Τμήμα Μηχανικών Η/Υ & Πληροφορικής,
Πανεπιστήμιο Ιωαννίνων
Ακαδημαϊκό έτος 2019-20

Μεταπτυχιακό μάθημα: "Εξόρυξη Δεδομένων"

1η Σειρά Ασκήσεων

(Ημερομηνία παράδοσης : 13/5/2020)

Ταξινόμηση συνόλου δεδομένων

Χρησιμοποιώντας το σύνολο δεδομένων ταξινόμησης δύο κατηγοριών που θα σας διανεμηθεί, να εξετάσετε τέσσερεις μεθόδους ταξινόμησης και να μετρήσετε τη γενίκευσή τους ως προς τη μετρική ***accuracy*** (ποσοστό επιτυχίας) με την μέθοδο ***10-fold cross validation***. Η κατηγορία αναφέρεται στην τελευταία στήλη του πίνακα δεδομένων.

Οι μέθοδοι ταξινόμησης που θα εξετάσετε είναι οι εξής:

- ***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*)

Να δημιουργήσετε μια σύντομη αναφορά σχετικά με τον τρόπο κατασκευής των ταξινομητών και τα αποτελέσματα των δοκιμών ανά μέθοδο, και να αναφέρετε την βέλτιστη μέθοδο που θα προκύψει από την σύγκριση μεταξύ των διαφορετικών μεθόδων και υπερπαραμέτρων. Να παραθέσετε επίσης τον κώδικα που χρησιμοποιήσατε με τη μορφή παραρτήματος.