**ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ**

**ΠΡΩΤΗ ΠΡΑΚΤΙΚΗ ΑΣΚΗΣΗ**

**Να λύσετε τουλάχιστον 4 από τα 6 υποθέματα του Θέματος 1 και 2 από τα 4 υποθέματα του Θέματος 2. Επιτυχής επίλυση περισσοτέρων υποθεμάτων δίνει bonus βαθμολογία.**

**Μπορείτε να χρησιμοποιήσετε όποια γλώσσα προγραμματισμού επιθυμείτε.**

**Να εργασθείτε σε ομάδες των 3 ατόμων**

**Ημερομηνία παράδοσης εργασίας: Τρίτη, 11 Δεκεμβρίου 2018 (εκτυπωμένες αναφορές)**

**Θέμα 1**

Ας θεωρήσουμε το πρόβλημα γενικευμένης γραμμικής παλινδρόμησης που ορίζεται από το εξής μοντέλο:

 (1)

Όπου τοαντιστοιχεί σε λευκό Γκαουσιανό θόρυβο και οι συνιστώσες του διανύσματος των παραμέτρων έχουν τις εξής τιμές:

, , , , . (2)

Σε κάθε μια από τις παρακάτω περιπτώσεις, λαμβάνουμεισαπέχοντα σημεία  στον άξονα των *x* και στο διάστημα [0,2] και τα χρησιμοποιούμε για να δημιουργήσουμε δείγματα του συνόλου εκπαίδευσης:

 (3)

όπου τα είναι i.i.d. δείγματα θορύβου, που προέρχονται από Γκαουσιανή κατανομή με μέσο 0 και διασπορά.

1. Χρησιμοποιώντας,=0.1 και τη δομή του σωστού μοντέλου (πολυώνυμο 5ου βαθμού με τον τεταρτοβάθμιο όρο μηδενικό), να εφαρμόσετε τη μέθοδο των Ελαχίστων Τετραγώνων (Least Squares) για να εκτιμήσετε το διάνυσμα των παραμέτρων. Υπολογίστε το Μέσο Τετραγωνικό Σφάλμα (Mean Square Error) για την εκτίμηση του *y* επί του συνόλου εκπαίδευσης και επί ενός συνόλου ελέγχου που αποτελείται από 1000 σημεία που λαμβάνονται τυχαία στο διάστημα [0,2].
2. Για και=0.1 να επιχειρήσετε παρεμβολή με τη μέθοδο των Ελαχίστων Τετραγώνων με χρήση πολυωνύμου 2ου βαθμού. Να εκτελέσετε 100 πειράματα χρησιμοποιώντας διαφορετικά δείγματα θορύβου στο σύνολο εκπαίδευσης για κάθε πείραμα. Για κάθε σημείο του εκπαιδευτικού συνόλου, να υπολογίσετε τη μέση τιμή και τη διασπορά της εκτίμησης για το *y* επί του συνόλου των 100 πειραμάτων και να τις απεικονίσετε στο επίπεδο (*x,y*) μαζί με την καμπύλη του πραγματικού μοντέλου.

Επαναλάβετε χρησιμοποιώντας πολυώνυμο 10ου βαθμού. Να συγκρίνετε τα αποτελέσματα από τις 2 περιπτώσεις (πολυώνυμο 2ου και πολυώνυμο 10ου βαθμού) με αναφορά στο δίλημμα bias-variance.

1. Να επαναλάβετε το 1) παραπάνω, εφαρμόζοντας τη μέθοδο της Παλινδρόμησης Κορυφογραμμής (Ridge Regression) με διαφορετικά *λ*, αντί για τη μέθοδο των Ελαχίστων Τετραγώνων. Να αναφέρετε αν για κάποιες από τις τιμές του *λ* παρατηρήσατε βελτίωση στην εκτίμηση του Μέσου Τετραγωνικού Σφάλματος.
2. Κωδικοποιούμε την πρότερη γνώση μας για το άγνωστο διάνυσμα παραμέτρων μέσω Γκαουσιανής κατανομής G(**θ**) με μέσο  ίσο με το πραγματικό διάνυσμα παραμέτρων της εξίσωσης (1) και πίνακα συνδιασποράς . Να χρησιμοποιήσετε τη δομή του πραγματικού μοντέλου και Μπεϋζιανό Συμπερασμό (Bayesian Inference) για να εκτιμήσετε το *y* για 20 τυχαία επιλεγμένα σημεία ελέγχου στο διάστημα [0,2] για δυο διαφορετικές τιμές του(0.05 and 0.15). Να απεικονίσετε τις εκτιμήσεις σας με τα σφάλματά τους στο επίπεδο (*x,y*).
3. Να επαναλάβετε το (4) χρησιμοποιώντας για την G(**θ**) μέσο διαφορετικό από το διάνυσμα παραμέτρων του πραγματικού μοντέλου:



Με =0.05, να εκτελέσετε το πείραμα τέσσερις φορές, χρησιμοποιώντας δύο διαφορετικές τιμές του (0.1 and 2) και δύο διαφορετικές τιμές του(20 and 500). Σχολιάστε τα αποτελέσματα.

1. Εδώ θα προσπαθήσουμε να ανακτήσουμε την (θεωρούμενη άγνωστη) διασπορά του θορύβου με τη μέθοδο της Αναμενόμενης Τιμής-Μεγιστοποίησης (Expectation-Maximization). Να κατασκευάσετε ένα σύνολο εκπαίδευσης μεκαι=0.05. Εκκινήστε τον αλγόριθμο με . Αφού επιτευχθεί σύγκλιση, να εκτιμήσετε τα *y* και το σφάλμα τους για 20 σημεία ελέγχου τυχαία επιλεγμένα στο διάστημα [0,2] και να τα απεικονίσετε στο επίπεδο (*x,y*) μαζί με την καμπύλη του πραγματικού μοντέλου.

**Θέμα 2**

1. Να υλοποιήσετε έναν ταξινομητή πλησιέστερων γειτόνων (k-ΝΝ). Να τον χρησιμοποιήσετε για να επιλύσετε τα εξής γνωστά από τη βιβλιογραφία προβλήματα:
2. IRIS PLANT DATABASE (ταξινόμηση φυτών Ιris σε τρία είδη).
3. PIMA INDIANS DIABETES DATABASE (ταξινόμηση εγγύων ινδιάνων της φυλής Pima σε άτομα που έχουν ή δεν έχουν διαβήτη).

Τα δεδομένα των προβλημάτων θα τα βρείτε στο αρχείο UCIdata-exercise1.rar.

Να μελετήσετε το ποσοστό ορθής ταξινόμησης ως συνάρτηση του αριθμού των πλησιέστερων γειτόνων. Η εκτίμηση του αποτελέσματος να γίνει με χρήση της μεθόδου tenfold crossvalidation.

1. Για το δεύτερο πρόβλημα, σχεδιάστε έναν ταξινομητή Bayes με υποκείμενες κανονικές κατανομές για τις πυκνότητες πιθανότητας των 2 κατηγοριών (με δικές σας παραδοχές για τους πίνακες συνδιασπορών, βασισμένες στα δεδομένα) και συγκρίνετε την επίδοσή του με αυτή του ταξινομητή k-ΝΝ χρησιμοποιώντας tenfold crossvalidation.
2. Να υλοποιήσετε έναν απλοϊκό ταξινομητή Bayes με υποκείμενες κανονικές κατανομές για το δεύτερο πρόβλημα, να αποτιμήσετε την επίδοσή του και να τη συγκρίνετε με τους προηγούμενους ταξινομητές (πάλι με tenfold crossvalidation).
3. Να υλοποιήσετε τον αλγόριθμο του perceptron και να τον χρησιμοποιήσετε στο πρόβλημα IRIS PLANT DATABASE ως εξής: Να εξετάσετε αν τα πρότυπα της κάθε κατηγορίας είναι γραμμικά διαχωρίσιμα από τα πρότυπα των συνδυασμένων άλλων δύο κατηγοριών (π.χ. αν τα πρότυπα της Iris Setosa είναι γραμμικά διαχωρίσιμα από τα ομαδοποιημένα πρότυπα της Iris Versicolor και της Iris Virginica).