## 

**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ**

**ΕΡΓΑΣΤΗΡΙΟ 1**

## WEKA: VISUALIZATION & FEATURE SELECTION (<HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA>)

1. Διαβάστε το αρχείο The Landsat Data.
2. Αποθηκεύστε σε έναν φάκελο τα αρχεία *sattrn.arff* και *sattst.arff*. Ανοίξτε το πρώτο με το Νotepad, για να δείτε τι περιέχει.
3. Από το ΕΝΑΡΞΗ επιλέξτε Weka 3.6.6 -> Weka 3.6. Από το παράθυρο του Weka επιλέξτε *Explorer*.
4. Από το μενού επιλέξτε *Preprocess* -> *Open File*. Ανοίξτε το αρχείο *sattrn.arff*.
5. Τα δεδομένα (τα χαρακτηριστικά) εμφανίζονται στα αριστερά της οθόνης. Κάνοντας κλικ πάνω σε ένα χαρακτηριστικό εμφανίζονται στατιστικές πληροφορίες για αυτό στο δεξί κάτω μέρος της οθόνης.

**Visualize**

1. Στην καρτέλα 'Visualize' δείτε διαφορετικά διαγράμματα των δεδομένων σας. Σε κάθε γράφημα μπορούν να συμμετέχουν τρία χαρακτηριστικά: δύο στους δυο άξονες και ένα με χρωματικό κώδικα. Οι ταινίες στα δεξιά είναι μονοδιάστατα γραφήματα των χαρακτηριστικών. Το Jitter είναι τυχαίος θόρυβος που μπορεί να προστεθεί στα δεδομένα για να αποφευχθεί η επικάλυψη στην απεικόνιση.
2. Μπορείτε να επιλέξετε χαρακτηριστικά για τα γραφήματα είτε μέσω των pull-down menus για τα X:, Y: και Colour, είτε με κλικ στα μονοδιάστατα γραφήματα κάθε χαρακτηριστικού (αριστερό κλικ για το Χ, δεξί για το Y).
3. Προσπαθήστε να βρείτε ένα διάγραμμα στο οποίο οι κλάσεις να διακρίνονται (διαχωρίζονται) καλά. Tip: Επικεντρωθείτε στο pixel 5, μια και αυτό είναι το κεντρικό, αυτό που στην πραγματικότητα πρέπει να ταξινομηθεί. Κρίνοντας από τα διαγράμματα, πρόκειται για δύσκολο πρόβλημα ταξινόμησης; Ποιες τιμές της κλάσης ταξινόμησης πιστεύετε θα είναι πιο εύκολο να ταξινομηθούν; Ποιές θα είναι πιο δύσκολες;

**Select attributes**

1. To Weka υποστηρίζει πολλές λειτουργίες για επιλογή χαρακτηριστικών. Πηγαίνετε στην καρτέλα 'Select attributes'. Επιλέξτε στο Attribute Evaluator το Information Gain attribute selection. Επιλέξτε την Ranker search method για αναζήτηση. Στο numToSelect επιλέξτε τον αριθμό των χαρακτηριστικών που θέλετε να κρατήσετε (πχ 10). Το Information Gain είναι ένα άλλο όνομα για την αμοιβαία πληροφορία (mutual information) μεταξύ δυο μεταβλητών. Στην επιλογή χαρακτηριστικών μας ενδιαφέρει η αμοιβαία πληροφορία ανάμεσα στην κλάση ταξινόμησης και το κάθε χαρακτηριστικό (δηλ. τη μείωση της εντροπίας της κλάσης ταξινόμησης, δεδομένων των τιμών του κάθε χαρακτηριστικού).
2. Ποια χαρακτηριστικά επιτυγχάνουν το μεγαλύτερο σκορ; Λογικά θα περιμένατε τα pixel6\_1 και pixel6\_2 να είναι πιο σημαντικά από το pixel5\_4? Πηγαίνετε πίσω στο Visualisation και φτιάξτε το διάγραμμα του pixel6\_1 με το pixel5\_1, και του pixel6\_2 με το pixel5\_2. Τι πρόβλημα έχει η επιλογή χαρακτηριστικών με information gain;
3. Όταν πραγματοποιείτε επιλογή χαρακτηριστικών μέσω του 'Select attributes', παίρνετε σαν έξοδο τη λίστα των επιλεγμένων χαρακτηριστικών. Το αρχείο δεδομένων σας, όμως παραμένει το ίδιο. Αν θέλετε η επιλογή να έχει επίδραση και στις υπόλοιπες καρτέλες (πχ να κάνετε ταξινόμηση με το σετ μειωμένων χαρακτηριστικών) θα πρέπει να την πραγματοποιήσετε μέσω της καρτέλας Preprocess. Στα αριστερά επιλέγετε Filter -> Choose ->Supervised->Features->Feature Selection. Διαλέγετε πχ. InfoGain και πατάτε 'Apply'. Τώρα το αρχείο δεδομένων σας αντικαταστάθηκε με αυτό των μειωμένων χαρακτηριστικών. Με Undo, μπορείτε να επανέλθετε στο αρχικό σας σετ δεδομένων.