

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ



Δέντρα αποφάσεων

Γιώργος Στάμου

Καθηγητής Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών ΕΜΠ
Διευθυντής Εργαστηρίου Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης - AILS Lab

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ



Πρόβλημα

X : σύνολο στιγμιότυπων

Y : σύνολο ετικετών

$f: X \rightarrow Y$: ιδανικός ταξινομητής

$H = \{h|h: X \rightarrow Y\}$: σύνολο ταξινομητών (υποθέσεις)

Είσοδος

Σύνολο δεδομένων $\{\langle x_i, y_i \rangle\}_{i=1}^n = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \dots, \langle x_n, y_n \rangle\}$

Έξοδος

Υπόθεση $h \in H$ που προσεγγίζει το f

Στήλες χαρακτηριστικών: x_i

Στήλη απόφασης: y_i

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

$\langle x_i, y_i \rangle$

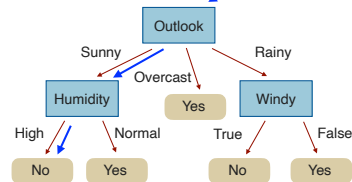
2

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΟΡΙΣΜΟΣ



Δέντρο απόφασης

Είσοδος $x_i = \langle \text{Sunny}, \text{Hot}, \text{High}, \text{True} \rangle$



έξοδος $y_i = \langle \text{No} \rangle$

Σε κάθε εσωτερικό κόμβο ελέγχεται η τιμή του χαρακτηριστικού x_i

Σε κάθε διακλάδωση επιλέγεται μία τιμή του χαρακτηριστικού x_i

Σε κάθε φύλλο αποδίδεται μία ετικέτα y στο στοιχείο

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

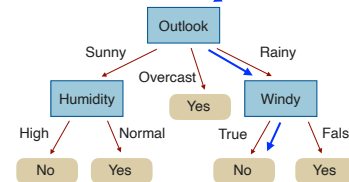
3

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΟΡΙΣΜΟΣ



Δέντρο απόφασης

Είσοδος $x_i = \langle \text{Rainy}, \text{Mild}, \text{Normal}, \text{True} \rangle$



έξοδος $y_i = \langle \text{No} \rangle$

Σε κάθε εσωτερικό κόμβο ελέγχεται η τιμή του χαρακτηριστικού x_i

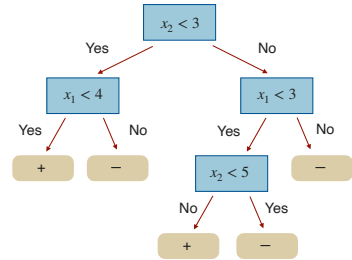
Σε κάθε διακλάδωση επιλέγεται μία τιμή του χαρακτηριστικού x_i

Σε κάθε φύλλο αποδίδεται μία ετικέτα y στο στοιχείο

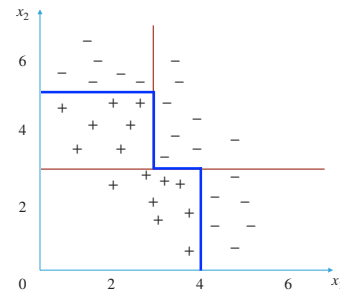
Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

4

Δέντρο απόφασης



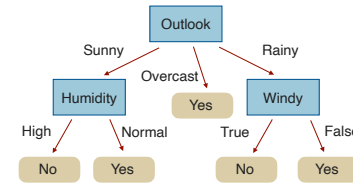
Καμπύλη διαχωρισμού



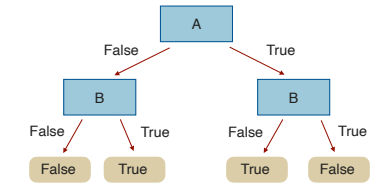
- ▶ Η καμπύλη διαχωρισμού χωρίζει το χώρο χαρακτηριστικών σε (υπερ-)κύβους παράλληλους των αξόνων
- ▶ Κάθε (υπερ-)κυβική επιφάνεια αντιστοιχίζεται σε μία ετικέτα - ή (στη γενική περίπτωση) σε μία κατανομή πιθανοτήτων πάνω στις ετικέτες

5

Ικανότητα αναπαράστασης



Δένδρο απόφασης για XOR



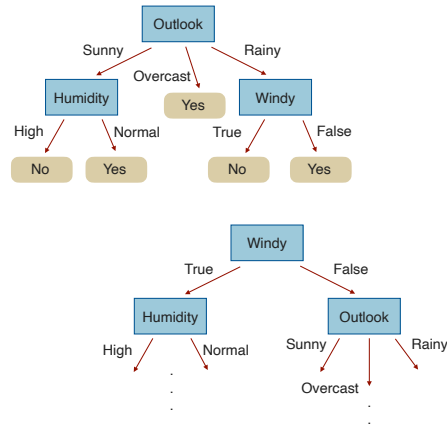
IF (Outlook IS Overcast) $\vee$ ((Outlook IS Sunny) $\wedge$ Humidity IS Normal) $\vee$ ((Outlook IS Rainy) $\wedge$ Windy IS False)
THEN (PlayTennis IS YES)

- ▶ Ένα δέντρο απόφασης αντιστοιχεί σε μία κανονική διαζευκτική μορφή (disjunctive normal form - DNF) μίας λογικής έκφρασης
- ▶ Τα δέντρα απόφασης μπορούν να αναπαραστήσουν οποιαδήποτε λογική συνάρτηση

6

Μοναδικότητα

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

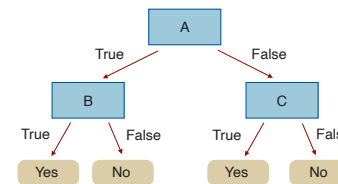


- ▶ Διαφορετικά δέντρα απόφασης μπορούν να είναι ισοδύναμα για το συγκεκριμένο σύνολο δεδομένων (να ταξινομούν στην ίδια κλάση κάθε στιγμότυπο)
- ▶ Πόσα διαφορετικά δέντρα απόφασης αναπαριστούν μία συγκεκριμένη λογική έκφραση;

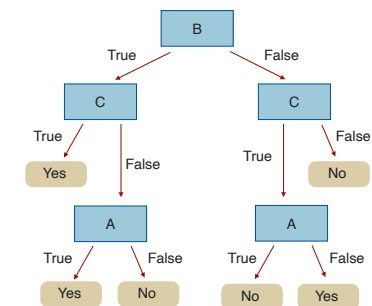
7

Λογική συνάρτηση $f(A, B, C) = (A \wedge B) \vee (\neg A \wedge C)$

Δέντρο απόφασης 1



Δέντρο απόφασης 2

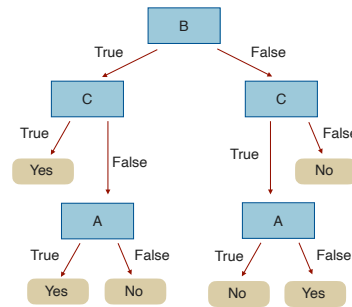


8

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΠΡΩΤΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρατηρήσεις

- Τα δέντρα απόφασης μπορούν να γίνουν πολύ μεγάλα σε μέγεθος (να έχουν εκθετικά πολλούς κόμβους σε σχέση με τα χαρακτηριστικά των στιγμιοτύπων)
- Τα δέντρα απόφασης είναι κατανοητά από τους ανθρώπους (interpretable), όταν είναι μικρά σε μέγεθος (ταξινομητές με χρήση υπερκύβων)
- Η απλούστερη συνεπής επεξήγηση είναι η βέλτιστη (Ockham's Razor) It is vain to do more what can be done with less... Entities should not be multiplied beyond necessity (William of Ockham - 1324) - Είναι σημαντικό να κατασκευάζουμε και να χρησιμοποιούμε απλά δέντρα απόφασης
- Το πρόβλημα εύρεσης του βέλτιστου δέντρου απόφασης είναι δυσεπίλυτο (Laurent Hyafil, Ronald L. Rivest, Constructing optimal binary decision trees is NP-complete, Information Processing Letters, Volume 5, Issue 1, May 1976, Pages 15-17) (Hancock T. R., Jiang T., Li M., and Tromp J., Lower bounds on learning decision lists and trees, Information and Computation 126(2):114-122, 1996)
- Ζητήματα που θα μας απασχολήσουν: αξιολόγηση, επιλογή χαρακτηριστικών, κλάδεμα, αλγόριθμοι εκπαίδευσης



Rokach, L. and Maimon, O.Z. (2008) Data Mining with Decision Trees: Theory and Applications. World Scientific Publishing Co., Inc., Singapore

9

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΞΙΟΛΟΓΗΣΗ

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

$\mathbf{x} \sim \mathcal{D}(X)$

$f: X \rightarrow Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

$\mathbf{x}_i \sim \mathcal{D}(X)$

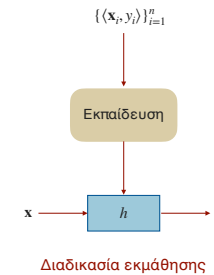
$H = \{h | h: X \rightarrow Y\}$: σύνολο δέντρων απόφασης $X \rightarrow Y$

Πρόβλημα

$h \leftarrow \text{DecisionTree.train}(\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$

$\text{minimal}(h)$

- Με την διαδικασία εκμάθησης κατασκευάζεται ένα δέντρο απόφασης που προσεγγίζει τον ιδανικό ταξινομητή
- Παρότι είναι υπολογιστικά δύσκολο, θα πρέπει το δέντρο απόφασης να είναι όσο πιο κοντά γίνεται στο ελάχιστο



10

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΣΦΑΛΜΑ ΕΚΜΑΘΗΣΗΣ (ΑΚΡΙΒΕΙΑ)

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

$\mathbf{x} \sim \mathcal{D}(X)$

$f: X \rightarrow Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

$\mathbf{x}_i \sim \mathcal{D}(X)$

$H = \{h | h: X \rightarrow Y\}$: σύνολο δέντρων απόφασης $X \rightarrow Y$

Πρόβλημα

$h \leftarrow \text{DecisionTree.train}(\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$

Δέντρο απόφασης $h \approx f$

$\text{minimal}(h)$

Σφάλμα (error)

Άγνωστο το f

$$L_f(h) = \frac{|X \in X : h(x) \neq f(x)|}{|X|}$$

Άγνωστα τα $f, \mathcal{D}$

$$L_{(\mathcal{D}, f)}(h) = \Pr_{\mathbf{x} \sim \mathcal{D}(X)} (h(\mathbf{x}) \neq f(\mathbf{x}))$$

Η πιθανότητα να επιλέξω ένα τυχαίο δείγμα $x \in X$ για το οποίο $h(x) \neq f(x)$

$$L_D(h) = \frac{|i \in \mathbb{N}_n : h(\mathbf{x}_i) \neq y_i|}{|n|}$$

Εμπειρικό σφάλμα - σφάλμα εκμάθησης (empirical error - training error)

11

ΕΠΙΠΛΕΟΝ ΜΕΤΡΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΑΞΙΝΟΜΗΤΩΝ

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

$\mathbf{x} \sim \mathcal{D}(X)$

$f: X \rightarrow Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

$\mathbf{x}_i \sim \mathcal{D}(X)$

$H = \{h | h: X \rightarrow Y\}$: σύνολο δέντρων απόφασης $X \rightarrow Y$

Πρόβλημα

$h \leftarrow \text{DecisionTree.train}(\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$

Δέντρο απόφασης $h \approx f$

$\text{minimal}(h)$

$$L_{\mathbb{D}_L}(h) = \frac{|i \in \mathbb{N}_k : h(\mathbf{x}_i) \neq y_i|}{|k|}$$

Σφάλμα εκμάθησης (training error)

$$L_{\mathbb{D}_T}(h) = \frac{|i \in \mathbb{N}_m : h(\mathbf{x}_i) \neq y_i|}{|m|}$$

Σφάλμα ελέγχου (testing error)

Μέθοδος απόκρυψης (hold-out)

Διαχωρισμός του $\mathbb{D}$ σε δύο υποσύνολα δεδομένων:

$\mathbb{D}_L \subset \mathbb{D}$: δεδομένα εκμάθησης

$\mathbb{D}_T \subset \mathbb{D}$: δεδομένα ελέγχου

Συνήθως: $\mathbb{D}_L \cup \mathbb{D}_T = \mathbb{D}$ και $|\mathbb{D}_L| > |\mathbb{D}_T|$

$$|\mathbb{D}_L| \approx \frac{2}{3} |\mathbb{D}|$$

12



Επιπλέον μέτρα αξιολόγησης

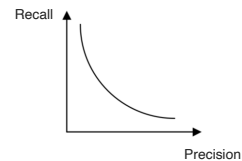
Συνήθως οι κλάσεις $\{+1, -1\}$ είναι σημαντικά διαφορετικές

Δηλαδή είναι διαφορετικό αν $h(x) \neq y$ ($y = +1$) ή αν $h(x) \neq y$ ($y = -1$)

$$\begin{aligned} \text{positive} &= |y_i|, y_i = +1 & \text{negative} &= |y_i|, y_i = -1 & \text{true-positive} &= |h(x_i) = +1|, y_i = +1 & \text{true-negative} &= |h(x_i) = -1|, y_i = -1 \\ & & & & \text{false-positive} &= |h(x_i) = +1|, y_i = -1 & \text{false-negative} &= |h(x_i) = -1|, y_i = +1 \end{aligned}$$

$$\text{Sensitivity} = \frac{\text{true-positive}}{\text{positive}} \quad (\text{λέγεται και Recall}) \quad \text{Specificity} = \frac{\text{true-negative}}{\text{negative}} \quad \text{Precision} = \frac{\text{true-positive}}{\text{true-positive} + \text{false-positive}}$$

$$\text{Accuracy} = \text{Sensitivity} \frac{\text{positive}}{\text{positive} + \text{negative}} + \text{Specificity} \frac{\text{negative}}{\text{positive} + \text{negative}}$$



13



DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree.train(D')

14



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree.train(D')

15



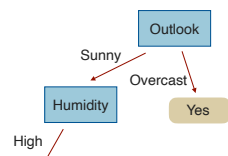
Outlook	Temperature	Humidity	Windy	PlayTennis
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree.train(D')

16

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Sunny	Mild	Normal	TRUE	Yes

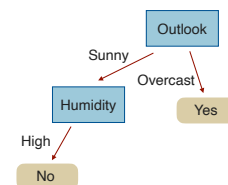
DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

17

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Sunny	Mild	High	FALSE	No

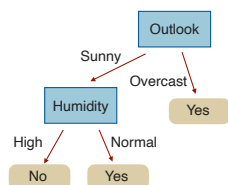
DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

18

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Cool	Normal	TRUE	Yes
Sunny	Mild	Normal	TRUE	Yes

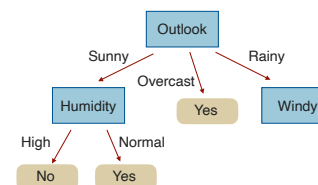
DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

19

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Rainy	Mild	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

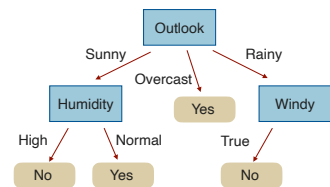
DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

20

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Cool	Normal	TRUE	No
Rainy	Mild	High	TRUE	No

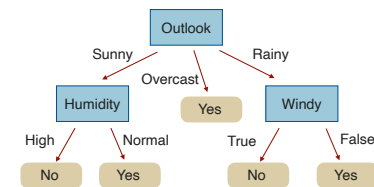
DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

21

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes

DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

22

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ



DecisionTree.train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
- Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
- Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή

Αλλιώς DecisionTree.train(D')

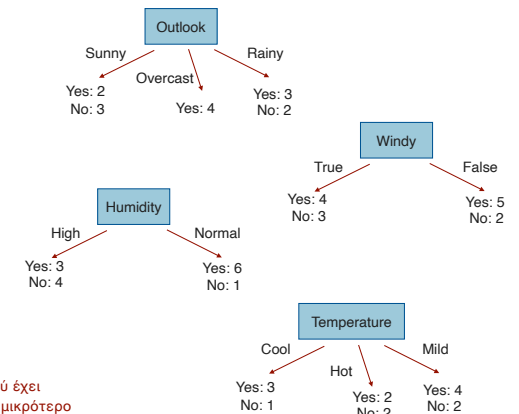
- Η επιλογή του χαρακτηριστικού εισόδου γίνεται με στόχο την αύξηση της πιθανότητας να οδηγηθούμε σε μικρότερο δέντρο
- Προφανώς, δεν μπορούμε να διασφαλίσουμε την κατασκευή του ελάχιστου δέντρου (υπενθυμίζουμε ότι το πρόβλημα είναι NP-complete)

23

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No



- Ποια από τις επιλογές χαρακτηριστικού έχει καλύτερη πιθανότητα να οδηγήσει σε μικρότερο δέντρο;

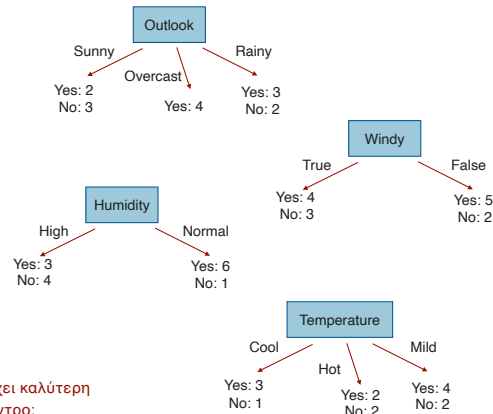
24

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ

Μέθοδος

- **Τυχαιά:** Επίλεξε ένα χαρακτηριστικό χωρίς κάποιο συγκεκριμένο κριτήριο
- **Λιγότερες τιμές:** Επίλεξε το χαρακτηριστικό με τη μικρότερη πληθικότητα του πεδίου τιμών
- **Περισσότερες τιμές:** Επίλεξε το χαρακτηριστικό με τη μεγαλύτερη πληθικότητα του πεδίου τιμών
- **Μεγαλύτερο όφελος:** Επίλεξε το χαρακτηριστικό με το μεγαλύτερο κέρδος πληροφορίας (information gain)

- Ποια από τις επιλογές χαρακτηριστικού έχει καλύτερη πιθανότητα να οδηγήσει σε μικρότερο δέντρο;



25

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ

Μέτρα μη-καθαρότητας (impurity measures)

Έστω μία τυχαιά μεταβλητή x με k διακριτές τιμές, με κατανομή πιθανότητας $P = (p_1, p_2, \dots, p_k)$

Μέτρο μη-καθαρότητας της μεταβλητής x είναι μία συνάρτηση $\phi : [0,1]^k \rightarrow \mathbb{R}$ που ικανοποιεί τις συνθήκες:

$$\phi(P) \geq 0$$

$$\phi(P) = 0 \text{ αν υπάρχει } i \in \mathbb{N}_k \text{ τέτοιο ώστε } p_i = 1$$

$$\phi(P) = 1 \text{ αν για κάθε } i \in \mathbb{N}_k \text{ ισχύει ότι } p_i = \frac{1}{k}$$

$$\phi \text{ συμμετρική στα } p_1, p_2, \dots, p_k$$

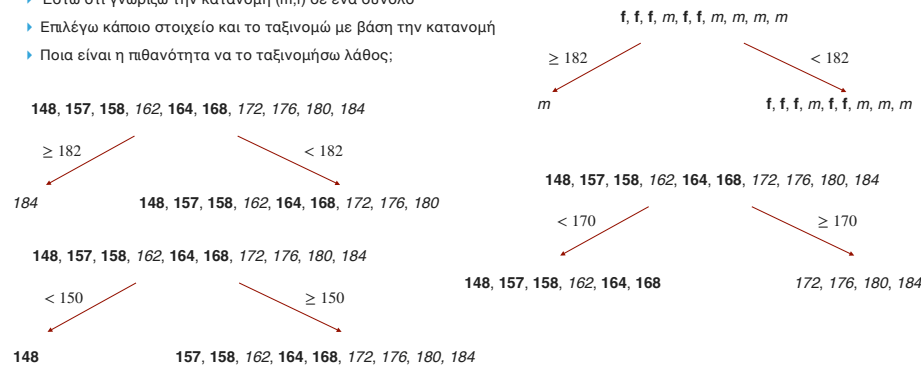
$$\phi \text{ ομαλή}$$

26

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ

Gini index

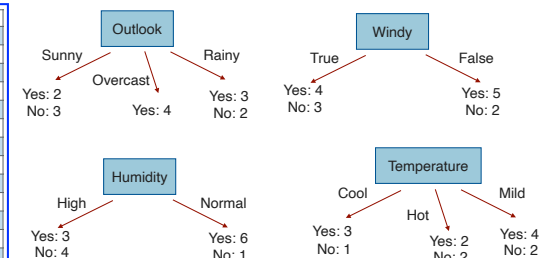
- Ταξινομώ σε male, female ανάλογα με το ύψος
- Έστω ότι γνωρίζω την κατανομή (m,f) σε ένα σύνολο
- Επιλέγω κάποιο στοιχείο και το ταξινομώ με βάση την κατανομή
- Ποια είναι η πιθανότητα να το ταξινομήσω λάθος;



27

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ - GINI INDEX

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No



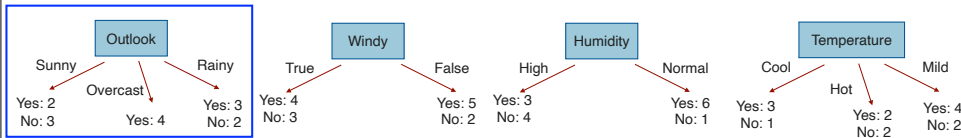
- Πόσο συχνά ένα τυχαίο δείγμα που επιλέγεται ταξινομείται λανθασμένα, αν του αποδοθεί μία τυχαία ετικέτα;

$$gini(D) = \sum_{i \in \text{labels}(D)} p_i(1 - p_i) = 1 - \sum_{i \in \text{labels}(D)} p_i^2$$

$$gini(\text{Root}) = 1 - \left(\frac{|\text{PlayTennis} = \text{Yes}|}{|\text{Root}|} \right)^2 - \left(\frac{|\text{PlayTennis} = \text{No}|}{|\text{Root}|} \right)^2 = 1 - \left(\frac{9}{14} \right)^2 - \left(\frac{5}{14} \right)^2$$

28

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ - GINI INDEX



$$\text{Information Gain: } ig(\text{Outlook}) = gini(\text{Root}) - gini(\text{Outlook}) = gini(\text{Root}) - \sum_{v \in \text{values}(\text{Outlook})} \frac{|\text{Outlook} = v|}{|\text{Root}|} gini(\text{Outlook} = v)$$

$$gini(\text{Root}) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$gini(\text{Outlook}) = \frac{5}{14} gini(\text{Outlook} = \text{Sunny}) + \frac{4}{14} gini(\text{Outlook} = \text{Overcast}) + \frac{5}{14} gini(\text{Outlook} = \text{Rainy})$$

$$gini(\text{Outlook} = \text{Sunny}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

$$gini(\text{Outlook} = \text{Overcast}) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2$$

$$gini(\text{Outlook} = \text{Rainy}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

29

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΛΓΟΡΙΘΜΟΣ CART



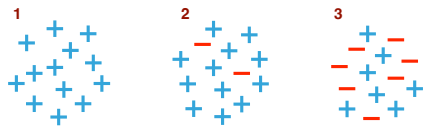
DecisionTree . CART(D)

- Επίλεξε το χαρακτηριστικό εισόδου a με το μεγαλύτερο κέρδος πληροφορίας στο D με βάση το gini
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree . CART(D')

Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen, *Classification and Regression Trees*, Chapman and Hall/CRC (1984)

30

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ - ΕΝΤΡΟΠΙΑ



Πόση πληροφορία λείπει;

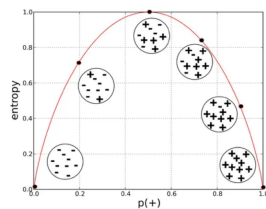
► Ποια από τις κατανομές έχει μεγαλύτερο κέρδος πληροφορίας;

$$E_1 = -1 \log_2 1 - 0 \log_2 0 = 0$$

$$E_2 = -0.133 \log_2 0.133 - 0.87 \log_2 0.87 \approx 0.565$$

$$E_3 = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

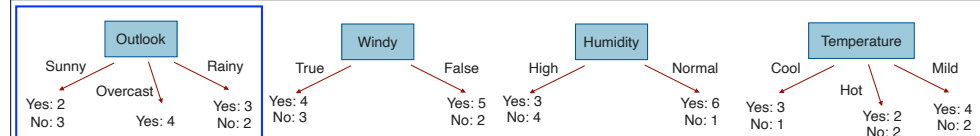
$$E = \sum_i (-p_i \log_2 p_i)$$



Provost, Foster, Fawcett, Tom. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*

31

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ - ΕΝΤΡΟΠΙΑ



$$\text{Information Gain: } ig(\text{Outlook}) = E(\text{Root}) - E(\text{Outlook}) = E(\text{Root}) - \sum_{v \in \text{values}(\text{Outlook})} \frac{|\text{Outlook} = v|}{|\text{Root}|} E(\text{Outlook} = v)$$

$$E(\text{Root}) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14}$$

$$E(\text{Outlook}) = \frac{5}{14} E(\text{Outlook} = \text{Sunny}) + \frac{4}{14} E(\text{Outlook} = \text{Overcast}) + \frac{5}{14} E(\text{Outlook} = \text{Rainy})$$

$$E(\text{Outlook} = \text{Sunny}) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5}$$

$$E(\text{Outlook} = \text{Overcast}) = -\frac{4}{4} \log_2 \frac{4}{4} - \frac{0}{4} \log_2 \frac{0}{4}$$

$$E(\text{Outlook} = \text{Rainy}) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5}$$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

32

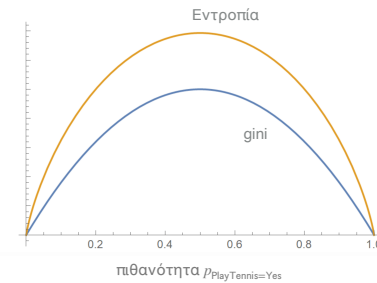


DecisionTree . ID3(D)

- Επίλεξε το χαρακτηριστικό εισόδου a με το μεγαλύτερο κέρδος πληροφορίας στο D με βάση την εντροπία
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα, τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree . ID3(D')

Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

33



- ▶ Παρόμοια αποτελέσματα στην πράξη
- ▶ Διαφωνούν σε ελάχιστες περιπτώσεις στην επιλογή χαρακτηριστικού
- ▶ Δυσκολότερος ο υπολογισμός για την εντροπία

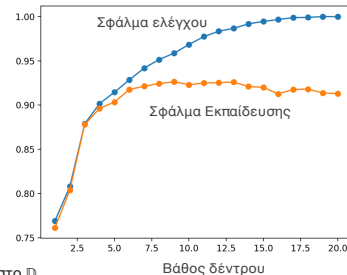
Laura Elena Raileanu and Kilian Stoffel, Theoretical comparison between the Gini Index and Information Gain criteria, Annals of Mathematics and Artificial Intelligence 41: 77–93, 2004

34



Υπερπροσαρμογή (overfitting)

- ▶ Τα δένδρα απόφασης μπορούν να ταξινομήσουν όλα τα δεδομένα εκμάθησης χωρίς σφάλμα
- ▶ στην περίπτωση αυτή θα καταλήξουμε με μεγάλα δένδρα απόφασης που θα έχουν δυσκολία γενίκευσης (σφάλμα στα δεδομένα ελέγχου)
- ▶ Απαιτείται συστηματική απλοποίηση του δένδρου



DecisionTree . train(D)

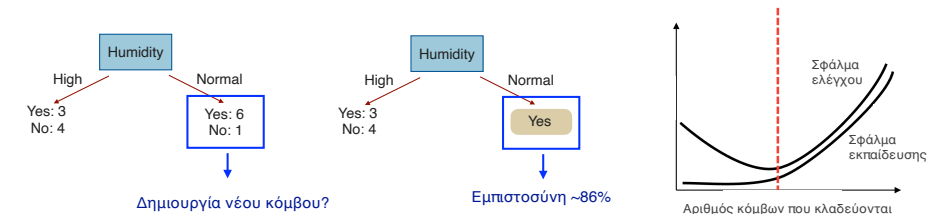
- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree . train(D')

35



DecisionTree . train(D)

- Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο D
 - Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
 - Για κάθε διαφορετική τιμή του a :
 - Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
 - Όρισε το D' ως το υποσύνολο του D με τα στοιχεία που έχουν τη τιμή αυτή για το a
 - Αν όλα τα $y \in D'$ έχουν την ίδια ετικέτα τότε κάνε τον A φύλλο με τιμή την ετικέτα αυτή
- Αλλιώς DecisionTree . train(D')



36



Reduced error pruning (REP)

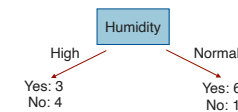
- Ορίζουμε ένα pruning set (μέρος του dataset)
- Κατασκευάζουμε το δέντρο απόφασης
- Εξετάζουμε όλους τους εσωτερικούς κόμβους από τα φύλλα προς τη ρίζα
- Ελέγχουμε τη μετατροπή του κόμβου σε φύλλο με ετικέτα την πιο πιθανή κλάση
- Αν επηρεάζεται η επίδοση του ταξινομητή στο pruning set συνεχίζουμε
- Αλλιώς, μετατρέπουμε τον κόμβο σε φύλλο
- Συνεχίζουμε μέχρι να μην υπάρχει κόμβος που η μετατροπή του να μην επηρεάζει την επίδοση του δέντρου στο pruning set

Quinlan, J. R. 1987. Simplifying Decision Trees. International Journal of Human-Computer Studies, Volume 51, Issue 2, August 1999, Pages 497-510

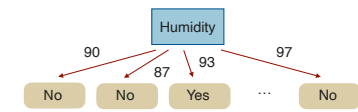
37



Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No



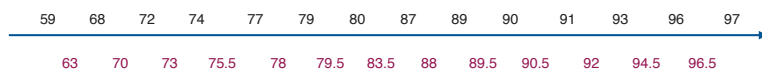
Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	90	FALSE	No
Sunny	Hot	87	TRUE	No
Overcast	Hot	93	FALSE	Yes
Rainy	Mild	89	FALSE	Yes
Rainy	Cool	79	FALSE	Yes
Rainy	Cool	59	TRUE	No
Overcast	Cool	77	TRUE	Yes
Sunny	Mild	91	FALSE	No
Sunny	Cool	68	TRUE	Yes
Rainy	Mild	80	FALSE	Yes
Sunny	Mild	72	TRUE	Yes
Overcast	Mild	96	TRUE	Yes
Overcast	Hot	74	FALSE	Yes
Rainy	Mild	97	TRUE	No



38

Επιλογή τιμής h_0

Ταξινομημένες τιμές humidity στο σύνολο δεδομένων



Πιθανές τιμές h_0

$$\text{Information Gain: } ig(h_0) = E(\text{Root}) - \frac{|\text{Humidity} \geq h_0|}{|\text{Root}|} E(\text{Humidity} \geq h_0) - \frac{|\text{Humidity} < h_0|}{|\text{Root}|} E(\text{Humidity} < h_0)$$

Βέλτιστη τιμή $h_0 = 83.5$ ($ig(83.5) = 0.94$)

39



$x_5^{\text{Outlook}} = \text{NULL}$

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

Διαχείριση άγνωστων τιμών κατά τη μάθηση

- Συμπληρώνεις με την πιθανότερη τιμή του χαρακτηριστικού
 $x_5^{\text{Outlook}} = \text{Sunny}$
- Συμπληρώνεις με την πιθανότερη τιμή του χαρακτηριστικού για τη συγκεκριμένη ετικέτα εξόδου
 $x_5^{\text{Outlook}} = \text{Overcast}$
- Προσθέτεις νέα στιγμιότυπα με όλες τις τιμές των χαρακτηριστικών
- Αγνοείς το συγκεκριμένο στιγμιότυπο όποτε εμπλέκεται στην επιλογή χαρακτηριστικού
- Κατασκευάζεις δέντρα απόφασης για την πρόβλεψη της τιμής των άγνωστων τιμών

40

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΑΓΝΩΣΤΕΣ ΤΙΜΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

- Κατασκευάζεις δέντρα απόφασης για την πρόβλεψη της τιμής των άγνωστων τιμών

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

Δέντρο απόφασης h

Temperature	Humidity	Windy	PlayTennis	Outlook
Hot	High	FALSE	No	Sunny
Hot	High	TRUE	No	Sunny
Hot	High	FALSE	Yes	Overcast
Mild	High	FALSE	Yes	Rainy
Cool	Normal	TRUE	No	Rainy
Cool	Normal	TRUE	Yes	Overcast
Mild	High	FALSE	No	Sunny
Cool	Normal	TRUE	Yes	Sunny
Mild	Normal	FALSE	Yes	Rainy
Mild	Normal	TRUE	Yes	Sunny
Mild	High	TRUE	Yes	Overcast
Hot	Normal	FALSE	Yes	Overcast
Mild	High	TRUE	No	Rainy

Cool	Normal	FALSE	Yes
------	--------	-------	-----

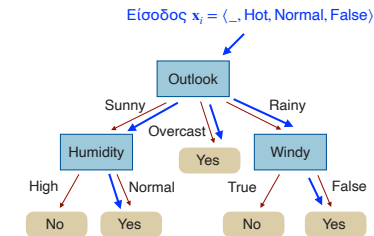
$h \rightarrow \text{Outlook} = ?$

41

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΑΓΝΩΣΤΕΣ ΤΙΜΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Διαχείριση άγνωστων τιμών κατά την πρόβλεψη

- Ελέγχει όλες τις διακλαδώσεις, δίνεις στην έξοδο την πιθανότερη τιμή φύλλου



έξοδος $y = (\text{Yes})$

42

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΛΓΟΡΙΘΜΟΣ C4.5

Βελτιώσεις στον αλγόριθμο ID3

- Χειρισμός ποιοτικών (κατηγορικών), διακριτών ποσοτικών και συνεχών ποσοτικών (αριθμητικών) χαρακτηριστικών
 - βρίσκεις τη βέλτιστη τιμή, με βάση την εντροπία
- Χειρισμός άγνωστων τιμών
 - αγνοείς τις τιμές των χαρακτηριστικών κατά τη μάθηση
 - εξετάζεις όλες τις διακλαδώσεις κατά την πρόβλεψη
- Κλάδεμα για ομαλοποίηση
- (Διαχείριση χαρακτηριστικών με διαφορετικά κόστη)
- Επιπλέον βελτιώσεις στον αλγόριθμο C5 (βελτιστοποιήσεις σε ταχύτητα, μνήμη, μικρότερα δέντρα)

Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993

Quinlan, J. R. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996

43

ΜΕΘΟΔΟΙ ENSEMBLE

Θεώρημα Condorcet's jury (προσαρμοσμένο)

Έστω $A = \{c_1, c_2, \dots, c_k\}$ ένα σύνολο από δυαδικούς ταξινομητές

$p(c_i)$ η πιθανότητα να δώσει ο ταξινομητής $c_i, i \in \mathbb{N}_k$ σωστή πρόβλεψη

M ο ταξινομητής που δίνει την πρόβλεψη της πλειοψηφίας των c_i

Αν $p(c_i) = p > 0.5$ για κάθε $c_i, i \in \mathbb{N}_k$ τότε $p(M) > p$

Αν $|A| \rightarrow \infty$ τότε $p(M) \rightarrow 1$

Nicolas de Condorcet (1743–1794), Application of Analysis to the Probability of Majority Decisions (Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix), 1785

44



Μπορούν όλα τα σύνολα ταξινομητών να οδηγήσουν σε καλούς πλειοψηφικούς ταξινομητές

Κριτήρια

- **Ποικιλία απόψεων** — Κάθε ταξινομητής-μέλος θα πρέπει να έχει ιδιωτική πληροφορία, ακόμα κι αν αυτή είναι απλώς μια εκκεντρική ερμηνεία των γνωστών γεγονότων
- **Ανεξαρτησία** — Οι απόψεις των ταξινομητών-μελών δεν καθορίζονται από τις απόψεις των άλλων
- **Διασπορά** — Τα μέλη εξειδικεύονται και εξαγουν συμπεράσματα με βάση μία τοπική γνώση
- **Σύνθεση απόψεων** — Πρέπει να καθοριστεί ένας έξυπνος μηχανισμός που συνδυάζει τις ατομικές κρίσεις και συνθέτει τη συλλογική απόφαση



Βασική ιδέα

- **Παρατήρηση:** Οι αλγόριθμοι εκμάθησης δέντρων αποφάσεων μπορούν να κατασκευάσουν δέντρα με ελαφρά διαφορετική δομή αλλά σημαντικά διαφορετικές προβλέψεις, ακόμα και για μικρές διαφορές στο σύνολο δεδομένων
- **Τεχνική:** Χρησιμοποιώντας **τμήματα** του συνόλου δεδομένων, κατασκευάσε πολλά διαφορετικά δέντρα αποφάσεων και **συνδύασε** τις προβλέψεις τους

Τμηματοποίηση συνόλου δεδομένων

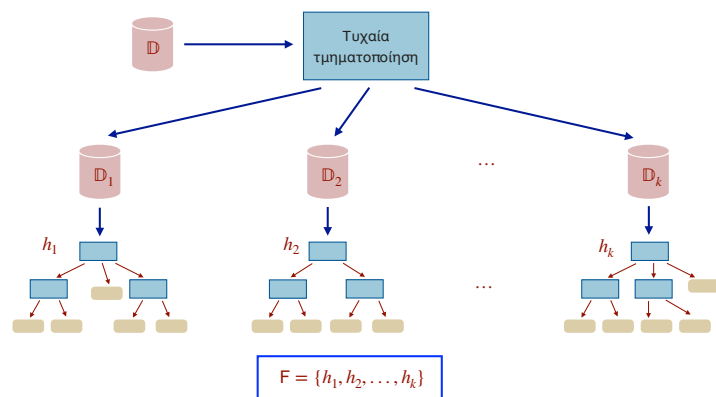
- **Bagging (Bootstrap aggregating):** Επίλεξε k διαφορετικά υποσύνολα του συνόλου δεδομένων
- **Feature Bagging (random suspace method):** Επίλεξε k διαφορετικά υποσύνολα του συνόλου χαρακτηριστικών και με βάση αυτά κατασκευάσε τα αντίστοιχα σύνολα δεδομένων (διαφορετικού χώρου εισόδου)

Εκμάθηση μοντέλου και πρόβλεψη

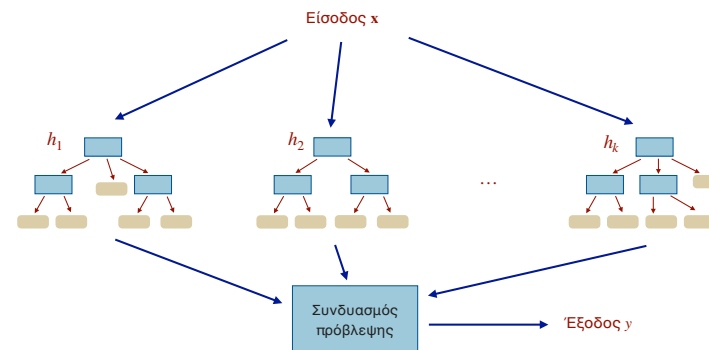
- Κατασκευάσε ένα δέντρο αποφάσεων για κάθε ένα από τα k διαφορετικά σύνολα δεδομένων
- Συνδύασε τα αποτελέσματα των k δέντρων αποφάσεων και δώσε στην έξοδο την πρόβλεψη ταξινόμησης της πλειοψηφίας



Εκμάθηση



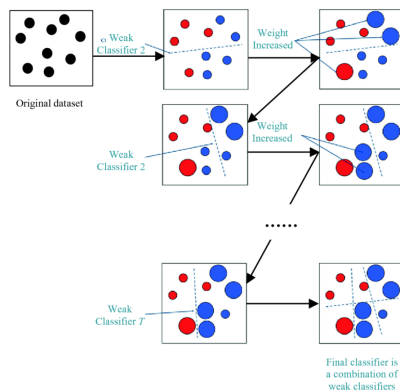
Ταξινόμηση





Βασική ιδέα

- ▶ **Παρατήρηση:** Όσο βελτιώνεται η απόδοση των δέντρων απόφασης σε μία περιοχή του πεδίου ορισμού, τόσο χειροτερεύει σε άλλες
- ▶ **Τεχνική:** Εκπαιδεύοντας ταξινομητές που έχουν καλύτερη απόδοση σε περιοχές που άλλοι έχουν χειρότερη, μπορούν οι ταξινομητές να εξειδικεύονται και έτσι ο συνδυασμός τους να αποδίδει καλύτερα



49



Βασική ιδέα

Σύνολο δεδομένων $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$

$H = \{h | h : X \rightarrow Y\}$: σύνολο δέντρων απόφασης $X \rightarrow Y$

$h \in H \leftarrow \text{DecisionTree.train}(\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$

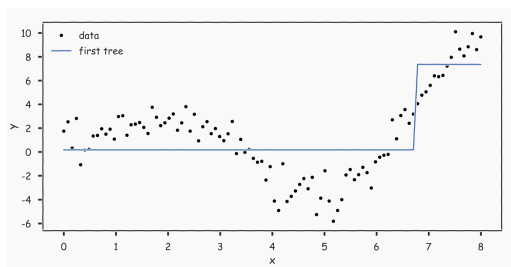
$\text{minimal}(h)$

$$h = \sum_{h_j \in H} \lambda_j h_j \leftarrow \text{DecisionTree.train}(\{(\mathbf{x}_i, y_i)\}_{i=1}^n)$$

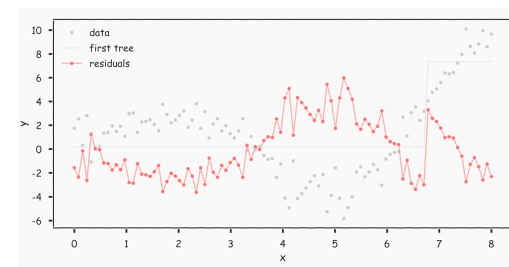
GradientBoostTree.train(D)

1. $h_0 = \text{DecisionTree.train}(D)$, όπου $h_0 \in H$, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
2. $h = h_0$
3. $D' = \{(\mathbf{x}_i, r_i)\}_{i=1}^n$, $r_i = y_i - h(\mathbf{x}_i)$
4. $h_k = \text{DecisionTree.train}(D')$, όπου $h_k \in H$, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
5. $h = h + \lambda h_k$
5. Επανάλαβε τα βήματα 3 έως 5, έως ότου φτάσεις το επιθυμητό σφάλμα

50

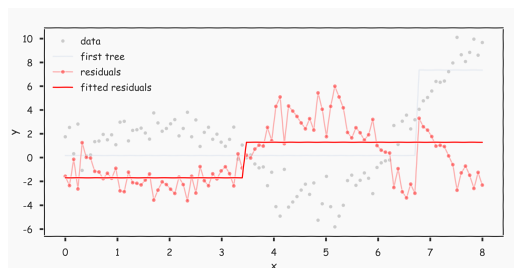


51



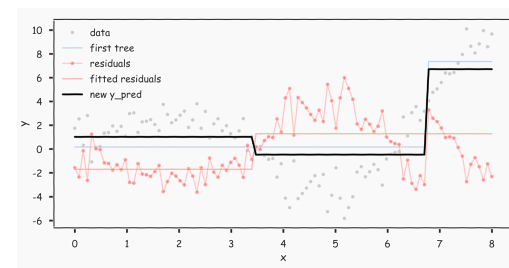
52

ΜΕΘΟΔΟΙ ENSEMBLE - ΕΝΙΣΧΥΣΗ (BOOSTING)



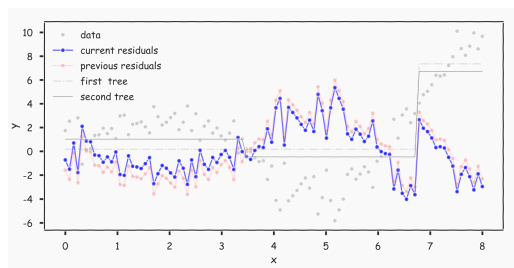
53

ΜΕΘΟΔΟΙ ENSEMBLE - ΕΝΙΣΧΥΣΗ (BOOSTING)



54

ΜΕΘΟΔΟΙ ENSEMBLE - ΕΝΙΣΧΥΣΗ (BOOSTING)



55