## 

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ**

**ΕΡΓΑΣΤΗΡΙΟ 3**

## WEKA: DECISION TREES (<HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA>)

1. Ξεκινήστε το Weka και φορτώστε το αρχείο *sattrn.arff*.
2. Στην καρτέλα Classify, στο Classifier και επιλέξτε στον φάκελο Trees τον αλγόριθμο *j48*.
3. H επιλογή 'minNumObj' καθορίζει τον ελάχιστο αριθμό αντικειμένων στα φύλλα του δέντρου. Ο αλγόριθμος υποστηρίζει δύο μορφές κλαδέματος:
   1. με διαστήματα εμπιστοσύνης (εξηγούνται παρακάτω). Αυτός ο τρόπος επιλέγεται με τιμή false στο πεδίο 'unpruned' και επιλογή μιας τιμής εμπιστοσύνης (confidence factor). H τιμή εμπιστοσύνης κυμαίνεται ανάμεσα στο 0 και το 1.
   2. με Reduced Error Pruning. Ξεχωρίζεται ένα σετ επικύρωσης (validation set) πριν την εκπαίδευση, και το δέντρο αξιολογείται σε αυτό μετά την εκπαίδευση. Στην συνέχεια αφαιρούνται κόμβοι ώστε να μεγιστοποιηθεί η απόδοση στο validation set.
4. Φτιάξτε δέντρο με την προεπιλεγμένη μέθοδο κλαδέματος (confidence intervals). Επιλέξτε τιμή εμπιστοσύνης 0.005 (πολύ ισχυρό κλάδεμα). Στο 'Test options' επιλέξτε 'Supplied test set', και επιλέξτε το *sattst.arff.* Παρατηρείστε το δέντρο απόφασης και τον πίνακα σύγχυσης. Ποιές κλάσεις είναι οι πιο προβληματικές στην ταξινόμηση; Συμβαδίζουν τα συμπεράσματά σας με αυτά που βγάλατε σε προηγούμενο εργαστήριο;
5. Πειραματιστείτε με τιμές εμπιστοσύνης 0.05 και 0.5. Απαντήστε αντίστοιχα στα προηγούμενα ερωτήματα.
6. Χρησιμοποιήστε πολύ μικρή τιμή εμπιστοσύνης, δηλ. 0.0005, και δώστε στο 'minNumObj' σχετικά μεγάλη τιμή, πχ 10. Αυτό θα οδηγήσει σε ένα απλό δέντρο που έχει παρόλα αυτά αξιοπρεπή απόδοση. Αυτό το δέντρο δείχνει την σημασία των δέντρων αποφάσεων: είναι κατανοητό και ερμηνεύσιμο.
7. Ενεργοποιήστε την επιλογή Reduced Error Pruning. Πώς είναι τα αποτελέσματα σε σχέση με πριν; Πώς είναι αυτό το δέντρο σε σχέση με τα προηγούμενα;

**Διαστήματα εμπιστοσύνης**

Σε περίπτωση που θέλουμε να αποφύγουμε την χρήση validation set για τον υπολογισμό του σφάλματος, μπορεί να γίνει εκτίμηση αυτού μέσω του σετ εκπαίδευσης. Δεδομένης μιας τιμής εμπιστοσύνης *c* (η προκαθορισμένη τιμή για τον C4.5 είναι *c* = 25%), τα διαστήματα εμπιστοσύνης *z* υπολογίζονται από τον τύπο



όπου f=E/N (το παρατηρηθέν σφάλμα, δηλ. το σφάλμα στα δεδομένα εκπαίδευσης), Ν, ο αριθμός των δειγμάτων, q το πραγματικό σφάλμα. Χρησιμοποιώντας το z (το άνω όριο εμπιστοσύνης) σαν εκτίμηση του σφάλματος σε έναν κόμβο, το σφάλμα αυτό υπολογίζεται ως εξής



Παράδειγμα: Χρησιμοποιούμε την προηγούμενη εξίσωση για c=0.25, δηλ. z=0.69. Για τον κόκκινο κυκλωμένο κόμβο ισχύει Ε=2, Ν=6, δηλ. f=0.33. Από τον τύπο e=0.47. Αντί να χρησιμοποιηθεί το 0.33 σαν σφάλμα θα χρησιμοποιηθεί το 0.47. Για τον μπλε κόμβο Ε=1, Ν=2, e=0.72. Για τον πράσινο κόμβο, έχω ίδια τιμή του e, όπως και για τον κόκκινο.



## Οι τρείς εκτιμήσεις σφάλματος συνδυάζονται με το ποσοστό των παραδειγμάτων που καλύπτουν, δηλ. 6:2:6. (Συνδ σφάλμα = (6/14)\*0.47 + (2/14)\*0.72 + (6/14)\*0.47 = 0.51).

## Στον μητρικό τους κόμβο έχω 9 κακά παραδείγματα και 5 καλά. f=5/14. e=0.46. Επειδή αυτό το σφάλμα είναι μικρότερο από το 0.51, οι τρεις θυγατρικοί κόμβοι κλαδεύονται.