Κεφάλαιο 4: Μη Παραμετρικές Τεχνικές

4.1 Εισαγωγή

Στο 3^{ο} κεφάλαιο περιγράφηκε η διαδικασία της μη επιβλεπόμενης μάθησης υπό τη θεώρηση ότι οι μορφές των αντίστοιχων συναρτήσεων πυκνότητας πιθανότητας είναι γνωστές. Όμως, στις περισσότερες εφαρμογές αναγνώρισης προτύπων αυτή η θεώρηση είναι υπό αμφισβήτηση. Οι συνήθεις παραμετρικές μορφές σπάνια ταιριάζουν με τις συναρτήσεις πυκνότητας πιθανότητας που παρατηρούνται στην πραγματικότητα. Πιο συγκεκριμένα, όλες οι κλασσικές παραμετρικές συναρτήσεις πυκνότητας πιθανότητας έχουν ένα μοναδικό τοπικό μέγιστο (unimodal), ενώ τα περισσότερα πρακτικά προβλήματα περιλαμβάνουν συναρτήσεις πυκνότητας πιθανότητας με πολλά τοπικά ακρότατα (multimodal). Επιπλέον, οι προσδοκίες ότι μία μεγάλης διάστασης συνάρτηση πυκνότητας πιθανότητας μπορεί να αντιπροσωπευθεί ικανοποιητικά από το γινόμενο πολλών μονοδιάστατων πολύ σπάνια πραγματοποιούνται. Στο κεφάλαιο αυτό, θα εξεταστούν διάφορες μη παραμετρικές τεχνικές που μπορούν να χρησιμοποιηθούν με τυχαίες κατανομές και χωρίς τη θεώρηση ότι οι μορφές των συναρτήσεων πυκνότητας πιθανότητας πρέπει να είναι γνωστές.

Υπάρχουν διάφοροι τύποι μη παραμετρικών μεθόδων που έχουν ιδιαίτερο ενδιαφέρον για τον τομέα της Αναγνώρισης Προτύπων. Μία κατηγορία αποτελείται από διαδικασίες που υπολογίζουν τις συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_j)$ από τυχαία δείγματα. Εάν οι υπολογισμοί αυτοί είναι ικανοποιητικοί, μπορούν να αντικαταστήσουν τις πραγματικές συναρτήσεις πυκνότητας πιθανότητας κατά το σχεδιασμό του ταξινομητή. Μία άλλη κατηγορία αποτελείται από διαδικασίες για άμεσο υπολογισμό των εκ των υστέρων πιθανοτήτων $P(\omega_j/x)$. Αυτή η κατηγορία σχετίζεται πολύ και με κάποιες μη παραμετρικές τεχνικές, όπως ο κανόνας του κοντινότερου γείτονα, οι οποίες προσπερνάνε τον υπολογισμό των τιμών των πιθανοτήτων και ασχολούνται άμεσα με τις συναρτήσεις απόφασης.

4.2 Υπολογισμός Συνάρτησης Πυκνότητας Πιθανότητας

Οι βασικές ιδέες που κρύβονται πίσω από πολλές μεθόδους υπολογισμού μιας άγνωστης συνάρτησης πυκνότητας πιθανότητας είναι πολύ απλές. Οι πιο σημαντικές τεχνικές βασίζονται στο ότι η πιθανότητα P ότι ένα διάνυσμα x θα βρίσκεται σε μία περιοχή R δίνεται από

$$P = \int_R p(x')dx' \tag{4.1}$$

Έτσι η P είναι μια εξομαλυσμένη έκδοση της συνάρτησης πυκνότητας πιθανότητας $p(x)$ και επομένως η τιμή της p μπορεί να βρεθεί υπολογίζοντας την τιμή της πιθανότητας P. Έστω n δείγματα $x_1, \ldots, x_n$ ανεξάρτητα και ομοιόμορφα κατανεμημένα σύμφωνα με την κατανομή πιθανότητας $p(x)$. Προφανώς η πιθανότητα ότι k από αυτά τα n δείγματα βρίσκονται μέσα στην περιοχή R δίνεται από τον παρακάτω τύπο

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \tag{4.2}$$

Και η αναμενόμενη τιμή για το k είναι

$$E[k] = nP \tag{4.3}$$

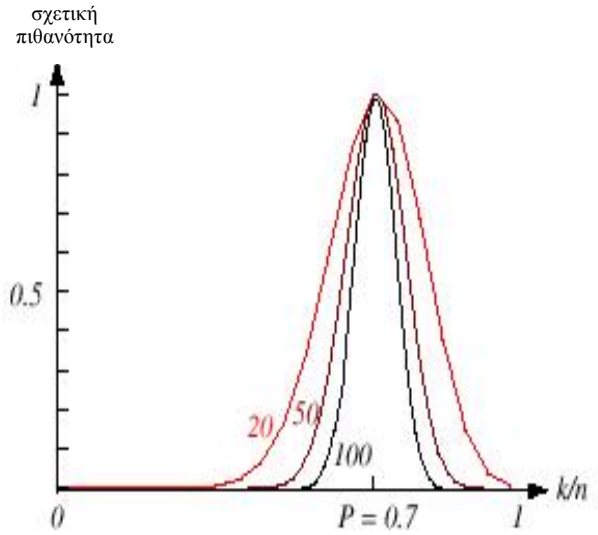
Επιπλέον αυτή η διωνυμική κατανομή για το k κορυφώνεται γύρω από τη μέση τιμή, έτσι ώστε να αναμένεται ότι ο λόγος k/n θα είναι μία πολύ καλή προσέγγιση της πιθανότητας P και επομένως και της εξομαλυσμένης συνάρτησης πυκνότητας πιθανότητας. Αυτή η προσέγγιση είναι ιδιαίτερα ακριβής όταν το n είναι πολύ μεγάλο. Εάν θεωρηθεί ότι η p(x) είναι συνεχής και ότι η περιοχή R είναι τόσο μικρή ώστε η τιμή της p να μην μεταβάλλεται σημαντικά μέσα σε αυτή μπορεί να γραφεί το εξής

$$\int_R p(x')dx' \approx p(x')V \tag{4.4}$$

όπου το x είναι ένα σημείο μέσα στην R και ο V είναι ο όγκος που περικλείεται από την R. Συνδυάζοντας τις εξισώσεις 4.1, 4.3 και 4.4 προκύπτει η ακόλουθη προφανής προσέγγιση για την p(x)

$$p(x) \approx \frac{k/n}{V} \tag{4.5}$$

όπως φαίνεται στην εικόνα 4.1.



Εικόνα 4.1: Η σχετική πιθανότητα, που υπολογίζεται από την εξίσωση 4.4, παράγει μια συγκεκριμένη τιμή για την πυκνότητα πιθανότητας, στο σημείο όπου η πραγματική τιμή της πιθανότητα έχει επιλεγεί να είναι ίση με 0.7. Κάθε καμπύλη έχει ετικέτα τον συνολικό αριθμό των δειγμάτων που επιλέχθηκαν και είναι κλιμακωμένη ώστε να δίνει το ίδιο μέγιστο (στην πραγματική τιμή της πιθανότητας). Η μορφή κάθε καμπύλης είναι διωνυμική, όπως φαίνεται από την εξίσωση 4.2. Για μεγάλο n, τέτοια διώνυμα έχουν έντονη κορυφή στην πραγματική τιμή της πιθανότητας. Στο όριο $n \to \infty$, η καμπύλη προσεγγίζει μία συνάρτηση δέλτα και υπάρχει εγγύηση ότι ο υπολογισμός θα δώσει την πραγματική τιμή της πιθανότητας.

Υπάρχουν πολλά προβλήματα που παραμένουν ακόμα αναπάντητα – κάποια πρακτικά και κάποια θεωρητικά. Εάν διορθωθεί ο όγκος V και χρησιμοποιηθούν όλο και περισσότερα δείγματα εκπαίδευσης, ο λόγος k/n συγκλίνει πιθανοτικά, όπως είναι επιθυμητό, αλλά τότε η τιμή της p(x) που υπολογίζεται δεν αποτελεί τίποτα παραπάνω από μία προσέγγιση του μέσου όρου της τιμής της,

$$\frac{P}{V} = \frac{\int_R p(x')dx'}{\int_R dx'} \tag{4.6}$$

Εάν πρέπει να υπολογιστεί η ακριβής τιμή της p(x) και όχι μόνο μια προσέγγιση του μέσου όρου της θα πρέπει ο όγκος V να μπορεί να τείνει στο μηδέν. Παρ' όλ' αυτά, εάν καθοριστεί ο αριθμός των δειγμάτων n και το V επιτρέπεται να τείνει στο μηδέν, η περιοχή θα γίνει τόσο μικρή που στην πραγματικότητα δεν θα περιλαμβάνει κανένα δείγμα και η προσέγγιση της p(x) θα είναι άχρηστη αφού θα ισχύει ότι $p(x) \approx 0$. Επίσης, εάν κατά τύχη ένα ή περισσότερα από τα δείγματα εκπαίδευσης πέφτουν πάνω στο x, η προσέγγιση τείνει στο άπειρο και επομένως είναι ομοίως άχρηστη.

Πρακτικά, ο αριθμός των δειγμάτων είναι πάντα πεπερασμένος. Έτσι, ο όγκος V δεν μπορεί να είναι οσοδήποτε μικρός. Εάν πρόκειται να χρησιμοποιηθεί αυτός ο υπολογισμός, θα πρέπει να ληφθεί υπόψη ένα συγκεκριμένο ποσοστό διακύμανσης στο λόγο k/n και ένα συγκεκριμένο ποσοστό μέσου όρου στην συνάρτηση πυκνότητας πιθανότητας p(x).

Από θεωρητικής άποψης, είναι ενδιαφέρον να βρεθεί πως αυτοί οι περιορισμοί μπορούν να παρακαμφθούν εάν θεωρηθεί ότι ο αριθμός των δειγμάτων τείνει στο άπειρο. Έστω ότι ακολουθείται η παρακάτω διαδικασία. Για τον υπολογισμό της p(x) στο x, καθορίζεται μία ακολουθία από περιοχές R_1, R_2, …, που περιέχουν το x – η πρώτη περιοχή για να χρησιμοποιηθεί με ένα δείγμα, η δεύτερη με δύο και ούτω καθεξής. Έστω ότι με V_n συμβολίζεται ο όγκος της R_n, ότι k_n είναι ο αριθμός των δειγμάτων που περιέχονται στην περιοχή R_n και ότι με $p_n(x)$ συμβολίζεται η n-οστή προσέγγιση της p(x):

$$p_n(x) = \frac{k_n/n}{V_n} \tag{4.7}$$

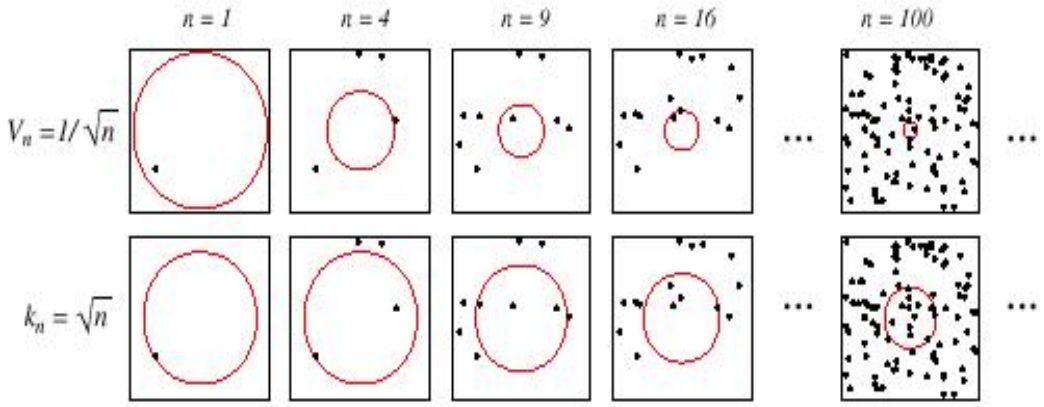
Για να συγκλίνει η $p_n(x)$ στην p(x), πρέπει να ισχύουν οι τρεις παρακάτω συνθήκες:

- $\lim_{n\to\infty} V_n = 0$
- $\lim_{n\to\infty} k_n = \infty$
- $\lim_{n\to\infty} k_n/n = 0$

Η πρώτη συνθήκη εγγυάται ότι ο ως προς το χώρο μέσος όρος του λόγου P/V τείνει στη p(x), δεδομένου ότι οι περιοχές μικραίνουν ομοιόμορφα και ότι η p(·) είναι συνεχής στο x. Η δεύτερη συνθήκη, η οποία έχει νόημα μόνο όταν $p(x)\neq 0$, εγγυάται ότι ο λόγος συχνότητας τείνει πιθανοτικά στην πιθανότητα P. Η τρίτη συνθήκη είναι προφανώς απαραίτητη στην περίπτωση που η $p_n(x)$, που δίνεται από την εξίσωση 4.7, συγκλίνει συνολικά. Εκφράζει επίσης το γεγονός ότι εάν και ένας τεράστιος αριθμός δειγμάτων θα περιέχεται σε μία μικρή περιοχή R_n, τα δείγματα αυτά θα αποτελούν μόλις ένα μικρό κλάσμα του συνολικού αριθμού των δειγμάτων.

Υπάρχουν δύο γνωστοί τρόποι για τη δημιουργία ακολουθιών από περιοχές που ικανοποιούν αυτές τις συνθήκες (εικόνα 4.2). Ο πρώτος είναι να μειώνεται διαρκώς μια αρχική περιοχή, ορίζοντας τον όγκο V_n ως μια συνάρτηση του n, όπως η $V_n = 1/\sqrt{n}$. Στη συνέχεια πρέπει να αποδειχθεί ότι οι τυχαίες μεταβλητές k_n και k_n/n συμπεριφέρονται σωστά, δηλαδή ότι η $p_n(x)$ τείνει στην p(x). Αυτή η μέθοδος είναι στην πραγματικότητα η τεχνική των παραθύρων Parzen η οποία θα περιγραφεί με λεπτομέρεια στην ενότητα 4.3. Η δεύτερη μέθοδος προσδιορίζει το k_n ως μια συνάρτηση του n, όπως η $k_n = \sqrt{n}$ Ο όγκος V_n αυξάνεται μέχρις ότου να περιέχει k_n

γείτονες του x. Αυτή αποτελεί την μέθοδο υπολογισμού των k_n κοντινότερων γειτόνων. Στην πραγματικότητα και οι δύο αυτές μέθοδοι συγκλίνουν, αν και είναι δύσκολο να γίνουν σημαντικές θεωρήσεις για την συμπεριφορά του τελικού δείγματός τους.



Εικόνα 4.2: Υπάρχουν δύο βασικές μέθοδοι για τον υπολογισμό της πυκνότητας σε ένα σημείο, στην περίπτωση που εξετάζεται, στο κέντρο κάθε τετραγώνου. Η πρώτη είναι να ξεκινήσει κανείς με ένα μεγάλο όγκο με κέντρο το ίδιο το σημείο και να τον ελαττώνει με βάση μια συνάρτηση όπως η $V_n = 1/\sqrt{n}$. Η άλλη μέθοδος είναι να ελαττώνεται ο όγκος με βάση ένα βασισμένο στα δεδομένα τρόπο, για παράδειγμα με το να πρέπει ο όγκος να περιλαμβάνει κάποιον αριθμό $k_n = \sqrt{n}$ από δείγματα. Οι ακολουθίες και στις δύο περιπτώσεις αντιπροσωπεύουν τυχαίες μεταβλητές που γενικά συγκλίνουν και επιτρέπουν τον υπολογισμό της πραγματικής τιμής της πυκνότητας στο σημείο.

4.3 Παράθυρα Parzen

Η περιγραφή της μεθόδου των παραθύρων Parzen για τον υπολογισμό των συναρτήσεων πυκνότητας πιθανότητας θα γίνει θεωρώντας προσωρινά ότι η περιοχή R_n είναι ένας d-διάστατος υπερκύβος. Εάν με h_n συμβολίζεται το μήκος μιας πλευράς αυτού του υπερκύβου, ο όγκος δίνεται από τον τύπο

$$V_n = h_n^d \tag{4.8}$$

Μπορεί να βρεθεί μια αναλυτική έκφραση για το k_n, δηλαδή τον αριθμό των δειγμάτων που περιέχονται στον υπερκύβο, ορίζοντας την ακόλουθη συνάρτηση παραθύρου:

$$\varphi(u) = \begin{cases} 1 & |u_j| \le 1/2 \quad j = 1,...,d \\ 0 & \textit{διαφορετικά} \end{cases} \tag{4.9}$$

Η φ(u) ορίζει έναν μοναδιαίο υπερκύβο με κέντρο την αρχή των αξόνων. Προφανώς, η έκφραση $\varphi((x - x_i)/h_n)$ είναι ίση με τη μονάδα εάν το x_i περιέχεται μέσα στον υπερκύβο όγκου V_n που έχει κέντρο το x και ίση με μηδέν οπουδήποτε αλλού. Ο αριθμός των δειγμάτων που περιέχονται σε αυτόν τον υπερκύβο δίνεται από τον τύπο

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h_n}\right) \tag{4.10}$$

και αντικαθιστώντας το παραπάνω στην εξίσωση 4.7 προκύπτει

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right) \tag{4.11}$$

Η παραπάνω εξίσωση παρουσιάζει μία γενικότερη προσέγγιση στον τρόπο υπολογισμού της συνάρτησης πυκνότητας πιθανότητας. Ας υποτεθεί τώρα ότι αντί για τη συνάρτηση παραθύρου του υπερκύβου (εξίσωση 4.9) χρησιμοποιείται μία πιο γενική κατηγορία συναρτήσεων παραθύρου. Σε μια τέτοια περίπτωση, η εξίσωση 4.11 εκφράζει την προσέγγιση για την p(x) ως ένα μέσο όρο των συναρτήσεων του x και των δειγμάτων x_i. Στην πραγματικότητα, η συνάρτηση παραθύρου χρησιμοποιείται για παρεμβολή – κάθε δείγμα συνεισφέρει στον υπολογισμό ανάλογα με την απόστασή του από το x.

Η $p_n(x)$ πρέπει να είναι μία νομιμοποιημένη συνάρτηση πυκνότητας πιθανότητας, δηλαδή να είναι μη αρνητική και το ολοκλήρωμά της να ισούται με τη μονάδα. Πιο συγκεκριμένα, εάν απαιτηθεί ότι

$$\varphi(x) \geq 0 \tag{4.12}$$

και

$$\int \varphi(u)du = 1 \tag{4.13}$$

και εάν διατηρηθεί η σχέση $V_n = h_n^d$, τότε προκύπτει άμεσα ότι και η $p_n(x)$ θα ικανοποιεί αυτές τις συνθήκες.

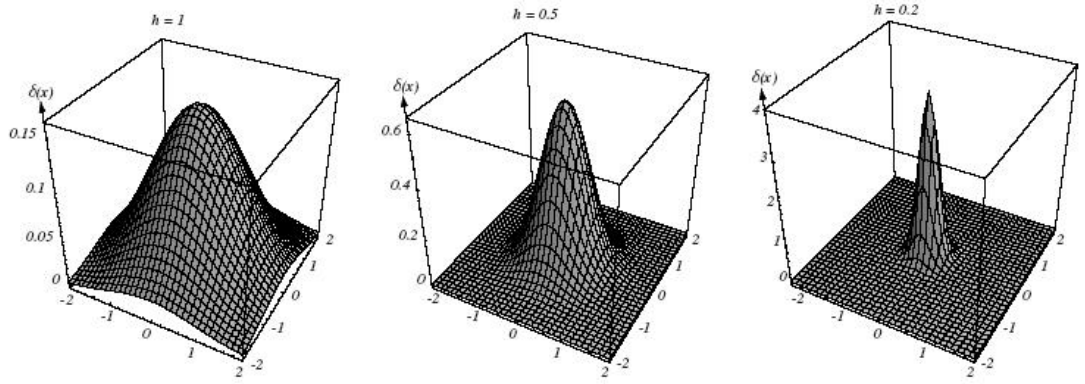
Ας εξεταστεί τώρα το αποτέλεσμα που έχει το πλάτος του παραθύρου h_n στην $p_n(x)$. Ας οριστεί πρώτα η συνάρτηση $\delta_n(x)$ ως εξής

$$\delta_n(x) = \frac{1}{V_n}\varphi\left(\frac{x}{h_n}\right) \tag{4.14}$$

τότε η $p_n(x)$ μπορεί να γραφεί ως ο μέσος όρος

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_n(x - x_i) \tag{4.15}$$

Επειδή $V_n = h_n^d$, το h_n επηρεάζει τόσο το εύρος όσο και το ύψος της $\delta_n(x)$ (εικόνα 4.3).



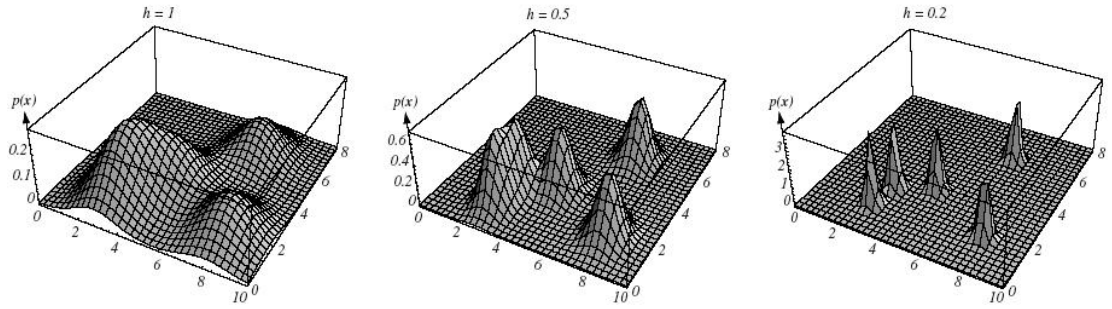
Εικόνα 4.3: Παραδείγματα δυσδιάστατων κυκλικά συμμετρικών κανονικών παραθύρων Parzen για τρείς διαφορετικές τιμές του h. Σημειώνεται ότι επειδή τα δ(x) είναι κλιμακούμενα πρέπει να χρησιμοποιηθούν διαφορετικές κατακόρυφες κλίμακες για να φανεί η δομή τους.

Εάν το h_n είναι πολύ μεγάλο τότε το ύψος της δ_n είναι μικρό και το x πρέπει να βρίσκεται μακριά από το x_i προτού το $\delta_n(x-x_i)$ μεταβληθεί σημαντικά από το $\delta_n(0)$. Σε αυτή την περίπτωση, η $p_n(x)$ είναι η υπερπροβολή n ευρέων, αργά μεταβαλλόμενων συναρτήσεων και αποτελεί ένα ομαλό "εκτός εμβέλειας" υπολογισμό της p(x). Από την άλλη πλευρά, εάν το h_n είναι πολύ μικρό, τότε η ακρότατη τιμή της $\delta_n(x-x_i)$ είναι μεγάλη και εμφανίζεται κοντά στο $x = x_i$. Στην περίπτωση αυτή η p(x) είναι η υπερπροβολή n απότομων παλμών με κέντρο τα δείγματα – ένας ακανόνιστος με θόρυβο υπολογισμός (εικόνα 4.4). Για οποιαδήποτε τιμή του h_n η κατανομή κανονικοποιείται, δηλαδή

$$\int \delta_n(x - x_i)dx = \int \frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right)dx = \int \varphi(u)du = 1 \tag{4.16}$$

Έτσι, εάν το h_n τείνει στο μηδέν, το $\delta_n(x-x_i)$ προσεγγίζει μία Dirac δέλτα συνάρτηση με κέντρο το x_i, και η $p_n(x)$ προσεγγίζει μία υπερπροβολή από συναρτήσεις δέλτα με κέντρα τα δείγματα.

Προφανώς, η επιλογή της τιμής του h_n (ή του V_n) επηρεάζει σημαντικά την $p_n(x)$. Εάν ο V_n είναι πολύ μεγάλος, ο υπολογισμός θα έχει πολύ μικρή ανάλυση. Εάν ο V_n είναι πολύ μικρός, ο υπολογισμός θα εμφανίζει πολύ μεγάλη στατιστική ποικιλία. Με περιορισμένο αριθμό δειγμάτων, το καλύτερο που μπορεί να γίνει είναι η εύρεση κάποιου αποδεκτού συμβιβασμού. Πάντως, στην περίπτωση απεριόριστου αριθμού δειγμάτων, είναι δυνατόν το V_n να προσεγγίζει το μηδέν καθώς το n αυξάνεται και έτσι η $p_n(x)$ να συγκλίνει στην άγνωστη συνάρτηση πυκνότητας πιθανότητας p(x).



Εικόνα 4.4: Τρεις υπολογισμοί πυκνοτήτων μέσω παραθύρων Parzen βασισμένοι στο ίδιο σύνολο πέντε δειγμάτων, που χρησιμοποιούν τις συναρτήσεις παραθύρου του εικόνατος 4.3. Όπως και πριν οι κατακόρυφοι άξονες έχουν διαφορετική κλίμακα για να φανεί η δομή κάθε κατανομής.

Αυτή η σύγκλιση αναφέρεται στη σύγκλιση μιας ακολουθίας από τυχαίες μεταβλητές, επειδή για κάθε συγκεκριμένο x η τιμή της $p_n(x)$ εξαρτάται από τα τυχαία δείγματα $x_1, \ldots, x_n$. Έτσι, η $p_n(x)$ έχει προφανώς μια μέση τιμή $\bar{p}_n(x)$ και μια διασπορά $\sigma_n^2(x)$. Ο υπολογισμός $p_n(x)$ συγκλίνει στην p(x) εάν

$$\lim_{n\to\infty} \bar{p}_n(x) = p(x) \tag{4.17}$$

και

$$\lim_{n\to\infty} \sigma_n^2(x) = 0 \tag{4.18}$$

Για να αποδειχθεί η σύγκλιση πρέπει να τεθούν περιορισμοί στην άγνωστη συνάρτηση πυκνότητας πιθανότητας p(x), στη συνάρτηση παραθύρου φ(u) και στο

πλάτος του παραθύρου h_n. Γενικότερα, απαιτείται συνέχεια της $p(\cdot)$ στο x και πρέπει να ισχύουν οι συνθήκες των εξισώσεων 4.12 και 4.13. Οι παρακάτω επιπλέον συνθήκες εγγυώνται τη σύγκλιση:

$$\sup_u \varphi(u) < \infty \tag{4.19}$$

$$\lim_{n\to\infty} \varphi(u)\prod_{i=1}^{d} u_i = 0 \tag{4.20}$$

$$\lim_{n\to\infty} V_n = 0 \tag{4.21}$$

και

$$\lim_{n\to\infty} nV_n = \infty \tag{4.22}$$

Οι εξισώσεις 4.19 και 4.20 φροντίζουν για την «καλή συμπεριφορά» της $\varphi(\cdot)$ και μπορούν να ικανοποιηθούν από τις περισσότερες συναρτήσεις πυκνότητας πιθανότητας που μπορούν να χρησιμοποιηθούν για συναρτήσεις παραθύρου. Οι εξισώσεις 4.21 και 4.22 δηλώνουν ότι ο όγκος V_n πρέπει να τείνει στο μηδέν, αλλά με ρυθμό μικρότερο από 1/n. Στις επόμενες υποενότητες θα αποδειχθεί γιατί οι παραπάνω συνθήκες αποτελούν τις βασικές συνθήκες για σύγκλιση.

4.3.1 Σύγκλιση της Μέσης Τιμής

Έστω ότι με $\bar{p}_n(x)$ συμβολίζεται η μέση τιμή της $p_n(x)$. Επειδή τα δείγματα x_i είναι στατιστικά ανεξάρτητα σε σχέση με την (άγνωστη) συνάρτηση πυκνότητας πιθανότητας p(x) ισχύει

$$\begin{aligned}\bar{p}_n(x) &= E[p_n(x)] \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{V_n}\varphi\left(\frac{x - x_i}{h_n}\right)\right] \\ &= \int \frac{1}{V_n}\varphi\left(\frac{x - v}{h_n}\right)p(v)dv \\ &= \int \delta_n(x - v)p(v)dv \end{aligned} \tag{4.23}$$

Η εξίσωση αυτή δείχνει ότι η αναμενόμενη τιμή του υπολογισμού είναι ένας μέσος όρος της άγνωστης συνάρτησης πυκνότητας πιθανότητας – μία συνέλιξη της άγνωστης συνάρτησης πυκνότητας πιθανότητας και της συνάρτησης παραθύρου. Έτσι, η $\bar{p}_n(x)$ είναι μία θολή εκδοχή της p(x) όπως φαίνεται μέσα από ένα παράθυρο μέσου όρου. Όμως, καθώς ο V_n πλησιάζει το μηδέν, το $\delta_n(x\text{-}v)$ προσεγγίζει μία συνάρτηση δέλτα με κέντρο το x. Έτσι, εάν η p είναι συνεχής στο x, η εξίσωση 4.21 εγγυάται ότι η $\bar{p}_n(x)$ θα προσεγγίσει την p(x) καθώς το n τείνει στο άπειρο.

4.3.2 Σύγκλιση της Διασποράς

Η εξίσωση 4.23 δείχνει ότι δεν απαιτείται απεριόριστος αριθμός δειγμάτων για να προσεγγίζει η $\bar{p}_n(x)$ την p(x). Αυτό μπορεί να επιτευχθεί για οποιοδήποτε n αφήνοντας το V_n να τείνει στο μηδέν. Φυσικά, για ένα συγκεκριμένο σύνολο n δειγμάτων ο «ακιδωτός» υπολογισμός που προκύπτει είναι άχρηστος. Αυτό δείχνει ότι πρέπει να ληφθεί σοβαρά υπόψη η διακύμανση του υπολογισμού. Επειδή η $p_n(x)$ αποτελεί το άθροισμα συναρτήσεων με στατιστικά ανεξάρτητες τυχαίες μεταβλητές, η διασπορά της είναι το άθροισμα των διασπορών των διαφορετικών όρων και επομένως

$$\sigma_n^2(x) = \sum_{i=1}^{n} E\left[\left(\frac{1}{nV_n}\varphi\left(\frac{x - x_i}{h_n}\right) - \frac{1}{n}\bar{p}_n(x)\right)^2\right]$$

$$= nE\left[\frac{1}{n^2V_n^2}\varphi^2\left(\frac{x - x_i}{h_n}\right) - \frac{1}{n}\bar{p}_n^2(x)\right]$$

$$= \frac{1}{nV_n}\int \frac{1}{V_n}\varphi^2\left(\frac{x - v}{h_n}\right)p(v)dv - \frac{1}{n}\bar{p}_n^2(x) \tag{4.24}$$

Αφαιρώντας το δεύτερο όρο, φράζοντας την φ(·) και χρησιμοποιώντας την εξίσωση 4.23 προκύπτει το εξής

$$\sigma_n^2(x) \le \frac{\sup(\varphi(\cdot))\bar{p}_n(x)}{nV_n} \tag{4.25}$$

Προφανώς, για να επιτευχθεί μια μικρή τιμή για τη διασπορά χρειάζεται μια μεγάλη τιμή για τον V_n. Μια τέτοια τιμή ελαττώνει τις τοπικές μεταβολές της συνάρτησης πυκνότητας πιθανότητας. Παρ' όλ' αυτά, επειδή ο απαριθμητής παραμένει σταθερός καθώς το n πλησιάζει το άπειρο, ο V_n μπορεί να τείνει στο μηδέν και πάλι να προκύπτει μηδενική διασπορά, με την προϋπόθεση ότι το $n \cdot V_n$ θα τείνει στο άπειρο. Για παράδειγμα, ο V_n μπορεί να είναι ίσος με $V_1/\sqrt{n}$, $V_1/\ln n$ ή οποιαδήποτε άλλη που ικανοποιεί τις εξισώσεις 4.21 και 4.22.

Το αποτέλεσμα αυτό εάν και αποτελεί το κύριο θεωρητικό αποτέλεσμα δεν αναφέρει τίποτα όσον αφορά την επιλογή της φ(·) και του V_n για να επιτευχθούν καλά αποτελέσματα στην περίπτωση πεπερασμένων δειγμάτων. Πραγματικά, εάν δεν υπάρχει περισσότερη γνώση για την p(x), εκτός από το ότι πρέπει να είναι συνεχής, δεν υπάρχει συγκεκριμένη μεθοδολογία για τη βελτιστοποίηση των αποτελεσμάτων πεπερασμένων δειγμάτων.

4.3.3 Εφαρμογές

Είναι ενδιαφέρον να εξεταστεί η συμπεριφορά της μεθόδου των παραθύρων Parzen σε κάποια απλά παραδείγματα και ιδιαίτερα το αποτέλεσμα που έχει η επιλογή της συνάρτησης παραθύρου. Αρχικά, θεωρείται η περίπτωση όπου η p(x) είναι μία κανονική συνάρτηση πυκνότητας πιθανότητας μιας μεταβλητής με μηδενική μέση τιμή και μοναδιαία διασπορά. Έστω ότι η συνάρτηση παραθύρου είναι της ίδιας μορφής:

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2} \tag{4.26}$$

Επιπλέον, έστω ότι $h_n = h_1/\sqrt{n}$ όπου το h_1 είναι μια παράμετρος επιλογής του χρήστη. Η $p_n(x)$ αποτελεί ένα μέσο όρο κανονικών συναρτήσεων πυκνότητας πιθανότητας με κέντρο στα δείγματα:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right) \tag{4.27}$$

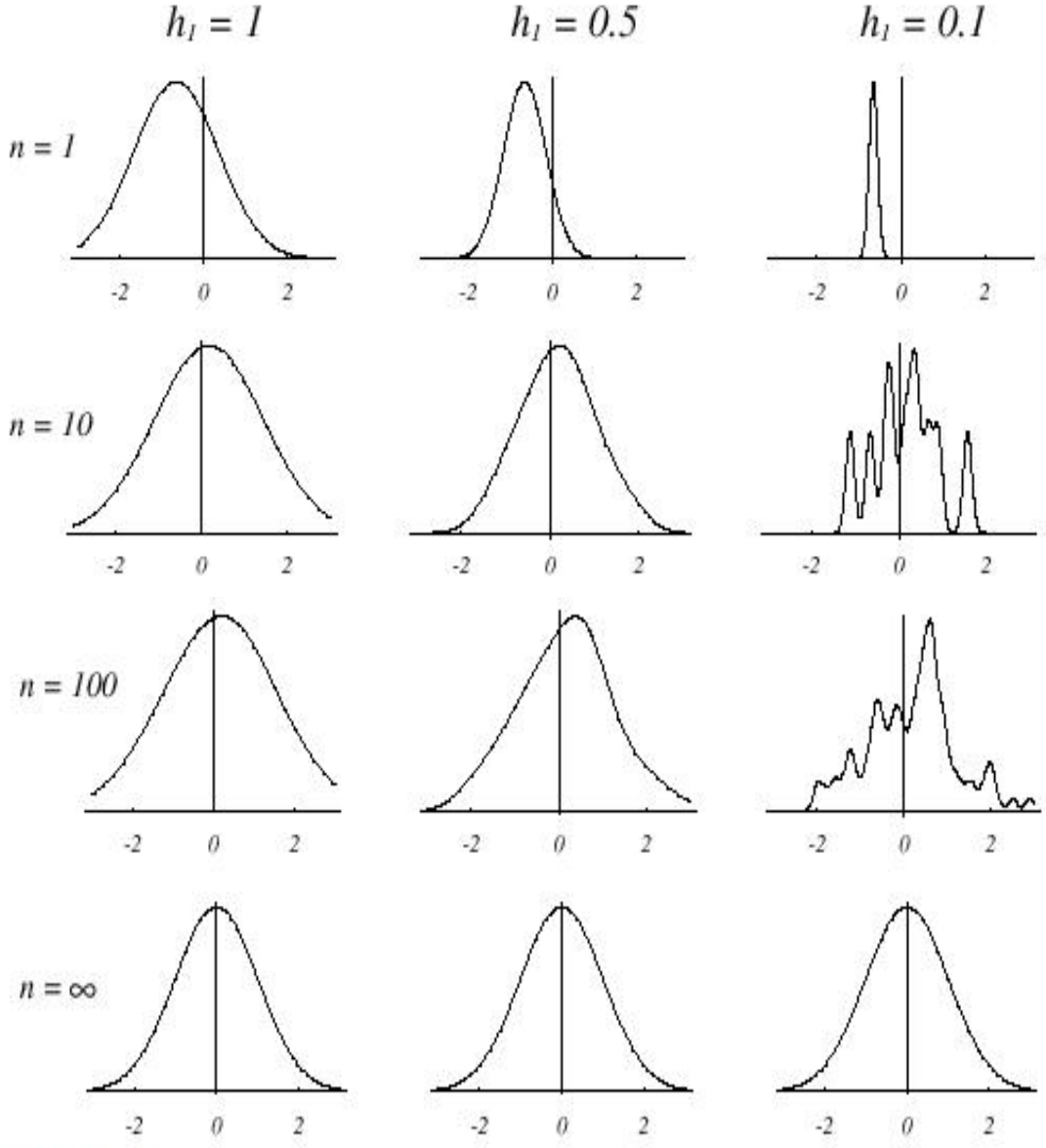
Εάν και δεν είναι δύσκολος ο υπολογισμός των εξισώσεων 4.23 και 4.24 για την εύρεση της μέσης τιμής και της διασποράς της $p_n(x)$, είναι πραγματικά ενδιαφέρον να παρουσιαστούν κάποια αριθμητικά αποτελέσματα. Τα αποτελέσματα που παρουσιάζονται στην εικόνα 4.5 προέρχονται από ένα σύνολο τυχαίων δειγμάτων κανονικής κατανομής τα οποία δημιουργήθηκαν και χρησιμοποιήθηκαν για τον υπολογισμό της $p_n(x)$. Τα αποτελέσματα αυτά εξαρτώνται τόσο από το n όσο και από

το h_1. Για n=1, η $p_n(x)$ μοιάζει περισσότερο με μία Gaussian κατανομή με κέντρο το πρώτο δείγμα η οποία όμως δεν έχει ούτε τη μέση τιμή ούτε τη διασπορά της πραγματικής κατανομής. Για n=10 και h_1=0.1, οι κατανομές των δειγμάτων είναι αρκετά ευδιάκριτες. Αυτό δεν ισχύει για h_1=1 και h_1=0.5. Όσο αυξάνεται το n η ικανότητα της $p_n(x)$ να αντιμετωπίσει τις μεταβολές στην p(x) μεγαλώνει. Αν και φαίνεται ότι η $p_n(x)$ συγκλίνει στην λεία κανονική καμπύλη καθώς το n τείνει στο άπειρο, είναι περισσότερο ευαίσθητη σε τοπικές δειγματοληπτικές ανωμαλίες για μεγάλες τιμές του n. Σύμφωνα λοιπόν με ότι φαίνεται στην εικόνα 4.5 απαιτείται μεγάλος αριθμός δειγμάτων για να επιτευχθεί ένας ακριβής υπολογισμός. Στην εικόνα 4.6 φαίνονται τα αντίστοιχα αποτελέσματα για την περίπτωση των δύο διαστάσεων.
Στο δεύτερο μονοδιάστατο παράδειγμα, η φ(x) και το h_n είναι ίδια με την εικόνα 4.5, αλλά η άγνωστη συνάρτηση πυκνότητας πιθανότητας είναι ένα μείγμα μιας κανονικής και μιας τριγωνικής συνάρτησης πυκνότητας πιθανότητας. Όπως και στο πρώτο παράδειγμα, η περίπτωση όπου n=1 δίνει περισσότερες πληροφορίες για τη συνάρτηση παραθύρου από ότι για την άγνωστη συνάρτηση πυκνότητας πιθανότητας. Για n=16, κανένας υπολογισμός δε φαίνεται να είναι ικανοποιητικός, ενώ για n=256 και h_1=1 τα αποτελέσματα αρχίζουν να γίνονται αποδεκτά.

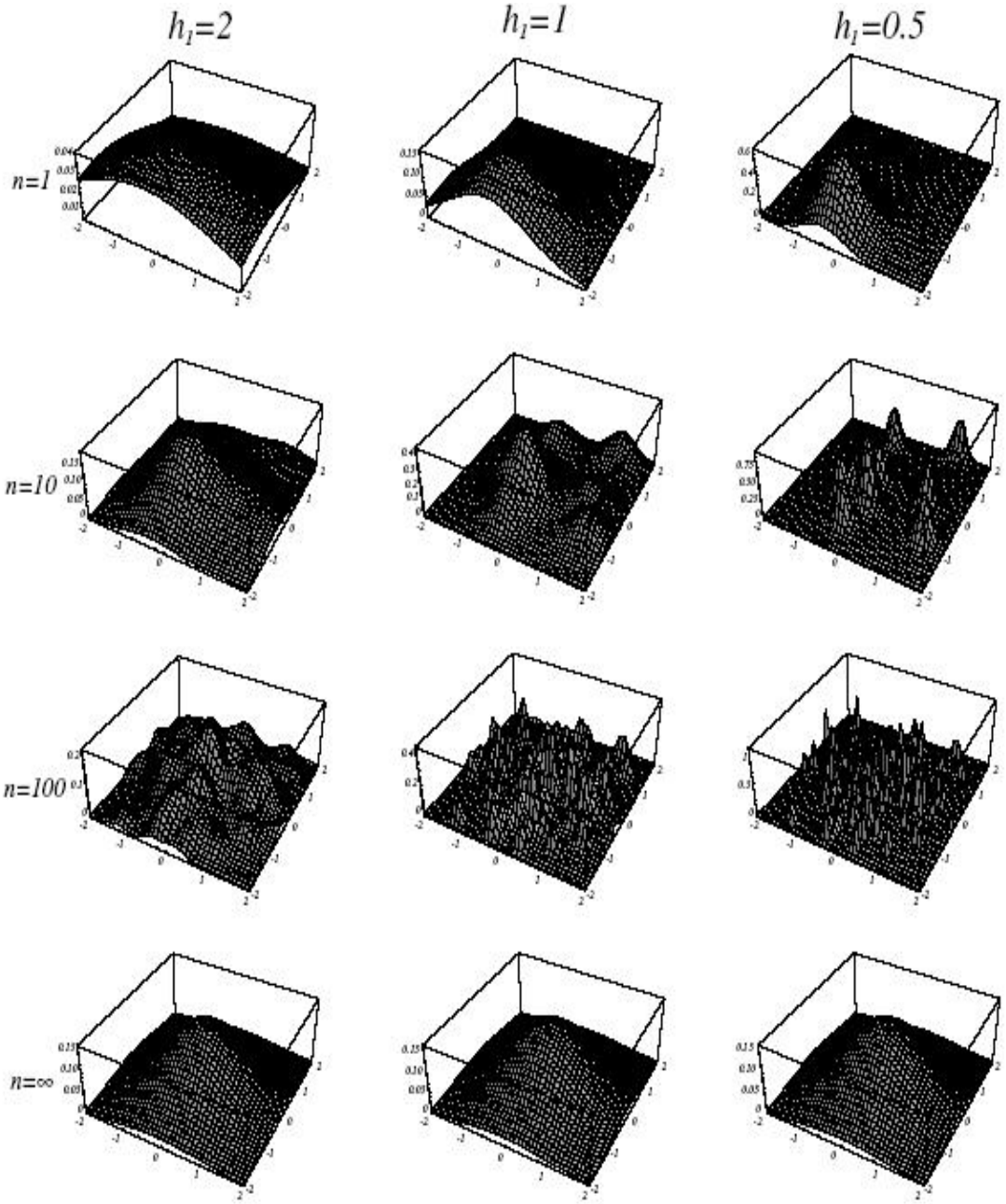
4.3.4 Παράδειγμα Ταξινόμησης

Στους ταξινομητές που βασίζονται σε υπολογισμούς μέσω παραθύρων Parzen, υπολογίζονται οι συναρτήσεις πυκνότητας πιθανότητας για κάθε κατηγορία και ταξινομείται ένα σημείο ελέγχου με βάση την ετικέτα που αντιστοιχεί στη μέγιστη εκ των υστέρων πιθανότητα. Εάν υπάρχουν πολλές κατηγορίες με άνισες εκ των υστέρων πιθανότητες μπορούν και αυτές κάλλιστα να συμπεριληφθούν. Οι περιοχές απόφασης για ένα ταξινομητή παραθύρων Parzen εξαρτώνται από την επιλογή της συνάρτησης παραθύρου, όπως φαίνεται στην εικόνα 4.8. Γενικότερα, το λάθος εκπαίδευσης, αυτό δηλαδή που υπολογίζεται με βάση τα επιλεγμένα σημεία εκπαίδευσης, μπορεί να γίνει αυθαίρετα μικρό εάν το πλάτος του παραθύρου γίνει αρκετά μικρό. Παρ' όλ' αυτά, ο βασικός σκοπός του σχεδιασμού και της δημιουργίας ενός ταξινομητή παραμένει η ταξινόμηση νέων προτύπων και ένα χαμηλό λάθος εκπαίδευσης δεν εγγυάται ένα αντίστοιχο μικρό λάθος ελέγχου. Αν και θα μπορούσε να χρησιμοποιηθεί ένα γενικής μορφής Gaussian παράθυρο, από τη στιγμή που δεν υπάρχει καμία άλλη πληροφορία για τις αντίστοιχες κατανομές δεν υπάρχει κάποιο θεωρητικό υπόβαθρο με βάση το οποίο να μπορεί να αποφασιστεί η «ανωτερότητα» ενός πλάτους παραθύρου σε σχέση με κάποιο άλλο. Αυτά τα παραδείγματα υπολογισμού συναρτήσεων πυκνότητας πιθανότητας και ταξινόμησης δείχνουν τα πλεονεκτήματα και τους περιορισμούς των μη παραμετρικών μεθόδων. Η πραγματική τους δύναμη είναι η γενικότητά τους. Όπως παρουσιάστηκε, χρησιμοποιείται η ίδια ακριβώς διαδικασία τόσο για την περίπτωση κανονικής κατανομής μιας μεταβλητής όσο και για την περίπτωση σύνθετης κατανομής δύο μεταβλητών. Επίσης δεν απαιτείται να γίνουν προβλέψεις για τις μελλοντικές τιμές των κατανομών. Εάν υπάρχουν αρκετά δείγματα, η σύγκλιση σε οποιαδήποτε συνάρτηση είναι σίγουρη ανεξάρτητα από την πολυπλοκότητά της. Από την άλλη πλευρά, ο αριθμός των δειγμάτων που μπορεί να απαιτηθεί υπάρχει περίπτωση να είναι πάρα πολύ μεγάλος, πολύ μεγαλύτερος προφανώς από εκείνον που θα χρειαζόταν εάν η μορφή της συνάρτησης πυκνότητας πιθανότητας ήταν γνωστή. Αυτό οδηγεί σε τεράστιες απαιτήσεις υπολογιστικού χρόνου και αποθηκευτικού χώρου. Επιπλέον, ο αριθμός των δειγμάτων που απαιτούνται αυξάνεται εκθετικά ως προς τη διάσταση του χώρου των χαρακτηριστικών. Αυτός ο περιορισμός είναι ευρύτερα γνωστός ως «η κατάρα

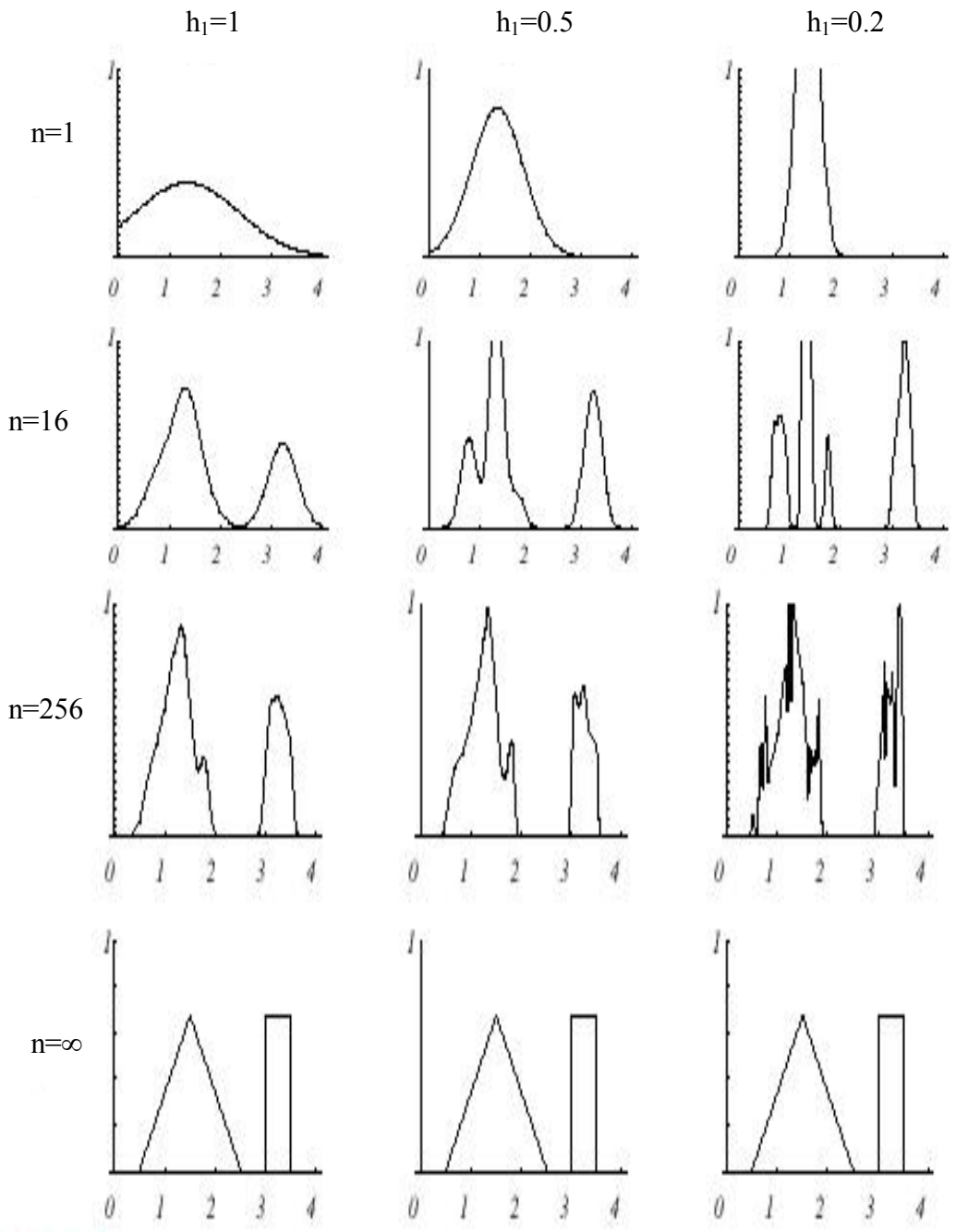
της διάστασης» και περιορίζει σημαντικά την πρακτική εφαρμογή τέτοιων μη παραμετρικών μεθόδων. Ο βασικός λόγος ύπαρξης αυτού του περιορισμού είναι ότι οι υψηλής διάστασης συναρτήσεις έχουν πολύ μεγαλύτερη πιθανότητα από τις χαμηλής διάστασης συναρτήσεις να είναι πολύπλοκες και να απαιτούν μεγάλο υπολογιστικό φόρτο. Ένας τρόπος για να αντιμετωπιστεί αυτός ο περιορισμός είναι να ενσωματωθεί στη μέθοδο εκ των προτέρων γνώση για τα δεδομένα, δηλαδή διαφόρων ειδών πληροφορίες που χαρακτηρίζουν τα δείγματα παρατήρησης.



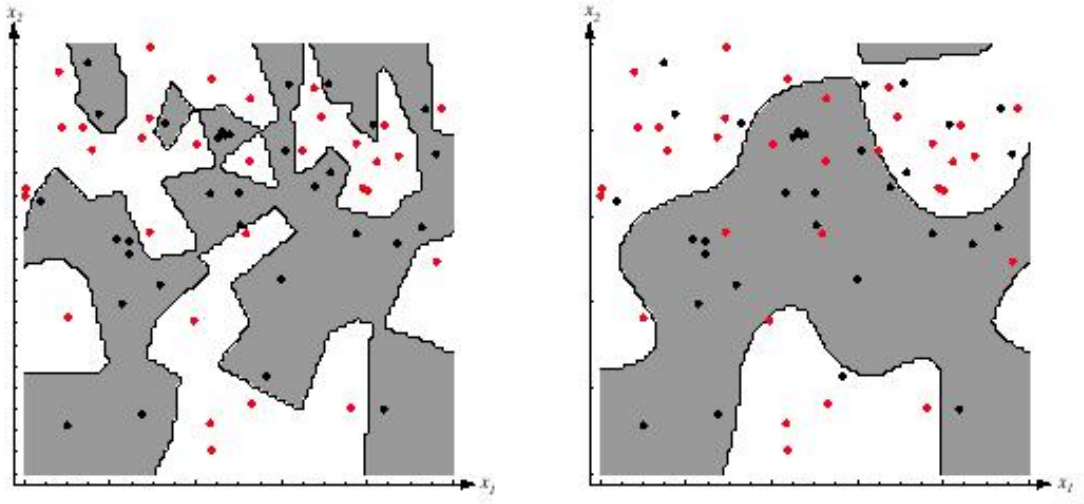
Εικόνα 4.5: Υπολογισμοί μέσω παραθύρων Parzen μιας κανονικής πυκνότητας μιας μεταβλητής χρησιμοποιώντας διαφορετικά πλάτη παραθύρων και αριθμούς δειγμάτων. Οι κατακόρυφοι άξονες έχουν διαφορετική κλίμακα ώστε να δείχνουν όσο το δυνατόν καλύτερα τη μορφή κάθε γραφικής παράστασης. Σημειώνεται ιδιαίτερα ότι οι υπολογισμοί για n = ∞ είναι ίδιοι (και βρίσκουν την πραγματική συνάρτηση πυκνότητας πιθανότητας) ανεξάρτητα από το πλάτος του παραθύρου.



Εικόνα 4.6: Υπολογισμοί μέσω παραθύρων Parzen μιας κανονικής πυκνότητας δύο μεταβλητών χρησιμοποιώντας διαφορετικά πλάτη παραθύρων και αριθμούς δειγμάτων. Οι κατακόρυφοι άξονες έχουν διαφορετική κλίμακα ώστε να δείχνουν όσο το δυνατόν καλύτερα τη μορφή κάθε γραφικής παράστασης. Σημειώνεται ιδιαίτερα ότι οι υπολογισμοί για n = ∞ είναι ίδιοι (και βρίσκουν την πραγματική συνάρτηση πυκνότητας πιθανότητας) ανεξάρτητα από το πλάτος του παραθύρου.



Εικόνα 4.7: Υπολογισμοί μέσω παραθύρων Parzen μιας κατανομής με δύο ακρότατα χρησιμοποιώντας διαφορετικά πλάτη παραθύρων και αριθμούς δειγμάτων. Οι κατακόρυφοι άξονες έχουν διαφορετικές κλίμακες ώστε να δείχνουν όσο το δυνατόν καλύτερα τη μορφή κάθε γραφικής παράστασης. Σημειώνεται ιδιαίτερα ότι οι υπολογισμοί για n = ∞ είναι ίδιοι (και βρίσκουν την πραγματική συνάρτηση πυκνότητας πιθανότητας) ανεξάρτητα από το πλάτος του παραθύρου.



Εικόνα 4.8: Τα όρια απόφασης σε ένα δυσδιάστατο διχοτόμο παραθύρων Parzen εξαρτώνται από το πλάτος του παραθύρου h. Στην αριστερή εικόνα, μία μικρή τιμή για το h οδηγεί σε πολύπλοκα όρια σε σχέση με αυτά που δημιουργούνται για μια μεγάλη τιμή του h, στη δεξιά εικόνα. Προφανώς για το παραπάνω παράδειγμα, μία μικρή τιμή για το h θα ήταν ιδανική για την πάνω περιοχή, ενώ για την κάτω θα χρειαζόταν μια μεγάλη τιμή.

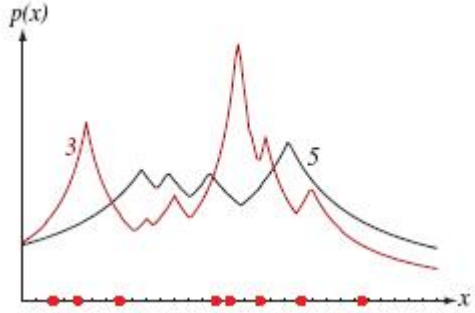
4.4 Μέθοδος Υπολογισμού k_n Πλησιέστερου Γείτονα

Μια πιθανή αντιμετώπιση του προβλήματος της εύρεσης της «βέλτιστης» συνάρτησης παραθύρου είναι να είναι ο όγκος του κελιού (cell) μία συνάρτηση των δεδομένων εκπαίδευσης αντί να είναι μια τυχαία συνάρτηση όλων των δειγμάτων. Για παράδειγμα, για τον υπολογισμό της p(x) από n δείγματα ή πρότυπα εκπαίδευσης μπορεί να τοποθετηθεί ένα κελί γύρω από το x και να μεγαλώνει μέχρις ότου περιλαμβάνει k_n δείγματα (όπου το k_n είναι μια καθορισμένη συνάρτηση του n). Αυτά τα δείγματα αποτελούν τους k_n πλησιέστερους γείτονες του x. Εάν η συνάρτηση πυκνότητας πιθανότητας έχει υψηλή τιμή γύρω από το x, το μέγεθος του κελιού θα είναι προφανώς μικρό, Εάν, από την άλλη, η τιμή της συνάρτησης πυκνότητας πιθανότητας είναι μικρή κοντά στο x, το κελί θα αυξήσει το μέγεθός του αλλά θα σταματήσει προφανώς όταν περιλάβει περιοχές με υψηλότερη πυκνότητα πιθανότητας. Σε κάθε περίπτωση, εάν

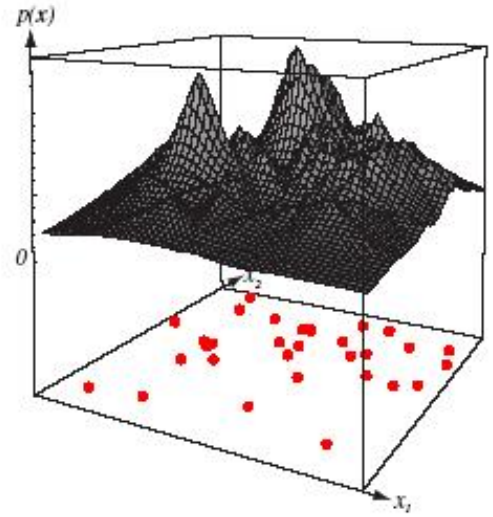
$$p_n(x) = \frac{k_n/n}{V_n} \tag{4.28}$$

το k_n θα πρέπει να τείνει στο άπειρο όταν το n τείνει στο άπειρο, αφού αυτή η συνθήκη εγγυάται ότι το k_n/n θα αποτελεί μια καλή προσέγγιση της πιθανότητας ότι ένα σημείο θα περιέχεται στο κελί όγκου V_n. Παρ' όλ' αυτά, το k_n πρέπει να αυξάνει σχετικά αργά, έτσι ώστε το μέγεθος του κελιού που απαιτείται για να περιέχονται σε αυτό k_n δείγματα εκπαίδευσης, να ελαχιστοποιείται εάν είναι δυνατόν στο μηδέν. Έτσι, είναι προφανές από την εξίσωση 4.28 ότι ο λόγος k_n/n πρέπει να τείνει στο μηδέν. Οι συνθήκες $\lim_{n\to\infty} k_n = \infty$ και $\lim_{n\to\infty} k_n/n = 0$ είναι ικανές και αναγκαίες για την $p_n(x)$ έτσι ώστε να τείνει πιθανοτικά στην p(x) για όλα τα σημεία στα οποία η p(x) είναι συνεχής. Εάν ισχύει $k_n = \sqrt{n}$ και θεωρηθεί ότι η $p_n(x)$ αποτελεί μία σχετικά καλή προσέγγιση της p(x), από την εξίσωση 4.28 προκύπτει ότι

$V_n \approx 1/\left(\sqrt{n}\,p(x)\right)$. Έτσι, ενώ ο V_n έχει ξανά τη μορφή $V_1/\sqrt{n}$, ο αρχικός όγκος V_1 καθορίζεται αυτή τη φορά από τη φύση των δεδομένων και όχι από κάποια τυχαία επιλογή. Είναι ενδιαφέρον να σημειωθεί ότι αν και η $p_n(x)$ είναι συνεχής, η κλίση της (gradient) δεν είναι. Επιπλέον, τα σημεία ασυνέχειας είναι σπάνια όμοια με τα σημεία των προτύπων (εικόνες 4.9 και 4.10).



Εικόνα 4.9: Έξι σημεία σε μία διάσταση και οι υπολογισμοί των πυκνοτήτων πιθανότητας σύμφωνα με τον κανόνα του k πλησιέστερου γείτονα, για k = 3 και k = 5. Σημειώνεται ότι οι ασυνέχειες στους υπολογισμούς βρίσκονται γενικά μακριά από τις θέσεις των σημείων των προτύπων.



Εικόνα 4.10: Ο υπολογισμός μιας δυσδιάστατης πυκνότητας πιθανότητας σύμφωνα με τον κανόνα του k πλησιέστερου γείτονα για k = 5. Σημειώνεται ότι οι ασυνέχειες στους υπολογισμούς βρίσκονται γενικά κατά μήκος γραμμών μακριά από τις θέσεις των σημείων των προτύπων.

4.4.1 Σύγκριση Μεθόδων Υπολογισμού k_n Πλησιέστερου Γείτονα και Παραθύρων Parzen

Είναι ενδιαφέρον να συγκριθεί η απόδοση της μεθόδου k_n Πλησιέστερου Γείτονα με τη μέθοδο των παραθύρων Parzen στα δεδομένα που χρησιμοποιήθηκαν στα προηγούμενα παραδείγματα. Για n=1 και $k_n = \sqrt{n} = 1$ ο υπολογισμός παίρνει τη μορφή

$$p_n(x) = \frac{1}{2|x - x_1|} \tag{4.29}$$

Αυτή η τιμή είναι προφανώς ένας κακός υπολογισμός για την p(x), ενώ το ολοκλήρωμά της προσεγγίζει το άπειρο. Όπως φαίνεται από την εικόνα 4.11, ο

υπολογισμός γίνεται σημαντικά καλύτερος όσο το n γίνεται μεγαλύτερο, εάν και το ολοκλήρωμά του συνεχίζει να τείνει στο άπειρο. Αυτό συμβαίνει διότι η $p_n(x)$ ποτέ δε μηδενίζεται, ακόμα και στην περίπτωση που κανένα δείγμα δεν περιλαμβάνεται μέσα σε κάποιο τυχαίο κελί ή παράθυρο. Αυτό το γεγονός, εάν και δε φαίνεται να έχει μεγάλη σημασία, μπορεί να αποκτήσει μεγάλη αξία σε χώρους υψηλότερων διαστάσεων. Όπως και στην προσέγγιση των παραθύρων Parzen, μπορεί να δημιουργηθεί μια ομάδα υπολογισμών θέτοντας $k_n = k_1\sqrt{n}$ και επιλέγοντας διαφορετικές τιμές για το k_1. Παρ' όλ' αυτά, εάν δεν υπάρχει κάποια επιπλέον πληροφορία, κάθε τιμή είναι τόσο καλή όσο και οποιαδήποτε άλλη, και σιγουριά ότι τα αποτελέσματα είναι σωστά υπάρχει μόνο στην περίπτωση άπειρων δεδομένων. Πρακτικά για ταξινόμηση, μια δημοφιλής μέθοδος είναι να προσαρμόζεται το πλάτος του παραθύρου μέχρι ο ταξινομητής να παρουσιάσει ελάχιστο λάθος για ένα ξεχωριστό σύνολο δειγμάτων, τα οποία προέρχονται επίσης από τις άγνωστες κατανομές.

4.4.2 Υπολογισμός των εκ των Υστέρων Πιθανοτήτων

Οι τεχνικές που παρουσιάστηκαν στις προηγούμενες ενότητες μπορούν να χρησιμοποιηθούν για τον υπολογισμό των εκ των υστέρων πιθανοτήτων $P(\omega_i/x)$ από ένα σύνολο n δειγμάτων (που έχουν ετικέτα, δηλαδή είναι γνωστή η κατηγορία στην οποία ανήκουν), τα οποία χρησιμοποιούνται επίσης για την εύρεση των αντίστοιχων συναρτήσεων πυκνότητας πιθανότητας. Έστω ότι τοποθετείται ένα κελί όγκου V γύρω από το x και καλύπτονται k δείγματα, k_i από τα οποία έχουν ετικέτα ω_i (δηλαδή ανήκουν στην κατηγορία ω_i). Τότε, ο προφανής υπολογισμός για την υπό συνθήκη πιθανότητα $p(x/\omega_i)$ προκύπτει από

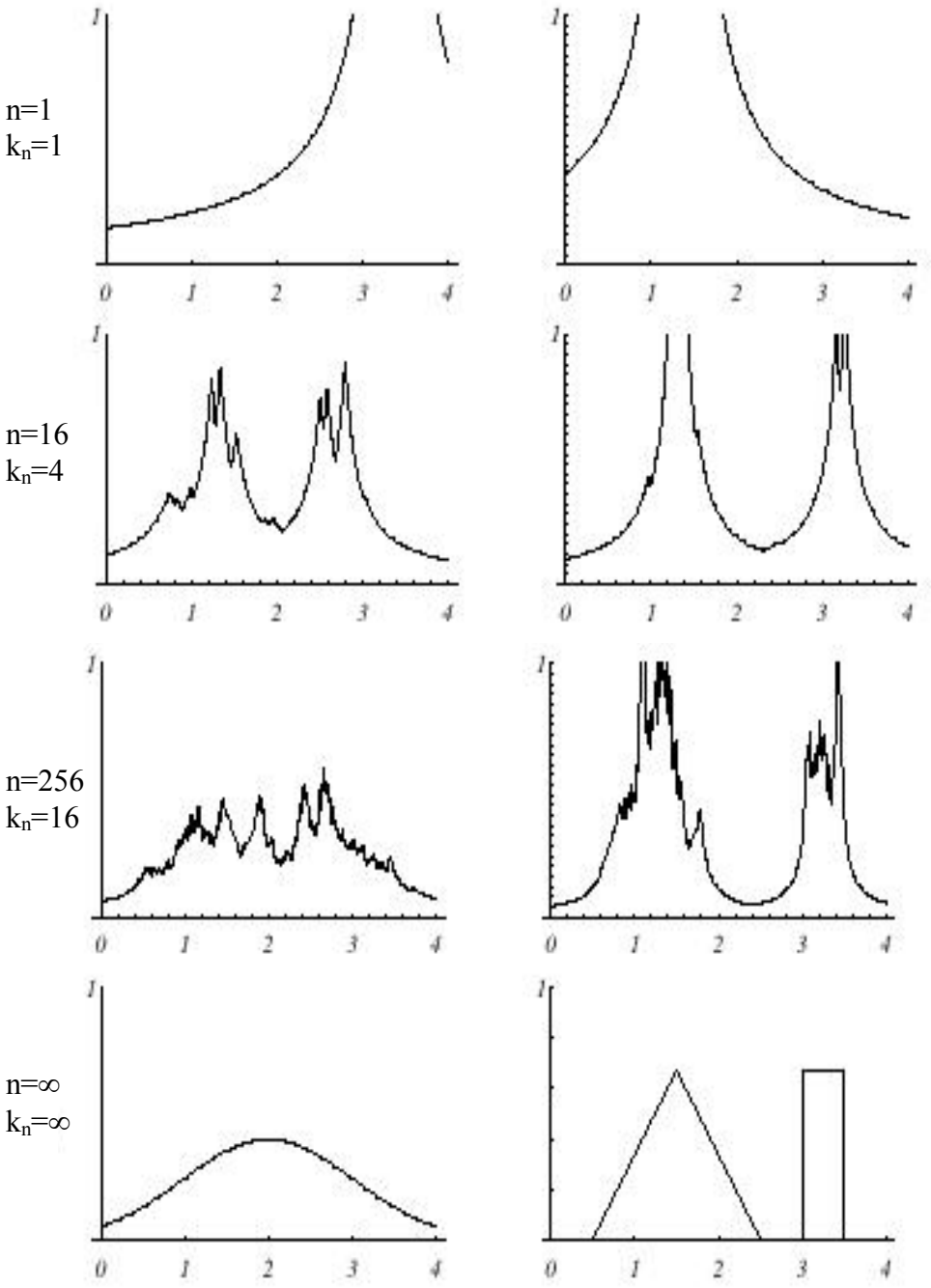
$$p_n(x,\omega_i) = \frac{k_i/n}{V} \tag{4.30}$$

και επομένως ένας λογικός υπολογισμός για την $P(\omega_i/x)$ είναι ο

$$P_n(\omega_i / x) = \frac{p_n(x,\omega_i)}{\sum_{j=1}^{c} p_n(x,\omega_j)} = \frac{k_i}{k} \tag{4.31}$$

Δηλαδή, ο υπολογισμός της εκ των υστέρων πιθανότητας ότι η ω_i είναι η κατάσταση της φύσης είναι περισσότερο ένα κλάσμα των δειγμάτων που βρίσκονται στο κελί και έχουν ετικέτα ω_i. Ως αποτέλεσμα, για ελάχιστο ρυθμό λάθους, επιλέγεται η κατηγορία που αντιπροσωπεύεται πιο πολύ μέσα στο συγκεκριμένο κελί. Εάν υπάρχουν αρκετά δείγματα και εάν το κελί είναι ικανοποιητικά μικρό, μπορεί να αποδειχθεί ότι σε αυτή την περίπτωση η απόδοση προσεγγίζει τη βέλτιστη δυνατή.

Όταν το θέμα είναι η επιλογή του μεγέθους του κελιού, είναι προφανές ότι μπορεί να χρησιμοποιηθεί είτε η προσέγγιση των παραθύρων Parzen είτε η προσέγγιση του k_n πλησιέστερου γείτονα. Στην πρώτη περίπτωση, ο V_n θα είναι κάποια συγκεκριμένη συνάρτηση του n, όπως η $V_n = 1/\sqrt{n}$. Στη δεύτερη περίπτωση, ο V_n θα επεκτείνεται μέχρι να καλυφθεί κάποιος συγκεκριμένος αριθμός δειγμάτων, όπως η $k = \sqrt{n}$. Σε κάθε περίπτωση πάντως, καθώς το n τείνει στο άπειρο ένας άπειρος αριθμός από δείγματα θα περιέχεται σε ένα απείρως μικρό κελί. Το γεγονός ότι ο όγκος του κελιού μπορεί να γίνει πάρα πολύ μικρός και παρ' όλ' αυτά να περιλαμβάνει ένα πάρα πολύ μεγάλο αριθμό δειγμάτων επιτρέπει την εύρεση των άγνωστων πιθανοτήτων με απόλυτη βεβαιότητα και επομένως επιτυγχάνεται μέγιστη απόδοση. Αρκετά ενδιαφέρουσα περίπτωση είναι αυτή στην οποία η απόφαση βασίζεται αποκλειστικά και μόνο στην ετικέτα (κατηγορία) του πιο κοντινού γείτονα του x.



Εικόνα 4.11: Διάφοροι υπολογισμοί δύο μονοδιάστατων πυκνοτήτων: μιας Gaussian και μιας με δύο ακρότατα.

4.5 Ο Κανόνας του Πλησιέστερου Γείτονα

Έστω ότι με $D^n = \{x_1, ..., x_n\}$ συμβολίζεται ένα σύνολο από n πρότυπα με ετικέτες και έστω ότι με $x' \in D^n$ συμβολίζεται το πρότυπο που βρίσκεται πιο κοντά σε ένα σημείο ελέγχου x. Ο κανόνας του πλησιέστερου γείτονα για την ταξινόμηση του x

είναι να του ανατεθεί η ετικέτα του x΄. Ο κανόνας του πλησιέστερου γείτονα είναι μία μη βέλτιστη διαδικασία. Η χρήση της οδηγεί συνήθως σε ένα ρυθμό λάθους μεγαλύτερο από τον ελάχιστο δυνατό, που είναι ο ρυθμός που προκύπτει από την εφαρμογή του κανόνα απόφασης του Bayes. Στη συνέχεια θα δειχθεί ότι για απεριόριστο αριθμό προτύπων ο ρυθμός λάθους δεν είναι ποτέ χειρότερος από το διπλάσιο ρυθμό που προκύπτει από την εφαρμογή του κανόνα απόφασης του Bayes.
Αρχικά, η ετικέτα θ΄ που σχετίζεται με τον πλησιέστερο γείτονα είναι μία τυχαία μεταβλητή ενώ η πιθανότητα ότι θ΄ = ω_i είναι ίση με την εκ των υστέρων πιθανότητα $P(\omega_i/x')$. Όταν ο αριθμός των δειγμάτων είναι πολύ μεγάλος, είναι λογικό να θεωρηθεί ότι το x΄ είναι αρκετά κοντά στο x έτσι ώστε $P(\omega_i/x') \approx P(\omega_i/x)$. Επειδή αυτή ακριβώς είναι η πιθανότητα ότι η κατάσταση της φύσης θα είναι η ω_i, ο κανόνας του πλησιέστερου γείτονα αντιστοιχεί αποδοτικά τις πιθανότητες με τις καταστάσεις της φύσης. Εάν οριστεί το $\omega_m(x)$ ως

$$P(\omega_m / x) = \max_i P(\omega_i / x) \tag{4.32}$$

τότε ο κανόνας απόφασης του Bayes επιλέγει πάντοτε το ω_m. Αυτός ο κανόνας επιτρέπει το διαχωρισμό του χώρου χαρακτηριστικών σε κελιά που αποτελούνται από όλα τα σημεία που βρίσκονται πλησιέστερα σε ένα δεδομένο σημείο εκπαίδευσης x΄ σε σχέση με οποιοδήποτε άλλο σημείο εκπαίδευσης. Ο διαχωρισμός αυτός είναι γνωστός ως κελιά Voronoi (εικόνα 4.12).
Όταν η $P(\omega_m/x)$ είναι κοντά στη μονάδα, η επιλογή με τον κανόνα του πλησιέστερου γείτονα είναι σχεδόν πάντα ίδια με την επιλογή κατά Bayes. Δηλαδή, όταν η ελάχιστη πιθανότητα λάθους είναι μικρή, η πιθανότητα λάθους του κανόνα του πλησιέστερου γείτονα είναι επίσης μικρή. Όταν η τιμή της $P(\omega_m/x)$ είναι κοντά στην 1/c, έτσι ώστε όλες οι κατηγορίες να είναι ισοπίθανες, οι επιλογές που γίνονται σύμφωνα με τον κανόνα του πλησιέστερου γείτονα είναι σπάνια οι ίδιες με αυτές που γίνονται σύμφωνα με τον κανόνα του Bayes. Παρ' όλ' αυτά, η πιθανότητα λάθους είναι περίπου ίση με 1-1/c και για τους δύο. Η ανάλυση της συμπεριφοράς του κανόνα του πλησιέστερου γείτονα θα επεκταθεί στη συνέχεια στην εύρεση της πεπερασμένων δειγμάτων μέσης υπό συνθήκη πιθανότητας λάθους P(e/x), όπου ο μέσος όρος είναι πάνω στα δείγματα εκπαίδευσης. Η χωρίς συνθήκη μέση πιθανότητα λάθους βρίσκεται από το μέσο όρο της P(e/x) πάνω σε όλα τα x:

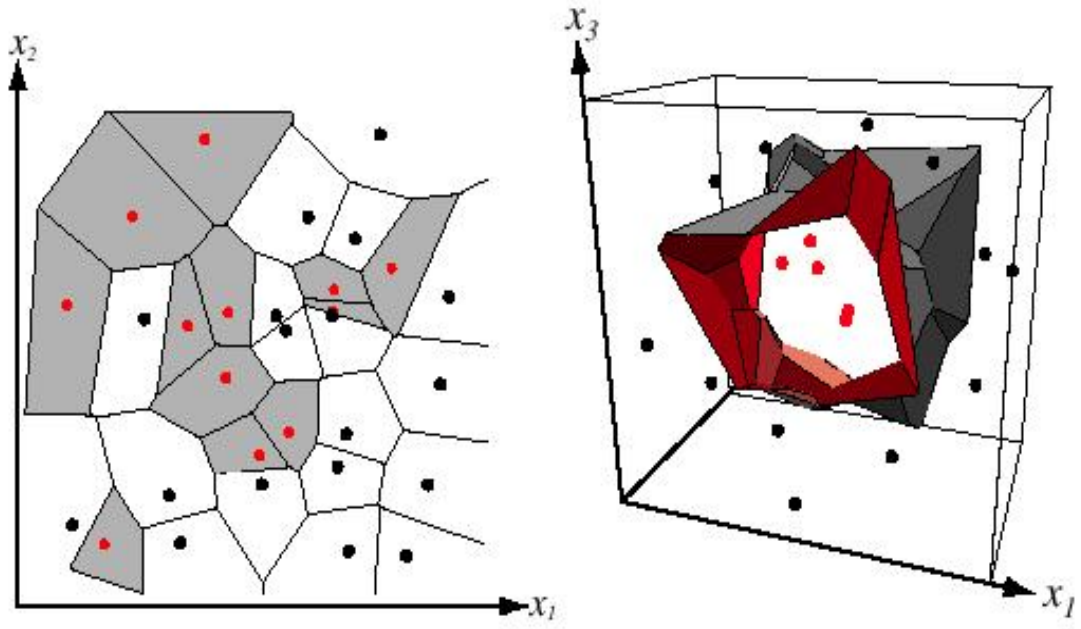
$$P(e) = \int P(e/x)p(x)dx \tag{4.33}$$

Στο σημείο αυτό πρέπει να σημειωθεί ότι ο κανόνας απόφασης του Bayes ελαχιστοποιεί την P(e) ελαχιστοποιώντας την P(e/x) για κάθε x. Επίσης (κεφάλαιο 2), ότι εάν με $P^*(e/x)$ συμβολιστεί η ελάχιστη δυνατή τιμή της P(e/x) και με P^* την ελάχιστη πιθανή τιμή της P(e) θα ισχύει

$$P^*(e/x) = 1 - P(\omega_m / x) \tag{4.34}$$

και

$$P^* = \int P^*(e/x)p(x)dx \tag{4.35}$$



Εικόνα 4.12: Στις δύο διαστάσεις, ο αλγόριθμος του πλησιέστερου γείτονα οδηγεί σε ένα διαχωρισμό του χώρου εισόδου σε κελιά Voronoi, καθένα από τα οποία έχει ως ετικέτα την κατηγορία του σημείου εκπαίδευσης που περιέχει. Στις τρεις διαστάσεις, τα κελιά είναι τρισδιάστατα και το όριο απόφασης μοιάζει με την επιφάνεια ενός κρυστάλλου.

4.5.1 Σύγκλιση του Κανόνα του Πλησιέστερου Γείτονα

Στην ενότητα αυτή θα υπολογιστεί η μέση πιθανότητα του λάθους για τον κανόνα του πλησιέστερου γείτονα. Πιο συγκεκριμένα, εάν $P_n(e)$ είναι ο ρυθμός λάθους του n-οστού δείγματος και αν

$$P = \lim_{n\to\infty} P_n(e) \tag{4.36}$$

θα πρέπει να δειχθεί ότι

$$P^* \le P \le P^*\left(2 - \frac{c}{c-1}P^*\right) \tag{4.37}$$

Αρχικά ας αναφερθεί ότι όταν χρησιμοποιείται ο κανόνας του πλησιέστερου γείτονα με ένα συγκεκριμένο σύνολο από n δείγματα, ο ρυθμός λάθους που προκύπτει εξαρτάται από τα τυχαία χαρακτηριστικά των δειγμάτων. Πιο συγκεκριμένα, εάν χρησιμοποιηθούν διαφορετικά σύνολα από n δείγματα για την ταξινόμηση του x, θα παραχθούν διαφορετικά διανύσματα x΄ για τον πλησιέστερο γείτονα του x. Επειδή ο κανόνας απόφασης εξαρτάται από τον πλησιέστερο γείτονα, ορίζεται μια υπό συνθήκη πιθανότητα του λάθους P(e/x, x΄) η οποία εξαρτάται εξίσου από τα x και x΄. Βρίσκοντας το μέσο όρο ως προς το x΄ προκύπτει

$$P(e/x) = \int P(e/x,x')p(x'/x)dx' \tag{4.38}$$

Είναι συνήθως πολύ δύσκολο να προκύψει μία συγκεκριμένη έκφραση για την υπό συνθήκη συνάρτηση πυκνότητας πιθανότητας p(x΄/x). Παρ' όλ' αυτά, επειδή το x΄ είναι εξ ορισμού ο πλησιέστερος γείτονας του x, αναμένεται αυτή η συνάρτηση πυκνότητας πιθανότητας να έχει υψηλές τιμές στην ενδιάμεση εγγύτητα του x και πολύ χαμηλές τιμές οπουδήποτε αλλού. Επιπλέον, καθώς το n τείνει στο άπειρο η

$p(x'/x)$ αναμένεται να προσεγγίζει μία συνάρτηση δέλτα με κέντρο το x, γεγονός που απλοποιεί τον υπολογισμό της εξίσωσης 4.38. Για να ισχύει αυτό, πρέπει να θεωρηθεί ότι στο δεδομένο x, η $p(\cdot)$ είναι συνεχής και διαφορετική από το μηδέν. Υπό αυτές τις συνθήκες, η πιθανότητα ότι οποιοδήποτε δείγμα περιέχεται μέσα σε μία υπερσφαίρα S με κέντρο το x είναι ένας θετικός αριθμός P_s:

$$P_s = \int_{x' \in S} p(x')dx' \tag{4.39}$$

Έτσι, η πιθανότητα ότι όλα τα n ανεξάρτητα επιλεγμένα δείγματα πέφτουν έξω από αυτές της υπερσφαίρες είναι $(1-P_s)^n$, η οποία τείνει στο μηδέν καθώς το n τείνει στο άπειρο. Έτσι το x΄ συγκλίνει στο x πιθανοτικά και η $p(x'/x)$ προσεγγίζει μια συνάρτηση δέλτα, όπως είναι αναμενόμενο.

4.5.2 Ρυθμός Λάθους για τον Κανόνα Πλησιέστερου Γείτονα

Στην ενότητα αυτή θα παρουσιαστεί ο υπολογισμός της υπό συνθήκη πιθανότητας του λάθους $P_n(e/x,x')$. Για να δειχθεί ξεκάθαρα ότι ο x΄, ο πλησιέστερος γείτονας του x, μπορεί να μεταβάλλεται καθώς αυξάνεται ο αριθμός n των δειγμάτων, ο πλησιέστερος γείτονας δηλώνεται με x'_n. Όταν υποτίθεται ότι υπάρχουν n ανεξάρτητα επιλεγμένα δείγματα με ετικέτα, θεωρείται ότι υπάρχουν n ζευγάρια από τυχαίες μεταβλητές (x_1,θ_1), (x_2,θ_2), …, (x_n,θ_n), όπου η θ_j μπορεί να είναι οποιαδήποτε από τις c καταστάσεις της φύσης ω_1, ω_2, …, ω_c. Θεωρείται ότι αυτά τα ζευγάρια δημιουργούνται από την επιλογή μιας κατάστασης της φύσης ω_j για κάθε θ_j με πιθανότητα $P(\omega_j)$ και στη συνέχεια επιλογή ενός x_j σύμφωνα με τον πιθανοτικό κανόνα $p(x/\omega_j)$. Τα ζευγάρια αυτά επιλέγονται ανεξάρτητα το ένα από το άλλο. Έστω ότι κατά τη διάρκεια της ταξινόμησης, η φύση επιλέγει ένα ζευγάρι (x,θ) και έστω επίσης ότι το x'_n, με ετικέτα θ'_n, είναι το πλησιέστερο στο x δείγμα εκπαίδευσης. Επειδή η κατάσταση της φύσης όταν επιλέχθηκε το x'_n είναι ανεξάρτητη από την κατάσταση της φύσης όταν επιλέγεται το x, ισχύει

$$P(\theta, \theta'_n / x, x'_n) = P(\theta / x)P(\theta'_n / x'_n) \tag{4.40}$$

Εάν χρησιμοποιηθεί ο κανόνας απόφασης του πλησιέστερου γείτονα, υπάρχει λάθος όταν $\theta \neq \theta'_n$. Έτσι, η υπό συνθήκη πιθανότητα του λάθους $P_n(e/x,x'_n)$ δίνεται από

$$\begin{aligned} P_n(e / x, x'_n) &= 1 - \sum_{i=1}^{c} P(\theta = \omega_i, \theta'_n = \omega_i / x, x'_n) \\ &= 1 - \sum_{i=1}^{c} P(\omega_i / x)P(\omega_i / x'_n) \end{aligned} \tag{4.41}$$

Για να υπολογιστεί η τιμή της $P_n(e)$ πρέπει να αντικατασταθεί η παραπάνω έκφραση στην εξίσωση 4.38 για την $P_n(e/x)$ και έπειτα να υπολογιστεί ο μέσος όρος του αποτελέσματος ως προς x. Αυτό γενικά είναι πολύ δύσκολο, αλλά όπως αναφέρθηκε και σε προηγούμενη ενότητα η ολοκλήρωση της εξίσωσης 4.38 γίνεται τετριμμένη καθώς το n τείνει στο άπειρο και η $p(x'_n/x)$ προσεγγίζει μία συνάρτηση δέλτα. Εάν η $P(\omega_i/x)$ είναι συνεχής στο x, προκύπτει το

$$\begin{aligned} \lim_{n\to\infty} P_n(e / x) &= \int \left[1 - \sum_{i=1}^{c} P(\omega_i / x)P(\omega_i / x'_n) \right] \delta(x'_n - x)dx'_n \\ &= 1 - \sum_{i=1}^{c} P^2(\omega_i / x) \end{aligned} \tag{4.42}$$

Επομένως, δεδομένου ότι μπορούν να μετατραπούν τα όρια και τα διαστήματα, ο ασυμπτωτικός ρυθμός λάθους του κανόνα πλησιέστερου γείτονα δίνεται από

$$P = \lim_{n\to\aleph} P_n(e) = \lim_{n\to\infty} \int P_n(e/x)p(x)dx$$
$$= \int \left[1 - \sum_{i=1}^{c} P^2(\omega_i / x)\right] p(x)dx \qquad (4.43)$$

4.5.3 Όρια Λάθους

Ενώ η εξίσωση 4.43 παρέχει ένα ακριβές αποτέλεσμα, είναι πιο χρήσιμο να προκύψουν όρια για την P σε όρους σχετικούς με το ρυθμό του κανόνα του Bayes P^*. Ένα προφανές κάτω όριο για την P είναι φυσικά το ίδιο το P^*. Επιπλέον, μπορεί να δειχθεί ότι για κάθε P^* υπάρχει ένα σύνολο από υπό συνθήκη και εκ των προτέρων πιθανότητες για τις οποίες μπορεί να επιτευχθεί το όριο. Έτσι, υπό αυτή την έννοια το όριο αυτό είναι ένα αυστηρό κάτω όριο.

Το πρόβλημα της εύρεσης ενός αυστηρού πάνω ορίου παρουσιάζει μεγαλύτερο ενδιαφέρον. Η ελπίδα για την ύπαρξη ενός χαμηλού πάνω ορίου προέρχεται από την παρατήρηση ότι εάν ο ρυθμός του κανόνα του Bayes είναι χαμηλός , η $P(\omega_i/x)$ είναι κοντά στο 1.0 για κάποιο i, έστω i = m. Έτσι, το ολοκλήρωμα της εξίσωσης 4.43 είναι περίπου $1 - P^2(\omega_m / x) \cong 2(1 - P(\omega_m / x))$, και αφού

$$P^*(e/x) = 1 - P(\omega_m / x) \qquad (4.44)$$

η ολοκλήρωση ως προς x μπορεί να οδηγήσει σε διπλάσιο ρυθμό λάθους από αυτόν του Bayes, ο οποίος είναι προφανώς αρκετά μικρός και αποδεκτός για αρκετές εφαρμογές. Για να υπολογιστεί ένα ακριβές πάνω όριο, πρέπει να βρεθεί πόσο μεγάλος μπορεί να γίνει ο ρυθμός λάθους του κανόνα πλησιέστερου γείτονα P για ένα δεδομένο ρυθμό κατά Bayes P^*. Έτσι, με βάση την εξίσωση 4.43 πρέπει να καθοριστεί πόσο μικρό μπορεί να γίνει το $\sum_{i=1}^{c} P^2(\omega_i / x)$ για μία δεδομένη τιμή της $P(\omega_m/x)$. Αρχικά ισχύει το εξής:

$$\sum_{i=1}^{c} P^2(\omega_i / x) = P^2(\omega_m / x) + \sum_{i=1}^{c} P^2(\omega_i / x) \qquad (4.45)$$

Στη συνέχεια πρέπει να φραγεί αυτό το άθροισμα ελαχιστοποιώντας το δεύτερο όρο με βάση τους παρακάτω περιορισμούς:

- $P(\omega_i / x) \geq 0$
- $\sum_{i\neq m} P(\omega_i / x) = 1 - P(\omega_m / x) = P^*(e/x)$

Εύκολα μπορεί κάποιος να δει ότι το $\sum_{i=1}^{c} P^2(\omega_i / x)$ ελαχιστοποιείται εάν όλες οι εκ των υστέρων πιθανότητες εκτός από τη m-οστή είναι ίσες. Ο δεύτερος περιορισμός οδηγεί στο εξής:

$$P(\omega_i / x) = \begin{cases} \dfrac{P^*(e/x)}{c-1} & i \neq m \\ 1 - P^*(e/x) & i = m \end{cases} \qquad (4.46)$$

Έτσι, προκύπτουν οι παρακάτω ανισώσεις

$$\sum_{i=1}^{c} P^2(\omega_i / x) \geq (1 - P^*(e/x))^2 + \frac{P^{*2}(e/x)}{c-1} \qquad (4.47)$$

και

$$1-\sum_{i=1}^{c} P^2(\omega_i / x) \le 2P^*(e/x) - \frac{c}{c-1} P^{*2}(e/x) \tag{4.48}$$

Αυτό άμεσα αποδεικνύει ότι $P \le 2P^*$, αφού το παραπάνω αποτέλεσμα μπορεί να αντικατασταθεί στην εξίσωση 4.43 και να απαλειφθεί ο δεύτερος όρος. Παρ' όλ' αυτά, ένα αυστηρότερο όριο μπορεί να αποκτηθεί με την παρατήρηση ότι η διασπορά ισούται με

$$\mathrm{Var}\left[P^*(e/x)\right] = \int \left[P^*(e/x) - P^*\right]^2 p(x)dx$$

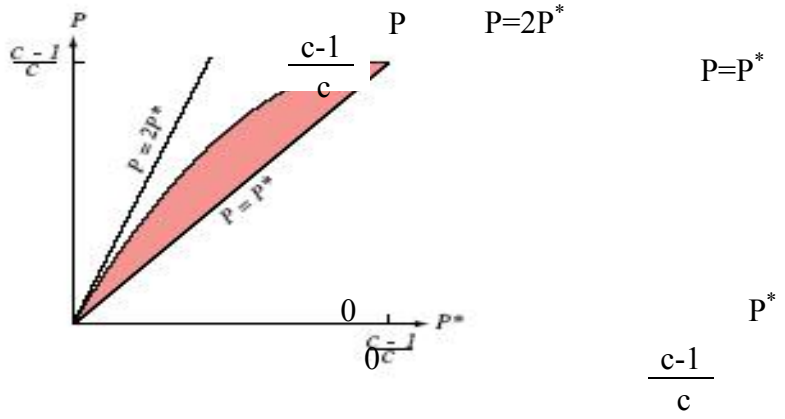
$$= \int P^{*2}(e/x)p(x)dx - P^{*2} \ge 0$$

Επομένως,

$$\int P^{*2}(e/x)p(x)dx \ge P^{*2} \tag{4.49}$$

με την ισότητα να ισχύει εάν και μόνο εάν η διασπορά της $P^*(e/x)$ είναι ίση με το μηδέν. Χρησιμοποιώντας το παραπάνω αποτέλεσμα και αντικαθιστώντας την εξίσωση 4.48 στην εξίσωση 4.43 προκύπτουν τα επιθυμητά όρια για το λάθος P, του κανόνα του πλησιέστερου γείτονα, στην περίπτωση άπειρου αριθμού δειγμάτων:

$$P^* \le P \le P^*\left(2 - \frac{c}{c-1} P^*\right) \tag{4.50}$$



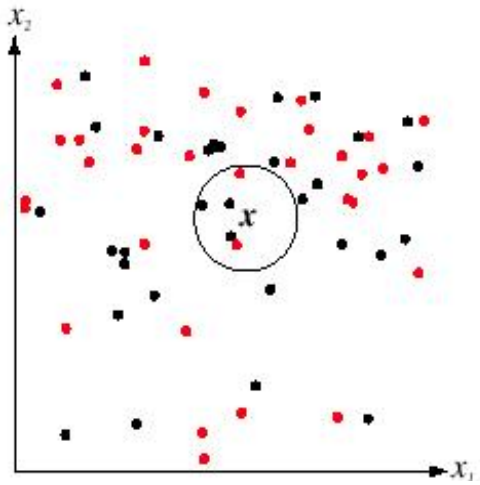
Εικόνα 4.13: Όρια φραγμού στο ρυθμό λάθους P του κανόνα του πλησιέστερου γείτονα, σε ένα πρόβλημα c κατηγοριών, με δεδομένο ότι υπάρχουν άπειρα δεδομένα εκπαίδευσης. Το P* είναι το λάθος κατά Bayes (εξίσωση 4.50). Για χαμηλούς ρυθμούς λάθους, ο ρυθμός λάθους του κανόνα του πλησιέστερου γείτονα φράσσεται από πάνω από το διπλάσιο του ρυθμού λάθους κατά Bayes.

Είναι εύκολο να δειχθεί ότι αυτό το πάνω όριο επιτυγχάνεται στην αποκαλούμενη περίπτωση της μηδενικής πληροφορίας, σύμφωνα με την οποία οι συναρτήσεις πυκνότητας πιθανότητας $p(x/\omega_i)$ είναι ίδιες, έτσι ώστε να ισχύει $P(\omega_i/x)=P(\omega_i)$ και η $P^*(e/x)$ να είναι ανεξάρτητη από το x. Επομένως, τα όρια που δίνονται από τη εξίσωση 4.50 είναι τα αυστηρότερα δυνατά, υπό την έννοια ότι για οποιοδήποτε P^* υπάρχουν οι υπό συνθήκη και εκ των προτέρων πιθανότητες για τις οποίες τα παραπάνω όρια μπορούν να επιτευχθούν. Πιο συγκεκριμένα, ο ρυθμός κατά Bayes P^* μπορεί να πάρει οποιαδήποτε τιμή μεταξύ 0 και (c-1)/c ενώ τα όρια θα βρίσκονται στις δύο ακραίες τιμές για τις πιθανότητες. Όταν ο ρυθμός κατά Bayes είναι μικρός, το άνω όριο είναι κατά προσέγγιση διπλάσιος από το ρυθμό κατά Bayes (εικόνα 4.13).

Επειδή το P είναι πάντοτε μικρότερο ή ίσο από το $2P^*$, εάν κάποιος είχε ένα απεριόριστο αριθμό δεδομένων και χρησιμοποιούσε έναν αυθαίρετα πολύπλοκο κανόνα απόφασης, το καλύτερο αποτέλεσμα που θα μπορούσε να πετύχει θα ήταν να περιορίσει το ρυθμό στο μισό της τιμής του. Υπό αυτή την έννοια, τουλάχιστον η μισή πληροφορία ταξινόμησης σε ένα άπειρο σύνολο δεδομένων βρίσκεται στον πλησιέστερο γείτονα.
Είναι φυσικό να αναρωτηθεί κανείς πώς να συμπεριφέρεται ο κανόνας του πλησιέστερου γείτονα στην περίπτωση ενός πεπερασμένου συνόλου δειγμάτων και πόσο γρήγορα συγκλίνει στην ασυμπτωτική τιμή. Δυστυχώς, έχει δειχθεί ότι η σύγκλιση είναι ιδιαίτερα αργή και ότι ο ρυθμός λάθους $P_n(e)$ δεν ελαττώνεται μονότονα ως προς το n. Όπως και με άλλες μη παραμετρικές μεθόδους, είναι δύσκολο να επιτευχθεί κάτι παραπάνω από ασυμπτωτικά αποτελέσματα χωρίς προηγουμένως να έχουν γίνει επιπλέον θεωρήσεις σχετικά με τη δομή των αντίστοιχων πιθανοτήτων.

4.5.4 Ο Κανόνας των k Πλησιέστερων Γειτόνων

Μία προφανής επέκταση του κανόνα του πλησιέστερου γείτονα είναι ο κανόνας των k πλησιέστερων γειτόνων. Όπως φαίνεται και από το όνομα, αυτός ο κανόνας ταξινομεί το x αποδίδοντάς του την ετικέτα που εμφανίζεται πιο συχνά ανάμεσα στα k πλησιέστερα δείγματα. Με άλλα λόγια, κάθε απόφαση για ταξινόμηση ενός δείγματος x παίρνεται αποκλειστικά με βάση τις ετικέτες των k πλησιέστερων γειτόνων του (εικόνα 4.14). Στη συνέχεια θα εξεταστεί ο κανόνας των k πλησιέστερων γειτόνων για την περίπτωση δύο κατηγοριών με το k περιττό αριθμό (για να αποφευχθούν ισοπαλίες) με σκοπό εξαχθούν περισσότερες πληροφορίες για τη λειτουργία του.



Εικόνα 4.14: Η αναζήτηση σύμφωνα με τον κανόνα του k πλησιέστερου γείτονα ξεκινάει από το σημείο ελέγχου x και μεγαλώνει μία σφαιρική περιοχή ώσπου να περιλαμβάνει k δείγματα εκπαίδευσης. Το σημείο ελέγχου παίρνει την ετικέτα που έχει η πλειοψηφία αυτών των δειγμάτων. Στην παραπάνω περίπτωση όπου k = 5, το σημείο ελέγχου x παίρνει την ετικέτα της κατηγορίας των μαύρων σημείων.

Το βασικό κίνητρο για την επινόηση του κανόνα του πλησιέστερου γείτονα βρίσκεται στην παρατήρηση που έγινε προηγουμένως σχετικά με την σχέση των πιθανοτήτων με την κατάσταση της φύσης. Παρατηρείται αρχικά ότι εάν το k είναι σταθερό και ο αριθμός των δειγμάτων n επιτρέπεται να πλησιάσει το άπειρο, τότε όλοι οι k πλησιέστεροι γείτονες θα συγκλίνουν στο x. Επίσης, όπως και στην περίπτωση του ενός πλησιέστερου γείτονα, οι ετικέτες σε καθένα από τους k πλησιέστερους γείτονες

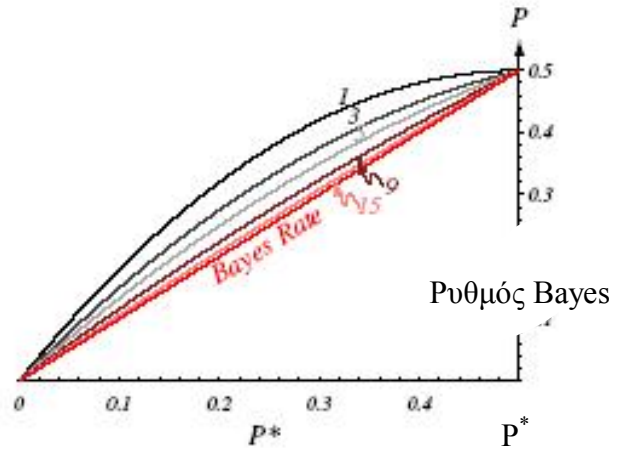
είναι τυχαίες μεταβλητές, οι οποίες υποδηλώνουν ανεξάρτητα τις τιμές ω_i με πιθανότητες $P(\omega_i/x)$, i =1, 2. Εάν η $P(\omega_m/x)$ είναι η μεγαλύτερη εκ των υστέρων πιθανότητα, τότε ο κανόνας απόφασης του Bayes επιλέγει πάντα την ω_m. Ο κανόνας του ενός πλησιέστερου γείτονα επιλέγει την ω_m με πιθανότητα $P(\omega_m/x)$. Ο κανόνας των k πλησιέστερων γειτόνων επιλέγει την ω_m εάν η πλειοψηφία των k πλησιέστερων γειτόνων έχει ετικέτα ω_m, ένα γεγονός που έχει πιθανότητα ίση με

$$\sum_{i=(k+1)/2}^{k} \binom{k}{i} P(\omega_m / x)^i \left[1 - P(\omega_m / x)\right]^{k-i} \tag{4.51}$$

Γενικά, όσο μεγαλύτερη είναι η τιμή του k, τόσο μεγαλύτερη είναι η πιθανότητα να επιλεγεί η κατάσταση ω_m.

Στη συνέχεια θα αναλυθεί ο κανόνας των k πλησιέστερων γειτόνων με παρόμοιο τρόπο με αυτόν που χρησιμοποιήθηκε για την ανάλυση του κανόνα του ενός πλησιέστερου γείτονα. Παρ' όλ' αυτά, θα δοθεί περισσότερη σημασία στα αποτελέσματα και όχι στις αποδείξεις τους. Μπορεί να αποδειχθεί ότι εάν το k είναι περιττό, ο ρυθμός λάθους για δύο κατηγορίες και μεγάλο αριθμό δειγμάτων φράσσεται από πάνω από τη συνάρτηση $C_k(P^*)$, όπου η $C_k(P^*)$ ορίζεται ως η μικρότερη κοίλη συνάρτηση του P^* που είναι μεγαλύτερη από

$$\sum_{i=0}^{(k-1)/2} \binom{k}{i} \left[(P^*)^{i+1} (1 - P^*)^{k-i} + (P^*)^{k-i} (1 - P^*)^{i+1} \right] \tag{4.52}$$



Ρυθμός Bayes

P^*

Εικόνα 4.15: Ο ρυθμός λάθους για τον κανόνα του k πλησιέστερου γείτονα για ένα πρόβλημα δύο κατηγοριών φράσσεται από το $C_k(P^*)$ στην εξίσωση 4.52. Κάθε καμπύλη έχει την ετικέτα του αντίστοιχου k. όταν k = ∞, οι τιμές που υπολογίζονται για τις πιθανότητες ταυτίζονται με τις πραγματικές τιμές και επομένως ο ρυθμός λάθους είναι ίσος με το ρυθμό κατά Bayes, δηλαδή $P = P^*$.

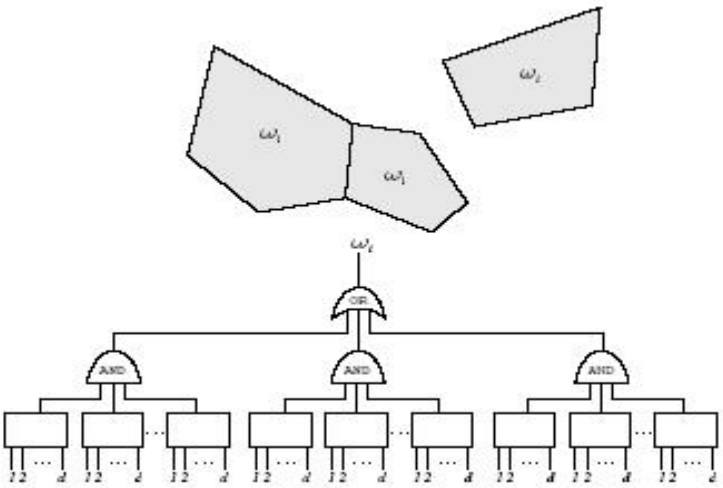
Η άθροιση στον πρώτο όρο της αγκύλης αντιπροσωπεύει την πιθανότητα του λάθους εξαιτίας των i σημείων που ανήκουν στην κατηγορία με τη μικρότερη πιθανότητα και των k – i > i σημείων που ανήκουν στην άλλη κατηγορία. Η άθροιση στο δεύτερο όρο της αγκύλης είναι η πιθανότητα ότι k – i σημεία είναι από την κατηγορία με τη μικρότερη πιθανότητα και i +1 < k – i σημεία είναι από την κατηγορία με τη μεγαλύτερη πιθανότητα. Και οι δύο αυτές περιπτώσεις αναφέρονται σε λάθη που αφορούν τον κανόνα των k πλησιέστερων γειτόνων και επομένως πρέπει να προστεθούν για την εύρεση της πλήρους πιθανότητας του λάθους.

Στην εικόνα 4.15 φαίνονται τα όρια για τους ρυθμούς λάθους των k πλησιέστερων γειτόνων για διάφορες τιμές του k. Καθώς το k αυξάνεται, τα άνω όρια πλησιάζουν

προοδευτικά στο κάτω όριο – το όριο κατά Bayes. Οριακά, καθώς το k τείνει στο άπειρο, τα δύο όρια συναντιούνται και ο κανόνας των k πλησιέστερων γειτόνων γίνεται βέλτιστος.
Ολοκληρώνοντας θα γίνει μια αναφορά στην περίπτωση πεπερασμένου αριθμού δειγμάτων που εμφανίζεται σε πρακτικές εφαρμογές. Ο κανόνας απόφασης των k πλησιέστερων γειτόνων μπορεί να θεωρηθεί ως άλλη μία προσπάθεια να υπολογιστούν οι εκ των υστέρων πιθανότητες $P(\omega_i/x)$ από τα δείγματα. Πρέπει να χρησιμοποιηθεί μία μεγάλη τιμή για το k για να επιτευχθεί ένας αξιόπιστος υπολογισμός. Από την άλλη πλευρά, πρέπει όλοι οι k πλησιέστεροι γείτονες x΄ να είναι πολύ κοντά στο x ώστε να είναι σίγουρο ότι οι $P(\omega_i/x')$ θα είναι περίπου ίδιες με τις $P(\omega_i/x)$. Αυτό αναγκάζει να επιλεγεί μία συμβιβαστική τιμή για το k που να αποτελεί ένα μικρό κλάσμα του αριθμού των δειγμάτων. Μόνο στην οριακή περίπτωση, όπου το n τείνει στο άπειρο μπορεί κάποιος να είναι σίγουρος για τη σχεδόν βέλτιστη συμπεριφορά του κανόνα των k πλησιέστερων γειτόνων.

4.5.5 Υπολογιστική Πολυπλοκότητα του Κανόνα των k Πλησιέστερων Γειτόνων

Η υπολογιστική πολυπλοκότητα του αλγορίθμου του πλησιέστερου γείτονα, τόσο όσον αφορά το χώρο (αποθήκευση των προτύπων) όσο και το χρόνο (αναζήτηση), έχει αναλυθεί σε αρκετά μεγάλο βαθμό. Υπάρχουν πολλά ενδιαφέροντα θεωρήματα από το χώρο της υπολογιστικής γεωμετρίας για την κατασκευή των Voronoi tessellations και των αναζητήσεων των πλησιέστερων γειτόνων τόσο για μονοδιάστατους όσο και για δισδιάστατους χώρους. Παρ' όλ' αυτά, επειδή οι τεχνικές του πλησιέστερου γείτονα χρησιμοποιούνται περισσότερο για προβλήματα με πολλά χαρακτηριστικά, το ενδιαφέρον επικεντρώνεται στην πιο γενική d-διάστατη περίπτωση.



Εικόνα 4.16: Ένα παράλληλο κύκλωμα υλοποίησης του κανόνα του πλησιέστερου γείτονα μπορεί να ολοκληρώσει την αναζήτηση σε σταθερό, δηλαδή O(1), χρόνο. Το d – διάστατο πρότυπο ελέγχου x δίνεται ως είσοδος σε κάθε κουτί το οποίο υπολογίζει σε ποια πλευρά της επιφάνειας ενός κελιού βρίσκεται το x. Εάν είναι στην «κλειστή» πλευρά κάθε επιφάνειας ενός κελιού, βρίσκεται στο κελί Voronoi του αποθηκευμένου προτύπου και παίρνει την ετικέτα του. Στην παραπάνω περίπτωση, καθεμιά από τις τρεις AND πύλες αντιστοιχεί σε ένα κελί Voronoi.

Έστω ότι υπάρχουν n δείγματα εκπαίδευσης με ετικέτες σε d διαστάσεις και γίνεται αναζήτηση για να βρεθεί το πλησιέστερο ως προς ένα σημείο ελέγχου x (k = 1). Σε μια πιο απλοϊκή προσέγγιση, ανιχνεύεται κάθε σημείο με τη σειρά του, υπολογίζεται η Ευκλείδεια απόστασή του από το x και διατηρείται η ταυτότητα μόνο του τρέχοντος

πιο κοντινού σημείου. Ο υπολογισμός κάθε απόστασης απαιτεί O(d) υπολογισμούς και επομένως συνολικά η αναζήτηση απαιτεί $O(dn^2)$. Μια εναλλακτική αλλά άμεση παράλληλη υλοποίηση παρουσιάζεται στην εικόνα 4.16. Αυτή η υλοποίηση έχει απαιτήσεις O(1) σε χρόνο και O(n) σε χώρο.

Υπάρχουν τρεις γενικές αλγοριθμικές τεχνικές για τη μείωση του υπολογιστικού φόρτου στις αναζητήσεις των πλησιέστερων γειτόνων: ο υπολογισμός των μερικών αποστάσεων, η προκατασκευή και η εμφάνιση των αποθηκευμένων προτύπων. Στην τεχνική των μερικών αποστάσεων, υπολογίζεται η απόσταση χρησιμοποιώντας ένα υποσύνολο r των συνολικών d διαστάσεων, και αν αυτή η μερική απόσταση είναι πολύ μεγάλη, ο υπολογισμός σταματάει. Η μερική απόσταση που βασίζεται σε r επιλεγμένες διαστάσεις είναι

$$D_r(a,b) = \left(\sum_{k=1}^{r} (\alpha_k - b_k)^2 \right)^{1/2} \tag{4.53}$$

όπου $r < d$. Οι μέθοδοι των μερικών αποστάσεων θεωρούν ότι η γνώση για την απόσταση σε ένα υποχώρο είναι αντιπροσωπευτική για όλο το χώρο. Φυσικά, η μερική απόφαση είναι αυστηρώς αύξουσα καθώς προστίθενται κατανομές από όλο και περισσότερες διαστάσεις. Ως αποτέλεσμα, η διαδικασία υπολογισμού της απόστασης ως προς οποιοδήποτε δείγμα μπορεί να τερματιστεί αν η μερική απόσταση γίνει μεγαλύτερη από την πλήρη $r = d$ Ευκλείδεια απόσταση ως προς το τρέχον πλησιέστερο πρότυπο.

Στις μεθόδους προκατασκευής, κατασκευάζεται κάποιας μορφής δέντρο αναζήτησης στο οποίο τα πρότυπα είναι επιλεκτικά συνδεδεμένα. Κατά τη διάρκεια της ταξινόμησης, υπολογίζεται η απόσταση του σημείου ελέγχου από ένα ή περισσότερα αποθηκευμένα πρότυπα (πρότυπα «εισόδου» ή πρότυπα «ρίζας») και στη συνέχεια μελετώνται μόνο τα πρότυπα που είναι συνδεδεμένα με αυτό. Εάν το δέντρο είναι σωστά δομημένο, ο συνολικός αριθμός των προτύπων που χρειάζεται να ανιχνευτούν μπορεί να μειωθεί.

Έστω μια τετριμμένη παρουσίαση της μεθόδου της προκατασκευής, κατά την οποία αποθηκεύεται ένας μεγάλος αριθμός από πρότυπα, τα οποία είναι ομοιόμορφα κατανεμημένα στο μοναδιαίο τετράγωνο, δηλαδή $p(x) \approx U\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}\right)$. Έστω ότι αυτό το σύνολο προκατασκευάζεται χρησιμοποιώντας τέσσερα πρότυπα εισόδου ή ρίζας τα $\begin{pmatrix}1/4\\1/4\end{pmatrix}$, $\begin{pmatrix}1/4\\3/4\end{pmatrix}$, $\begin{pmatrix}3/4\\1/4\end{pmatrix}$ και $\begin{pmatrix}3/4\\3/4\end{pmatrix}$, το καθένα συνδεδεμένο μόνο με σημεία που βρίσκονται στο αντίστοιχο τεταρτημόριο. Όταν παρουσιάζεται ένα πρότυπο ελέγχου x, καθορίζεται αρχικά το πλησιέστερο σε αυτό από τα τέσσερα πρότυπα εισόδου και στη συνέχεια η αναζήτηση περιορίζεται στα πρότυπα του αντίστοιχου τεταρτημόριο. Με αυτόν τον τρόπο, τα 3/4 των προτύπων δε θα χρειαστεί να ανιχνευθούν ποτέ.

Σημειώνεται ότι σε αυτή τη μέθοδο δεν υπάρχει εγγύηση για την εύρεση του πραγματικά πλησιέστερου προτύπου. Για παράδειγμα, έστω ότι το σημείο ελέγχου βρίσκεται κοντά στο σύνορο των τεταρτημορίων, για παράδειγμα $x = \begin{pmatrix}0.499\\0.499\end{pmatrix}$. Σε αυτή τη συγκεκριμένη περίπτωση, θα ανιχνευθούν μόνο τα πρότυπα που ανήκουν στο πρώτο τεταρτημόριο. Να σημειωθεί όμως ότι το πλησιέστερο πρότυπο μπορεί να βρίσκεται σε κάποιο από τα υπόλοιπα τρία τεταρτημόρια, κάπου κοντά στο $\begin{pmatrix}0.5\\0.5\end{pmatrix}$.

Αυτό υποδηλώνει ένα κλασσικό γενικό χαρακτηριστικό της αναγνώρισης προτύπων: το ισοζύγιο ανάμεσα στην πολυπλοκότητα και την ακρίβεια της αναζήτησης.
Πιο εξελιγμένα δέντρα αναζήτησης συνδέουν τα αποθηκευμένα πρότυπα με μικρό αριθμό άλλων προτύπων. Όμως, όσο ο αριθμός των προτύπων που ανιχνεύεται είναι σχετικά μικρός, η μέθοδος δεν εγγυάται ότι θα βρεθεί το πραγματικά πλησιέστερο πρότυπο.
Η τρίτη μέθοδος για τη μείωση της πολυπλοκότητας της αναζήτησης του πλησιέστερου γείτονα είναι η απαλοιφή των «άχρηστων» προτύπων κατά τη διάρκεια της εκπαίδευσης, μια τεχνική που είναι ευρύτερα γνωστή ως περικοπή ή συμπύκνωση. Μία απλή μέθοδος για να ελαττωθεί η O(n) πολυπλοκότητα του χώρου είναι να απαλειφθούν πρότυπα τα οποία είναι περιτριγυρισμένα από σημεία εκπαίδευσης που ανήκουν στην ίδια κατηγορία. Αυτό αφήνει τα όρια απόφασης, και επομένως και το λάθος, αδιάφορα, ενώ ταυτόχρονα μειώνει τους χρόνους απόκρισης. Ένας απλός αλγόριθμος περικοπής είναι ο παρακάτω:

Αλγόριθμος 1: Περικοπή Πλησιέστερου Γείτονα
1 αρχή αρχικοποίησε $j \leftarrow 0$, D ← σύνολο δεδομένων, n ← αριθμός προτύπων
2 κατασκεύασε το πλήρες διάγραμμα Voronoi του D
3 κάνε $j \leftarrow j + 1$; για κάθε πρότυπο x'_j
4 βρες τους Voronoi γείτονες του x'_j
5 Εάν κάποιος γείτονας δεν ανήκει στην ίδια κλάση με το x'_j τότε σημείωσε το x'_j
6 μέχρι $j = n$
7 απέρριψε όλα τα σημεία που δεν είναι σημειωμένα
8 κατασκεύασε το διάγραμμα Voronoi των υπόλοιπων (σημειωμένων) προτύπων
9 τέλος

Η πολυπλοκότητα του αλγόριθμου περικοπής είναι $O\left(d^3 n^{\lfloor d/2 \rfloor} \ln n\right)$, όπου ο τελεστής $\lfloor \cdot \rfloor$ υπονοεί ότι $\lfloor d/2 \rfloor = k$ εάν το d είναι άρτιο και $2k - 1$ εάν το d είναι περιττό. Με βάση τον αλγόριθμο 3, εάν ένα πρότυπο συνεισφέρει σε ένα όριο απόφασης, δηλαδή τουλάχιστον ένας από τους γείτονές του ανήκει σε διαφορετική κατηγορία, τότε παραμένει στο σύνολο, διαφορετικά περικόπτεται. Ο αλγόριθμος αυτός δεν εγγυάται ότι θα βρεθεί το ελάχιστο σύνολο σημείων. Παρ' όλ' αυτά, αποτελεί ένα από τα κλασσικά παραδείγματα της ταξινόμησης προτύπων στα οποία η υπολογιστική πολυπλοκότητα μπορεί να μειωθεί, κάποιες φορές σημαντικά, χωρίς να επηρεαστεί η ακρίβεια. Ένα μειονέκτημα αυτών των συστημάτων πλησιέστερου γείτονα είναι το ότι γενικά δεν μπορούν να προστεθούν εκπαιδευτικά δεδομένα αργότερα, επειδή το βήμα περικοπής απαιτεί εκ των προτέρων γνώση όλων των δεδομένων εκπαίδευσης. Καταλήγοντας, σημειώνεται ότι, προφανώς, αυτές οι τρεις μέθοδοι μείωσης της πολυπλοκότητας μπορούν να συνδυαστούν. Για παράδειγμα, μπορεί αρχικά να γίνει edit στα πρότυπα, έπειτα να κατασκευαστεί ένα δέντρο αναζήτησης κατά τη διάρκεια της εκπαίδευσης και τελικά να υπολογιστούν οι μερικές αποστάσεις κατά την ταξινόμηση.

4.6 Ταξινόμηση Πλησιέστερου Γείτονα και Μέτρα Απόδοσης

Ο ταξινομητής πλησιέστερου γείτονα βασίζεται σε μία συνάρτηση μέτρου ή απόστασης ανάμεσα στα πρότυπα. Αν και μέχρι το σημείο αυτό, αναφέρθηκε μόνο το Ευκλείδειο μέτρο σε d διαστάσεις, η έννοια του μέτρου είναι πολύ πιο γενική και στη συνέχεια θα γίνει μια προσπάθεια να χρησιμοποιηθούν εναλλακτικά μέτρα απόστασης για να αντιμετωπιστούν σημαντικά προβλήματα στην ταξινόμηση. Αρχικά θα γίνει μια συνοπτική παρουσίαση των ιδιοτήτων ενός μέτρου. Ένα μέτρο $D(\cdot,\cdot)$ αποτελεί περισσότερο μια συνάρτηση η οποία παρέχει μια γενικευμένη βαθμωτή απόσταση ανάμεσα σε δύο πρότυπα.

4.6.1 Ιδιότητες των Μέτρων

Ένα μέτρο πρέπει να έχει τέσσερις ιδιότητες: Για οποιαδήποτε διανύσματα a, b και c αυτές οι ιδιότητες είναι οι εξής:

Θετικότητα: $D(a, b) \geq 0$.

Αντανακλαστικότητα: $D(a, b) = 0$ εάν και μόνο εάν $a = b$.

Συμμετρικότητα: $D(a, b) = D(b, a)$.

Τριγωνική Ανισότητα: $D(a, b) + D(b, c) \geq D(a, c)$.

Είναι εύκολο να αποδειχθεί ότι ο τύπος για την Ευκλείδεια απόσταση σε d διαστάσεις,

$$D(a,b) = \left(\sum_{k=1}^{d} (a_k - b_k)^2 \right)^{1/2} \tag{4.54}$$

έχει τις ιδιότητες του μέτρου. Αν και κάποιος μπορεί πάντα να υπολογίσει την Ευκλείδεια απόσταση ανάμεσα σε δύο διανύσματα, τα αποτελέσματα μπορεί να έχουν αλλά μπορεί και να μην έχουν κανένα νόημα. Για παράδειγμα, εάν ο χώρος μετασχηματίζεται με πολλαπλασιασμό κάθε συντεταγμένης με μία αυθαίρετα επιλεγμένη μεταβλητή, οι σχέσεις της Ευκλείδειας απόστασης στο μετασχηματισμένο χώρο μπορεί να είναι πολύ διαφορετικές από τις αρχικές (προ μετασχηματισμού), ακόμα και αν ο μετασχηματισμός που έγινε περισσότερο αφορά σε μια διαφορετική επιλογή των μονάδων σύγκρισης για τα χαρακτηριστικά. Μια τέτοια αλλαγή στις μονάδες μπορεί να επηρεάσει πολύ σημαντικά την απόδοση των ταξινομητών πλησιέστερου γείτονα.

Μια γενική κατηγορία μέτρων για d-διάστατα πρότυπα είναι το μέτρο του Minkowski

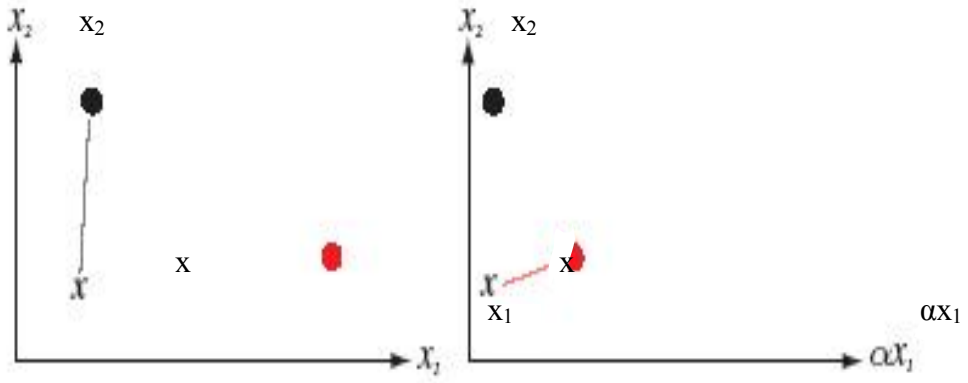
$$L_k(a,b) = \left(\sum_{i=1}^{d} |a_i - b_i|^k \right)^{1/k} \tag{4.55}$$

το οποίο αναφέρεται και ως η L_k νόρμα. Άρα, η Ευκλείδεια απόσταση είναι η L_2 νόρμα. Η L_1 νόρμα καλείται απόσταση Manhattan ή απόσταση αστικού τετραγώνου, το συντομότερο μονοπάτι ανάμεσα στο a και στο b, κάθε τμήμα του οποίου είναι παράλληλο σε ένα άξονα συντεταγμένων (το όνομα προκύπτει από το ότι οι δρόμοι στο Manhattan έχουν κατεύθυνση βορρά-νότου και ανατολής-δύσης). Έστω ότι υπολογίζονται οι αποστάσεις ανάμεσα στις προβολές του a και του b σε καθένα από τους d άξονες συντεταγμένων. Η L_∞ απόσταση ανάμεσα στο a και στο b αντιστοιχεί στη μέγιστη απόσταση ανάμεσα στις προβολές (εικόνα 4.18).

Το μέτρο του Tanimoto βρίσκει μεγαλύτερη εφαρμογή στην ταξινομία, όπου η απόσταση ανάμεσα σε δύο σύνολα ορίζεται ως

$$D_{\text{Tanimoto}}(S_1, S_2) = \frac{n_1 + n_2 + 2n_{12}}{n_1 + n_2 - n_{12}} \quad (4.56)$$

όπου τα n_1 και n_2 είναι οι αριθμοί των στοιχείων που ανήκουν στα σύνολα S_1 και S_2 αντίστοιχα και το n_{12} είναι ο αριθμός των στοιχείων που ανήκουν και στα δύο σύνολα. Το μέτρο Tanimoto βρίσκει πολύ μεγάλη εφαρμογή σε προβλήματα στα οποία δύο πρότυπα ή χαρακτηριστικά – στοιχεία του συνόλου – είναι είτε τα ίδια είτε διαφορετικά και δεν υπάρχει καμία φυσική έννοια διαβαθμισμένης ομοιότητας.
Η επιλογή ανάμεσα σε αυτά ή και άλλα μέτρα συνήθως υπαγορεύεται από τις υπολογιστικές δυνατότητες. Είναι γενικά δύσκολο η επιλογή να βασίζεται σε εκ των προτέρων γνώση για τις κατανομές. Μία εξαίρεση είναι όταν υπάρχει μεγάλη διαφορά στη διακύμανση των δεδομένων σε διαφορετικούς άξονες σε ένα πολυδιάστατο χώρο δεδομένων. Εδώ, θα scale τα δεδομένα – ή αντίστοιχα θα μεταβληθεί το μέτρο – όπως φαίνεται και στο εικόνα 4.17.

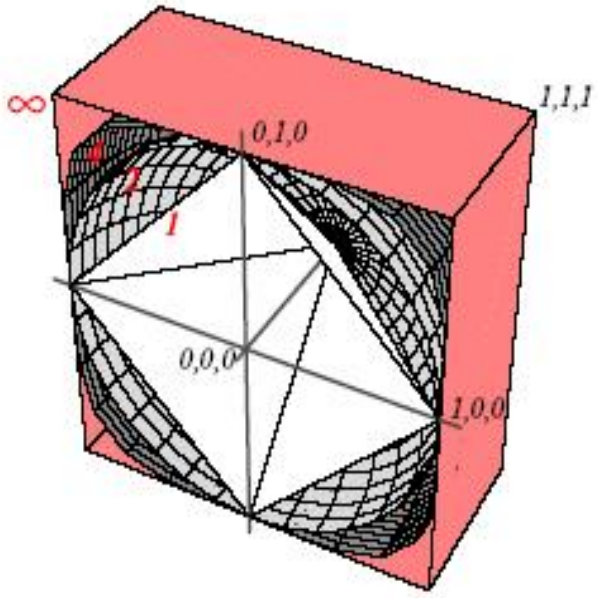


Εικόνα 4.17: Η κλιμάκωση των συντεταγμένων ενός χώρου χαρακτηριστικών μπορεί να αλλάξει τις σχέσεις των αποστάσεων που υπολογίζονται από το Ευκλείδειο μέτρο. Παραπάνω φαίνεται πως μία τέτοια κλιμάκωση μπορεί να επηρεάσει τη συμπεριφορά ενός ταξινομητή πλησιέστερου γείτονα. Έστω το σημείο ελέγχου x και ο πλησιέστερος γείτονάς του. Στον αρχικό χώρο (αριστερά), το πλησιέστερο πρότυπο είναι το μαύρο. Στο εικόνα στα δεξιά, ο x_1 άξονας έχει κλιμακωθεί από ένα παράγοντα α = 1/3 και ο πλησιέστερος γείτονας του x είναι το γκρι πρότυπο.

4.6.2 Απόσταση Εφαπτομένης

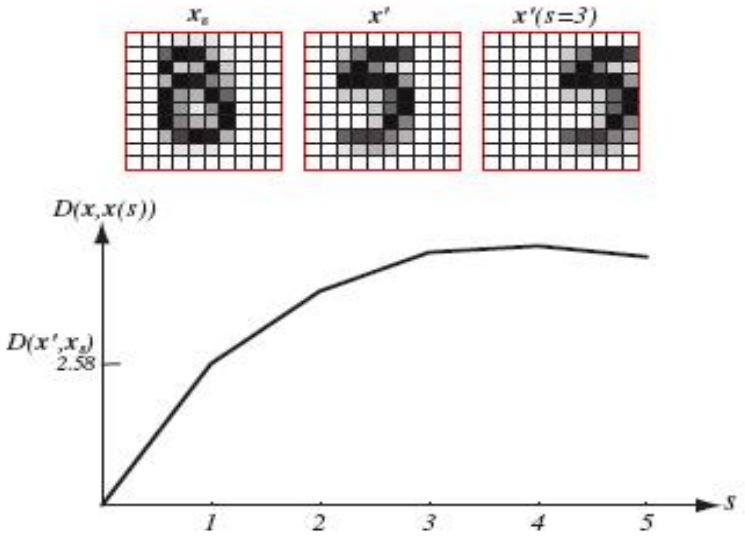
Μπορεί να υπάρξουν σημαντικά μειονεκτήματα εάν γίνει άκριτη χρήση ενός συγκεκριμένου μέτρου σε ταξινομητές πλησιέστερου γείτονα. Αυτά τα μειονεκτήματα μπορούν να αντιμετωπιστούν με προσεκτική χρήση περισσότερο γενικών μέτρων της απόστασης. Ένα πολύ σημαντικό τέτοιο πρόβλημα είναι αυτό της σταθερότητας. Έστω ένα πρότυπο 100 διαστάσεων x΄ που αντιπροσωπεύει μία ασπρόμαυρη εικόνα με 10×10 εικονοστοιχεία ενός συμβόλου «5» γραμμένου με το χέρι. Έστω επίσης, η Ευκλείδεια απόσταση από το x΄ προς το πρότυπο που αντιπροσωπεύει μία εικόνα η οποία είναι ίδια με το x΄ αλλά μετατοπισμένη οριζόντια (εικόνα 4.19). Ακόμα και αν η σχετική μετατόπιση s είναι μικρότερη από τρία εικονοστοιχεία η Ευκλείδεια απόσταση γίνεται πολύ μεγάλη – πολύ μεγαλύτερη από την απόσταση από ένα μη μετατοπισμένο σύμβολο «8». Προφανώς, το Ευκλείδειο μέτρο έχει ελάχιστη σημασία σε ένα ταξινομητή πλησιέστερου γείτονα ο οποίος πρέπει να είναι ανεπηρέαστος από τέτοιου είδους μετασχηματισμούς.

Ομοίως, άλλοι μετασχηματισμοί, όπως η ολική περιστροφή ή κλιμάκωση της εικόνας, δεν αντιμετωπίζονται ικανοποιητικά από την Ευκλείδεια απόσταση με αυτόν τον τρόπο. Τέτοιου είδους μειονεκτήματα εμφανίζονται κυρίως εάν πρέπει ο ταξινομητής να είναι στιγμιαία ανεπηρέαστος από διάφορους μετασχηματισμούς, όπως ο οριζόντιος μετασχηματισμός, ο κάθετος μετασχηματισμός, η συνολική κλιμάκωση, η περιστροφή, η αύξηση του πλάτους των γραμμών, κλπ. Θα μπορούσε κάποιος να προεπεξεργαστεί τις εικόνες διατάσσοντας τα κέντρα τους στο ίδιο σημείο, μετασχηματίζοντάς τα ώστε να περιέχονται στο ίδιο πλαίσιο, και ούτω καθεξής. Μια τέτοια προσέγγιση όμως έχει τις δυσκολίες της, όπως την ευαισθησία στα περιφερειακά εικονοστοιχεία ή/και στο θόρυβο. Στη συνέχεια θα ερευνηθούν εναλλακτικές προσεγγίσεις στην τεχνική της προεπεξεργασίας.



Εικόνα 4.18: Κάθε χρωματισμένη επιφάνεια αποτελείται από σημεία που απέχουν απόσταση 1.0 από την αρχή των αξόνων, μετρημένα χρησιμοποιώντας διαφορετικές τιμές για το k στο μέτρο του Minkowski. Έτσι, οι άσπρες επιφάνειες αντιστοιχούν στη L_1 νόρμα (απόσταση Manhattan), η ανοικτή γκρίζα σφαίρα αντιστοιχεί στη L_2 νόρμα (Ευκλείδεια απόσταση), η σκούρες γκρίζες επιφάνειες αντιστοιχούν στη L_4 νόρμα και οι επιφάνειες του εξωτερικού κουτιού αντιστοιχούν στη L_∞ νόρμα.

Στην ιδανική περίπτωση, δε θα υπολογιστεί η απόσταση μεταξύ δύο προτύπων μέχρις ότου να μετασχηματιστούν με τέτοιο τρόπο ώστε να είναι όσο το δυνατόν πιο όμοια το ένα με το άλλο. Όμως, η υπολογιστική πολυπλοκότητα τέτοιων μετασχηματισμών είναι συχνά πολύ υψηλή. Επιπλέον, η περιστροφή μιας k× k εικόνας κατά ένα γνωστό ποσό και η παρεμβολή σε ένα νέο πλέγμα απαιτεί $O(k^2)$. Φυσικά όμως, δεν είναι γνωστή η κατάλληλη γωνία περιστροφής και πρέπει να γίνει αναζήτηση σε πολλές τιμές, καθεμιά από τις οποίες απαιτεί τον υπολογισμό μιας απόστασης για να ελεγχθεί εάν έχει βρεθεί ή όχι ή βέλτιστη τιμή.
Εάν πρέπει να αναζητηθεί το βέλτιστο σύνολο παραμέτρων για διάφορους μετασχηματισμούς για κάθε αποθηκευμένο πρότυπο κατά τη διάρκεια της ταξινόμησης, ο υπολογιστικός φόρτος είναι απαγορευτικός.

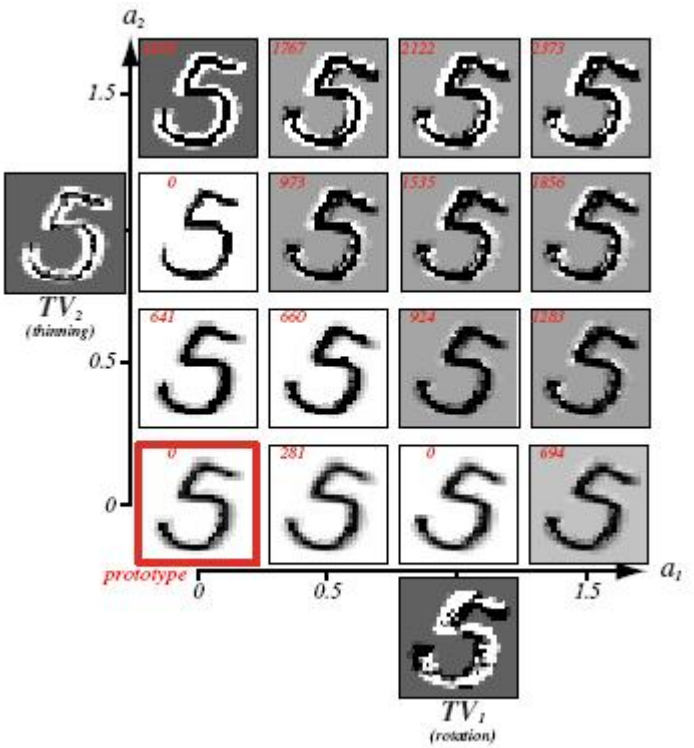


Εικόνα 4.19: Το πρότυπο x΄ αντιπροσωπεύει ένα χαρακτήρα «5» γραμμένο με το χέρι και το πρότυπο x΄(s=3) αντιπροσωπεύει τον ίδιο χαρακτήρα μετατοπισμένο τρία εικονοστοιχεία προς τα δεξιά. Η Ευκλείδεια απόσταση D(x΄, x΄(s=3)) είναι πολύ μεγαλύτερη από την απόσταση D(x΄, x_8), όπου το x_8 αντιπροσωπεύει ένα χαρακτήρα «8» γραμμένο με το χέρι. Η ταξινόμηση πλησιέστερου γείτονα που βασίζεται στην Ευκλείδεια απόσταση σε αυτή την περίπτωση οδηγεί σε πολύ μεγάλα λάθη. Στην πραγματικότητα, πρέπει κανείς να χρησιμοποιεί ένα μέτρο της απόστασης το οποίο να μην είναι ευαίσθητο σε τέτοιους μετασχηματισμούς (μετατόπιση, κλιμάκωση, περιστροφή).

Η γενική προσέγγιση στους ταξινομητές tangent απόστασης είναι η χρήση ενός πρωτοποριακού μέτρου της απόστασης και μιας γραμμικής προσέγγισης των διαφόρων μετασχηματισμών. Έστω ότι υπάρχουν r μετασχηματισμοί που μπορούν να εφαρμοστούν σε ένα πρόβλημα, όπως ο οριζόντιος μετασχηματισμός, ο κάθετος μετασχηματισμός, η περιστροφή, η κλιμάκωση και η λέπτυνση των γραμμών. Κατά τη διάρκεια της κατασκευής του ταξινομητή, σε κάθε αποθηκευμένο πρότυπο x΄ εφαρμόζεται καθένας από τους μετασχηματισμούς $F_i(x'; \alpha_i)$. Έτσι, το $F_i(x'; \alpha_i)$ αντιπροσωπεύει την εικόνα που περιγράφεται από το x΄ περιστραμμένο υπό μια μικρή γωνία α_i. Έπειτα, κατασκευάζεται ένα tangent διάνυσμα TV_i για κάθε μετασχηματισμό

$$TV_i = F_i(x'; a_i) - x' \tag{4.57}$$

Αν και κάθε τέτοιος μετασχηματισμός μπορεί να είναι υπολογιστικά ακριβός – όπως για παράδειγμα ο μετασχηματισμός της λέπτυνσης των γραμμών – χρειάζεται να εφαρμοστεί μόνο μία φορά, κατά τη διάρκεια της εκπαίδευσης όταν οι υπολογιστικοί περιορισμοί είναι ελαστικοί. Με αυτόν τον τρόπο κατασκευάζεται για κάθε πρότυπο x΄ ένας $r \times d$ πίνακας T, που αποτελείται από τα εφαπτόμενα διανύσματα στο x΄. (Τέτοια διανύσματα μπορούν να είναι ορθοκανονικά, αλλά εδώ χρειάζεται απλώς να θεωρηθούν ως γραμμικώς ανεξάρτητα. Πρέπει επίσης να γίνει ξεκάθαρο ότι αυτή η μέθοδος δεν θα λειτουργήσει για δυαδικές εικόνες, διότι αυτές δεν έχουν την απαιτούμενη έννοια της παραγώγου. Εάν τα δεδομένα είναι δυαδικά, είναι σύνηθες να εισάγεται θολούρα στις εικόνες πριν κατασκευαστεί ένας ταξινομητής βασισμένος στην απόσταση της εφαπτομένης.



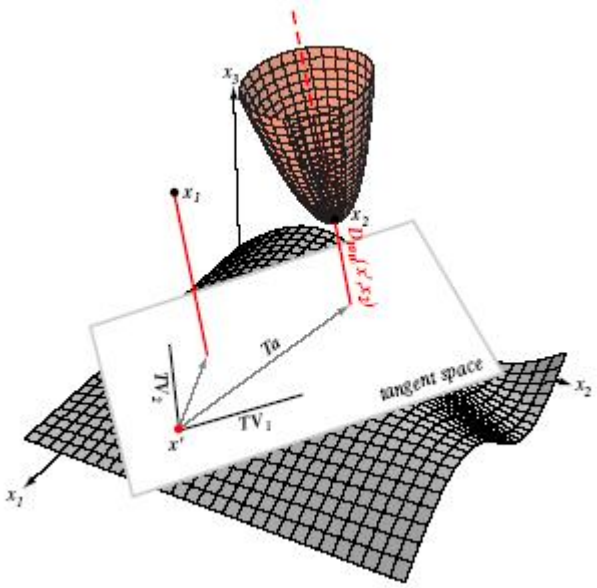
Εικόνα 4.20: Η εικόνα του γραμμένου στο χέρι χαρακτήρα «5» που βρίσκεται κάτω αριστερά υφίσταται δύο μετασχηματισμούς, περιστροφή και λέπτυνση των γραμμών για να προκύψουν τα εφαπτόμενα διανύσματα TV_1 και TV_2. Οι εικόνες που αντιστοιχούν σε αυτά τα διανύσματα φαίνονται έξω από τους άξονες. Καθεμία από τις 16 εικόνες μέσα στα όρια των αξόνων αντιστοιχούν σε ένα πρότυπο που έχει υποστεί ένα γραμμικό μετασχηματισμό των δύο εφαπτόμενων διανυσμάτων με συντελεστές a1 και a2. Ο μικρός αριθμός πάνω δεξία από κάθε εικόνα είναι η Ευκλείδεια απόσταση ανάμεσα στη εφαπτόμενη προσέγγιση και στην εικόνα που προκύπτει από τους μη προσεγγισμένους μετασχηματισμούς. Φυσικά, αυτή η Ευκλείδεια απόσταση ισούται με το 0 για το ίδιο το πρότυπο και για τους μετασχηματισμούς όπου a1 = 1, a2 = 0 και a1 = 0, a2 = 1. Τα πρότυπα που δημιουργούνται με a1 + a2 > 1 έχουν σκούρο φόντο λόγω της αυτόματης εφαρμογής του γκρι μετασχηματισμού στις εικόνες που έχουν αρνητικές τιμές εικονοστοιχείων.

Κάθε σημείο στον υποχώρο που εκτείνεται από τα r εφαπτόμενα διανύσματα που περνάνε από το x΄ αντιπροσωπεύει μία γραμμική προσέγγιση του πλήρους συνδυασμού των μετασχηματισμών, όπως φαίνεται στο εικόνα 4.21. Κατά τη διάρκεια της ταξινόμησης αναζητείται το σημείο στο tangent χώρο το οποίο είναι πλησιέστερο ως προς ένα σημείο ελέγχου x – που αποτελεί το γραμμικό μετασχηματισμό. Όπως θα δειχθεί στη συνέχεια, αυτή η αναζήτηση μπορεί να είναι ιδιαίτερα γρήγορη.

Στη συνέχεια θα υπολογιστεί η απόσταση της εφαπτομένης από ένα σημείο ελέγχου x προς ένα συγκεκριμένο αποθηκευμένο πρότυπο x΄. Δεδομένου λοιπόν ενός πίνακα T ο οποίος αποτελείται από τα r εφαπτόμενα διανύσματα στο x΄, η απόσταση της εφαπτομένης από το x΄ στο x είναι:

$$D_{\tan}(x', x) = \min_{a}\left[\left\|(x' + Ta) - x\right\|\right] \tag{4.58}$$

δηλαδή, η Ευκλείδεια απόσταση από το x ως προς τον εφαπτόμενο χώρο του x΄. Η εξίσωση 4.60 περιγράφει την αποκαλούμενη «μονόπλευρη» απόσταση της εφαπτομένης, επειδή μόνο ένα πρότυπο, το x΄, μετασχηματίζεται. Η «από δύο πλευρές» απόσταση της εφαπτομένης επιτρέπει τόσο στο x όσο και στο x΄ να μετασχηματίζονται αλλά βελτιώνει ελάχιστα την ακρίβεια προσθέτοντας σημαντικό υπολογιστικό φόρτο. Για αυτό το λόγο το ενδιαφέρον επικεντρώνεται στην μονόπλευρή εκδοχή.
Κατά τη διάρκεια της ταξινόμησης του x η απόσταση της εφαπτομένης ως προς το x΄ προκύπτει από την εύρεση της βέλτιστης τιμής για το a που απαιτείται από την εξίσωση 4.58. Αυτή η ελαχιστοποίηση είναι στην πραγματικότητα σχετικά απλή, επειδή η τετραγωνική απόσταση που πρέπει να ελαχιστοποιηθεί είναι μία τετραγωνική συνάρτηση του a, όπως φαίνεται και στο εικόνα 4.21. Η βέλτιστη τιμή για το a βρίσκεται χρησιμοποιώντας μια απλή τεχνική αναζήτησης, όπως η επαναληπτική gradient descent ή κάποιες μεθόδους πινάκων.



Εικόνα 4.21: Ένα αποθηκευμένο πρότυπο x΄, εάν μετασχηματιστεί με συνδυασμό δύο από τους βασικούς μετασχηματισμούς, θα βρεθεί κάπου μέσα σε μία πολύπλοκη καμπυλωτή επιφάνεια στον πλήρη d – διάστατο χώρο. Ο χώρος των εφαπτομένων στο x΄ είναι ένας r – διάστατος Ευκλείδειος χώρος, που εκτείνεται από τα εφαπτόμενα διανύσματα (TV_1 και TV_2). Η απόσταση της εφαπτομένης $D_{tan}(x', x)$ είναι η ελάχιστη Ευκλείδεια απόσταση από το x στο χώρο των εφαπτομένων του x΄, όπως φαίνεται από τις δύο έντονες γραμμές για δύο σημεία x_1 και x_2. Έτσι, εάν και η απόσταση από το x΄ στο x_1 είναι μικρότερη από αυτή ως προς το x_2, για την απόσταση των εφαπτόμενων η κατάσταση είναι αντίθετη. Η Ευκλείδεια απόσταση από το x_2 στο χώρο των εφαπτόμενων του x΄ είναι μία τετραγωνική συνάρτηση του διανύσματος των παραμέτρων a. Επομένως, απλές gradient descent μέθοδοι μπορούν να βρουν τη βέλτιστη τιμή για το διάνυσμα a και προφανώς και την απόσταση της εφαπτομένης $D_{tan}(x', x_2)$.

4.7 ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] David W. Aha, editor. Lazy Learning. Kluwer, Boston, MA, 1997.

[2] Mark A. Aizerman, Emmanuil M. Braverman. and Leo I, Rozonoer. The Robbins-Monro process and the method of potential functions. Automation and Remote Control, 26:1882-1885, 1965.

[3] Claudi Alsina, Enric Trillas, and Llorenc Valverde. On some logical connectives for fuzzy set theory. Journal of Mathematical Analysis and Applications, 93{l):15-26,

[4] David Avis and Binay K. Bhattacharya. Algorithms for computing d -dimensional Voronoi diagrams and their duals. In Franco P. Preparata, editor, Advances in Computing Research: Computational Geometry, pages 159-180, JAI Press, Greenwich, CT, 1983.

[5] James C. Bezdek and Sankar K. Pal, editors. Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data, IEEE Press, New York, 1992.

[6] Emmanuil M. Braverman. On the potential function method. Automation and Remote Control, 26:2130-2138, 1965.

[7] Peter Cheeseman. In defense of probability. In Proceedings of the Ninth International Joint Conference on Artificial Intelligence, pages 1002-1009. Morgan Kauf mann. San Mateo, CA, 1985.

[8] Peter Cheeseman. Probabilistic versus fuzzy reasoning. In Laveen N. Kanal and John F. Lemmer. editors. Uncertainty in Artificial Intelligence, pages 85-102, Else-vier Science Publishers. Amsterdam, 1986.

[9] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, IT-13(l):21-27 1967

[10] Richard T. Cox. Probability frequency and reasonable expectation. American Journal of Physic 14(1) 1-13 1946.

[11] Belur V. Dasarath, editor Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques IEEE Computer Society, Washington DC 1991

[12] Luc P. Devroye. On the inequality of Cover and Hart in nearest neighbor discrimination IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 3(1):75-78, 1981.

[13] Evelyn Fix and Joseph L Hodge Jr Discriminatory analysis: Nonparametric discrimination Consistency properties. USAF School of Aviation Medicine 4:261-279, 1951.

[14] Evelyn Fix and Joseph L Hodges Jr. Discrimina tory analysis: Nonparametric discrimination Small sample performance. USAF School of Aviation Medicine II, 280-322, 1952.

[15] James D. Foley, Andne Van Dam Steven K. Feiner and John F. Hughes. Fundamental of Interactive Computer Graphics: Principle and Practice Addisson Wesley Reading, MA, second edition 1990

[16] Jerome H. Friedman. An overview of predictive learning and function approximation. In Vladimir Cherkassky, Jerome H. Friedman, and Harry Wechsler, editors. From Statistics to Neural Networks: Theory and Pattern Recognition Applications, pages 1-61, Springer-Verlag, NATO AS1. New York. 1994.

[17] Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. ACM Transactions on Mathematical Software, 3(3):209-226, 1977.

[18] Allen Gersho and Robert M. Gray. Vector Quantization and Signal Processing. Kluwer Academic Publishers, Boston, MA, 1992.

[19] Richard M. Golden. Mathematical Methods for Neural Network Analysis and Design. MIT Press, Cambridge, MA, 1996.
[20] Peter Hart. The condensed nearest neighbor rule. IEEE Transactions on Information Theory, IT-14(3):515-516, 1968.
[21] Trevor Hastie, Patrice Simard, and Eduard Sackinger. Learning prototype models for tangent distance. In Gerald Tesauro, David S. Touretzky, and Todd K. Leen, editors. Advances in Neural Information Processing Sys-terns, volume 7, pages 999-1006. Cambridge, MA, 1995. MIT Press.
[22] Anil K. Jain and Madras D. Ramaswami. Classifier design with Parzen windows. In Edzard S. Gelsema and Laveen N. Kanal. editors, Pattern Recognition and Artificial Intelligence, pages 211-227. Elsevier Science Publishers, New York, 1988.
[23] Edwin T. Jaynes. Probability Theory: The Logic of Science (unpublished manuscript), unpublished edition, 1994.
[24] Abraham Kandel. Fuzzy Techniques in Pattern Recognition. Wiley, New York, 1982.
[25] John Maynard Keynes. A Treatise on Probability. Macmillan, New York, 1929.
[26] Donald E. Knuth. The An of Computer Programming, volume 1. Addison-Wesley, Reading, MA, first edition. 1973.
[27] Bart Kosko. Fuzziness vs. probability. International Journal of General Systems, 17(2):211-240. 1990.
[28] Jan Lukasiewicz. Logical foundations of probability theory. In Ludwik Borkowski. editor. Jan Lukasiewicz: Selected Works, pages 16-43. North-Holland, Amsterdam, 1970.
[29] Joseph L. Mundy and Andrews Zisserman, editors. Geometric Invariance in Computer Vision. MIT Press, Cambridge, MA, 1992.
[30] Elizbar A. Nadaraya. On estimating regression. Theory of Probability and Its Applications, 9(1): 141-142, 1964.
[31] Emanuel Parzen. On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33(3):1065-1076, 1962.
[32] Edward A. Patrick and Frederick P. Fischer, ID. A generalized k-nearest neighbor rule. Information and Control, 16(2):128-152, 1970.
[33] Witold Pedrycz and Fernando Gomide. An Introduction to Fuzzy Sets. MIT Press, Cambridge, MA, 1998.
[34] Joseph S. Perkell and Dennis H. Klatt, editors. Invariance and Variability in Speech Processes. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.
[35] Franco P. Preparata and Michael Ian Shamos. Computational Geometry: An Introduction. Springer-Verlag, New York, 1985.
[36] Douglas L. Reilly and Leon N. Cooper. An overview of neural networks: Early models to real world systems. In Steven F. Zometzer, Joel L. Davis, Clifford Lau, and Thomas McKenna, editors. An Introduction to Neural and Electronic Networks, pages 229-250. Academic Press, New York, second edition, 1995.
[37] Douglas L. Reilly, Leon N. Cooper, and Charles Elbaum, A neural model for category learning. Biological Cybernetics, 45(1)35-41, 1982.
[38] Bernhard Scholkopf, Christopher I. C. Burges, and Alexander J. Smola, editors. Advances in Kernel Methods: Support Vector Learning. MTT Press, Cambridge, MA, 1999.
[39] Bernard W. Silverman and M. Christopher Jones. E. Fix and J. L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation. International Statistical Review, 57 (3):233-247, 1989.
[40] Patrice Simard, Yann Le Cun, and John Danker. Efficient pattern recognition using a new transformation distance. In Stephen J. Hanson, Jack D. Cowan, and C, Lee Giles, editors. Advances in Neural Information Processing Systems, volume 5, pages 50-58, Morgan Kaufmann, San Mateo, CA, 1993.
[41] Donald F. Specht. Generation of polynomial discriminant functions for pattern recognition. IEEE Transactions on Electronic Computers, EC-16(3):308-319, 1967.

[42] Alessandro Sperduti and David G. Stork. A rapid graph-based method for arbitrary transformation-invariant pattern classification. In Gerald Tesauro, David S. Touretzky, and Todd K. Leen, editors, Advances in Neural Information Processing Systems, volume 7, pages 665-672, MIT Press, Cambridge, MA, 1995.

[43] Godfried T. Toussaint, Binay K. Bhattacharya, and Ronald S. Poulsen. Application of Voronoi diagrams to nonparametric decision rules. In Proceedings of Computer Science and Statistics: The 16th Symposium on the Interface, pages 97-108, North-Holland, Amsterdam. 1984.

[44] Geoffrey S. Watson. Smooth regression analysis. Sankhya: The Indian Journal of Statistics, Series A, 26:359-372, 1964.

[45] Lofti Zadeh. Fuzzy Sets. Information and Control, 8(3): 338-353, 1965.