## 

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ**

**ΕΡΓΑΣΤΗΡΙΟ 4**

## WEKA: NAÏVE BAYES & ΔΙΚΤΥΑ BAYES (<HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA>)

1. Ξεκινήστε το Weka και φορτώστε το αρχείο *weather.nominal*.
2. Στην καρτέλα Classify, στο Classifier και επιλέξτε στον φάκελο Bayes τον αλγόριθμο *NaiveBayesSimple*.
3. Τρέξτε τον αλγόριθμο με 10-fold cross validation. Στα αποτελέσματα βλέπετε τις πιθανότητες της τιμής του κάθε χαρακτηριστικού για κάθε τιμή της κλάσης ταξινόμησης (προσαρμοσμένες με add-1 smoothing). Παρατηρήστε τα αποτελέσματα και τον πίνακα σύγχυσης.
4. Στον φάκελο Bayes επιλέξτε τον αλγόριθμο *BayesNet.* Στο πεδίο *estimator* επιλέξτε τον *SimpleEstimator*. Aυτός υπολογίζει τα conditional probability tables για το κάθε χαρακτηριστικό.
5. Στο πεδίο *searchAlgorithm* επιλέξτε Κ2. Είναι ο αλγόριθμος που θα μάθει την δομή του δικτύου από τα δεδομένα. Αφήστε τις άλλες παραμέτρους ως έχουν.
6. Τρέξτε τον αλγόριθμο με 10-fold cross validation. Στα αποτελέσματα βλέπετε την δομή του δικτύου. Την δομή την βλέπετε και στο Result List, κάνοντας δεξί κλικ πάνω στο όνομα του πειράματος (Visualize graph). Τι δίκτυο είναι αυτό;
7. Στον K2 επιλέξτε maxNrOfParents=2. Αφήστε τις υπόλοιπες παραμέτρους ως έχουν. Τρέξτε τον αλγόριθμο με 10-fold cross validation. Παρατηρήστε την δομή του δικτύου και συγκρίνετέ την με πριν. Τι διαφορά έχει; Παρατηρήστε τα αποτελέσματα. Πώς είναι σε σχέση με πριν;
8. Στον K2 επιλέξτε maxNrOfParents=3. Αφήστε τις υπόλοιπες παραμέτρους ως έχουν. Τρέξτε τον αλγόριθμο με 10-fold cross validation. Παρατηρήστε την δομή του δικτύου και συγκρίνετέ την με πριν. Τι διαφορά έχει; Παρατηρήστε τα αποτελέσματα. Πώς είναι σε σχέση με πριν;
9. Στο πεδίο *searchAlgorithm* επιλέξτε ΤΑΝ. Τρέξτε τον αλγόριθμο με 10-fold cross validation. Παρατηρήστε την δομή του δικτύου και συγκρίνετέ την με πριν. Τι διαφορά έχει; Παρατηρήστε τα αποτελέσματα. Πώς είναι σε σχέση με πριν;

**Add-1 Smoothing ή Laplace Smoothing**

Όταν η κλάση PLAY παίρνει τιμή ΥES (9 φορές στα δεδομένα), το OUTLOOK παίρνει τιμή SUNNY 2, OVERCAST 4 και RAINY 3 φορές αντίστοιχα. Δηλαδή οι πιθανότητες P(SUNNY|C), P(OVERCAST|C) και P(RAINY|C) είναι 2/9, 4/9 και 3/9 αντίστοιχα. Αυτές ονομάζονται maximum likelihood estimates (MLEs).

To πρόβλημα με τις MLEs είναι ότι μια τιμή ενός χαρακτηριστικού μπορεί να μην εμφανίζεται με όλες τις τιμές της κλάσης ταξινόμησης στα δεδομένα. Τότε το αντίστοιχο MLE μηδενίζεται, και θα μηδενίζει και όλο το γινόμενο του κανόνα του Bayes.

Μια τεχνική για να αντιμετωπιστεί αυτό είναι να προστεθεί σε κάθε αριθμητή των παραπάνω κλασμάτων +1, και για εξισορρόπηση να προστεθεί και σε κάθε παρανομαστή +3 (όσες είναι οι τιμές του χαρακτηριστικού). Έτσι τα παραπάνω κλάσματα γίνονται P(SUNNY|C) = (2+1)/(9+3), P(OVERCAST|C) = (4+1)/(9+3) και P(RAINY|C) = (3+1)/(9+3).