγουμε μια κατάλληλη εκ των προτέρων κατανομή $p(\Theta)$ και να μεγιστοποιούμε τη συνάρτηση (2.5) συν ένα ξεχωριστό όρο $\log p(\Theta)$. Μια τέτοια προσέγγιση προτείνεται στο κεφάλαιο 5 ως μελλοντική έρευνα.

Κεφάλαιο 4

Πειραματικά αποτελέσματα

4.1 Μέθοδος αξιολόγησης των αλγορίθμων

Στο παρόν κεφάλαιο παρουσιάζουμε πειραματικά αποτελέσματα από την εφαρμογή των αλγορίθμων που περιγράψαμε στα προηγούμενα δύο κεφάλαια. Ο σκοπός των πειραμάτων είναι εκτίμηση της γενικευτικής ικανότητας των αλγορίθμων. Στη ενότητα αυτή περιγράφουμε συνοπτικά την μέθοδο εκτίμησης της γενικευτικής ικανότητας που θα χρησιμοποιήσουμε στη συνέχεια.

Έστω ότι διαθέτουμε ένα σύνολο X δεδομένων. Ο απλούστερος τρόπος εκτίμησης της γενικευτικής ικανότητας μιας μεθόδου είναι να διαχωρίσουμε το αρχικό σύνολο σε δύο ξένα υποσύνολα X_1 και X_2. Στη συνέχεια χρησιμοποιούμε το X_1 για εκπαίδευση και το X_2 για έλεγχο. Αν X_2^{err} είναι το σύνολο των δεδομένων που ανήκουν στο X_2 και ταξινομούνται λάθος, το σφάλμα γενίκευσης εκτιμάται από την σχέση

$$error = \frac{|X_2^{err}|}{|X_2|}. \tag{4.1}$$

Ωστόσο η παραπάνω μέθοδος εκτίμησης της σφάλματος γενίκευσης εξαρτάται σημαντικά από την επιλογή των συνόλων X_1 και X_2. Μια ακριβέστερη μέθοδος εκτίμησης του σφάλματος γενίκευσης είναι αυτή της διασταυρωμένης επικύρωσης (cross validation) [3]. Σύμφωνα με την μέθοδο αυτή διαχωρίζουμε το αρχικό σύνολο δεδομένων X σε k ξένα υποσύνολα X_i έτσι ώστε τα υποσύνολα αυτά να περιέχουν κατά το δυνατόν ίσο αριθμό δεδομένων. Σχηματίζουμε σύνολα εκπαίδευσης και ελέγχου ως εξής: Θεωρούμε για κάθε i ως σύνολο εκπαίδευσης το $X - X_i$, ως σύνολο ελέγχου το X_i και υπολογίζουμε το σφάλμα γενίκευσης $error_i$ σύμφωνα με την σχέση (4.1). Έπειτα εκτιμούμε το σφάλμα γενίκευσης με βάση τον ακόλουθο μέσο όρο

$$error = \frac{1}{k}\sum_{i=1}^{k} error_i. \tag{4.2}$$

Με την παραπάνω μέθοδο παίρνουμε επίσης μια εκτίμηση της τυπικής απόκλισης των επιμέρους σφαλμάτων με βάση την σχέση

$$std = \sqrt{\frac{\sum_{i=1}^{k}(error_i - error)^2}{k}}. \tag{4.3}$$

Στη συνέχεια ως μέθοδο αξιολόγησης των αλγορίθμων χρησιμοποιούμε την διασταυρωμένη επικύρωση με $k = 5$.

4.2 Προβλήματα ταξινόμησης

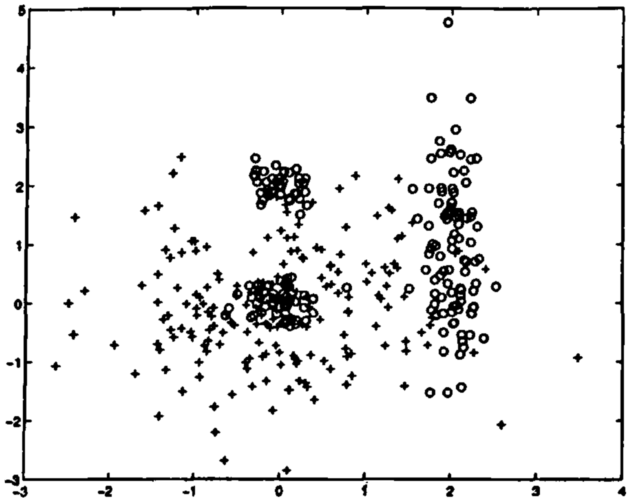
Οι μέθοδοι εφαρμόστηκαν σε πέντε γνωστά σύνολα δεδομένων. Συγκεκριμένα στα σύνολα Clouds, Satimage και Phoneme που προέρχονται από την βάση ELENA [7] καθώς και το Pima Indians και Ionosphere που προέρχονται την UCI [6]. Ακολουθεί μια σύντομη περιγραφή των δεδομένων.

Clouds: Πρόκειται για ένα τεχνητό σύνολο δεδομένων με σημαντικό βαθμό επικάλυψης μεταξύ των δεδομένων των δυο κατηγοριών. Αποτελείται συνολικά από 5000 πρότυπα του διδιάστατου χώρου (Σχήμα 4.1). Τα πρότυπα της κατηγορίας C_1 ακολουθούν μια μικτή κανονική κατανομή με τρεις πυρήνες. Τα πρότυπα της κατηγορίας C_2 ακολουθούν μια απλή κανονική κατανομή.

Phoneme: Πρόκειται για ένα σύνολο πραγματικών δεδομένων που δημιουργήθηκε για την ανάπτυξη συστημάτων αναγνώρισης ομιλίας σε πραγματικό χρόνο, ιδιαίτερα για την διάκριση ρινικών και στοματικών φωνηέντων που προέρχονται από διάφορες συλλαβές. Αποτελείται από 5404 πρότυπα του 5-διάστατου χώρου που ανήκουν σε δύο κατηγορίες.

Satimage: Είναι ένα σύνολο πραγματικών δεδομένων που ελήφθησαν από δορυφορικές εικόνες. Περιγράφουν τις τιμές των pixels σε τέσσερις φασματικές ζώνες για μια γειτονιά 3 x 3 pixels με σκοπό την ταξινόμηση του μεσαίου pixel της γειτονιάς. Αποτελείται από 6435 πρότυπα του 36-διάστατου χώρου που ανήκουν σε 6 κατηγορίες.

Pima Indians: Είναι ένα σύνολο πραγματικών δεδομένων που συγκεντρώθηκαν από γυναίκες άνω των 21 ετών του πληθυσμού των ινδιάνων Pima. Τα χαρακτηριστικά των ασθενών αφορούν ιατρικές εξετάσεις, ηλικία των ασθενών



Σχήμα 4.1: Απεικόνιση των δεδομένων του συνόλου Clouds.

και πληροφορίες εγχυμοσύνης. Χρησιμοποιήθηκαν για την διάγνωση του διαβήτη. Αποτελείται από 768 πρότυπα του 8-διάστατου χώρου που ανήκουν σε δύο κατηγορίες.

Ionosphere: Αποτελεί ένα σύνολο πραγματικών δεδομένων που συγκεντρώθηκαν από παρατηρήσεις ηλεκτρονίων στη ιονόσφαιρα. Το πρόβλημα είναι δύο κατηγοριών ενώ τα δεδομένα ανήκουν στον 35-διάστατο χώρο και είναι συνολικά 351.

4.3 Αποτελέσματα

Οι αλγόριθμοι που αξιολογούνται ως προς την επίδοση τους είναι: i) Το μοντέλο των κοινών πυρήνων (CCM) ii) Οι ανεξάρτητες μικτές κατανομές (SM) iii) Το Z^*-μοντέλο και iv) Το λPRBF μοντέλο.

Εφαρμόσαμε τις μεθόδους στα πέντε σύνολα δεδομένων που περιγράψαμε προηγουμένως χρησιμοποιώντας κανονικούς πυρήνες και για διάφορες τιμές του συνολικού αριθμού πυρήνων. Ορισμένα σύνολα δεδομένων, όπως για παράδειγμα το Clouds (Σχήμα 4.1), παρουσιάζουν σημαντικό βαθμό επικάλυψης μεταξύ των δεδομένων διαφορετικών κατηγοριών, ενώ κάποια άλλα παρουσιάζουν ασθενή επικάλυψη ή και τα δύο. Με βάση την ανάλυση που πραγματοποιήθηκε κυρίως στο κεφάλαιο 2 και δεδομένου ότι χρησιμοποιείται ένας επαρχής αριθμός πυρήνων για αναπαράσταση των δεδομένων, σε προβλήματα όπου έχουμε σημαντική επικάλυψη το μοντέλο των κοινών πυρήνων αναμένεται να έχει καλύτερη επίδοση.

	4 πυρήνες		6 πυρήνες		8 πυρήνες		10 πυρήνες	
	error	std	error	std	error	std	error	std
Z^*-μοντέλο	18.82	4.94	12.4	0.93	11.42	0.51	10.82	0.85
CCM	**13.06**	0.8	**11.12**	0.84	**11.32**	0.89	**10.42**	0.89
SM	24.24	2.03	20.44	4.45	11.86	0.85	11.36	0.98
λPRBF	19.4	3.04	13.12	1.47	11.46	0.96	11.16	0.91

Πίνακας 4.1: Σφάλμα γενίκευσης για το σύνολο δεδομένων Clouds.

Επίσης οι ανεξάρτητες μικτές κατανομές αναμένεται να έχουν καλύτερη επίδοση σε περιπτώσεις ασθενούς επικάλυψης. Το Z^*-μοντέλο λόγω της ιδιότητας του να προσαρμόζεται στη γεωμετρία του συνόλου δεδομένων αναμένεται να έχει σε κάθε περίπτωση καλή απόδοση (είτε την καλύτερη από τις υπόλοιπες μεθόδους είτε πολύ κοντά στη καλύτερη). Το μοντέλο λPRBF αποτελεί μια άμεση γενίκευση του μοντέλου των κοινών πυρήνων και των ανεξάρτητων μικτών κατανομών[1] υπό την έννοια ότι παίρνουμε ενδιάμεσες αναπαραστάσεις των δεσμευμένων κατανομών. Με βάση την ιδιότητα αυτή του μοντέλου λPRBF αναμένουμε με την μέθοδο του μέσου όρου (ενότητα 3.3.1) να λαμβάνουμε πιο ευσταθή συστήματα ταξινόμησης (σε σύγκριση με το μοντέλο των κοινών πυρήνων και των ανεξάρτητων κατανομών) ως προς την ικανότητα γενίκευσης.

Η επιλογή του συνολικού αριθμού πυρήνων γίνεται έτσι ώστε σε κάθε περίπτωση ο αριθμός αυτός να είναι πολλαπλάσιος του αριθμού των κατηγοριών. Μια τέτοια σύμβαση έγινε για το λόγο ότι θα θέλαμε κατά την εφαρμογή των ανεξάρτητων μικτών κατανομών η κάθε ανεξάρτητη δεσμευμένη μικτή κατανομή να χρησιμοποιεί ίσο αριθμό πυρήνων, αφού υποτίθεται ότι δεν διαθέτουμε καμιά εκ των προτέρων πληροφορία για την πολυπλοκότητα της κάθε κατηγορίας. Επίσης κατά την εφαρμογή του Z^*-μοντέλου η αρχικοποίηση των παραμέτρων περιορισμού που χρησιμοποιήθηκε ήταν $r_{jk} = 1/K$ για όλα τα j και k. Τέλος όλοι οι αλγόριθμοι EM ήταν εύκολα υλοποιήσιμοι και με μικρό απαιτούμενο χρόνο εκτέλεσης.

Στους πίνακες 1-5 εμφανίζονται τα αποτελέσματα επίδοσης (σφάλμα γενίκευσης και η αντίστοιχη τιμή της τυπικής απόκλισης χρησιμοποιώντας την διασταυρωμένη επικύρωση με $k = 5$) για τα πέντε προβλήματα ταξινόμησης καθώς και για διάφορες τιμές του αριθμού πυρήνων.

[1] Οι ανεξάρτητες μικτές κατανομές αντιστοιχούν σε μια συγκεκριμένη επιλογή του αριθμού των πυρήνων που χρησιμοποιεί η κάθε μικτή κατανομή.

	12 πυρήνες		18 πυρήνες		24 πυρήνες	
	error	std	error	std	error	std
Z^*-μοντέλο	12.33	0.5	11.4	0.74	11.1	0.75
CCM	13.23	0.56	12.28	0.79	11.52	0.75
SM	12.05	0.53	11.21	0.75	10.98	0.71
λPRBF	**11.90**	0.54	**11.2**	0.65	**10.72**	0.56

Πίνακας 4.2: Σφάλμα γενίκευσης για το σύνολο δεδομένων Satimage.

	10 πυρήνες		12 πυρήνες		14 πυρήνες	
	error	std	error	std	error	std
Z^*-μοντέλο	17.96	1.14	**17.07**	1.01	**15.85**	1.19
CCM	20.62	0.75	20.03	0.75	20.98	1.04
SM	**17.85**	1.4	17.37	0.75	16.88	1.15
λPRBF	18.52	1.25	17.33	1.19	17.24	1.1

Πίνακας 4.3: Σφάλμα γενίκευσης για το σύνολο δεδομένων Phoneme.

	10 πυρήνες		12 πυρήνες		14 πυρήνες	
	error	std	error	std	error	std
Z^*-μοντέλο	27.08	2.6	26.92	3.26	25.94	2.27
CCM	29.95	3.06	28.12	2.21	28.25	1.97
SM	26.69	3.58	26.43	1.34	27.08	2.22
λPRBF	**24.71**	2.9	**24.44**	1.88	**24.44**	2.2

Πίνακας 4.4: Σφάλμα γενίκευσης για το σύνολο δεδομένων Pima Indians.

	8 πυρήνες		10 πυρήνες		12 πυρήνες	
	error	std	error	std	error	std
Z^*-μοντέλο	11.11	2.3	**8.55**	2.4	9.13	3.92
CCM	15.11	3.85	9.41	3.35	9.27	3.21
SM	11.82	1.89	12.24	3.77	9.39	3
λPRBF	**9.41**	2.19	8.56	4.06	**7.97**	3.27

Πίνακας 4.5: Σφάλμα γενίκευσης για το σύνολο δεδομένων Ionosphere.

4.4 Συμπεράσματα

Από τα πειραματικά αποτελέσματα συμπεραίνουμε τα ακόλουθα:

- Ανάλογα με την γεωμετρία του συνόλου δεδομένων καθώς και τον αριθμό των διαθέσιμων πυρήνων το μοντέλο των κοινών πυρήνων μπορεί να έχει καλύτερη επίδοση από τις ανεξάρτητες μικτές κατανομές και αντιστρόφως.

- Το μοντέλο λPRBF έχει καλύτερη επίδοση και από το μοντέλο των κοινών πυρήνων και από το αντίστοιχο των ανεξάρτητων μικτών κατανομών.

- Σε όλα τα προβλήματα που εξετάστηκαν το Z^*-μοντέλο είτε υπερτερεί των υπολοίπων μεθόδων είτε η απόδοση του είναι κοντά στη απόδοση του καλύτερου. Πρέπει να τονισθεί ότι σε καμιά από τις περιπτώσεις των προβλημάτων ταξινόμησης η επίδοση του Z^*-μοντέλο δεν ήταν κατώτερη των υπολοίπων. Αυτό δείχνει έστω και πειραματικά την ικανότητα του Z-μοντέλου (ειδικότερα του αντίστοιχου αλγορίθμου εκπαίδευσης) να προσαρμόζεται στη γεωμετρία των δεδομένων χρησιμοποιώντας ικανοποιητικά το διαθέσιμο αριθμό πυρήνων δίνοντας ένα ικανοποιητικό σύστημα ταξινόμησης.

Επίσης πρέπει να σημειωθεί ότι σε όλες τις περιπτώσεις οι λύσεις παραμέτρων που έδωσε το Z^*-μοντέλο ήταν προσεγγιστικά τύπου 1 με μεγάλη πιθανοφάνεια. Το γεγονός αυτό αποτελεί μια πειραματική απόδειξη της ορθότητας των επιχειρημάτων που αναπτύξαμε στην ενότητα 2.2 σχετικά με το ποιες λύσεις οδηγούν σε καλύτερη επίδοση ταξινόμησης.

Κεφάλαιο 5

Επίλογος

5.1 Τι προτάθηκε στην εργασία

Έχουμε γενικεύσει τις μέχρι τώρα γνωστές μεθόδους εκτίμησης δεσμευμένων κατανομών με μικτές κατανομές δηλαδή του μοντέλου των κοινών πυρήνων και των ανεξάρτητων μικτών κατανομών έτσι ώστε να συμπεριλάβουμε μοντέλα όπου ο κάθε πυρήνας χρησιμοποιείται μόνο από ένα μη κενό υποσύνολο των κατηγοριών. Στο κεφάλαιο 2 δείξαμε με παραδείγματα ότι για σταθερό συνολικό αριθμό πυρήνων το Z-μοντέλο έχει την δυνατότητα να δώσει καλύτερο σύστημα ταξινόμησης σε σύγκριση με το μοντέλο των κοινών πυρήνων και αυτό των ανεξάρτητων μικτών κατανομών. Επίσης βασιζόμενοι στη μέθοδο της μέγιστης πιθανοφάνειας παρουσιάσαμε μια ανάλυση (συγκρίνοντας το μοντέλο των κοινών πυρήνων με το αντίστοιχο των ανεξάρτητων μικτών κατανομών) των περιπτώσεων που η χρήση κοινών πυρήνων είναι ωφέλιμη από την σκοπιά της ταξινόμησης και αντιστρόφως. Τα συμπεράσματα της ανάλυσης είναι ιδιαίτερα σημαντικά αφού προσφέρουν βασικές κατευθύνσεις για την εύρεση των Z-μοντέλων με την αναμενόμενη καλύτερη επίδοση ταξινόμησης.

Στο κεφάλαιο 3 περιγράψαμε μια υπολογιστικά αποδοτική μέθοδο εκπαίδευσης του Z-μοντέλου (επιλογή του πίνακα Z) που βασίζεται στη χρήση ειδικών παραμέτρων περιορισμού και στην εισαγωγή μιας κατάλληλης αντικειμενικής συνάρτησης. Από τις τιμές των περιορισμών εξαρτάται ο βαθμός της χρήσης του κάθε πυρήνα από μια κατηγορία. Επίσης ο αλγόριθμος εκπαίδευσης μπορεί να χρησιμοποιηθεί με σταθερές παραμέτρους περιορισμού. Στην περίπτωση αυτή μπορούν να προκύψουν διάφορες τεχνικές εκπαίδευσης δεσμευμένων μικτών κατανομών, όπως το μοντέλο λPRBF. Επίσης πρέπει να σημειωθεί ότι σε όλες τις περιπτώσεις εκπαίδευσης και ιδιαίτερα κατά την επιλογή του Z-μοντέλου, χρησιμοποιήθηκε ο αλγόριθμος EM. Κάτι τέτοιο αποτελεί πλεονέκτημα διότι ο γενικά αλγόριθμος EM συγκλίνει σε λίγες επαναλήψεις, ενώ επιπλέον ήταν εύκολα υλοποιήσιμος.

Πέρα από την θεωρητική προσφορά της εργασίας στη ανάπτυξη νέων μεθόδων, ο τελικός σκοπός ήταν η βελτίωση της επίδοσης προγενέστερων τεχνικών και κυρίως της τεχνικής των ανεξάρτητων μικτών κατανομών που είναι η επικρατέστερη στη περιοχή της στατιστικής αναγνώρισης προτύπων. Για τα προβλήματα ταξινόμησης που εξετάσθηκαν στο κεφάλαιο 4 το παραπάνω επιτυγχάνεται σε ικανοποιητικό βαθμό.

5.2 Μελλοντική έρευνα

Ως κατευθύνσεις μελλοντικής έρευνας προτείνουμε τα ακόλουθα:

1. *Τεχνικές εκπαίδευσης του Z-μοντέλου.* Εναλλακτικά της μεθόδου εκπαίδευσης που παρουσιάσαμε στη εργασία φαίνονται πιθανές και οι ακόλουθες κατευθύνσεις:

   - **Διακριτή βελτιστοποίηση του πίνακα Z.** Μια τέτοια προσέγγιση μπορεί να έχει την ακόλουθη μορφή. Αρχικά υποθέτουμε την πιο γενική τιμή του πίνακα Z (το μοντέλο των κοινών πυρήνων) και σταδιακά αλλάζουμε τις τιμές του πίνακα κατά την εκπαίδευση. Κάτι τέτοιο μπορεί να βασιστεί στη ανάλυση που παρουσιάστηκε στην ενότητα 2.2. Σύμφωνα με την ανάλυση αυτή γνωρίζουμε ότι επιθυμητές λύσεις είναι αυτές που είναι τύπου 1 και που επιπλέον δίνουν όσο το δυνατόν υψηλότερη τιμή πιθανοφάνειας. Επομένως μπορούμε να μεγιστοποιούμε την πιθανοφάνεια και εν συνεχεία να ελέγχουμε αν η λύση παραμέτρων που πήραμε είναι τύπου 1 ή 2. Σε περίπτωση που είναι τύπου 2 μπορούμε να αλλάζουμε κατάλληλα τον πίνακα Z ώστε με μια νέα εφαρμογή του EM να οδηγηθούμε σε μια λύση που θα είναι πιο κοντά σε λύση τύπου 1. Ωστόσο η αλλαγή του πίνακα Z θα πρέπει να γίνεται προσεχτικά προχειμένου να διατηρούμε όσο το δυνατόν υψηλότερη την τιμή της πιθανοφάνειας.

   - **Μπεϋζιανή προσέγγιση.** Όπως είδαμε στα κεφάλαιο 2 και 3 ένα συγκεκριμένο Z-μοντέλο (δηλ. ο πίνακας Z έχει δεδομένη τιμή) προκύπτει αν θέσουμε κατάλληλους περιορισμούς στο χώρο παραμέτρων που αντιστοιχεί στο μοντέλο των κοινών πυρήνων ($z_{jk} = 1$ για κάθε j

και k). Επιπλέον ένα Z-μοντέλο με καλές ιδιότητες ταξινόμησης πρέπει να επιλέγεται με βάση την ανάλυση της ενότητας 2.2. Επομένως αντί να αναζητούμε ένα συγκεκριμένο Z-μοντέλο μπορούμε να ορίσουμε μια εκ των προτέρων κατανομή $p(\Theta)$ που να θέτει περιορισμούς στο χώρο παραμέτρων του μοντέλου των κοινών πυρήνων ευνοώντας λύσεις με καλές ιδιότητες από τη σκοπιά της ταξινόμησης.

2. *Αύξηση του αριθμού των πυρήνων* με σκοπό την βελτίωση της εκτίμησης των δεσμευμένων κατανομών σε περιοχές των δεδομένων που βρίσκονται κοντά στα όρια απόφασης.

Για την επίτευξη του παραπάνω στόχου προϋποτίθεται ότι υπάρχει η δυνατότητα ανίχνευσης μιας περιοχής δεδομένων με επικάλυψη (περιοχές ορίων απόφασης) καθώς και ο ακόλουθος χαρακτηρισμός της περιοχής αυτής ως προς τον βαθμό επικάλυψης. Με την βοήθεια των κοινών πυρήνων μπορούν να ανιχνεύονται περιοχές με επικάλυψη. Πράγματι ένας κοινός πυρήνα j αναπαριστά δεδομένα όλων των κατηγοριών C_k, με $k \in \{i : \pi_{ji} > 0\}$ που ενδεχομένως βρίσκονται σε μια περιοχή επικάλυψης. Μια επικαλυπτόμενη περιοχή μπορεί να χαρακτηρίζεται από ασθενή ή σημαντικό βαθμό επικάλυψης. Μια τέτοια πληροφορία ενδεχομένως να μπορεί να εξαχθεί μέσω της ανάλυσης των στάσιμων σημείων που παρουσιάστηκε στην ενότητα 2.2. Όπως έχει εξηγηθεί στην ίδια ενότητα, όταν ένας πυρήνας αναπαριστά μια περιοχή με ασθενή βαθμό επικάλυψης, τότε αυτό έχει σημαντικές αρνητικές συνέπειες στο σφάλμα γενίκευσης. Επομένως αναγνωρίζοντας τέτοιες περιοχές μπορούμε εν συνεχεία να λύσουμε το πρόβλημα προσθέτοντας (ή διαιρώντας τον αρχικό) τοπικά νέους πυρήνες. Για παράδειγμα στο παράδειγμα του σχήματος 2.3 (κεφάλαιο 2) και για το μοντέλο των κοινών πυρήνων (α) θα μπορούσαμε τον πυρήνα που βρίσκεται πάνω στο όριο απόφασης να τον διαχωρίσουμε σε δύο έτσι ώστε ο καθένας να αναγνωρίζει δεδομένα μιας μόνο κατηγορίας.

3. *Εφαρμογή των τεχνικών σε άλλα μοντέλα μάθησης χωρίς επίβλεψη* που βασίζονται σε μικτές κατανομές. Για παράδειγμα μια άμεση γενίκευση θα ήταν να χρησιμοποιήσουμε ιεραρχικές μικτές κατανομές [5] ως μοντέλα για τις δεσμευμένες κατανομές. Επίσης οι τεχνικές που αναπτύχθηκαν θα μπορούσαν να εφαρμοστούν σε μεθόδους που ελαττώνουν την διάσταση του χώρου δεδομένων και βασίζονται σε μικτές κατανομές [28, 11].

Παράρτημα Α

Αλγόριθμος EM για το Z-μοντέλο

Για να ορίσουμε το σύνολο των χρυμμένων μεταβλητών παρατηρούμε ότι για κάθε κατηγορία C_k το σύνολο των πυρήνων που χρησιμοποιούνται από το αντίστοιχο μοντέλο $p(x|C_k)$ της δεσμευμένης κατανομής είναι $S_k = \{j : z_{jk} = 1\}$. Επομένως για κάθε δεδομένο x της κατηγορίας C_k ορίζουμε ένα $|S_k|$-διάστατο διάνυσμα $w(x)$ το οποίο υποδειχνύει τον πυρήνα που παρήγαγε το x ($w_j(x) = 1$ αν ο πυρήνας j παρήγαγε το x και $w_i(x) = 0$, $i \neq j$)). Η λογαριθμική πιθανοφάνεια του πλήρους συνόλου δεδομένων δίνεται από

$$L_C(\Theta) = \sum_{k=1}^{K} \sum_{x \in X_k} \sum_{j:z_{jk}=1} w_j(x) \log\{\pi_{jk} p(x|j;\theta_j)\}. \tag{A.1}$$

Κατά την επανάληψη $t+1$ του αλγορίθμου EM υπολογίζεται η αναμενόμενη τιμή της ποσότητας $L_C(\Theta)$ η οποία δίνεται από τη σχέση

$$Q(\Theta;\Theta^{(t)}) = \sum_{k=1}^{K} \sum_{x \in X_k} \sum_{j:z_{jk}=1} P(j|x, C_k; z_k, \pi_k^{(t)}, \theta^{(t)}) \log\{\pi_{jk} p(j|x;\theta_j)\}, \tag{A.2}$$

όπου έχουμε αντικαταστήσει $w_j(x)$ με την αναμενόμενη τιμή $P(j|x, C_k; z_k, \pi_k^{(t)}, \theta^{(t)})$ (με βάση την εξίσωση (2.20)). Η συνάρτηση Q μπορεί να γραφεί ως εξής:

$$Q(\Theta;\Theta^{(t)}) = Q_1(\pi;\Theta^{(t)}) + Q_2(\theta;\Theta^{(t)}), \tag{A.3}$$

όπου με π συμβολίζουμε όλες τις εκ των προτέρων πιθανότητες και

$$Q_1(\pi;\Theta^{(t)}) = \sum_{k=1}^{K} \sum_{x \in X_k} \sum_{j:z_{jk}=1} P(j|x, C_k; z_k, \pi_k^{(t)}, \theta^{(t)}) \log \pi_{jk}, \tag{A.4}$$

$$Q_2(\theta;\Theta^{(t)}) = \sum_{k=1}^{K} \sum_{x \in X_k} \sum_{j:z_{jk}=1} P(j|x, C_k; z_k, \pi_k^{(t)}, \theta^{(t)}) \log p(x|j;\theta_j). \tag{A.5}$$

Οι δύο παραπάνω όροι μεγιστοποιούνται ανεξάρτητα λόγω του ότι δεν περιέχουν κοινές παραμέτρους. Αν υποθέσουμε ότι οι κατανομές των πυρήνων δίνονται από κανονικές κατανομές της ακόλουθης γενικής μορφής

$$p(x|j;\mu_j,\Sigma_j)=\frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right\}, \tag{A.6}$$

τότε μεγιστοποιώντας την $Q_1(\theta;\Theta^{(t)})$ (παίρνοντας μερικές παραγώγους και εξισώνοντας με το μηδέν) καταλήγουμε στις εξής εξισώσεις ενημέρωσης των παραμέτρων

$$\mu_j^{(t+1)}=\frac{\sum_{k:z_{jk}=1}\sum_{x\in X_k}P(j|x,C_k;z_k,\pi_k^{(t)},\theta^{(t)})x}{\sum_{k:z_{jk}=1}\sum_{x\in X_k}P(j|x,C_k;z_k,\pi_k^{(t)},\theta^{(t)})}, \tag{A.7}$$

$$\Sigma_j^{(t+1)}=\frac{\sum_{k:z_{jk}=1}\sum_{x\in X_k}P(j|x,C_k;z_k,\pi_k^{(t)},\theta^{(t)})(x-\mu_j^{(t+1)})(x-\mu_j^{(t+1)})^T}{\sum_{k:z_{jk}=1}\sum_{x\in X_k}P(j|x,C_k;z_k,\pi_k^{(t)},\theta^{(t)})}, \tag{A.8}$$

για κάθε $j=1,\ldots,M$.

Προκειμένου να μεγιστοποιήσουμε την ποσότητα $Q_2(\pi;\Theta^{(t)})$ λαμβάνοντας υπόψη τον περιορισμό (2.15) εισάγουμε K πολλαπλασιαστές Lagrange λ_k. Η συνάρτηση που μεγιστοποιούμε παίρνει τη μορφή

$$\tilde{Q}_2(\pi;\Theta^{(t)})=Q_2(\pi;\Theta^{(t)})-\sum_{k=1}^{K}\lambda_k\left(\sum_{j:z_{jk}=1}\pi_{jk}-1\right). \tag{A.9}$$

Τώρα, παίρνοντας παραγώγους και εξισώνοντας με το μηδέν παίρνουμε τελικά

$$\pi_{jk}^{(t+1)}=\frac{1}{|X_k|}\sum_{x\in X_k}P(j|x,C_k;z_k,\pi_k^{(t)},\theta^{(t)}), \tag{A.10}$$

για κάθε j και k τέτοια ώστε $z_{jk}=1$.

Παράρτημα Β

Αλγόριθμος EM για επιλογή Z^*-μοντέλου

Καταρχήν βρίσκουμε την μορφή της συνάρτησης Q που υπολογίζεται στο E-βήμα. Η αντικειμενική συνάρτηση που μεγιστοποιείται είναι:

$$L(\Theta, r) = \log P(X; \Theta, r) = \log \prod_{k=1}^{K} \prod_{x \in X_k} \sum_{j=1}^{M} r_{jk} \pi_{jk} p(x|j; \theta_j). \tag{B.1}$$

Σύμφωνα με τα όσα είπαμε στην ενότητα 3.4 η $P(X; \Theta, r)$ δεν αντιστοιχεί σε κατανομή (ως προς το X), και οπότε η αντιχειμενική συνάρτηση (B.1) μπορεί να θεωρηθεί ως η 'λογαριθμική πιθανοφάνεια του ελλιπούς συνόλου' με μια ευρύ έννοια. Για κάθε δεδομένο $x \in X$ θεωρούμε την χρυμμένη μεταβλητή $y(x) \in \{1, \dots, M\}$ η οποία υποδειχνύει τον πυρήνα με βάση τον οποίο έχει παραχθεί το x. Το συνολικό διάνυσμα των χρυμμένων μεταβλητών είναι $Y = (y(x^1), \dots, y(x^N))$. Χρησιμοποιώντας τις χρυμμένες μεταβλητές η λογαριθμική πιθανοφάνεια του πλήρους συνόλου ορίζεται ως εξής:

$$L_C(\Theta, r) = \log P(X, Y; \Theta, r) = \log \prod_{k=1}^{K} \prod_{x \in X_k} \{r_{y(x)k} \pi_{y(x)k} p(x|y(x); \theta_{y(x)})\}. \tag{B.2}$$

Μπορεί να δειχθεί ότι οι συναρτήσεις $P(X; \Theta, r)$ και $P(X, Y; \Theta, r)$ συνδέονται με βάση την ακόλουθη σχέση:

$$P(X; \Theta, r) = \sum_{Y} P(X, Y; \Theta, r), \tag{B.3}$$

όπου το άθροισμα είναι ως προς όλες τιμές του διανύσματος Y. Επιπλέον, ορίζουμε την συνάρτηση

$$P(Y; X, \Theta, r) = \frac{P(X, Y; \Theta, r)}{P(X; \Theta, r)}, \tag{B.4}$$

η οποία είναι *κατανομή πιθανότητας* του διανύσματος των κρυμμένων μεταβλητών Y δοθέντος των παρατηρήσιμων δεδομένων X και των παραμέτρων Θ (αφού $\sum_Y P(Y;X,\Theta,r) = 1$ λόγω της (B.3)). Τώρα, χρησιμοποιώντας τις σχέσεις (B.1), (B.2) και (3.6) η $P(Y;X,\Theta,r)$ μπορεί να γραφεί

$$P(Y;X,\Theta,r) = \prod_{k=1}^{K}\prod_{x\in X_k} \frac{r_{y(x)k}\pi_{y(x)k}p(x|y(x);\theta_{y(x)})}{\sum_{j=1}^{M} r_{jk}\pi_{jk}p(x|j;\theta_j)} = \prod_{k=1}^{K}\prod_{x\in X_k} \Phi_{y(x)}(x;C_k,r_k,\pi_k,\theta). \tag{B.5}$$

Έστω ότι ο αλγόριθμος ΕΜ βρίσκεται στην επανάληψη $t+1$ και οι τρέχουσα τιμή του διανύσματος παραμέτρων είναι $(\Theta^{(t)}, r^{(t)})$. Τότε ορίζουμε τη συνάρτηση Q ως της αναμενόμενη τιμή της $\log P(X,Y;\Theta,r)$ ως προς την κατανομή $P(Y;X,\Theta^{(t)},r^{(t)})$

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \sum_Y \log\{P(X,Y;\Theta,r)\}P(Y;X,\Theta^{(t)},r^{(t)}) \tag{B.6}$$

και με βάση την (B.5)

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \sum_Y \log\{P(X,Y;\Theta,r)\}\prod_{k=1}^{K}\prod_{x\in X_k} \Phi_{y(x)}(x;C_k,r_k^{(t)},\pi_k^{(t)},\theta^{(t)}). \tag{B.7}$$

Αντικαθιστώντας την ποσότητα $\log P(X,Y;\Theta,r)$ σύμφωνα με την σχέση (B.2) βρίσκουμε

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \tag{B.8}$$
$$\sum_Y\left\{\sum_{k=1}^{K}\sum_{x\in X_k}\log\{r_{y(x)k}\pi_{y(x)k}p(x|y(x),\theta_{y(x)})\}\prod_{i=1}^{K}\prod_{x\in X_i}\Phi_{y(x)}(x;C_i,r_i^{(t)},\pi_i^{(t)},\theta^{(t)})\right\}$$

και χρησιμοποιώντας τον συμβολισμό του Kronecker

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \tag{B.9}$$
$$\sum_Y\left\{\sum_{k=1}^{K}\sum_{x\in X_k}\sum_{j=1}^{M}\delta_{jy(x)}\log\{r_{jk}\pi_{jk}p(x|j,\theta_j)\}\prod_{i=1}^{K}\prod_{x\in X_i}\Phi_{y(x)}(x;C_i,r_i^{(t)},\pi_i^{(t)},\theta^{(t)})\right\}.$$

λόγω του γεγονότος ότι $\sum_{j=1}^{M}\Phi_j(x;C_k,r_k,\pi_k,\theta) = 1$ η εξίσωση (B.10) παίρνει τελικά την μορφή[1]

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \sum_{k=1}^{K}\sum_{x\in X_k}\sum_{j=1}^{M}\Phi_j(x;C_k,r_k^{(t)},\pi_k^{(t)},\theta^{(t)})\log\{r_{jk}\pi_{jk}p(x|j;\theta_j)\}. \tag{B.10}$$

[1]Μια ανάλογη απόδειξη στη περίπτωση των μικτών κατανομών περιγράφεται στο [3], σελ. 69-72.

Στο Παράρτημα Γ δείχνεται ότι ο αλγόριθμος EM που σε κάθε επανάληψη στο E-βήμα υπολογίζει την συνάρτηση Q και ενώ στο M-βήμα την μεγιστοποιεί ως προς τις παραμέτρους Θ και r, εγγυάται την μονότονη αύξηση της αντικειμενικής συνάρτησης $L(\Theta, r)$. Παρακάτω, δίνουμε τις αναλυτικές εξισώσεις ενημέρωσης στο M-βήμα για την περίπτωση κανονικών πυρήνων.

Η συνάρτηση Q (B.10) μπορεί να γραφεί ως το άθροισμα τριών όρων

$$Q(\Theta, r; \Theta^{(t)}, r^{(t)}) = Q_1(r; \Theta^{(t)}) + Q_2(\pi; \Theta^{(t)}) + Q_3(\theta; \Theta^{(t)}), \qquad \text{(B.11)}$$

όπου, ακριβώς ανάλογα με το Παράρτημα Α, οι μόνες προσαρμοζόμενοι παράμετροι στη $Q_1(r;\Theta^{(t)})$ είναι οι περιορισμοί r, στη $Q_2(\pi;\Theta^{(t)})$ οι εκ των προτέρων πιθανότητες π και στη $Q_3(\theta;\Theta^{(t)})$ οι παράμετροι των πυρήνων θ. Προφανώς, κάθε όρος μπορεί να μεγιστοποιηθεί ανεξάρτητα. Υποθέτοντας κανονικούς πυρήνες οι όροι $Q_3(\theta;\Theta^{(t)})$ και $Q_2(\pi;\Theta^{(t)})$ μεγιστοποιούνται ακριβώς ανάλογα με τη περίπτωση του Παραρτήματος Α και τελικά οι εξισώσεις που παίρνουμε για κάθε j είναι

$$\mu_j^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)}) x}{\sum_{k=1}^{K} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)})}, \qquad \text{(B.12)}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{k=1}^{K} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)})(x - \mu_j^{(t+1)})(x - \mu_j^{(t+1)})^T}{\sum_{k=1}^{K} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)})}, \qquad \text{(B.13)}$$

$$\pi_{jk}^{(t+1)} = \frac{1}{|X_k|} \sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)}) \quad k = 1, \ldots, K. \qquad \text{(B.14)}$$

Προκειμένου να μεγιστοποιήσουμε τον όρο $Q_1(r;\Theta^{(t)})$ εισάγουμε M πολλαπλασιαστές Lagrange λ_j (για την ικανοποίηση του περιορισμού (3.1)), οπότε η ποσότητα που μεγιστοποιείται είναι

$$\tilde{Q}_3(r; \Theta^{(t)}) = Q_3(r; \Theta^{(t)}) - \sum_{j=1}^{M} \lambda_j \left(\sum_{k=1}^{K} r_{jk} - 1 \right). \qquad \text{(B.15)}$$

Παίρνοντας μερικές παραγώγους, μηδενίζοντας και κάνοντας πράξεις τελικά καταλήγουμε στις ακόλουθες σχέσεις ενημέρωσης

$$r_{jk}^{(t+1)} = \frac{\sum_{x \in X_k} \Phi_j(x; C_k, r_k^{(t)}, \pi_k^{(t)}, \theta^{(t)})}{\sum_{i=1}^{K} \sum_{x \in X_i} \Phi_j(x; C_i, r_i^{(t)}, \pi_i^{(t)}, \theta^{(t)})}, \qquad \text{(B.16)}$$

όπου $j = 1, \ldots, M$ και $k = 1, \ldots, K$.

Παράρτημα Γ

Απόδειξη της μονότονης αύξησης της $L(\Theta, r)$

Δίνουμε μια απλή απόδειξη ότι ο αλγόριθμος EM της ενότητας 3.4 εγγυάται σε κάθε επανάληψη την μονότονη αύξηση της αντικειμενικής συνάρτησης (3.4) έως ότου βρεθεί ένα τοπικό ελάχιστο. Εάν πολλαπλασιάσουμε και διαιρέσουμε το όρισμα του λογαρίθμου της εξίσωσης (B.6) με $P(X;\Theta,r)$, η συνάρτηση Q που υπολογίζεται στο E-βήμα γράφεται ως

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \sum_{Y} \log\left\{\frac{P(X,Y;\Theta,r)P(X;\Theta,r)}{P(X;\Theta,r)}\right\} P(Y;X,\Theta^{(t)},r^{(t)}). \tag{Γ.1}$$

Χωρίζοντας το λογάριθμο και με βάση την (B.4) παίρνουμε

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = \sum_{Y} \log\{P(X;\Theta,r)\}P(Y;X,\Theta,r) + \sum_{Y} \log\{P(Y;X,\Theta,r)\}P(Y;X,\Theta^{(t)},r^{(t)}). \tag{Γ.2}$$

Ακολούθως χρησιμοποιώντας την (B.1) καθώς και το γεγονός ότι $\sum_Y P(Y;X,\Theta,r) = 1$ έχουμε

$$Q(\Theta,r;\Theta^{(t)},r^{(t)}) = L(\Theta,r) + \sum_{Y} \log\{P(Y;X,\Theta,r)\}P(Y;X,\Theta^{(t)},r^{(t)}). \tag{Γ.3}$$

Έστω τώρα ότι στο M-βήμα του αλγορίθμου βρίσκουμε ένα διάνυσμα παραμέτρων $(\Theta^{(t+1)},r^{(t+1)})$ τέτοιο ώστε $Q(\Theta^{(t+1)},r^{(t+1)};\Theta^{(t)},r^{(t)}) \geq Q(\Theta^{(t)},r^{(t)};\Theta^{(t)},r^{(t)})$, (μια τέτοια υπόθεση αφορά την γενικότερη περίπτωση αλγόριθμου GEM [21]). Τότε μπορούμε να γράψουμε ότι

$$L(\Theta^{(t+1)},r^{(t+1)}) - L(\Theta^{(t)},r^{(t)}) + \sum_{Y} \log\left\{\frac{P(Y;X,\Theta^{(t+1)},r^{(t+1)})}{P(Y;X,\Theta^{(t)},r^{(t)})}\right\} P(Y;X,\Theta^{(t)},r^{(t)}) \geq 0 \tag{Γ.4}$$

Το άθροισμα στη παραπάνω εξίσωση σύμφωνα με την ανισότητα Jensen δεν μπορεί να πάρει θετική τιμή (δες [3], σελ. 66). Έτσι, καταλήγουμε στο ότι $L(\Theta^{(t+1)}, r^{(t+1)}) \geq L(\Theta^{(t)}, r^{(t)})$.

Παράρτημα Δ

Αλγόριθμος EM για το μοντέλο λPRBF

Ο αλγόριθμος EM για την μεγιστοποίηση της ποσότητας (3.12) μπορεί να προκύψει ως ειδική περίπτωση του αλγορίθμου που περιγράφτηκε στο Παράρτημα B με μόνη διαφορά ότι οι τιμές των περιορισμών r παραμένουν σταθερές κατά την βελτιστοποίηση. 'Οπως είδαμε στη ενότητα **3.3** οι περιορισμοί r που αντιστοιχούν στο μοντέλο λPRBF ορίζονται από την σχέση:

$$r_{jk} = \begin{cases} \frac{1}{1+\lambda(K-1)} & j \in T_k \\ \frac{\lambda}{1+\lambda(K-1)} & j \notin T_k \end{cases} \tag{Δ.1}$$

Για τις παραπάνω τιμές των r οι ποσότητες Φ_j που εμφανίζονται κατα την $t+1$ επανάληψη του αλγορίθμου EM γράφονταϊ σε μια πιο βολική μορφή ως εξής:

$$\Phi_j(x; C_k, \lambda, \pi_k^{(t)}, \theta^{(t)}) = \begin{cases} \frac{\pi_{jk}^{(t)} p(x|j,\theta_j^{(t)})}{\sum_{i\in T_k} \pi_{ik}^{(t)} p(x|i,\theta_i^{(t)}) + \lambda \sum_{i\notin T_k} \pi_{ik}^{(t)} p(x|i,\theta_i^{(t)})} = h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}), & j \in T_k \\ \frac{\lambda \pi_{jk}^{(t)} p(x|j,\theta_j^{(t)})}{\sum_{i\in T_k} \pi_{ik}^{(t)} p(x|i,\theta_i^{(t)}) + \lambda \sum_{i\notin T_k} \pi_{ik}^{(t)} p(x|i,\theta_i^{(t)})} = \lambda h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}), & j \notin T_k \end{cases} \tag{Δ.2}$$

Επομένως οι εξισώσεις ενημέρωσης των παραμέτρων είναι οι ακόλουθες

$$\mu_j^{(t+1)} = \frac{\sum_{x\in X_k} h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}) x + \lambda \sum_{\ell \neq k} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)}) x}{\sum_{x\in X_k} h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}) + \lambda \sum_{\ell \neq k} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)})} \tag{Δ.3}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{x\in X_k} h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}) w(x) + \lambda \sum_{\ell \neq k} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)}) w(x)}{\sum_{x\in X_k} h_{jk}(x; \pi_k^{(t)}, \theta^{(t)}) + \lambda \sum_{\ell \neq k} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)})} \tag{Δ.4}$$

$$\pi_{j\ell} = \begin{cases} \frac{1}{|X_\ell|} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)}), & \ell = k \\ \frac{\lambda}{|X_\ell|} \sum_{x \in X_\ell} h_{j\ell}(x; \pi_\ell^{(t)}, \theta^{(t)}), & \ell \neq k \end{cases} \tag{Δ.5}$$

όπου $j \in T_k$, $k = 1, \ldots, K$ και $w(x)$ αντιστοιχεί στη ποσότητα $(x - \mu_j^{(t+1)})(x - \mu_j^{(t+1)})^T$.

Βιβλιογραφία

[1] Alba J. L., Docio L., Docampo D., Marquez O. W., *Growing Gaussian mixtures network for classification.* Signal Processing 76: (1), 43-60, JUL, 1999.

[2] Bengio Y. Gingras F. Gouland B. Lina J. M. and Scott K., *Gaussian mixture densities for classification of nuclear power plant data.* Computer and Artificial Intelligence, 17: (2-3) 189-209, 1998.

[3] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[4] C. M. Bishop, *Latent variables models.* In Learning in Graphical Models, M.I. Jordan (Ed.) Dordrecht: Kluwer, pp. 371-403.

[5] C. M. Bishop and M. E. Tipping, *A hierarchical latent variable model for data visualization.* IEEE transactions on Pattern Analysis and Machine intelligence, 20, 281-293.

[6] C. L. Blake and C. J. Merz, *UCI repository of machine learning databases*, University of California, Irvine, Dept. of Computer and Information Sciences, 1998.

[7] Datasets and technical reports available via anonymous ftp from: ftp.dice.ucl.ac.be/pub/neural-nets/ELENA/databases.

[8] A. P. Dempster, N. M. Laird and D. B. Rubin, *Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society B*, vol. 39, pp. 1-38, 1977.

[9] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[10] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.

[11] Z. Ghahramani and G. E. Hinton, *The Em algorithm for factor analyzers* Technical report No. CRG-TR-96-1. Toronto: the University of Toronto.

[12] Z. Ghahramani and M. I. Jordan *Learning from incomplete data.* Technical Report CBCL 108,Massachusetts Institute of Technology.

[13] W. R. Gilks, S. Richardson and D.J. Spiegelhalter, (eds), *Markov Chain monte carlo in Practice.* London, Chapman & Hall.

[14] P. J. Green, *On use of the EM algorithm for penalized likelihood estimation*, Journal of the Royal Statistical Society B, 52, 443-452, 1990.

[15] T. J. Hastie and R. J. Tibshirani, *Discriminant Analysis by Gaussian Mixtures*, Journal of the Royal Statistical Society B, vol. 58, pp. 155-176, 1996.

[16] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton, *Adaptive mixtures of local experts*, Neural Computation, vol.3, pp. 79-87, 1991.

[17] M. I. Jordan and R. A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm*, Neural Computation, vol.6, pp. 181-214, 1994.

[18] D. J. C. MacKay, *Bayesian interpolation*, Neural computation 4, 415447.

[19] G. J. McLachlan and K. Basford, *Mixture models: Inference and applications to clustering.* Wiley, 1988.

[20] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

[21] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Marcel Dekker, 1997.

[22] A. C. P. Miguel and S. Renals, *Practical Identifiability of Finite mixtures of Bernouli Distributions.* Neural Computation.

[23] R. M. Neal, *Bayesian Learning for Neural Networks.* Lecture Notes in Statistics (no. 118), Springer, New York, 1996.

[24] R. M. Neal and G. E. Hinton, *A view of the EM algortihm that justifies incremental, sparse and other variants.* In Jordan M. I. (Ed.), learning in graphical Models. Kluwer Academic Press, 1998.

[25] R. Redner and H. Walker, 'Mixture densities, Maximum Likelihood and the EM Algorithm', *SIAM Review*, vol. 26, no. 2, pp. 195-239, 1984.

[26] S. Richardson and P. Green, *On Bayesian analysis of mixtures with an unknown number of components*, Journal of the Royal Statistical Society, B 59, 731-792, 1997.

[27] A. Sharkey, *Combining Artificial Neural nets*, Springer, London, 1999.

[28] M. E. Tipping and C. M. Bishop, *mixtures of probabilistic principal component analysis*. Neural computation, 11, 443-482.

[29] M. K. Titsias and A. Likas, 'A Probabilistic RBF network for Classification', *Proc. of International Joint Conference on Neural Networks*, Como, Italy, July 2000.

[30] M. K. Titsias and A. Likas, 'Shared Kernel Models for Class Conditional Density Estimation', *IEEE Trans. on Neural Networks*, to appear.

[31] M. K. Titsias and A. Likas, 'Class Conditional Density Estimation using Mixrures with Constrained Component Sharing;, Technical Report No 08-2001, Department of Computer Science, University of Ioannina, March 2001.

[32] D. M. Titterington, A. F. Smith and U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, Wiley, 1985.

[33] N. A. Vlassis and A. Likas, "A Kurtosis-Based Dynamic Approach to Gaussian Mixture Modeling", *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 29, no. 4, pp. 393-399, 1999.

[34] C. F. J. Wu, *on the convergence properties of the EM algorihtm*. Annals of Statistics 11, 95-103.

[35] X. H. Zhuang, Y. Huang, K. Palaniappan and Y. X. Zhao, *Gauusian mixture density modeling, decomposition, and applications*. IEEE Transactions on Image Processing, 5: 9, 1293-1302, 1996.