ΕΡΓΑΣΙΑ 4

ΚΑΡΑΠΕΠΕΡΑ ΕΛΠΙΔΑ

57423

—

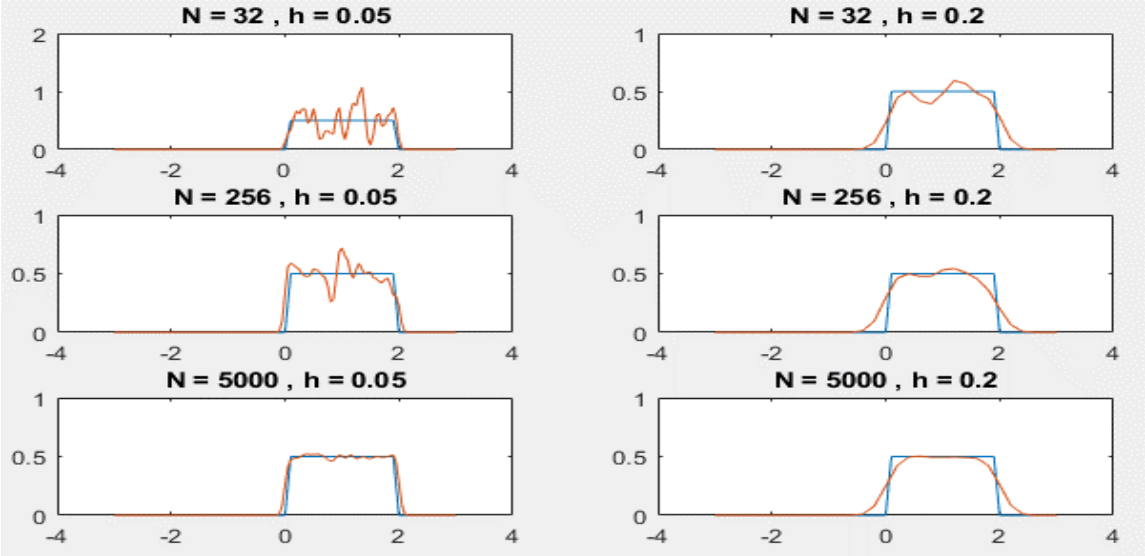
ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

—

2020

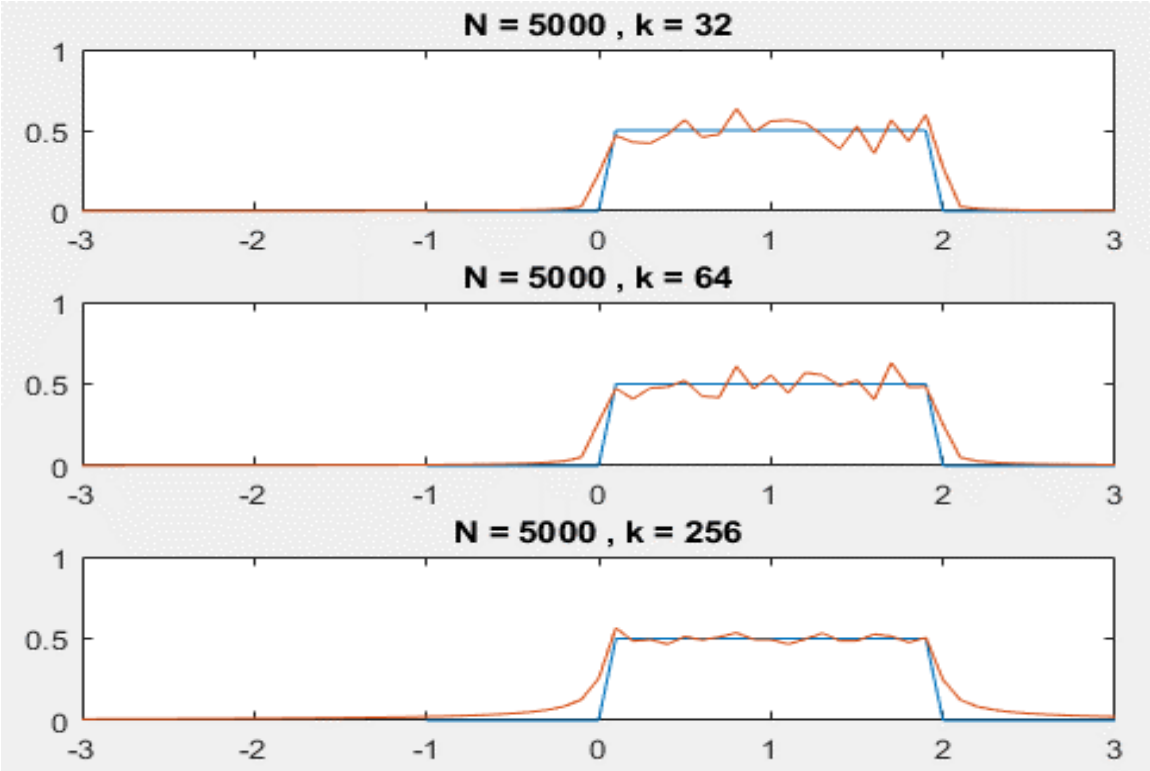
Άσκηση 4.1

α)



Το N είναι ο αριθμός των γκαουσιανών που χρησιμοποιήθηκαν (μία γκαουσιανή για κάθε σημείο). Όπως φαίνεται και στην παραπάνω εικόνα, όσο μεγαλύτερο είναι το N, τόσο περισσότερα “spikes” θα έχουμε (τόσο πιο πολλές γκαουσιανές), αλλά και τόσο πιο ακριβής θα είναι η καμπύλη που θα σχεδιαστεί, αφού θα έχουν χρησιμοποιηθεί περισσότερα σημεία για την κατασκευή της. Το h είναι παράγοντας εξομάλυνσης (είναι το σ της κάθε γκαουσιανής, δηλαδή το πόσο θα εξαπλωθεί). Οπότε όσο το h μεγαλώνει, τόσο θα εξομαλύνεται η καμπύλη. Αυτό φαίνεται και στην εικόνα, όπου φαίνεται τα spikes της καμπύλης με h=0.05 να εξομαλύνονται σημαντικά με h=0.2. Η εξομάλυνση είναι σημαντική για να μην συμβεί overfit στο μοντέλο αν αυτή είναι πολύ μικρή, ή υπεργενίκευση αν αυτή είναι πολύ μεγάλη.

β)



Το knn βρίσκει τις εκ των υστέρων πιθανότητες. Σε κάθε μικρή περιοχή x η εκ των υστέρων πιθανότητα βρίσκεται από το ποσοστό των δειγμάτων εντός της περιοχής που έχουν ετικέτα ω_i. k είναι ο αριθμός των δειγμάτων που περιλαμβάνει η περιοχή γύρω από το x. Όσο μεγαλύτερο είναι το k, άρα και ο αριθμός των δειγμάτων που λαμβάνουμε υπόψιν, τόσο πιο ακριβές είναι το αποτέλεσμα. Αυτό επιβεβαιώνεται και από το παραπάνω σχήμα.

Φυσικά, από ένα σημείο και μετά, όταν το k μεγαλώσει αρκετά, χειροτερεύει το αποτέλεσμα, αφού πλέον δε λαμβάνει υπόψιν μόνο γειτονικά δείγματα.

ΕΚΤΙΜΗΣΗ ΚΝΝ

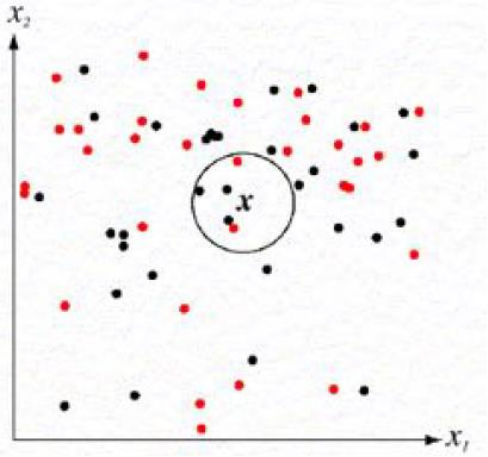
Το πλάτος του παραθύρου επιλέγεται ως μία συνάρτηση των δεδομένων εκπαίδευσης.
Ξεκινάω από μία εκτίμηση με ένα παράθυρο με λίγα στοιχεία και μεγαλώνω το παράθυρο μέχρι αυτό να περιέχει ένα προκαθορισμένο πλήθος δειγμάτων k_n.
Η πυκνότητα υπολογίζεται από τον τύπο:

$$\frac{k_n/n}{V_n}$$

Προσπαθούμε να ικανοποιήσουμε τις συνθήκες:

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

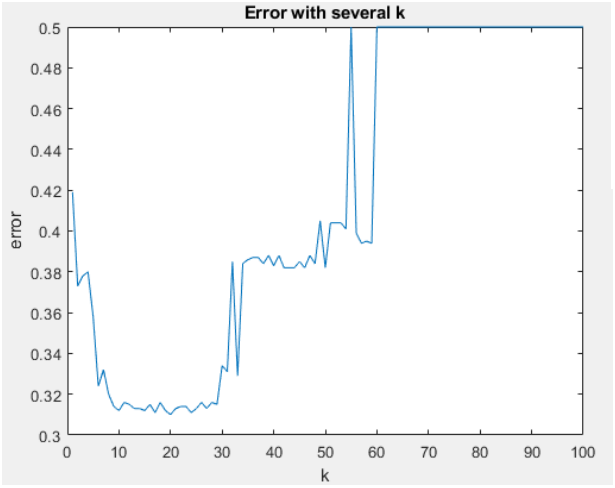
Το σημείο x που εξετάζεται τοποθετείται στην κλάση στην οποία ανήκει η πλειοψηφία των k κοντινότερων γειτόνων του.



Άσκηση 4.2

B)

Τα αποτελέσματα που περιμένουμε είναι όσο αυξάνεται το k τόσο να μειώνεται το σφάλμα. Πράγματι, φαίνεται για k=2 να έχει τη διαφορά που περιμένουμε από το k=1. Το k=3 έχει περίπου το ίδιο σφάλμα με το k=2.




```
For k = 1 Error = 0.419
For k = 2 Error = 0.373
For k = 3 Error = 0.378
With Bayesian classifier = 0.317
minimum error: 0.31 with k = 20
```

Σε σχέση με τον Bayesian ταξινομητή παρατηρούμε ότι το μικρότερο σφάλμα που μπορεί ν πετύχει ο knn (0.31 για k = 20) είναι καλύτερο από αυτό του Bayesian.

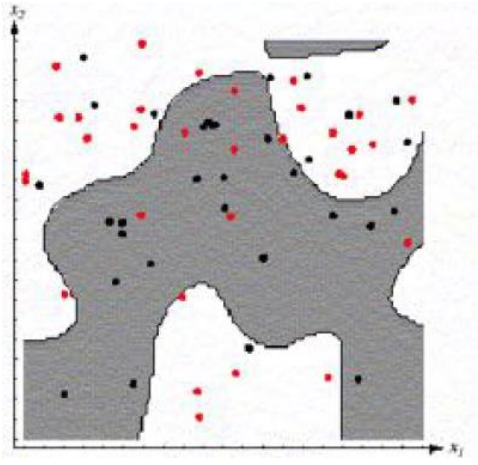
Στο σχήμα φαίνεται και αυτό που αναφέρθηκε στο ερώτημα 4.1.β) : ότι από ένα σημείο και μετά, όσο αυξάνει το k η πρόβλεψη γίνεται όλο και λιγότερο αξιόπιστη, γιατί οι γείτονες που λαμβάνει υπόψιν του ο ταξινομητής knn είναι πολύ μακρινοί.

ΠΑΡΑΘΥΡΑ PARZEN

Για να βρω την πιθανότητα P(x=x'|ω2) ολοκληρώνω την πραγματική συνάρτηση πυκνότητας πιθανότητας σε μια περιοχή R που περιέχει το x'.
Όσο αυξάνεται το data set μικραίνει το παράθυρο R.
Ξεκινάω από μία συγκεκριμένη περιοχή R_1 με όγκο V_n και καθώς αυξάνω το n μειώνεται ο όγκος, προσπαθώντας να ικανοποιήσω τη σχέση:

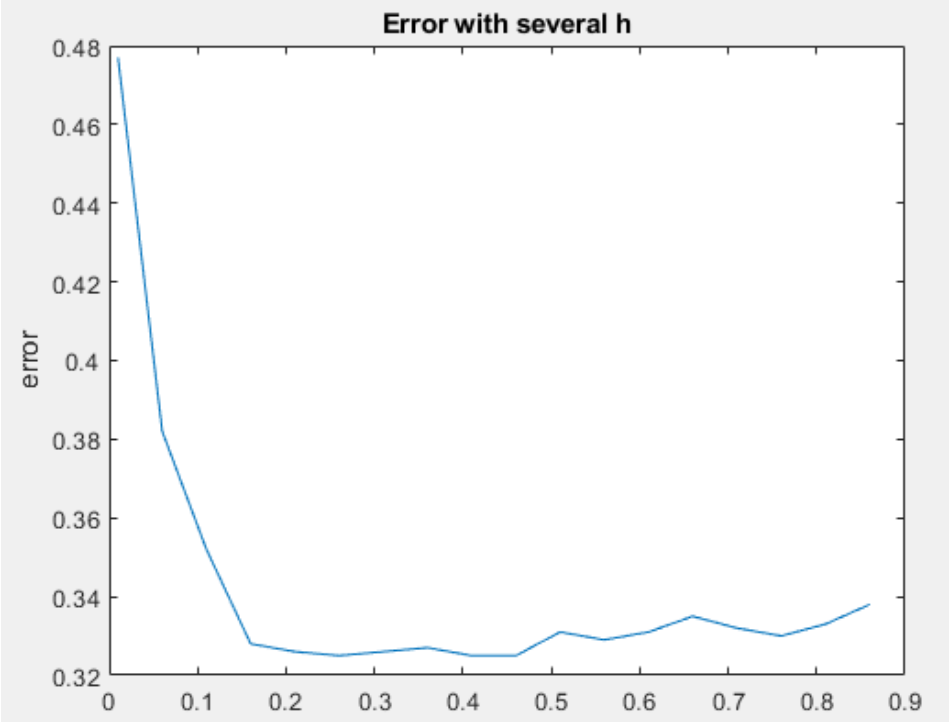
$$\lim_{n\to\infty} V_n = 0$$

Με τον τρόπο αυτό γίνεται εκτίμηση της πιθανοφάνειας $p(x|\omega_i)$. Στη συνέχεια, για την ταξινόμηση ενός σημείου x χρησιμοποιείται ο κανόνας Bayes : γίνεται υπολογισμός των εκ των υστέρων πιθανοτήτων και το σημείο x κατατάσσεται στην κλάση με την μεγαλύτερη πιθανότητα.



Γ)

Τα h που ελέγχθηκαν είναι τα h = [0.01:0.05:0.9].



Παρατηρούμε τη διακύμανση του Error σε σχέση με το h. Το μικρότερο error που επιτεύχθηκε είναι 0.325, για h = 0.26.

PARZEN PNN

PPNN είναι συνδυασμός της εκτίμησης της pdf με το παράθυρο Parzen και την ταξινόμηση Bayesian. Για ένα διάνυσμα χαρακτηριστικών x επιλέγεται η κλάση ω_i όπου το P (ω_i | x) είναι το μέγιστο.

Ένα PPNN είναι ένα νευρωνικό δίκτυο δύο επιπέδων (NN) όπου τα δεδομένα εισόδου είναι πλήρως συνδεδεμένα με το πρώτο στρώμα νευρώνων και το τελευταίο είναι αραιά συνδεδεμένο με το δεύτερο στρώμα (και που είναι και το ouput).

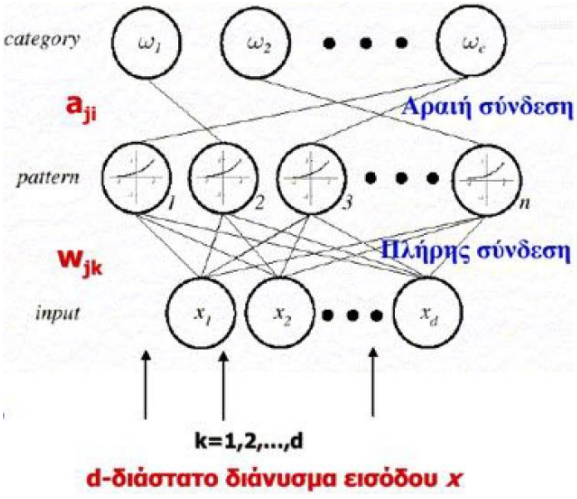
Το επίπεδο εξόδου αποτελείται από c νευρώνες, όπου c είναι ο αριθμός των κατηγοριών του ταξινομητή.

Τα weights στο πρώτο επίπεδο εκπαιδεύονται ως εξής: κάθε δείγμα δεδομένων κανονικοποιείται έτσι ώστε το μήκος του να είναι ενιαίο, κάθε δείγμα δεδομένων γίνονται νευρώνες με τις ομαλοποιημένες τιμές ως βάρη w.

Τα δεδομένα εισαγωγής x πολλαπλασιάζονται με τα βάρη που λαμβάνουν την ενεργοποίηση του δικτύου $net = w^{Tx}$.

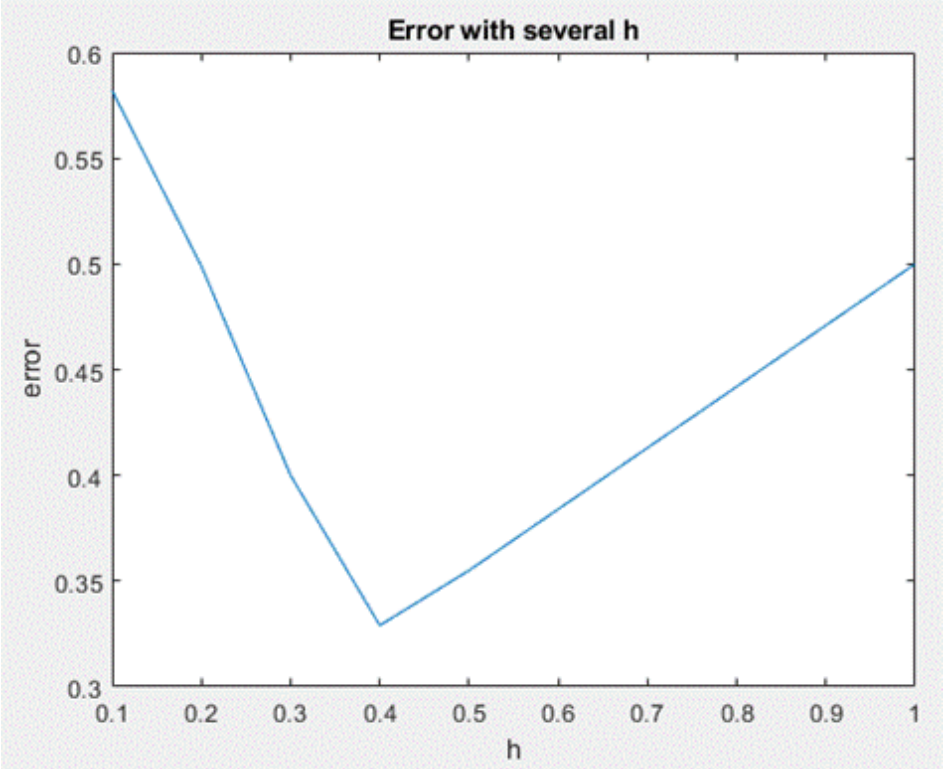
PARZEN PNN

Στη συνέχεια υπολογίζεται η εκθετική μη γραμμικότητα για να ληφθούν τα σήματα συναπτικής ενεργοποίησης. Κατά τη διάρκεια της διαδικασίας εκμάθησης, κάθε νευρώνας πρώτου στρώματος συνδέεται με τον νευρώνα στρώματος εξόδου που σχετίζεται με την τάξη του με weight 1. Κατά τη διαδικασία ταξινόμησης, ο νευρώνας εξόδου κάθε τάξης αθροίζει τα σήματα ενεργοποίησης από όλους τους νευρώνες του πρώτου στρώματος. Η υψηλότερη τιμή εξόδου επιλέγει την κλάση των δεδομένων εισαγωγής.



Δ)

Τα h που ελέγθηκαν είναι τα h = [0.1 0.2 0.3 0.4 0.5 1].



Παρατηρούμε τη διακύμανση του Error σε σχέση με το h. Το μικρότερο error που επιτεύχθηκε είναι 0.329, για h = 0.4.