Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

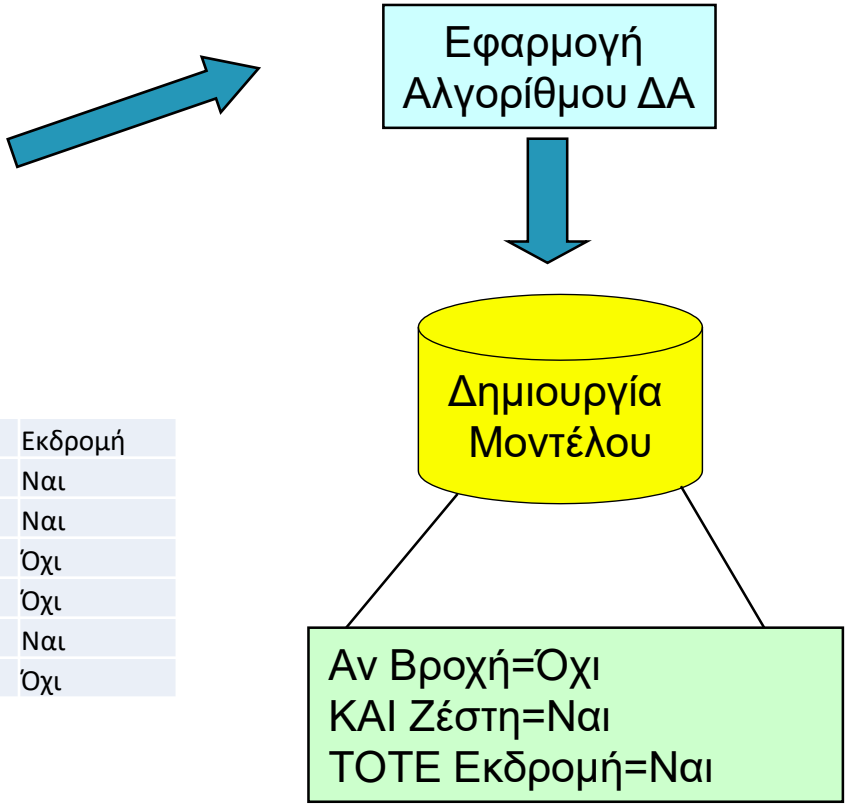
Δέντρα Αποφάσεων

Δέντρα Αποφάσεων

Οι λειτουργίες που υποστηρίζει ένα δέντρο απόφασης είναι:

- Κατηγοριοποιεί τα δεδομένα χτίζοντας ΔΑ βασισμένα στο training set και
- Προβλέπει την κατηγορία ενός νέου αντικειμένου με βάση το ΔΑ που χτίστηκε στο προηγούμενο βήμα.

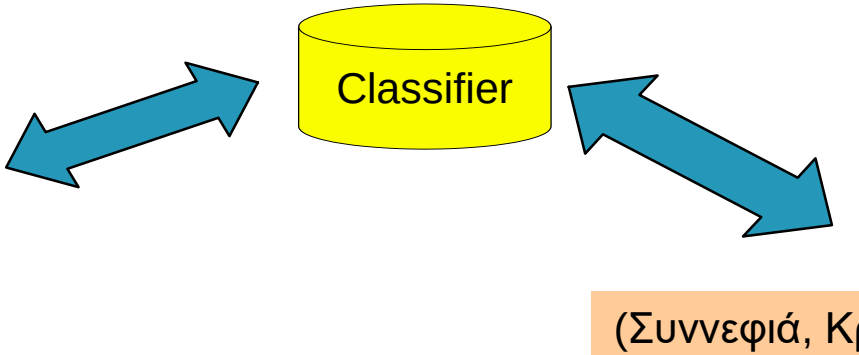
ΔΑ: Κατασκευή Μοντέλου



Training Data

Α	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
1	Ηλιοφάνεια	Ζέστη	Πολλές	Όχι	Ναι
2	Ηλιοφάνεια	Ζέστη	Πολλές	Ναι	Ναι
3	Συννεφιά	Ζέστη	Πολλές	Όχι	Όχι
4	Βροχή	Κανονική	Πολλές	Όχι	Όχι
5	Βροχή	Κρύο	Λίγες	Ναι	Ναι
6	Συννεφιά	Κρύο	Λίγες	Όχι	Όχι

ΔΑ: Χρήση Μοντέλου για Πρόβλεψη



	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
	Συννεφιά	Κρύο	Λίγες	Όχι	?

(Συννεφιά, Κρύο, Λίγες, Όχι)

Εκδρομή;

Όχι

Παράδειγμα Δημιουργίας ΔΑ

Στο παρακάτω παράδειγμα έχουμε όλα τα στοιχεία σχετικά με την πραγματοποίηση ή μη μιας εκδρομής.

A	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
1	Ηλιοφάνεια	Ζέστη	Πολλές	Όχι	Ναι
2	Ηλιοφάνεια	Ζέστη	Πολλές	Ναι	Ναι
3	Συννεφιά	Ζέστη	Πολλές	Όχι	Όχι
4	Βροχή	Κανονική	Πολλές	Όχι	Όχι
5	Βροχή	Κρύο	Λίγες	Ναι	Ναι
6	Συννεφιά	Κρύο	Λίγες	Όχι	Όχι
7	Βροχή	Κανονική	Λίγες	Όχι	Όχι
8	Συννεφιά	Ζέστη	Λίγες	Όχι	Όχι

Για τα παραπάνω δεδομένα θα χτίσουμε το ΔΑ χρησιμοποιώντας τους αλγορίθμους ID3 και Cart.

Παράδειγμα ID3

Για την εφαρμογή του αλγορίθμου ID3 θα πρέπει να υπολογίσουμε

- την εντροπία $E(S) = -\sum_{i=1}^{k} P_i log_2(Pi)$ και
- το Πληροφοριακό κέρδος $G(S, A) = E(S) - I(S, A)$.

Παράδειγμα ID3

- Προκειμένου να υπολογίσουμε την εντροπία E(S), παρατηρούμε ότι για την μεταβλητή κλάσης έχουμε την εμφάνιση του NAI 3 φορές και του Όχι 5 φορές. Συνεπώς η εντροπία είναι

$$E(S) = -\frac{3}{8}log_2(\frac{3}{8}) - \frac{5}{8}log_2(\frac{5}{8})$$=0.53+0.42=0.95

Παράδειγμα ID3

Στη συνέχεια υπολογίζουμε το πληροφοριακό κέρδος για κάθε μεταβλητή.

<u>Καιρός.</u> Συνολικά 8 δείγματα, Ηλιοφάνεια*2 (ΝΑΙ), Συννεφιά*3(ΟΧΙ) και Βροχή*3(1 ΝΑΙ,2 ΟΧΙ). Συνεπώς, έχουμε:

$$G(S,\text{Καιρός})=E(S)-I(S,\text{Καιρός})=E(S)-\frac{2}{8}E(S_{\text{Ηλιοφάνεια}})-\frac{3}{8}E(S_{\text{Συμμετοχές}})-\frac{3}{8}E(S_{\text{Βροχή}})$$

Συνεπώς θα πρέπει να υπολογίσουμε τα παρακάτω

$E(S_{\text{Ηλιοφάνεια}})$, $E(S_{\text{Συμμετοχές}})$ και $E(S_{\text{Βροχή}})$.

Παράδειγμα ID3

- $E(S_{\text{Ηλιοφάνεια}}) = -\frac{2}{2}log_2(\frac{2}{2}) - \frac{0}{2}log_2(\frac{0}{2})$=0
- $E(S_{\text{Συμμετοχές}}) = -\frac{0}{3}log_2(\frac{0}{3}) - \frac{3}{3}log_2(\frac{3}{3})$=0
- $E(S_{\text{Βροχή}}) = -\frac{1}{3}log_2(\frac{1}{3}) - \frac{2}{3}log_2(\frac{2}{3})$=0.53+0.39=0.92
- Οπότε G(S,Καιρός)=E(S)-I(S, Καιρός)=0.95 - $\frac{2}{8}$E (0) - $\frac{3}{8}$E (0) - $\frac{3}{8}$E (0.92)=0.345

Παράδειγμα ID3

Στη συνέχεια υπολογίζουμε το πληροφοριακό κέρδος για κάθε μεταβλητή.

<u>Θερμοκρασία</u>. Συνολικά 8 δείγματα, Ζέστη*4 (2 ΝΑΙ,2 ΟΧΙ), Κανονική*2(ΟΧΙ) και Κρύο*2(1 ΝΑΙ,1 ΟΧΙ). Συνεπώς, έχουμε:

$$G(S,\text{Θερμοκρασία})=E(S)-I(S,\text{Ζέστη})=E(S)-\frac{4}{8}E(S_{\text{Ηλιοφάνεια}})-\frac{2}{8}E(S_{\text{Συμμετοχές}})-\frac{2}{8}E(S_{\text{Βροχή}})$$

Παράδειγμα ID3

- $E(S_{\text{Ζέστη}}) = -\frac{2}{4}log_2(\frac{2}{4}) - \frac{2}{4}log_2(\frac{2}{2})$=1

- $E(S_{\text{Κανονική}}) = -\frac{0}{2}log_2(\frac{0}{2}) - \frac{2}{2}log_2(\frac{2}{2})$=0

- $E(S_{\text{Κρύο}}) = -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2})$=1

Οπότε

G(S,Θερμοκρασία)=E(S)-I(S,Θερμοκρασία)=0.95 - $\frac{4}{8}$ *1 - $\frac{2}{8}$*0- $\frac{2}{8}$*1=0.2

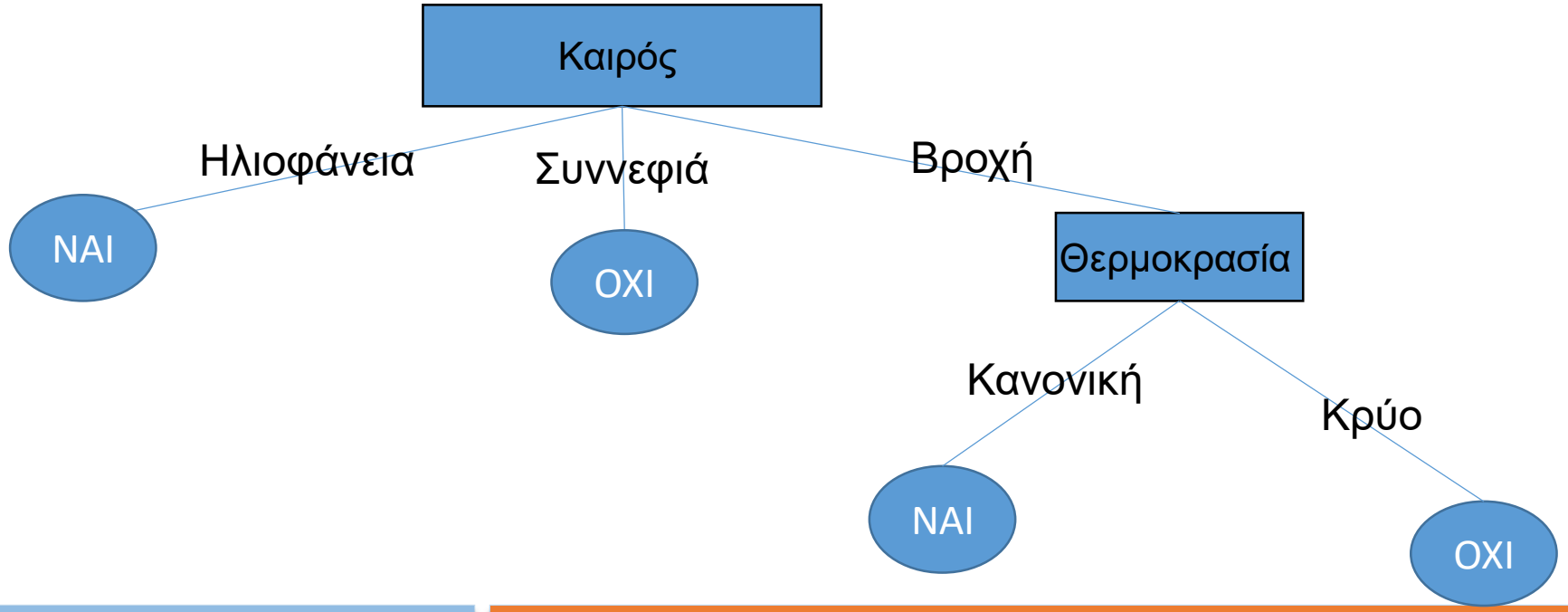
Παράδειγμα ID3

Επαναλαμβάνουμε την ίδια διαδικασία για τις υπόλοιπες μεταβλητές, οπότε τα πληροφοριακά κέρδη διαμορφώνονται ως εξής

- G(S,Καιρός)=0.345
- G(S,Θερμοκρασία.)=0.2
- G(S,Συμμετοχές)=0.045
- G(S,Αέρας.)=0.125

Παράδειγμα ID3

Το ΔΑ που προκύπτει για τις παραπάνω μεταβλητές έχει την παρακάτω μορφή.



Cart ΔΑ

Για την δημιουργία ενός ΔΑ με την τεχνική του CART χρησιμοποιούμε τον δείκτη Gini Index του οποίου ο τρόπος υπολογισμού περιγράφεται από την παρακάτω σχέση

$$gini(s) = 1 - \sum_{j=1}^{K} \mathrm{P}_j^2$$

- S=σύνολο δεδομένων,
- K=κλάσεις και
- Pj είναι η πιθανότητα εμφάνισης της κλάσης j στο σύνολο δεδομένων S.

Δένδρα Απόφασης – Gini Index

A	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
1	Ηλιοφάνεια	Ζέστη	Πολλές	Όχι	Ναι
2	Ηλιοφάνεια	Ζέστη	Πολλές	Ναι	Ναι
3	Συννεφιά	Ζέστη	Πολλές	Όχι	Όχι
4	Βροχή	Κανονική	Πολλές	Όχι	Όχι
5	Βροχή	Κρύο	Λίγες	Ναι	Ναι
6	Συννεφιά	Κρύο	Λίγες	Όχι	Όχι

- $gini(\text{Ηλιοφάνεια}) = 1 - (P_{NAI}^2 + P_{OXI}^2) = 1 - (\frac{2}{2}^2 + \frac{0}{2}^2) = 1 - (1+0) = 0$
- $gini(\text{Συννεφιά}) = 1 - (P_{NAI}^2 + P_{OXI}^2) = 1 - (\frac{0}{2}^2 + \frac{2}{2}^2) = 1 - (0+1) = 0$
- $gini(\text{Βροχή}) = 1 - (P_{NAI}^2 + P_{OXI}^2) = 1 - (\frac{1}{2}^2 + \frac{1}{2}^2) = 1 - (\frac{1}{2}) = 0.5$

Οπότε για την μεταβλητή Καιρός καταλήγουμε στο συμπέρασμα

$gini(\text{Καιρός}) = \frac{2}{6} gini(\text{Ηλιοφάνεια}) + \frac{2}{6} gini(\text{Συννεφιά}) + \frac{2}{6} gini(\text{Βροχή}) =$

$= \frac{2}{6} * 0 + \frac{2}{6} * 0 + \frac{2}{6} * 1/2 = 0.16$

Δένδρα Απόφασης – Gini Index

A	Καιρός	Θερμοκρασία	Συμμετοχές	Βροχή	Εκδρομή
1	Ηλιοφάνεια	Ζέστη	Πολλές	Όχι	Ναι
2	Ηλιοφάνεια	Ζέστη	Πολλές	Ναι	Ναι
3	Συννεφιά	Ζέστη	Πολλές	Όχι	Όχι
4	Βροχή	Κανονική	Πολλές	Όχι	Όχι
5	Βροχή	Κρύο	Λίγες	Ναι	Ναι
6	Συννεφιά	Κρύο	Λίγες	Όχι	Όχι

Κάνω το ίδιο για τις υπόλοιπες , οπότε έχουμε

- $gini$(Καιρός) =0.16.
- $gini$(Θερμοκρασία) =0.35.
- $gini$(Συμμετοχές) =0.5.
- $gini$(Βροχή) =0.25.

ΤΕΛΟΣ ΕΝΟΤΗΤΑΣ