# Μηχανική μάθηση

Ιωάννης Γ. Τσούλος

Τμήμα Πληροφορικής και τηλεπικοινωνιών
Πανεπιστήμιο Ιωαννίνων

2022

# Περίληψη

# Βασικά στοιχεία

1. Κατάλληλα για δημιουργία κανόνων απόφασης
2. Κατάλληλα για Data mining
3. Απαιτούν την ύπαρξη θετικών και αρνητικών περιπτώσεων για την δημιουργία κανόνων απόφασης.

# Αναπαράσταση

1. Τα δένδρα απόφασης είναι μοντέλα της μηχανικής μάθησης που χρησιμοποιούνται στην ταξινόμηση δεδομένων.
2. Σε αυτά τα μοντέλα δημιουργείται ένα σύνολο κανόνων απόφασης σε δενδρική δομή
3. Στους εσωτερικούς κόμβους του δένδρου βρίσκονται χαρακτηριστικά από το πρόβλημα
4. Στα φύλλα βρίσκονται αποφάσεις, δηλαδή η κατηγορία που θα επιλεχθεί.

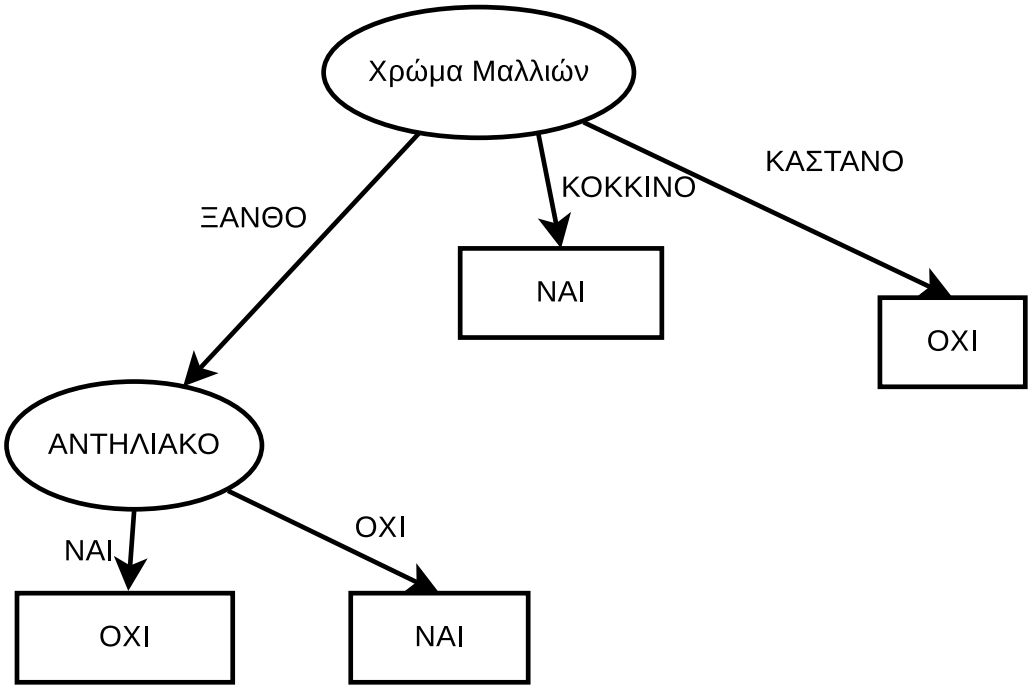
# Παράδειγμα κατηγορίας εγκαυμάτων

| Όνομα | Μαλλιά | Ύψος | Βάρος | Αντηλιακό | Κάηκε |
|---|---|---|---|---|---|
| Σάρα | Ξανθά | Μέτριο | Ελαφρύ | Όχι | Ναι |
| Άννα | Ξανθά | Ψηλό | Μέτριο | Ναι | Όχι |
| Νίκος | Καστανά | Κοντό | Μέτριο | Ναι | Όχι |
| Άλεξανδρος | Ξανθά | Κοντό | Μέτριο | Όχι | Ναι |
| Νίκη | Κόκκινα | Μέτριο | Βαρύ | Όχι | Ναι |
| Τάκης | Καστανά | Ψηλό | Βαρύ | Όχι | Όχι |
| Καίτη | Καστανά | Μέτριο | Βαρύ | Όχι | Όχι |
| Γιάννης | Ξανθά | Κοντό | Ελαφρύ | Ναι | Όχι |

# Παράδειγμα κατηγορίας εγκαυμάτων

1. Οι στήλες ΟΝΟΜΑ, ΜΑΛΛΙΑ, ΥΨΟΣ, ΒΑΡΟΣ, ΑΝΤΗΛΙΑΚΟ είναι χαρακτηριστικά.
2. Η στήλη ΚΑΗΚΕ είναι η επιθυμητή κατηγορία.
3. Η πρώτη στήλη ΟΝΟΜΑ δεν αναμένενεται να έχει κάποιο αποτέλεσμα στην μάθηση και μπορεί να αφαιρεθεί.

# Σχήμα δένδρου για εγκαύματα



# Δένδρο για εγκαύματα

Από το δένδρο αυτό μπορούν να εξαχθούν οι επόμενοι κανόνες:

1. Αν κάποιος έχει κόκκινο χρώμα μαλλιών **καίγεται**.
2. Αν κάποιος έχει καστανό χρώμμα μαλλιών **δεν καίγεται**.
3. Αν κάποιος έχει ξανθό χρώμα μαλλιών και δεν φορά αντηλιακο **καίγεται**.
4. Αν κάποιος έχει ξανθό χρώμα μαλλιών και φορά αντηλιακό τότε **δεν καίγεται**.

Προφανώς σε αυτό το δένδρο απόφασης δεν έχουν ληφθεί υπόψη χαρακτηριστικά τα οποία δεν έχουν σχέση με το αποτέλεσμα όπως το βάρος και το ύψος κάτι που στις περισσότερες περιπτώσεις δεν συμβαίνει.

# Τυπικός αλγόριθμος κατασκευής δένδρου

Ο τυπικός αλγόριθμος κατασκευής δένδρων έχει ως ακολούθως:

1. Έστω $x_1, x_d, ..., x_n$ τα χαρακτηριστικά του προβλήματος.
2. Έστω $S$ τα δεδομένα εκπαίδευσης.
3. Επιλογή ενός χαρακτηριστικού $x_k$ από το σύνολο των χαρακτηριστικών του προβλήματος
4. Για κάθε μία επιλογή $f_1, f_2, .., f_M$ του χαρακτηριστικού $x_k$ κάνε
   1. Αν τα δεδομένα που περιέχουν μόνον στο χαρακτηριστικό $x_k$ το $f_i$ κατατάσσονται σε μια κατηγορία, τότε βάλε στο δένδρο σαν τερματικό φύλλο αυτή την τιμή της κατηγορίας
   2. Αλλιώς δημίουργησε νέο υποδένδρο με κάποιο άλλο χαρακτηριστικό (που δεν το έχουμε λάβει ήδη) και θέσε σαν ρίζα του την τιμή $f_i$
5. Η παραπάνω διαδικασία συνεχίζεται μέχρι να δημιουργηθούν μόνο τερματικοί κόμβοι.

# Ο αλγόριθμος ID3

1. Ο παραπάνω αλγόριθμος επιλέγει τα χαρακτηριστικά με τυχαίο τρόπο και πολλές φορές αυτό δεν είναι αποδοτικό.
2. Για το παράδειγμα των εγκαυμάτων θα οδηγούσε στην επιλογή όλων των χαρακτηριστικών ακόμα και αυτών που δεν συνεισφέρουν κάτι στην επιλογή της σωστής κατηγορίας.
3. Σε αυτήν την περίπτωση θα οδηγούσε σε μεγάλο δένδρο με αργό χρόνο εκπαίδευσης και απόκρισης.
4. Ο αλγόριθμος ID3 επιλέγει κάθε φορά το καλύτερο χαρακτηριστικό βρίσκοντας την συνεισφορά του στο τελικό αποτέλεσμα μέσω της εντροπίας (entropy) και του κέρδους πληροφορίας (information gain).

# Εντροπία

- Η εντροπία για τα δεδομένα στο $S$ δίνεται από την εξίσωση:
- $$H(S) = \sum_{c \in C} -p(c) \log_2 p(c) \quad (1)$$

  όπου
  - $S$ είναι τα δεδομένα στο σύνολο εκπαίδευσης
  - C είναι το σύνολο των κατηγορίων, στο παράδειγμα με το έγκαυμα είναι (NAI,OXI)
  - p(c) είναι το κλάσμα των δεδομένων εκπαίδευσης που ανήκουν στην κατηγορία c από το σύνολο S.
- Για το παράδειγμα με τα εγκαύματα η εντροπία είναι

$$-\frac{3}{8} \log_2 \left(\frac{3}{8}\right) - \frac{5}{8} \log_2 \left(\frac{5}{8}\right) = 0.954434$$

καθώς 3 από τα 8 ανήκουν στην κατηγορία NAI και 5 από τα 8 στην κατηγορία OXI.

# Κέρδος πληροφορίας

- Το κέρδος πληροφορίας για ένα χαρακτηριστικό A από το σύνολο S ορίζεται ως:

$$IG(A,S) = H(S) - \sum_{t \in T} p(t)H(t)$$

όπου

  - H(S) είναι η εντροπία του συνόλου S, όπως υπολογίστηκε παραπάνω.
  - T είναι τα υποσύνολα του S αν το διαχωρίσουμε με βάση τις διαφορετικές τιμές που λαμβάνει το χαρακτηριστικό A, $S = \cup_{t \in T} t$
  - p(t) το κλάσμα των δεδομένων που βρίσκονται στο t προς τον συνολικό αριθμό δεδομένων που είναι στο S.
  - H(t) είναι η εντροπία του υποσυνόλου t.

# Κέρδος πληφορορίας για εγκαύματα

- Για παράδειγμα ας υπολογίσουμε στο παράδειγμα το κέρδος από το χαρακτηριστικό **Βάρος**
  - IG(Βάρος)=H(S)-p(Βάρος=ελαφρύ)H(Βάρος=ελαφρύ)-p(Βάρος=Μέτριο)H(Βάρος=Μέτριο)-p(Βάρος=Βαρύ)H(Βάρος=Βαρύ) = 0.9544-0.25-0.344361-0.344361=0.015678
- Το κέρδος από το χαρακτηριστικό **Μαλλιά** είναι
  - IG(Μαλλιά)=H(S)-p(Μαλλιά=Ξανθά)H(Μαλλιά=Ξανθά)-p(Μαλλιά=Καστανά)H(Μαλλιά=Καστανά)-p(Μαλλιά=Κόκκινα)H(Μαλλιά=Κόκκινα)=0.9544-0.5-0.0-0.0=0.50
- Επομένως το χαρακτηριστικό **Μαλλιά** είναι πιθανότερο να επιλεγεί από ότι το χαρακτηριστικό **βάρος**.

# Ο αλγόριθμος ID3 σε βήματα

1. Υπολόγισε το κέρδος πληροφορίας κάθε μεταβλητής.
2. Θέσε ως ρίζα την μεταβλητή με το μεγαλύτερο κέρδος.
3. Κάνε τόσα κλαδιά όσες και οι τιμες της μεταβλητής.
4. Διαχωρισμός του dataset σε τόσα υποσύνολα όσες και οι τιμες της μεταβλητής.
5. Επέλεξε ένα υποσύνολο. Αν στο υποσύνολο αντιστοιχεί μια τιμή κατηγορίας μετάβαση στο 6 αλλιώς μετάβαση στο 7.
6. Βάλε την τιμή κατηγορίας ως φύλλο και μετάβαση στο 5.
7. Υπολογισμός κέρδους πληροφορίας για τις υπόλοιπες μεταβλητές για το τρέχον υποσύνολο.
8. Επιλογή μεταβλητής με το μεγαλύτερο κέρδος και δημιουργία κόμβου για αυτήν την μεταβλητή.
9. Μετάβαση στο 3, μέχρι να μην μπορούν να δημιουργηθούν άλλα φύλλα.

# Παράδειγμα χρήσης ID3

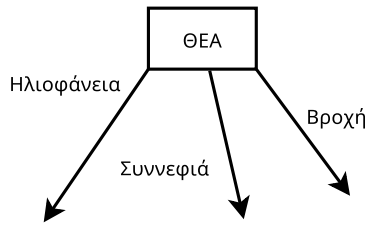
Το σύνολο

S=

| **θεα** | **θερμοκρασια** | **υγρασια** | **αερας** | **κατηγορια** |
|---|---|---|---|---|
| ηλιοφανεια | υψηλη | υψηλη | ασθενης | μεσα |
| ηλιοφανεια | υψηλη | υψηλη | δυνατος | μεσα |
| συννεφια | υψηλη | υψηλη | ασθενης | εξω |
| βροχη | κανονικη | υψηλη | ασθενης | εξω |
| βροχη | χαμηλη | κανονικη | δυνατος | μεσα |
| συννεφια | χαμηλη | κανονικη | ασθενης | εξω |
| βροχη | κανονικη | κανονικη | ασθενης | εξω |
| συννεφια | υψηλη | κανονικη | ασθενης | εξω |

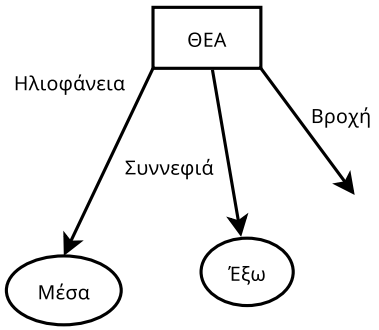
# Υπολογισμός κερδών

1. G(S,Θέα)=0.345
2. G(S,Θερμοκρασία)=0.20
3. G(S,Υγρασία)=0.045
4. G(S,Αέρας)=0.125
5. Δημιουργείται το εξής αρχικό δένδρο



# Δεύτερο βήμα δημιουργίας δένδρου

1. Για τις τιμές Ηλιοφάνεια και Συννεφιά όλα τα πρότυπα ανατίθενται στις κατηγορίες Μέσα και έξω αντίστοιχα, οπότε το νέο δένδρο έχει ως εξής:



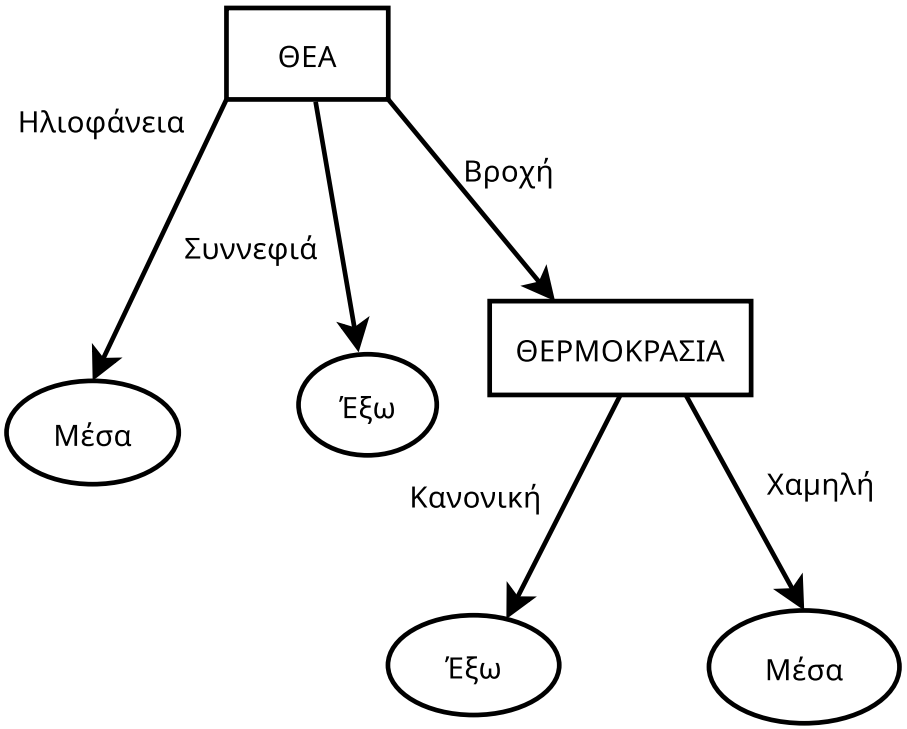

2. Για το νέο υπόδενδρο δημιουργείται το υποσύνολο S(ΒΡΟΧΗ)

3.

| Θέα | Θερμοκρασία | Υγρασία | Αέρας | Κατηγορία |
|---|---|---|---|---|
| Βροχή | Κανονική | Υψηλή | Ασθενής | Έξω |
| Βροχή | Κανονική | Κανονική | Ασθενής | Έξω |
| Βροχή | Χαμηλή | Κανονική | Δυνατός | Μέσα |

# Τρίτο βήμα κατασκευής δένδρου

1. Η κατηγορία με το μεγαλύτερο κέρδος είναι η θερμοκρασια με: G(S(ΒΡΟΧΗ),Θερμοκρασία)=0.92 και επομένως το τελικό δένδρο θα είναι:



# Ο αλγόριθμος CART

1. Gini index: Μετράει την ανισότητα μεταξύ τιμών μιας κατανομής.
2. Gini index=0, πλήρης ισότητα
3. Gini index=1, πλήρης ανισότητα.
4. Υπολογίζεται ως: $\text{gini}(S) = 1 - \sum_{i=1}^{k} p_i^2$ όπου $p_i$ είναι η πιθανότητα εμφάνισης της κατηγορίας $i$ στο σύνολο $S$.
5. Αν το σύνολο $S$ χωριστεί σε $S_1$ και $S_2$ τότε ο δείκτης υπολογίζεται ως

$$gini(S) = \frac{n_1}{n} gini(S_1) + \frac{n_2}{n} gini(S_2)$$

# Παράδειγμα υπολογισμού GINI

1. Gini(Ηλιοφάνεια)=1-(p(Μέσα)^2+p(έξω)^2)=1-(1^2+0)=1-1=0
2. Gini(Συννεφιά)=1-(p(Μέσα)^2+p('Εξω)^2)=1-(0+1^2)=1-1=0
3. Gini(Βροχή)=1-(p(Μέσα)^2+p('Εξω)^2)=1-(0.5^2+0.5^2)=0.5
4. Συνολικά Gini(Θέα)=2/6*Gini(Ηλιοφάνεια)+2/6*gini(Συννεφιά)+2/6*gini(Β
5. Σε κάθε φάση προτιμάται το χαρακτηριστικό με το μικρότερο GINI, καθώς είναι πιθανόν να τερματίσει τον αλγόριθμο κατασκευής ταχύτερα.

# Πλεονεκτήματα δένδρων απόφασης

1. Είναι γρήγορα στην λήψη απόφασης (σε αντίθεση πχ με τα KNN).
2. Συνήθως εξάγουν κανόνες σε μορφή κατανοητή.
3. Μπορούν να χρησιμοποιηθούν σε μεγάλες βάσεις δεδομένων.

# Μειονεκτήματα δένδρων απόφασης

1. Δεν μπορούν να εφαρμοστούν σε συνεχή δεδομένα (εκτός και αν βάλουμε όρια τιμών)
2. Προυποθέτουν ότι υπάρχουν κανόνες ταξινόμησης, αλλά αυτό δεν ισχύει πάντα
3. Μπορεί να προκύψουν δένδρα με μεγάλο βάθος
4. Δεν μπορούν να χειριστούν ελλειπή δεδομένα

# Βιβλιογραφία I



Rosenblatt, Frank (1958), The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Cornell Aeronautical Laboratory, Psychological Review, v65, No. 6, pp. 386–408.


Freund, Y. and Schapire, R. E. 1998. Large margin classification using the perceptron algorithm. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT' 98). ACM Press.