ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΠΟΘΗΚΕΣ ΚΑΙ ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ

WEKA (HTTP://WWW.CS.WAIKATO.AC.NZ/ML/WEKA)

Meta-Learning

Bagging

- Ξεκινήστε το Weka και φορτώστε το αρχείο *iris.arff*
- Στην καρτέλα Classify, στο Classifier επιλέξτε *Meta --> Bagging*.
- Κάντε κλικ πάνω στο όνομα του αλγορίθμου προκειμένου να εμφανιστούν οι παράμετροί του.
- Οι παράμετροι είναι οι εξής:
    - BagSizePercent: ποσοστό των παραδειγμάτων εκπαίδευσης που θα χρησιμοποιεί σε κάθε επανάληψη ο αλγόριθμος. Επιλέξτε 100.
    - Classifier: ο αλγόριθμος μάθησης που θα τρέξει επαναληπτικά (base learner). Επιλέξτε *trees --> j48*.
    - numIterations: ο αριθμός των επαναλήψεων που θα πραγματοποιηθούν. Επιλέξτε 10.
    - Seed: ένας τυχαίος αριθμός που καθορίζει την επιλογή του σετ παραδειγμάτων εκπαίδευσης σε κάθε bag. Αφήστε το όπως έχει.
- Επιλέξτε 10-fold cross validation στο Test options.
- Τρέξτε τον αλγόριθμο. Πόσα δέντρα απόφασης δημιουργούνται;
- Τρέξτε τον απλό j48 (όχι με bagging) με τις ίδιες παραμέτρους όπως πριν.
- Τι παρατηρείτε; Τα αποτελέσματα με το bagging είναι καλύτερα ή χειρότερα;
- Τρέξτε πειράματα με BagSizePercent 40, 60, 80. Τι παρατηρείτε;

Boosting

- Στην καρτέλα Classify, στο Classifier επιλέξτε *Meta --> AdaBoostM1*.
- Κάντε κλικ πάνω στο όνομα του αλγορίθμου προκειμένου να εμφανιστούν οι παράμετροί του.
- Οι παράμετροι είναι οι εξής:
    - Classifier: ο αλγόριθμος μάθησης που θα τρέξει επαναληπτικά (base learner). Επιλέξτε *trees --> j48*.
    - numIterations: ο αριθμός των επαναλήψεων που θα πραγματοποιηθούν. Επιλέξτε 10.
    - Seed: ένας τυχαίος αριθμός που καθορίζει την επιλογή του σετ παραδειγμάτων εκπαίδευσης σε κάθε bag. Αφήστε το όπως έχει.
    - UseResampling: Όταν ο base classifier δεν μπορεί να αντιμετωπίσει βάρη άμεσα, τα δεδομένα εκπαίδευσης επανα-δειγματοληπτούνται (με επανατοποθέτηση) βάσει της κατανομής βαρών ώστε να δημιουργηθεί ένα καινούριο σετ δεδομένων εκπαίδευσης. Παραδείγματα με υψηλότερο βάρος είναι πιο πιθανό να επιλεγούν ενώ αυτά με χαμηλότερο βάρος είναι λιγότερο πιθανό να επιλεγούν. Το μοντέλο που προκύπτει είναι πιο πιθανό να ταξινομήσει σωστά παραδείγματα που εμφανίζονται

περισσότερες φορές στα δεδομένα σε σχέση με παραδείγματα που εμφανίζονται μία μόνο φορά.

- weightThreshold: Όσο περισσότερες επαναλήψεις λαμβάνουν χώρα, ο συνολικός ταξινομητής γίνεται όλο και πιο σίγουρος για τις ταξινομήσεις των παραδειγμάτων που ταξινομεί σωστά. Έτσι, τα βάρη αυτών των παραδειγμάτων γίνονται όλο και πιο μικρά. Αντί να αναλώνονται ακόλουθοι base learners (σε επόμενες επαναλήψεις) να μαθαίνουν από αυτά τα πολύ μικρού βάρους παραδείγματα, μπορούν τα παραδείγματα αυτά να διαγραφούν από τα δεδομένα εκπαίδευσης. Αυτό το όριο βάρους καθορίζεται εδώ.

- Επιλέξτε 10-fold cross validation στο Test options.
- Τρέξτε τον αλγόριθμο. Πόσα δέντρα απόφασης δημιουργούνται; Τι είναι το weight που εμφανίζεται κάτω από κάθε δέντρο;
- Τι παρατηρείτε σε σχέση με τον απλό j48; Τα αποτελέσματα είναι καλύτερα ή χειρότερα;
- Τι παρατηρείτε σε σχέση με το bagging; Τα αποτελέσματα είναι καλύτερα ή χειρότερα;

Stacking

- Στην καρτέλα Classify, στο Classifier επιλέξτε *Meta --> Stacking*.
- Κάντε κλικ πάνω στο όνομα του αλγορίθμου προκειμένου να εμφανιστούν οι παράμετροί του.
- Οι παράμετροι είναι οι εξής:
  - Classifiers: οι αλγόριθμοι μάθησης που θα τρέξουν (level-0 learners).
    - Πατώντας πάνω στο text box, ανοίγει παράθυρο με το οποίο προσθέτετε ταξινομητή.
    - Με το Choose τον επιλέγετε, με το Add τον προσθέτετε.
    - Όπως πάντα, με κλικ πάνω στο όνομα του αλγορίθμου ανοίγουν οι παράμετροί του.
    - Με Delete σβήνετε έναν αλγόριθμο που είχατε προσθέσει.
    - Με τα Up και Down αλλάζω την κατάταξη των αλγορίθμων.
  - Επιλέξτε *j48, Naive Bayes, 5-NN* .
  - metaClassifier: εδώ επιλέγετε τον level-1 ταξινομητή. Επιλέξτε *trees --> j48*.
  - numFolds: ο αριθμός των folds για την διασταυρωτική αξιολόγηση. Επιλέξτε 10.
- Επιλέξτε 10-fold cross validation στο Test options.
- Τρέξτε τον αλγόριθμο.
- Τι παρατηρείτε σε σχέση με τον απλό j48; Τα αποτελέσματα είναι καλύτερα ή χειρότερα;
- Τι παρατηρείτε σε σχέση με το bagging; Τα αποτελέσματα είναι καλύτερα ή χειρότερα;
- Τι παρατηρείτε σε σχέση με τον AdaBoost; Τα αποτελέσματα είναι καλύτερα ή χειρότερα;