

ΑΛΓΟΡΙΘΜΟΙ ΕΚΜΑΘΗΣΗΣ (Α΄ ΜΕΡΟΣ)

ΔΕΝΔΡΑ ΚΑΙ ΚΑΝΟΝΕΣ

- Αρχικός πειραματισμός: εφαρμογή απλών κανόνων.
- Απλοϊκοί κανόνες: δημιουργία κανόνων ενός και μόνο χαρακτηριστικού.
- Ο πιο απλός 1R: αλγόριθμος εκμάθησης δέντρου απόφασης ενός επιπέδου.
 - Δημιουργία κανόνων ελέγχου ενός ξεχωριστού χαρακτηριστικού.
 - Βασική έκδοση: (ονομαστικά χαρακτηριστικά)
 - Ένας κλάδος για κάθε τιμή
 - Κάθε κλάδος εκχωρεί την τάξη με τη μεγαλύτερη συχνότητα
 - Τιμή σφάλματος: ποσοστό υποδειγμάτων που δεν ανήκουν στην πλειψηφούσα τάξη του αντίστοιχου κλάδου
 - Επιλογή χαρακτηριστικού με τη μικρότερη τιμή σφάλματος.

ΚΑΤΑΣΚΕΥΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ

- **Στρατηγική:** από πάνω προς τα κάτω - επαναληπτική υλοποίηση της μεθόδου διαίρει και βασίλευε.
- Επιλογή χαρακτηριστικού ως ρίζα.
- Δημιουργία ενός κλάδου για κάθε πιθανή τιμή του χαρακτηριστικού.
- Διάσπαση υποδειγμάτων σε υποσύνολα, ένα για κάθε τιμή του χαρακτηριστικού.
- Επανάληψη αυτών για κάθε κλάδο με χρήση μόνο του υποσυνόλου των υποδειγμάτων κάθε κλάδου.
- Ολοκλήρωση όταν όλα τα υποδείγματα ανήκουν στην ίδια τάξη.
- **Στόχος:** δημιουργία του μικρότερου δυνατού δέντρου, γι' αυτό επιλέγεται το χαρακτηριστικό που δημιουργεί τα περισσότερα ομοιογενή υποσύνολα κλάδων.
- Σύνηθες κριτήριο ανομοιογένειας είναι το κέρδος πληροφορίας το οποίο αυξάνεται με την αύξηση της μέσης ομοιογένειας των υποσυνόλων. Έτσι επιλέγεται το χαρακτηριστικό που δίνει μέγιστο κέρδος πληροφορίας.
- **Πρόβλημα:** το κριτήριο κέρδους μεροληπτεί υπέρ της επιλογής χαρακτηριστικών με μεγάλο αριθμό τιμών. Πιθανή συνέπεια υπερπροσαρμογή.
- Το πρόβλημα της μεροληψίας αντιμετωπίζεται με το λόγο του κέρδους

- Από τη μετατροπή δένδρου απόφασης σε σύνολο κανόνων προκύπτει υπερβολικά σύνθετο σύνολο κανόνων με πολύπλοκες μεθόδους.
- Εναλλακτικά: απευθείας δημιουργία συνόλου κανόνων με εφαρμογή αλγορίθμων κάλυψης. Δηλαδή, για κάθε τάξη εύρεση συνόλου κανόνων που καλύπτουν το σύνολο των υποδειγμάτων.
- Απλοϊκός αλγόριθμος κάλυψης: δημιουργία κανόνα με την προσθήκη ελέγχων που μεγιστοποιούν την ακρίβεια του κανόνα. Αυτά υλοποιούνται με τον αλγόριθμο PRISM.
- ID3: Η μέθοδος που περιγράψαμε για την κατασκευή δένδρου με κριτήριο το κέρδος πληροφορίας. Επέκταση: C4.5 (αριθμητικά χαρακτηριστικά, χειρισμό άγνωστων τιμών, ανθεκτικότητα στην ύπαρξη θορύβου.)
- Κλάδεμα: πρόληψη υπερπροσαρμογής δένδρου σε θόρυβο δεδομένων, με 2 στρατηγικές: μετακλάδεμα και πρόκλάδεμα.
- Μετατροπή δένδρων σε κανόνες:
 - Ένας κανόνας για κάθε φύλλο.
 - Κλάδεμα των συνθηκών που επηρεάζουν αρνητικά το εκτιμώμενο σφάλμα των κανόνων.
 - Εύρεση όλων των κανόνων για κάθε τάξη
 - Επιλογή υποσυνόλων με οδηγό την αρχή MDL
 - Ταξινόμηση των υποσυνόλων για αποφυγή αντικρουόμενων υποδείξεων
 - Αποκοπή κλάδων σε περίπτωση που μειώνει το συνολικό σφάλμα επί των δεδομένων εκπαίδευσης.
 - WEKA: PART