Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων
Ακαδημαϊκό έτος 2017-18

Μεταπτυχιακό μάθημα: "Εξόρυξη Δεδομένων"

1η Σειρά Ασκήσεων

(Ημερομηνία παράδοσης : 9/5/2018)

Πρόβλημα: Ταξινόμηση (κατηγοριοποίηση) δεδομένων

Χρησιμοποιήστε τα δύο πειραματικά σύνολα δεδομένων που θα σας δοθούν (το όνομα των δύο συνόλων θα είναι το επίθετο των μελών της ομάδας – το πρώτο με κατάληξη *.txt* και το δεύτερο με κατάληξη *.mat*). Στα σύνολα αυτά θα εφαρμόσετε τέσσερις (4) εναλλακτικές μεθόδους ταξινόμησης και θα μετρήσετε την απόδοση (***accuracy*** - ποσοστό επιτυχίας) με την μέθοδο ***cross validation***:

Θα χωρίσετε με τυχαίο τρόπο το σύνολο δεδομένων σε *10* ξένα μεταξύ τους υποσύνολα (*10 fold cross validation*) και σε κάθε ένα θα μετρήσετε την επίδοση της μεθόδου (*accuracy*) κάνοντας εκπαίδευση στα υπόλοιπα *9 folds*. Η συνολική επίδοση κάθε μεθόδου ταξινόμησης θα προκύψει από την μέση επίδοσή του στα *10 folds*.

Οι ταξινομητές που θα χρησιμοποιήσετε είναι οι:

- ***kNN*** *– Nearest Neighbor classifier*: δοκιμάστε διάφορες τιμές του *k=[1, 9]* και επιλέξτε τον καλύτερο ταξινομητή,
- ***Naïve Bayes*** *classifier* υποθέτοντας κανονική κατανομή,
- ***SVM* (*Support Vector Machines*)** ταξινομητή με *RBF kernel function* και με *linear kernel*. Ειδικά για την πρώτη περίπτωση (RBF) δοκιμάστε διαφορετικές τιμές πλάτους του πυρήνα (σ) επιλέγοντας κάθε φορά την καλύτερη τιμή.
- ***Decision Trees,*** *ρυθμίζοντας με διάφορους τρόπους την πολυπλοκότητα του δέντρου που προκύπτει (number of nodes ή leaf size*)

Δώστε ένα σύντομο report με τον τρόπο κατασκευής των μεθόδων και τα αποτελέσματα των δοκιμών ανά μέθοδο, όπως επίσης και την βέλτιστη μέθοδο που θα προκύψει από την σύγκριση μεταξύ των διαφορετικών μεθόδων. Να δοθεί επίσης και ο κώδικας σε Matlab που χρησιμοποιήθηκε με τη μορφή παραρτήματος.