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 στιγμιοτύπων

Y : σύνολο ετικετών

$f : X \rightarrow Y$: ιδανικός ταξινομητής

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

Είσοδος

Σύνολο δεδομένων $\{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle\}$

Έξοδος

Υπόθεση $h \in H$ που προσεγγίζει το f

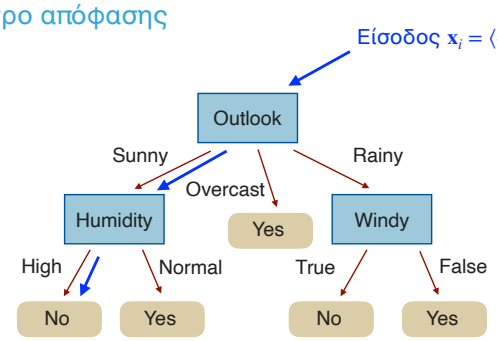
Στήλες χαρακτηριστικών: $\mathbf{x}_i$ Παίζεις τένις ή όχι: y_i

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

$\langle \mathbf{x}_i, y_i \rangle$

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΟΡΙΣΜΟΣ

Δέντρο απόφασης



έξοδος $y_i = \langle$No$\rangle$

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

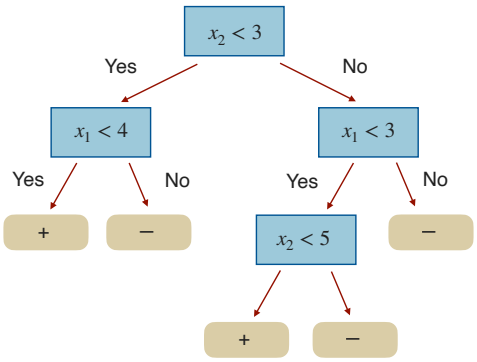
Σε κάθε εσωτερικό κόμβο ελέγχεται η τιμή του χαρακτηριστικού x_i

Σε κάθε διακλάδωση επιλέγεται μία τιμή του χαρακτηριστικού x_i

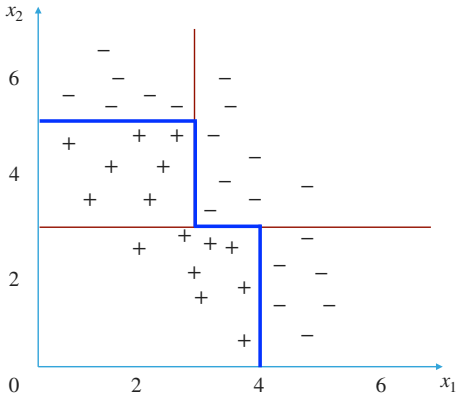
Σε κάθε φύλλο αποδίδεται μία ετικέτα y στο στοιχείο

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΚΑΜΠΥΛΕΣ ΔΙΑΧΩΡΙΣΜΟΥ

Δέντρο απόφασης

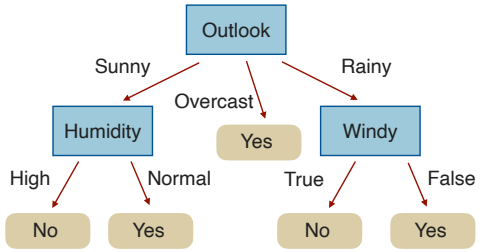


Καμπύλη διαχωρισμού

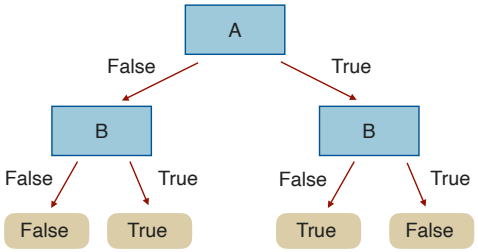


- Η καμπύλη διαχωρισμού χωρίζει το χώρο χαρακτηριστικών σε (υπερ-)κύβους παράλληλους των αξόνων
- Κάθε (υπερ-)κυβική επιφάνεια αντιστοιχίζεται σε μία ετικέτα - ή (στη γενική περίπτωση) σε μία κατανομή πιθανοτήτων πάνω στις ετικέτες

Ικανότητα αναπαράστασης



Δένδρο απόφασης για XOR

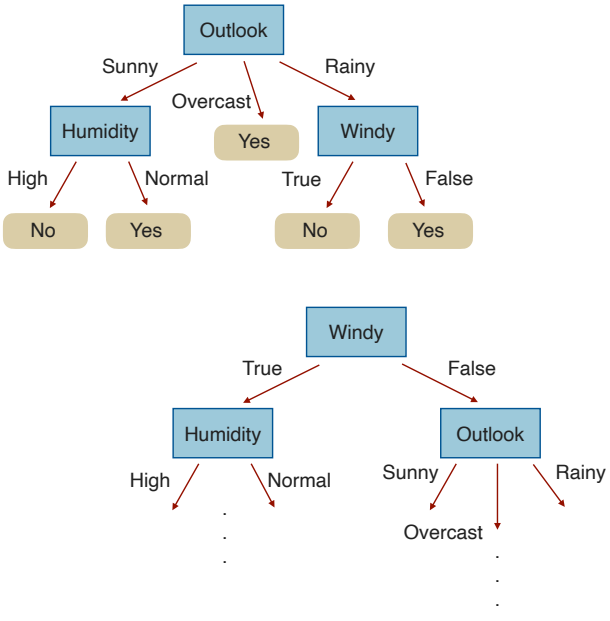


IF (Outlook IS Overcast) ∨ ((Outlook IS Sunny) ∧ Humidity IS Normal)) ∨ ((Outlook IS Rainy) ∧ Windy IS False))
THEN (PlayTennis IS YES)

- 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

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

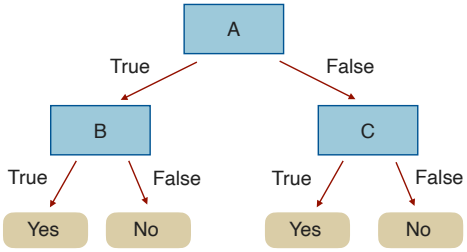


- Διαφορετικά δέντρα απόφασης μπορούν να είναι ισοδύναμα (να ταξινομούν στην ίδια κλάση ένα στιγμιότυπο)
- *Πόσα διαφορετικά δέντρα απόφασης αναπαριστούν μία συγκεκριμένη λογική έκφραση;*

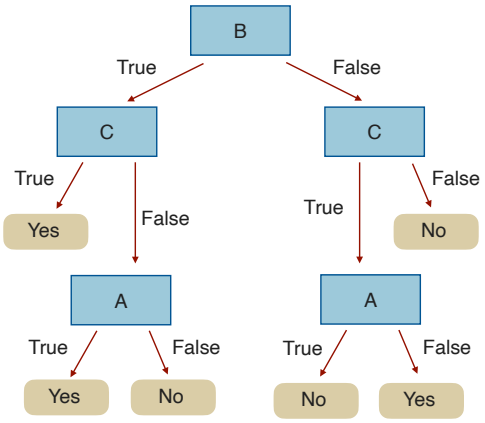
ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΚΦΡΑΣΤΙΚΟΤΗΤΑ

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

Δέντρο απόφασης 1



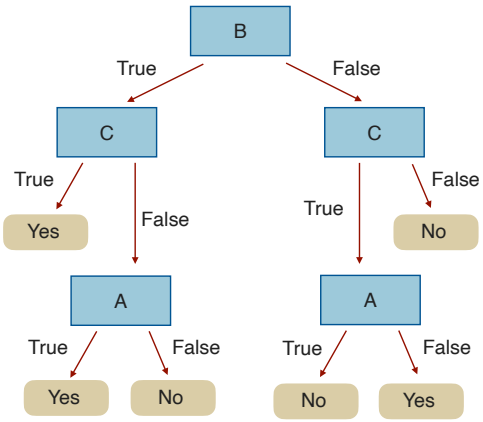
Δέντρο απόφασης 2



ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΠΡΩΤΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρατηρήσεις - Θέματα για μελέτη

- Το βάθος των δέντρων απόφασης δεν μπορεί να είναι μεγαλύτερο από το πλήθος των χαρακτηριστικών των στιγμιοτύπων
- 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 (human-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 Occam - 1324)
- Είναι σημαντικό να κατασκευάζουμε και να χρησιμοποιούμε *απλά* δέντρα απόφασης
- Το πρόβλημα εύρεσης του ελάχιστου δέντρου απόφασης είναι δισεπίλυτο, έχει αποδειχθεί NP-complete (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)



ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ

Δεδομένα

X : σύνολο στιγμιοτύπων

Y : σύνολο ετικετών

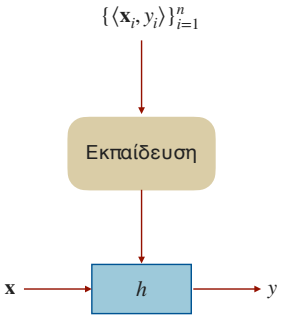
$\mathbf{x} \sim \mathcal{D}(X)$

$f : X \to Y$: ιδανικός ταξινομητής

Σύνολο δεδομένων $\mathbb{D} = \{\langle \mathbf{x}_i, y_i \rangle\}_{i=1}^{n} = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_n, y_n \rangle\}$

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

$H = \{h \mid h : X \to Y\}$: σύνολο δέντρων απόφασης $X \to Y$



Διαδικασία εκμάθησης

Πρόβλημα

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

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

DesicionTree . train($\mathbb{D}$)

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

     Αλλιώς DecisionTree . train($\mathbb{D}'$)

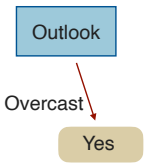
ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

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

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

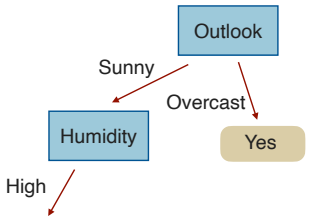


Outlook	Temperature	Humidity	Windy	PlayTennis
Overcast	Hot	High	FALSE	Yes
Overcast	Cool	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

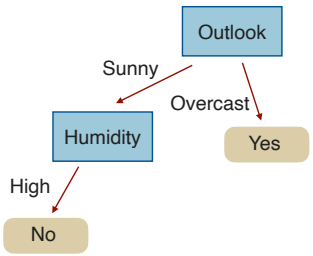


Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Sunny	Mild	Normal	TRUE	Yes

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

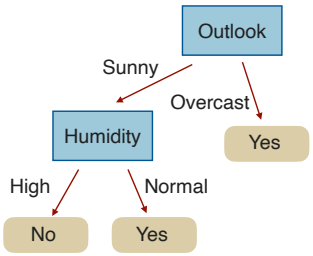


Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Sunny	Mild	High	FALSE	No

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

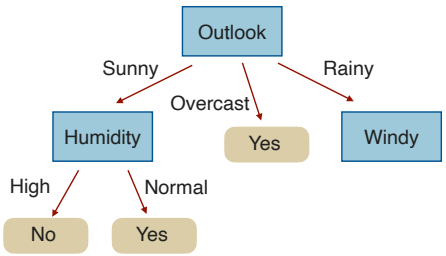


Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Cool	Normal	TRUE	Yes
Sunny	Mild	Normal	TRUE	Yes

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

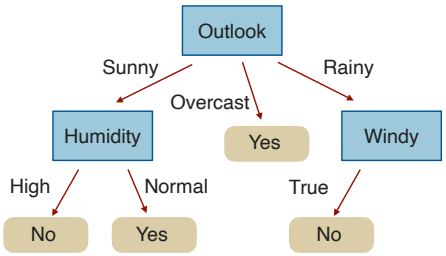


Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Rainy	Mild	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

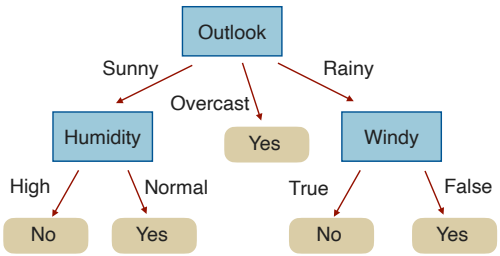


Outlook	Temperature	Humidity	Windy	PlayTennis
Rainy	Cool	Normal	TRUE	No
Rainy	Mild	High	TRUE	No

DesicionTree . train($\mathbb{D}$)

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

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

DesicionTree . train($\mathbb{D}$)

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΦΕΛΗΣ ΑΛΓΟΡΙΘΜΟΣ

DesicionTree . train($\mathbb{D}$)

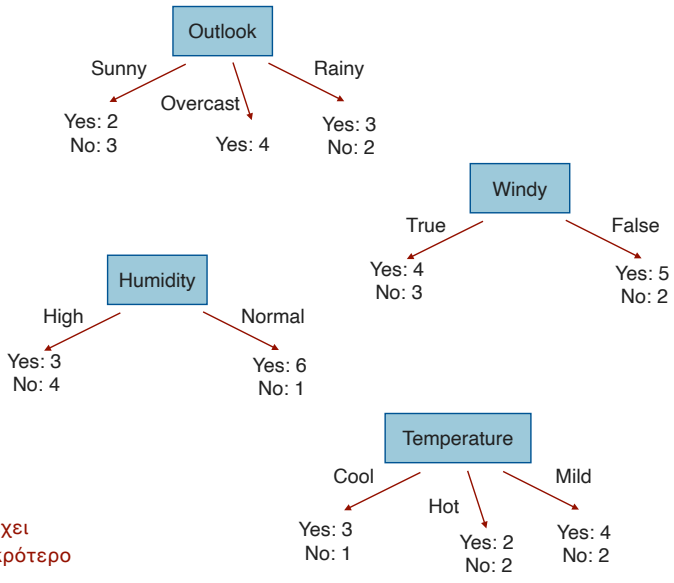
1. Επίλεξε ένα χαρακτηριστικό εισόδου a που παίρνει διαφορετικές τιμές στο $\mathbb{D}$
2. Φτιάξε ένα νέο κόμβο A και όρισε το a ως χαρακτηριστικό απόφασης
3. Για κάθε διαφορετική τιμή του a:
   - 3.1 Φτιάξε ένα νέο κόμβο ως παιδί του τρέχοντος κόμβου
   - 3.2 Όρισε το $\mathbb{D}'$ ως το υποσύνολο του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

     Αλλιώς DecisionTree . train($\mathbb{D}'$)

- 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)

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;

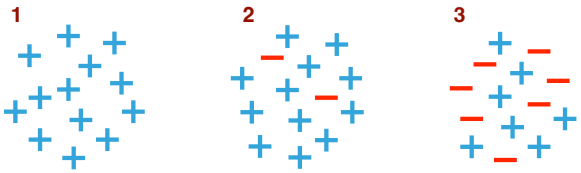
Μέθοδος

- **Τυχαία**: 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)

Εντροπία

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

- Μέτρο της καθαρότητας ενός συνόλου παραδειγμάτων
  - μικρότερη εντροπία, μεγαλύτερη καθαρότητα
- Μείωση εντροπίας, κέρδος πληροφορίας



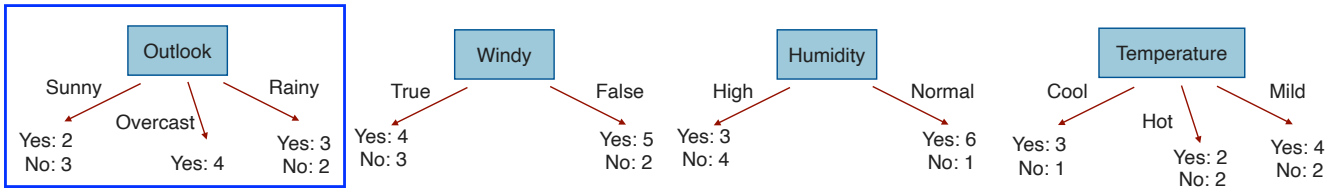
- Ποια από τις κατανομές έχει μεγαλύτερο κέρδος πληροφορίας;

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

$$E_2 = -0.133 \ \log_2 0.133 - 0.87 \ \log_2 0.87 \simeq 0{,}565$$

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

Επιλέγουμε το χαρακτηριστικό που γνωρίζοντας την τιμή του πετυχαίνουμε τη μεγαλύτερη μείωση της εντροπίας



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

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

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

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

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

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

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

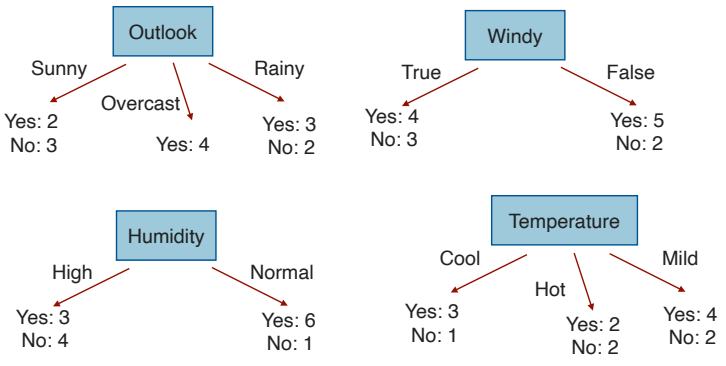
DesicionTree . ID3($\mathbb{D}$)

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

     Αλλιώς DecisionTree . ID3($\mathbb{D}'$)

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

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



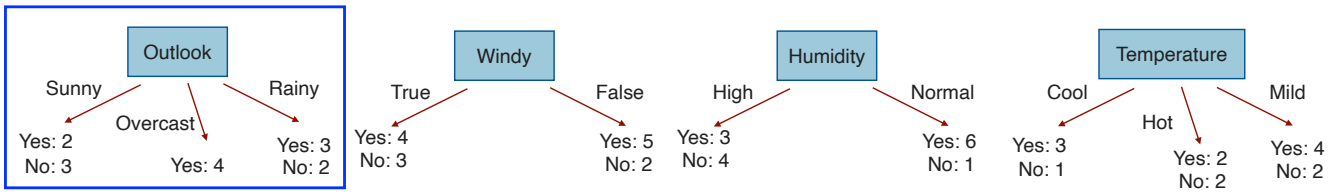
- Πόσο συχνά ένα τυχαίο δείγμα που επιλέγεται ταξινομείται λανθασμένα, αν του αποδοθεί μία τυχαία ετικέτα;

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

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

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

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΟΥ ΑΠΟΦΑΣΗΣ - GINI INDEX



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

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

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

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

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

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

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

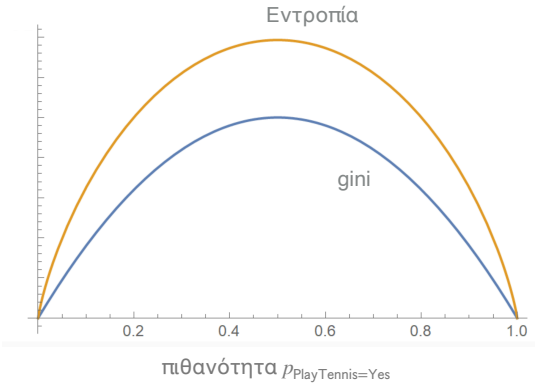
ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΑΛΓΟΡΙΘΜΟΣ CART

DesicionTree . CART($\mathbb{D}$)

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

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΣΥΓΚΡΙΣΗ ΚΡΙΤΗΡΙΩΝ ΕΠΙΛΟΓΗΣ



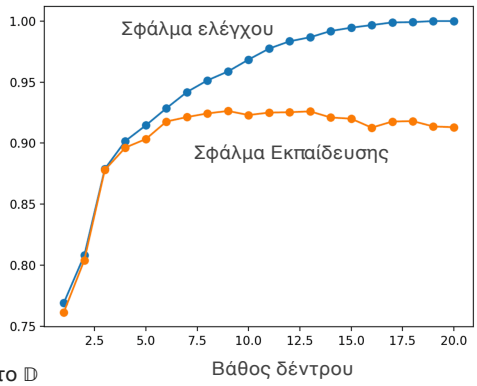
- Παρόμοια αποτελέσματα στην πράξη
  - διαφωνούν σε ελάχιστες περιπτώσεις στην επιλογή χαρακτηριστικού
- Δυσκολότερος ο υπολογισμός για την εντροπία

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

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΥΠΕΡΠΡΟΣΑΡΜΟΓΗ

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



DesicionTree . train($\mathbb{D}$)

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

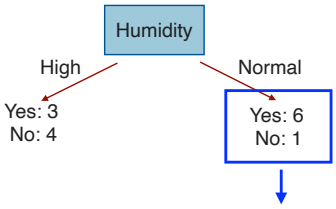
     Αλλιώς DecisionTree . train($\mathbb{D}'$)

ΕΚΜΑΘΗΣΗ ΔΕΝΤΡΩΝ ΑΠΟΦΑΣΗΣ - ΚΛΑΔΕΜΑ

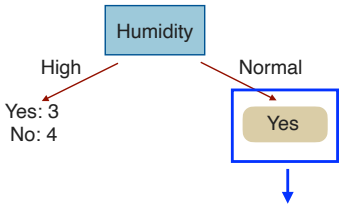
DesicionTree . train($\mathbb{D}$)

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

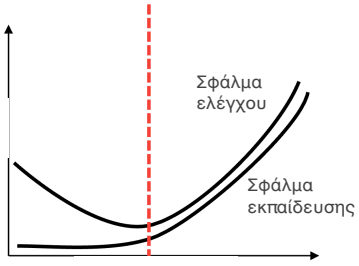
   Αλλιώς DecisionTree . train($\mathbb{D}'$)



Δημιουργία νέου κόμβου?



Εμπιστοσύνη ~86%

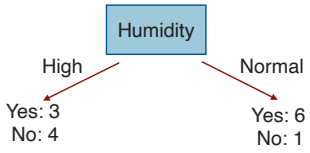


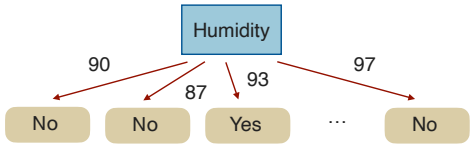
Αριθμός κόμβων που κλαδεύονται

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕ ΑΡΙΘΜΗΤΙΚΕΣ ΤΙΜΕΣ

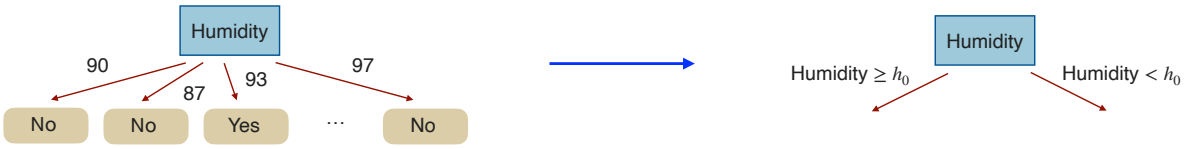
Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	TRUE	Yes
Rainy	Mild	Normal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





ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕ ΑΡΙΘΜΗΤΙΚΕΣ ΤΙΜΕΣ



Επιλογή τιμής h_0

Ταξινομημένες τιμές humidity στο σύνολο δεδομένων

59 68 72 74 77 79 80 87 89 90 91 93 96 97

63 70 73 75.5 78 79.5 83.5 88 89.5 90.5 92 94.5 96.5

Πιθανές τιμές h_0

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

Βέλτιστη τιμή $h_0 = 83.5$ $(\text{ig}(83.5) = 0.94)$

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΑΓΝΩΣΤΕΣ ΤΙΜΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

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

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
	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

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

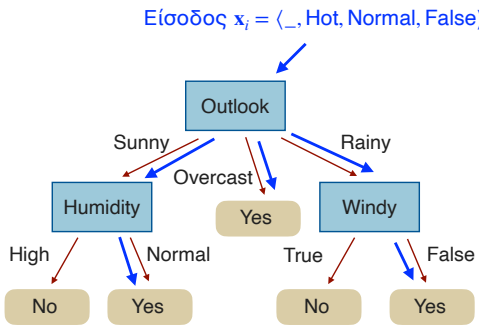
Cool	Normal	FALSE	Yes

$h \rightarrow$ Outlook = ?

ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ - ΑΓΝΩΣΤΕΣ ΤΙΜΕΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

Διαχείριση άγνωστων τιμών κατά την πρόβλεψη

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



έξοδος $y = \langle$Yes$\rangle$

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 αγνώστων 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

ΜΕΘΟΔΟΙ ENSEMBLE - ΤΥΧΑΙΑ ΔΑΣΗ (RANDOM FORESTS)

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, κατασκεύασε πολλά διαφορετικά δέντρα αποφάσεων και *συνδύασε* τις προβλέψεις τους

Τμηματοποίηση συνόλου δεδομένων

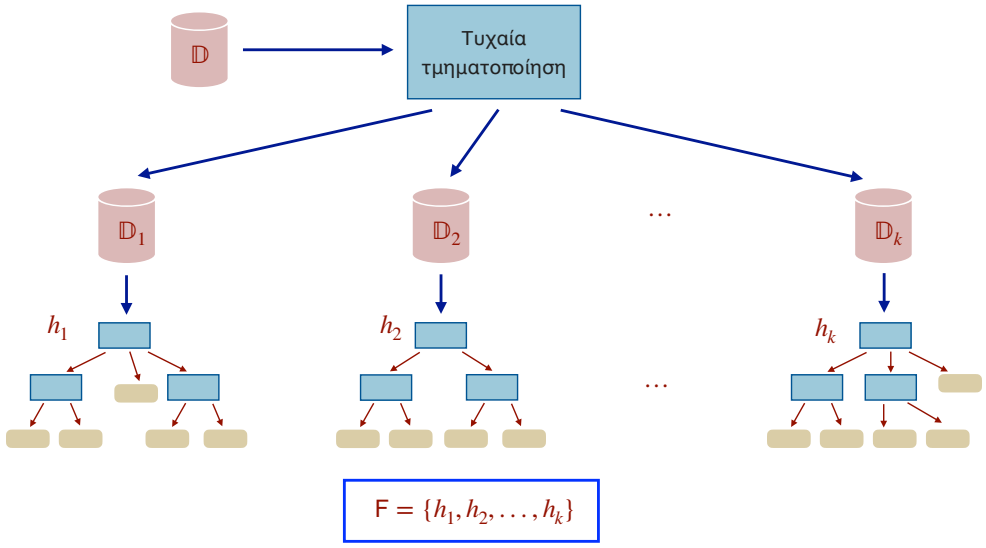
- **Bagging (Bootstrap aggregating):** Επίλεξε *k* διαφορετικά υποσύνολα του συνόλου δεδομένων
- **Feature Bagging (random suspace method):** Επίλεξε *k* διαφορετικά υποσύνολα του συνόλου χαρακτηριστικών και με βάση αυτά κατασκεύασαμε τα αντίστοιχα σύνολα δεδομένων (διαφορετικού χώρου εισόδου)

Εκμάθηση μοντέλου και πρόβλεψη

- Κατασκεύασε 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

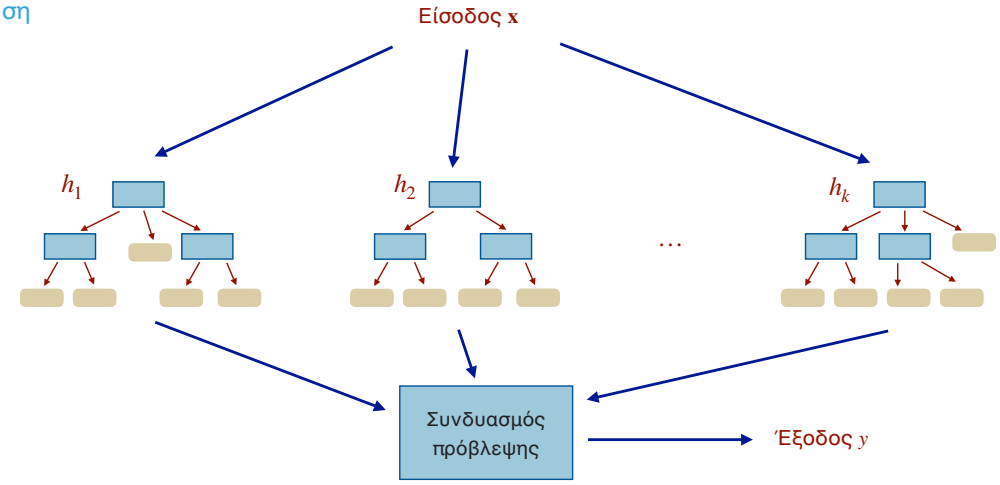
ΜΕΘΟΔΟΙ ENSEMBLE - ΤΥΧΑΙΑ ΔΑΣΗ (RANDOM FORESTS)

Εκμάθηση



ΜΕΘΟΔΟΙ ENSEMBLE - ΤΥΧΑΙΑ ΔΑΣΗ (RANDOM FORESTS)

Ταξινόμηση



Breiman, L. Random Forests. *Machine Learning* **45**, 5–32 (2001)