

Προπτυχιακό μάθημα: “Αναγνώριση Προτύπων”

1^η Σειρά Ασκήσεων

(Ημερομηνία παράδοσης : έως Παρασκευή 7/4/2017)

Πρόβλημα: Ταξινόμηση (κατηγοριοποίηση) δεδομένων

Χρησιμοποιήστε τα παρακάτω δύο πειραματικά σύνολα δεδομένων 2 κατηγοριών που βρίσκονται στην ιστοσελίδα του μαθήματος:

<http://www.cs.uoi.gr/~kblekas/courses/PR/tmp/>

Στα σύνολα αυτά θα εφαρμόσετε τέσσερις (4) μεθόδους ταξινόμησης και θα μετρήσετε την απόδοση (*accuracy* - ποσοστό επιτυχίας) με την μέθοδο *cross validation*: Συγκεκριμένα, 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 υποσύνολα (*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 με Ευκλείδεια απόσταση*: δοκιμάστε διάφορες τιμές του $k=[1, 9]$ και επιλέξτε τον καλύτερο ταξινομητή,
- *Naïve Bayes classifier* υποθέτοντας κανονική κατανομή,
- *Least-Squares Linear Classifier*,
- *Gradient Descent Linear Classifier*: υποθέστε σιγμοειδή συνάρτηση και δοκιμάστε διαφορετικές τιμές για την παράμετρο του ρυθμού μάθησης (η) για τη μέθοδο ελαχιστοποίησης *gradient descent*.

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.