



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΟΛΟΓΙΕΣ ΕΥΦΥΟΥΣ ΔΙΑΧΕΙΡΙΣΗΣ
ΑΝΘΡΩΠΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

ΑΣΚΗΣΗ: WEKA: INSTANCE-BASED LEARNING
([HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA](http://www.cs.waikato.ac.nz/ml/weka))

Στόχος αυτού του εργαστηρίου είναι

- η εξοικείωση με πειράματα ταξινόμησης στο Weka
- η εξοικείωση με ταξινόμηση βάσει παραδειγμάτων

1. Φορτώστε στο Weka το αρχείο adult.arff.

2. **Ταξινόμηση:** Επιλέξτε το μενού *Classify*. Στο *Choose* επιλέξτε *classifiers* -> *lazy* -> *IB1*.

IB1 classifier: *Uses a simple distance measure to find the training instance closest to the given test instance, and predicts the same class as this training instance. If multiple instances are the same (smallest) distance to the test instance, the first one found is used.*

Στο *Test Options* επιλέγω τον τρόπο με τον οποίο θέλω να ελέγξω την απόδοση του αλγορίθμου.

Use training set: Ελέγχεται η απόδοση του αλγορίθμου σε όλα τα δεδομένα εκπαίδευσης (δεν χρησιμοποιούνται καθόλου καινούρια, άγνωστα παραδείγματα) – δεν ενδείκνυται!

Supplied test set: Παρέχω ένα αρχείο ελέγχου με καινούρια, άγνωστα παραδείγματα. Το αρχείο ελέγχου έχει ακριβώς την ίδια μορφή με το αρχείο εκπαίδευσης.

Cross validation: Χωρίζονται τα δεδομένα του αρχείου δεδομένων σε *X* ίσα τμήματα. Πραγματοποιούνται *X* πειράματα (folds), όπου σε κάθε ένα διατηρώ το ένα τμήμα του αρχείου δεδομένων για έλεγχο και τα υπόλοιπα για εκπαίδευση.

Percentage Split: Διαιρώ το αρχείο δεδομένων σε δύο τμήματα (δίνοντας ποσοστό): το ένα χρησιμοποιείται για εκπαίδευση και το άλλο για έλεγχο.

Επιλέξτε *Cross Validation* με 10 folds. Πατήστε το *Start*. Στο δεξί τμήμα του παραθύρου θα εμφανιστούν τα αποτελέσματα της ταξινόμησης.

3. Απόδοση ταξινόμησης: Παρατηρήστε στο δεξί παράθυρο τις ενδείξεις *Correctly classified instances* και *Incorrectly classified instances*. Η πρώτη ονομάζεται **Accuracy (Ορθότητα)**.

Η ορθότητα δεν αποτελεί αξιόπιστο μέτρο απόδοσης από μόνη της. Γιατί; _____

4. **Αξιολόγηση των αποτελεσμάτων:** Κάνοντας δεξί κλικ πάνω στο *id3* στο πεδίο *result list* (κάτω αριστερά), εμφανίζονται μια σειρά από ενέργειες που μπορώ να επιλέξω για το μοντέλο ταξινομητή που έφτιαξα. Μπορώ να σώσω τα αποτελέσματά μου ώστε να αποθηκευτούν σε κάποιο αρχείο, μπορώ να σώσω το μοντέλο του ταξινομητή, ώστε να μπορώ να το

ξαναχρησιμοποιήσω στο μέλλον για να ταξινομήσω καινούρια παραδείγματα, μπορώ να δω τα λάθη ταξινόμησης.

Επιλέξτε *Visualize classifier errors*. Εμφανίζεται ένα διάγραμμα σε δύο άξονες. Μπορείτε να επιλέξετε οποιοδήποτε από τα χαρακτηριστικά να απεικονίζεται στον ένα ή στον άλλο άξονα, από τα αντίστοιχα πεδία (X και Y). Κάθε τιμή της κλάσης ταξινόμησης έχει το δικό της χρώμα (φαίνεται στο υπόμνημα στο κάτω μέρος του διαγράμματος). Οι σωστές ταξινομήσεις φαίνονται ως σταυροί, οι λάθος φαίνονται ως τετραγωνάκια. Με κλικ πάνω σε ένα παράδειγμα μπορείτε να δείτε πληροφορίες σχετικές με το παράδειγμα (τον αριθμό του, τις τιμές των χαρ/κων του, την πραγματική του κλάση και την κλάση στην οποία ταξινομήθηκε).

5. **Πειράματα:** Επανέλθετε στο αρχικό σεντ χαρακτηριστικών. Από το μενού *Classify*, στο *Choose* επιλέξτε *classifiers* -> *lazy* -> *IBk*.

Difference between IB1 and IBk: The biggest one is probably that IBk will keep extending the list of neighbours when several instances are equally far away. This happens quite frequently on datasets with nominal attributes (e.g. more than one neighbour might be used for prediction if $k=1$). IB1 doesn't do this and uses the first nearest neighbour it finds.

Τρέξτε τον *IBk* για διαφορετικές τιμές του k . Τι παρατηρείτε; Ποιά τιμή του k οδηγεί στην καλύτερη ορθότητα;

Για την βέλτιστη τιμή του k τρέξτε πειράματα με διαφορετική μέθοδο αξιολόγησης (πάνω στο σώμα εκπαίδευσης, 10-fold cross validation, percentage split). Τι παρατηρείτε;
