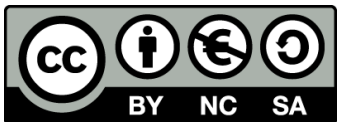
ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 5: Κατηγοριοποίηση – Μέρος Α΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ



Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Κατηγοριοποίηση – Μέρος Α΄

Λειτουργία της κατηγοριοποίησης, μέθοδος δένδρων αποφάσεων

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Εισαγωγικές Έννοιες..
2. Δένδρα Απόφασης.

Σκοποί ενότητας

- Παρουσίαση της λειτουργίας της κατηγοριοποίησης.
- Κατηγοριοποίηση με τη μέθοδο των δένδρων αποφάσεων.

Κατηγοριοποίηση

- Η ανάθεση αντικειμένων σε προκαθορισμένες κλάσεις.
- Ιδιότητες $X_1, X_2, \ldots X_\kappa$.
- Μοντέλο κατηγοριοποίησης:

$$f: D(X_1) \times \ldots \times D(X_k) \rightarrow D(C)$$

- Εκπαίδευση από υπάρχοντα δεδομένα (σύνολο εκμάθησης).

Παράδειγμα

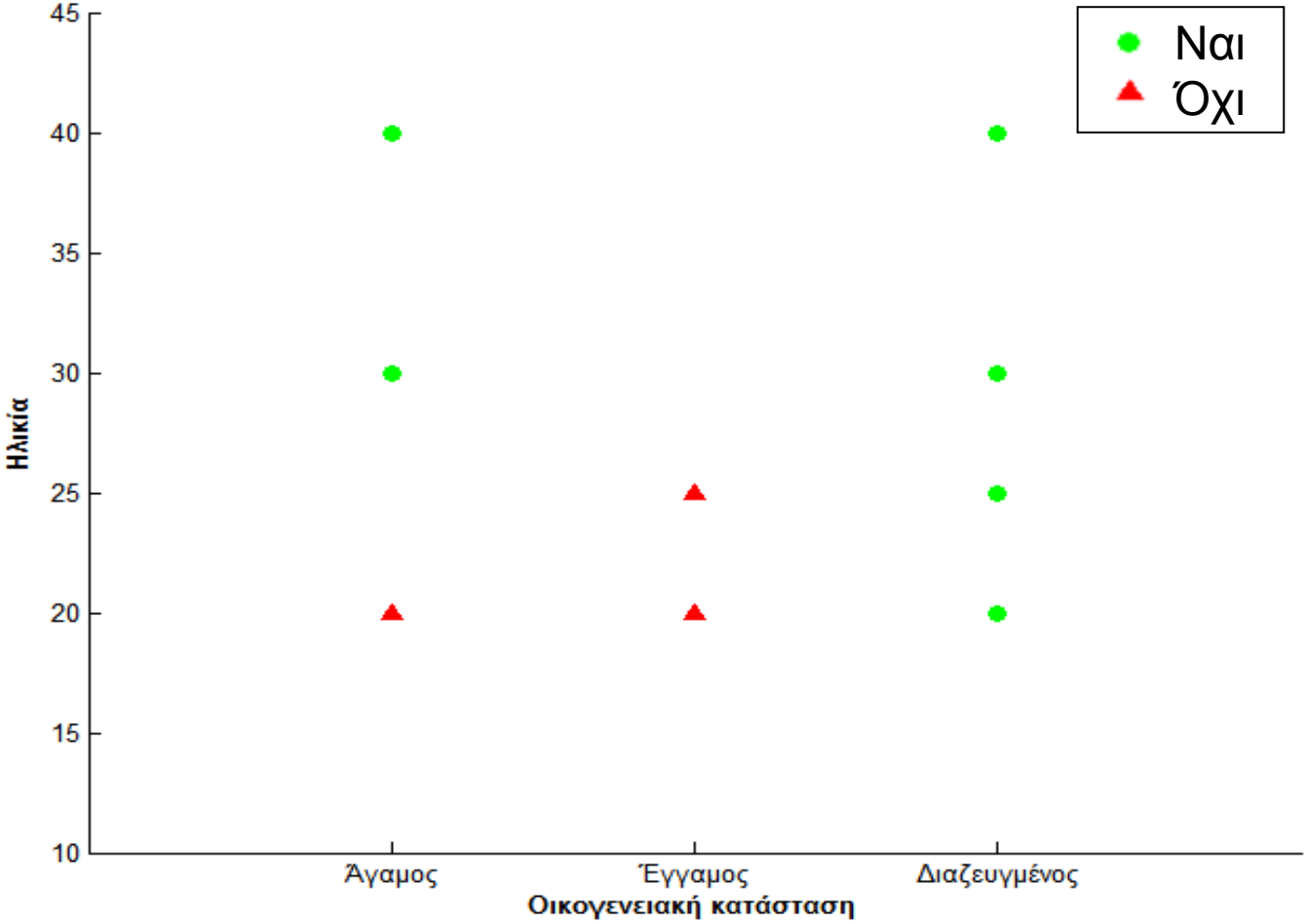
Ηλικία	Οικογενειακή Κατάσταση	Αγοραστής
20	Διαζευγμένος	ΝΑΙ
30	Διαζευγμένος	ΝΑΙ
25	Έγγαμος	ΟΧΙ
30	Άγαμος	ΝΑΙ
40	Άγαμος	ΝΑΙ
20	Έγγαμος	ΟΧΙ
30	Διαζευγμένος	ΝΑΙ
25	Διαζευγμένος	ΝΑΙ
40	Διαζευγμένος	ΝΑΙ
20	Άγαμος	ΟΧΙ

f: [20...40] x {Άγαμος, Έγγαμος, Διαζευγμένος} → {ΝΑΙ, ΟΧΙ}

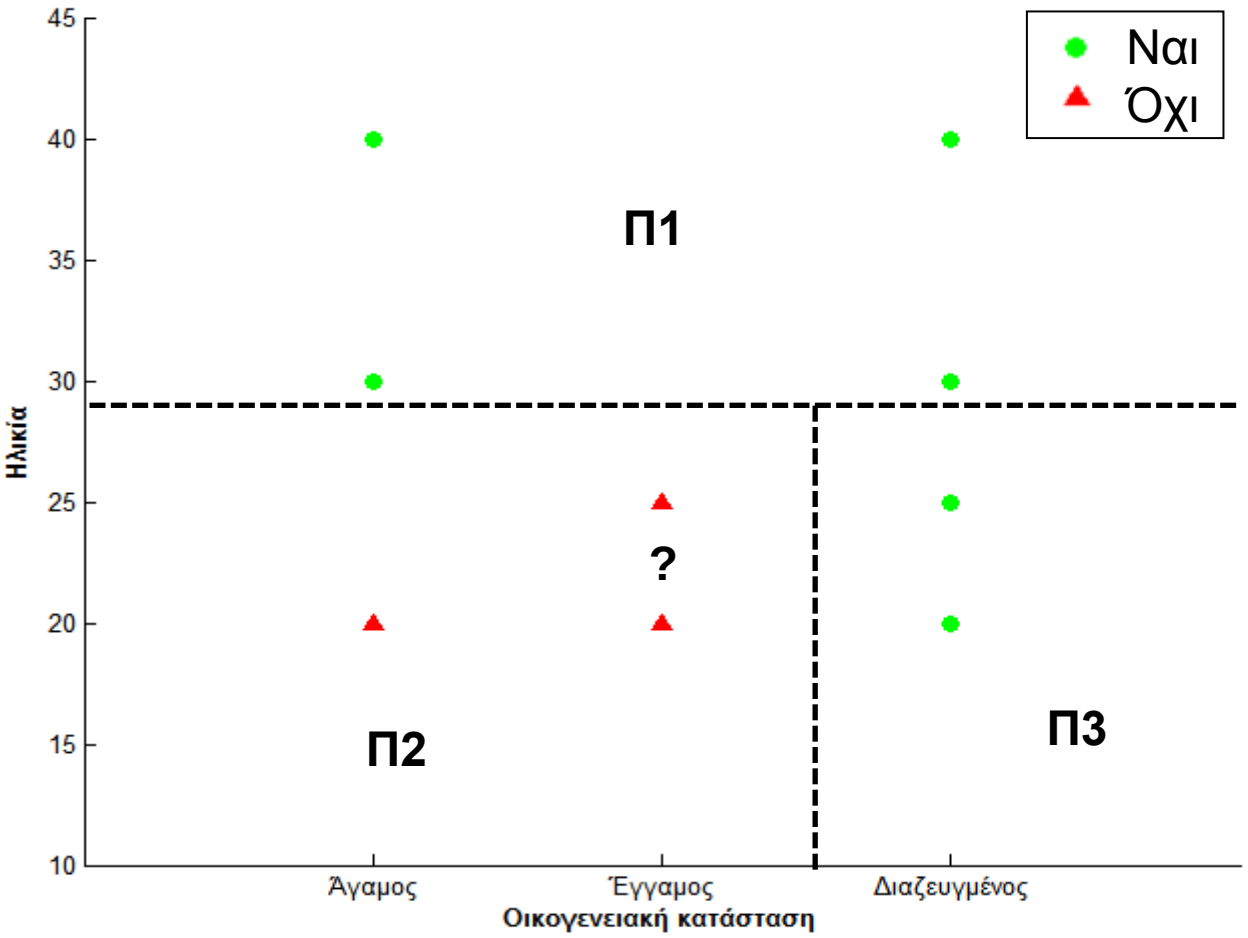
Κατηγοριοποιητής

- Αλγόριθμος κατασκευής μοντέλου:
  - Διαμερισμός σε περιοχές.
  - Εξέταση κατανομών πιθανότητας.
  - Εξέταση πλησιέστερων αντικειμένων.

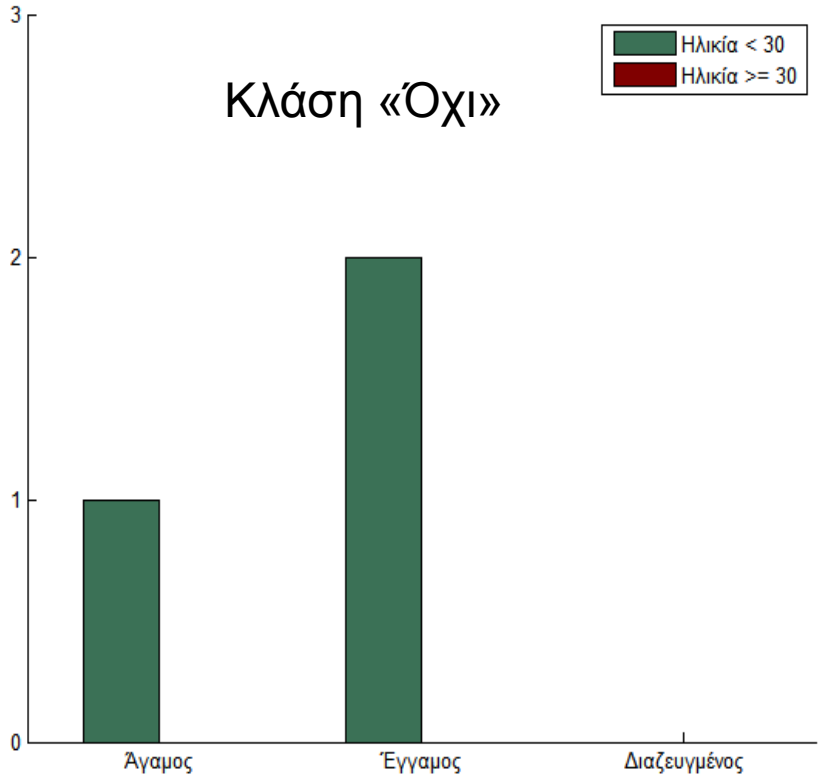
Παράδειγμα

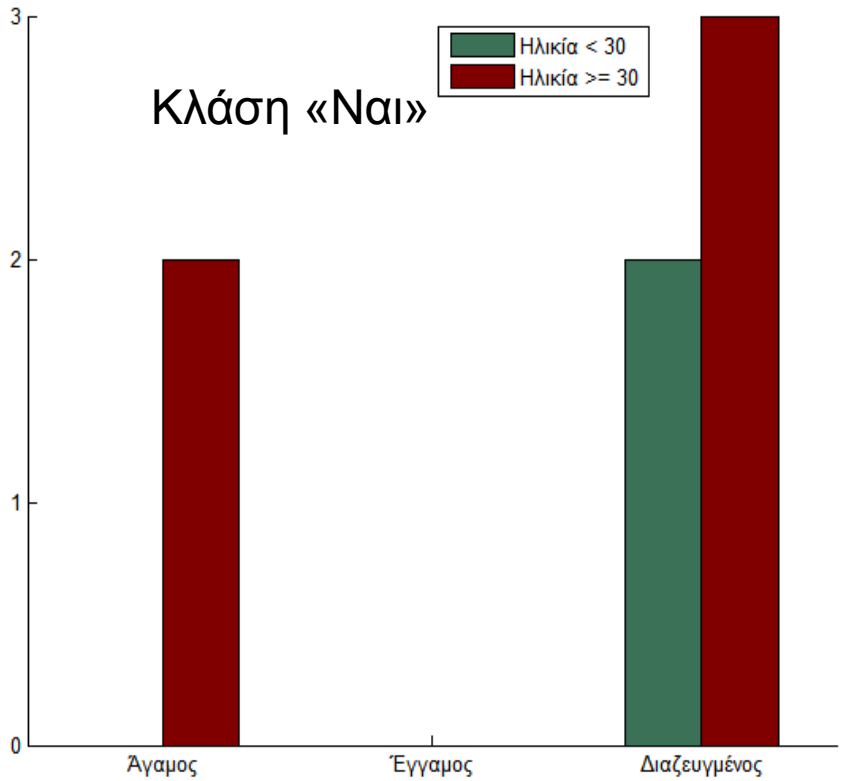

45
40
35
30
25
20
15
10
Ηλικία
Ναι
Όχι
Άγαμος
Έγγαμος
Διαζευγμένος
Οικογενειακή κατάσταση

Διαμερισμός σε περιοχές

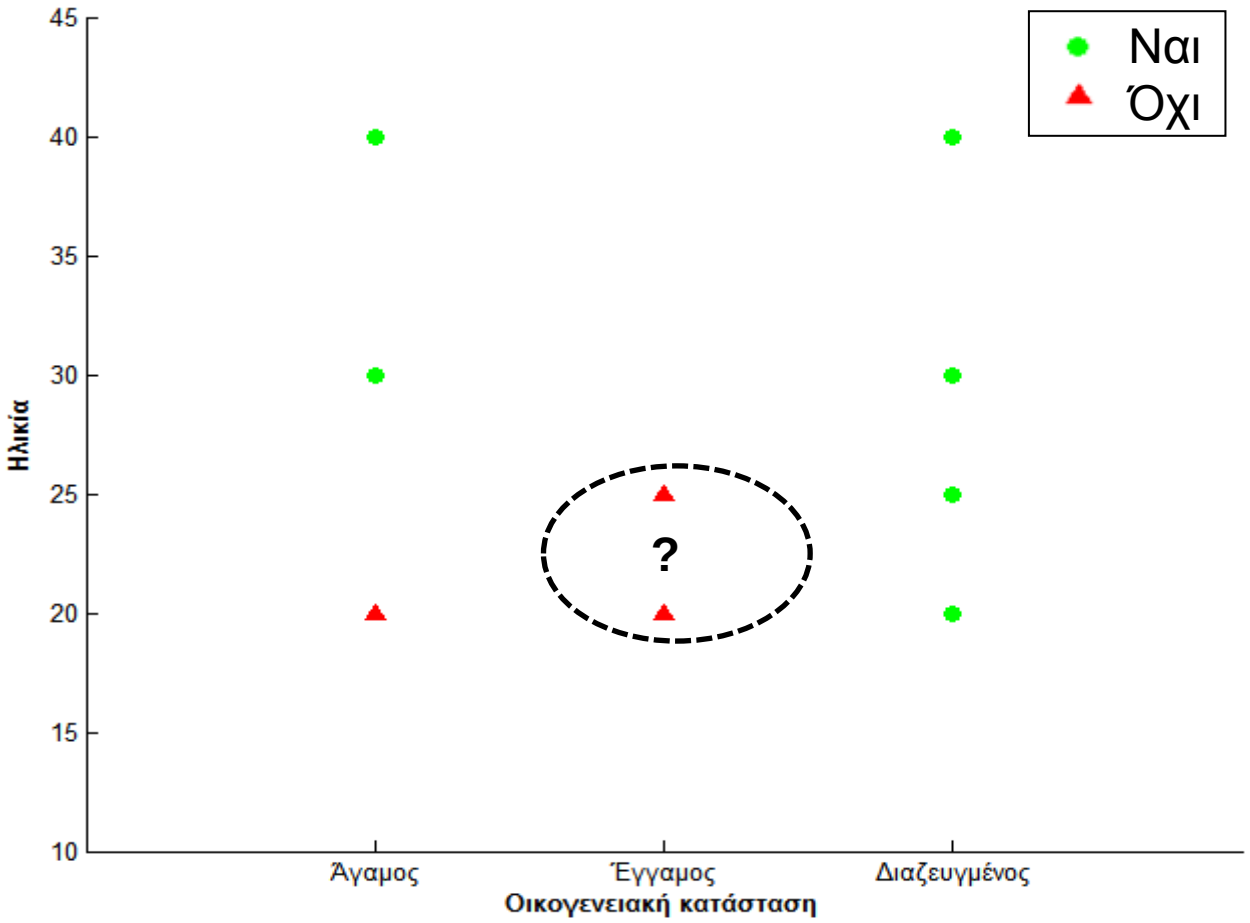


Εξέταση κατανομών πιθανότητας





Εξέταση πλησιέστερων αντικειμένων



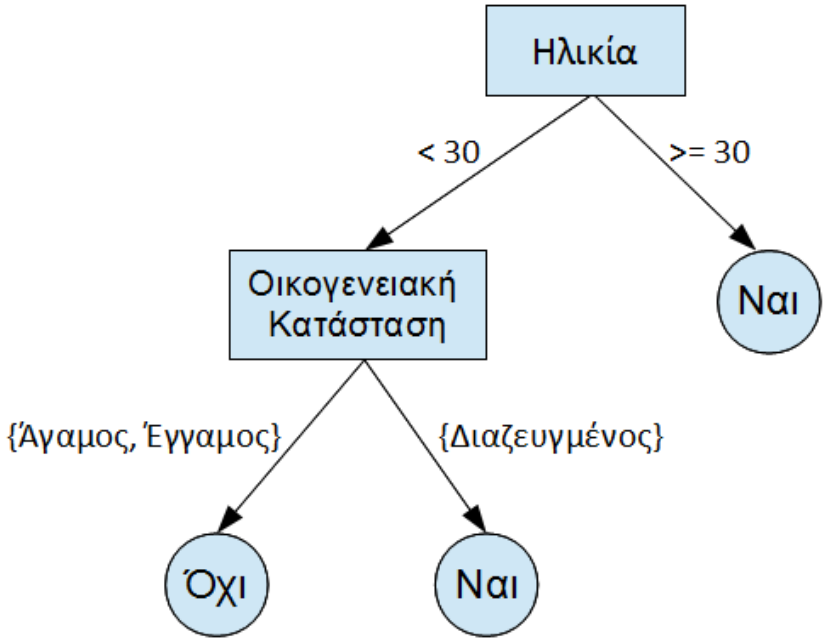
Κριτήρια Αξιολόγησης Κατηγοριοποιητών

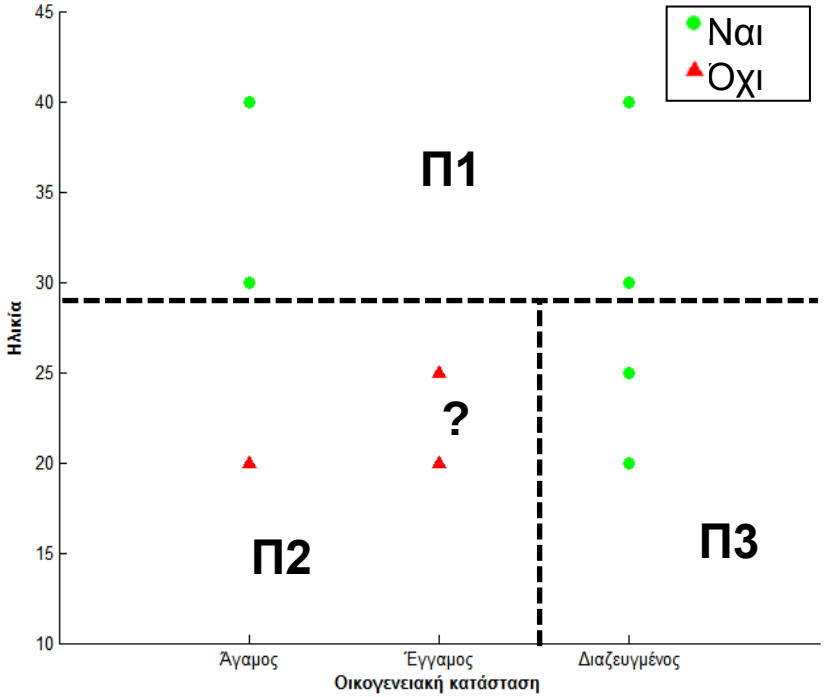
- Ακρίβεια πρόβλεψης του μοντέλου:

$$ακρίβεια = \frac{\#σωστές - προβλέψεις}{\#προβλέψεις}$$

- Ευκολία στην κατανόηση του μοντέλου.
- Κλιμάκωση στο μέγεθος του συνόλου εκμάθησης.
- Ανοχή στο θόρυβο και στις ελλιπείς τιμές.

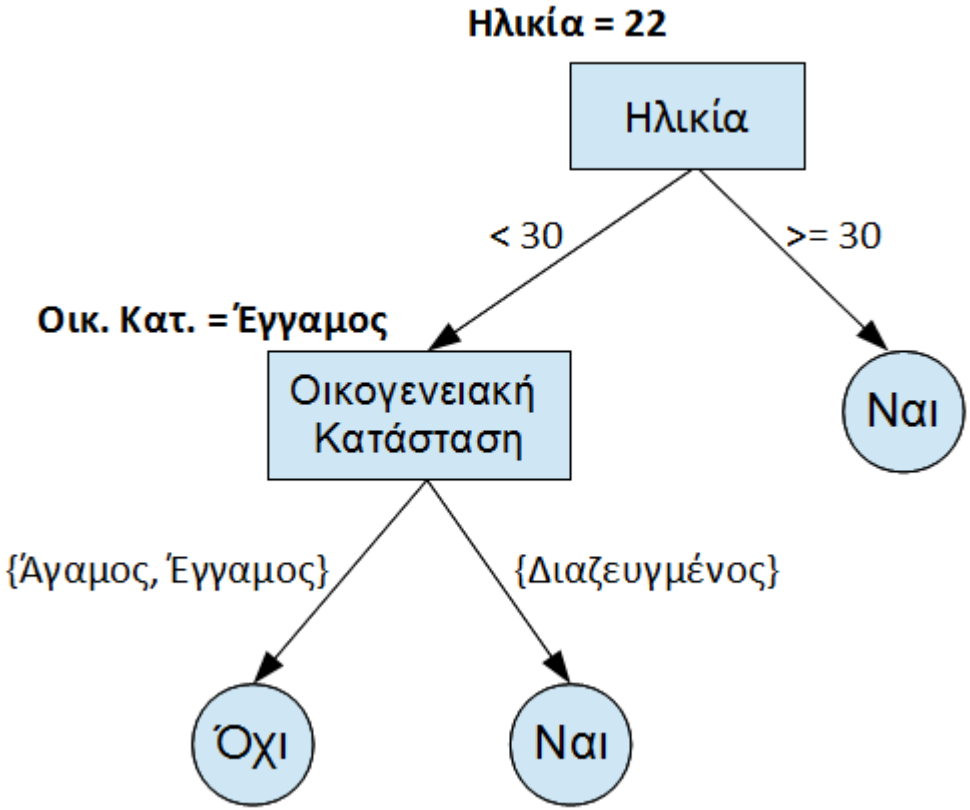
Δένδρα απόφασης





Κατηγοριοποίηση με δένδρο

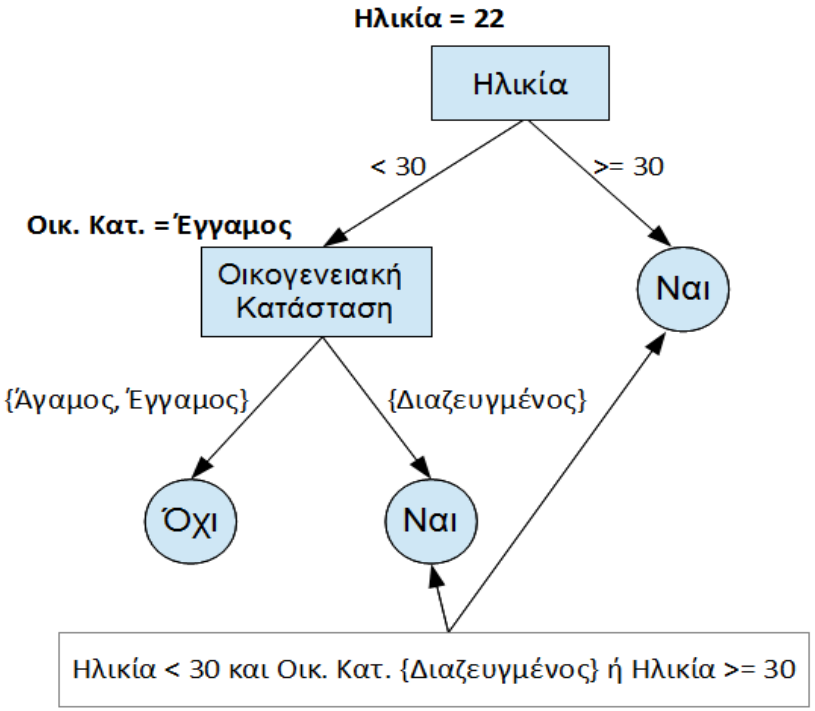
- {Έγγαμος, 22 ετών}. Πιθανός αγοραστής;



Εξαγωγή κανόνων

- Μονοπάτι = συζεύξεις
- Κλάση = διάζευξη συζεύξεων

1. Ηλικία < 30 και Οικ. Κατ. {Α, Ε}
2. Ηλικία < 30 και Οικ. Κατ. {Δ}
3. Ηλικία >= 30



Κατασκευή δένδρου απόφασης (brute-force)

- Κατασκευή κάθε δυνατού πιθανού δένδρου.
- Επιλογή του ακριβέστερου.
- NP-complete.

Κατασκευή δένδρου απόφασης (greedy)

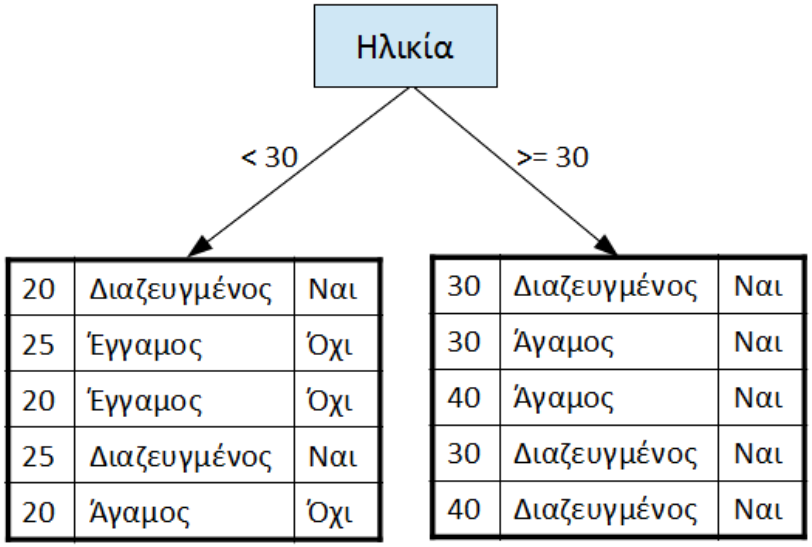
Κατασκευή (κόμβος N, σύνολο D)

1. **If** όλα τα αντικείμενα στο D ανήκουν στην ίδια κλάση C.
2. Κάνε τον *N* φύλλο και ανάθεσέ του την κλάση C.
3. **Else**
4. Επέλεξε μία διαχωριστική ιδιότητα για τον κόμβο *N*.
5. Καθόρισε *k* συνθήκες ελέγχου για τη διαχωριστική ιδιότητα.
6. Δημιούργησε *k* κόμβους, n_1, ..., n_k, και θέσε τους ως παιδιά του *N*.
7. Διαμοίρασε με τη συνθήκη ελέγχου το D σε *k* ομάδες d_1, ..., d_k.
8. **For** $i = 1$ έως k
9. **Call** *Κατασκευή*(n_i, d_i)
10. **End**

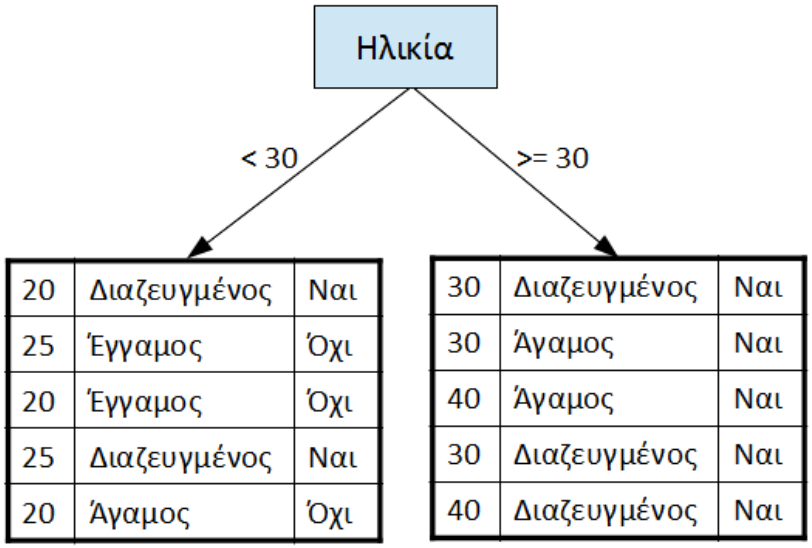
Παράδειγμα (1/3)

20	Διαζευγμένος	Ναι
30	Διαζευγμένος	Ναι
25	Έγγαμος	Όχι
30	Άγαμος	Ναι
40	Άγαμος	Ναι
20	Έγγαμος	Όχι
30	Διαζευγμένος	Ναι
25	Διαζευγμένος	Ναι
40	Διαζευγμένος	Ναι
20	Άγαμος	Όχι

Ø

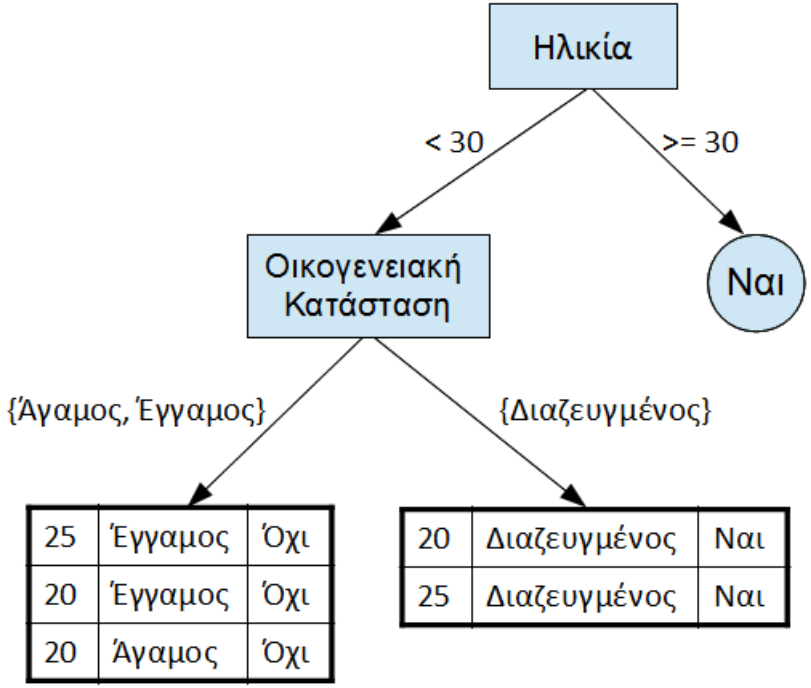


Παράδειγμα (2/3)



20	Διαζευγμένος	Ναι
25	Έγγαμος	Όχι
20	Έγγαμος	Όχι
25	Διαζευγμένος	Ναι
20	Άγαμος	Όχι

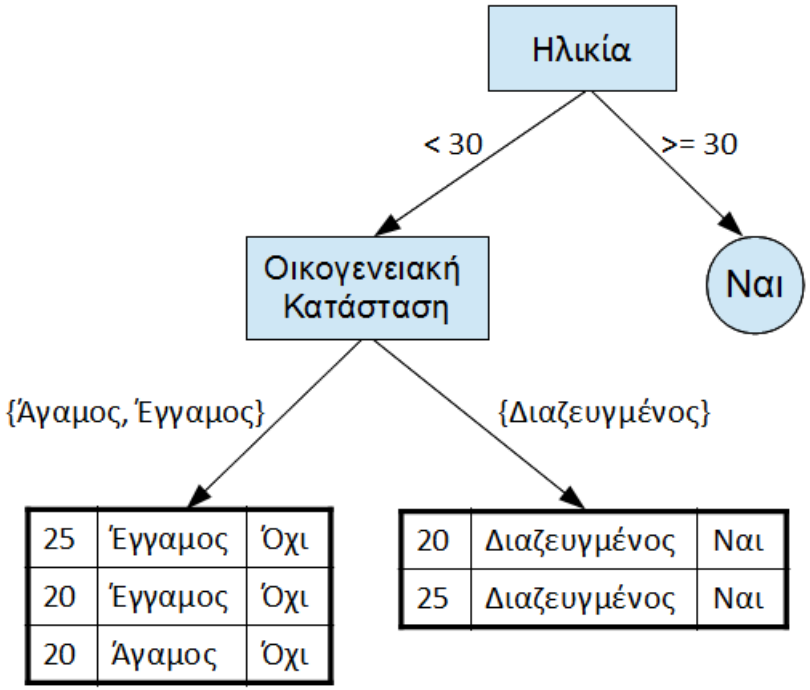
30	Διαζευγμένος	Ναι
30	Άγαμος	Ναι
40	Άγαμος	Ναι
30	Διαζευγμένος	Ναι
40	Διαζευγμένος	Ναι

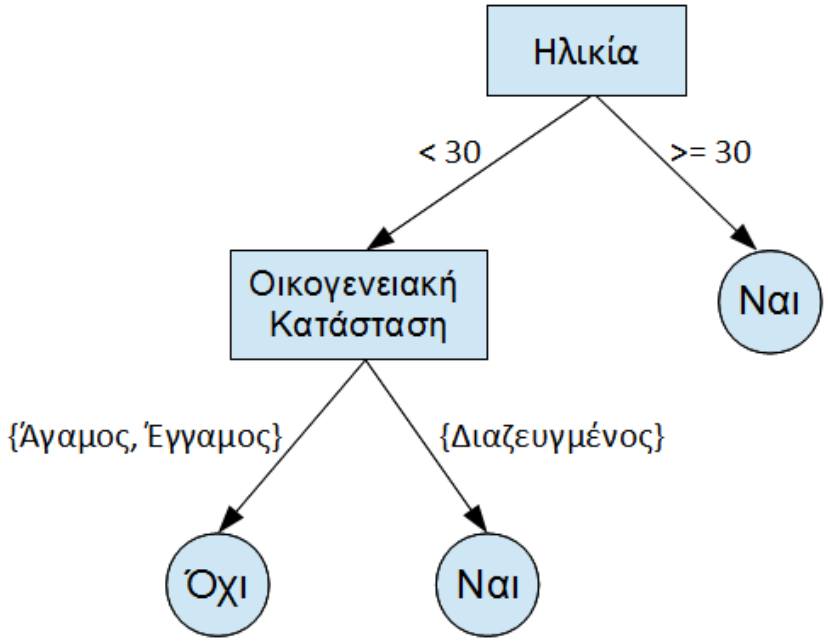


25	Έγγαμος	Όχι
20	Έγγαμος	Όχι
20	Άγαμος	Όχι

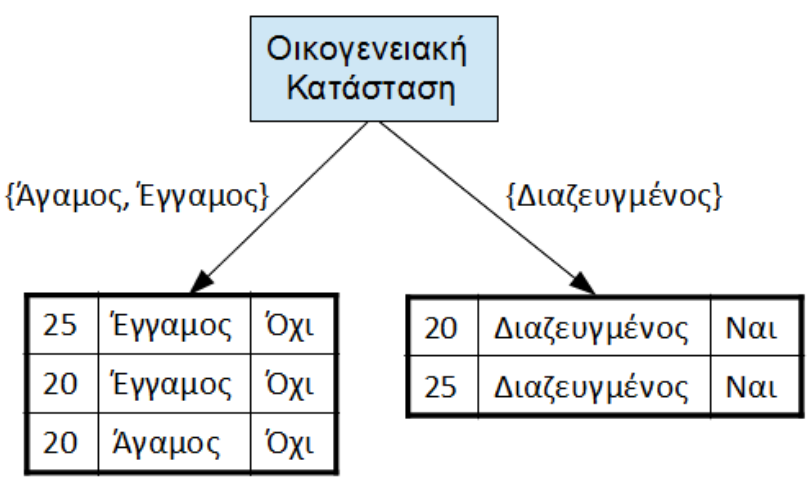
20	Διαζευγμένος	Ναι
25	Διαζευγμένος	Ναι

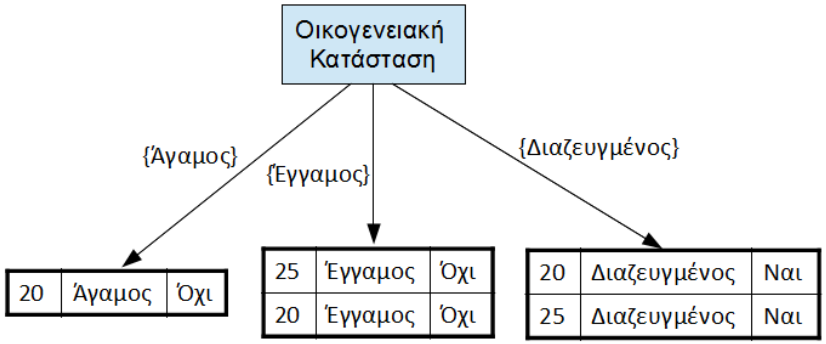
Παράδειγμα (3/3)





Πόσα παιδιά;





- 2 ή 3 (όσες οι διαφορετικές Οικ. Καταστάσεις);

Πλεονεκτήματα - Μειονεκτήματα

Πολλά παιδιά (=όλα)

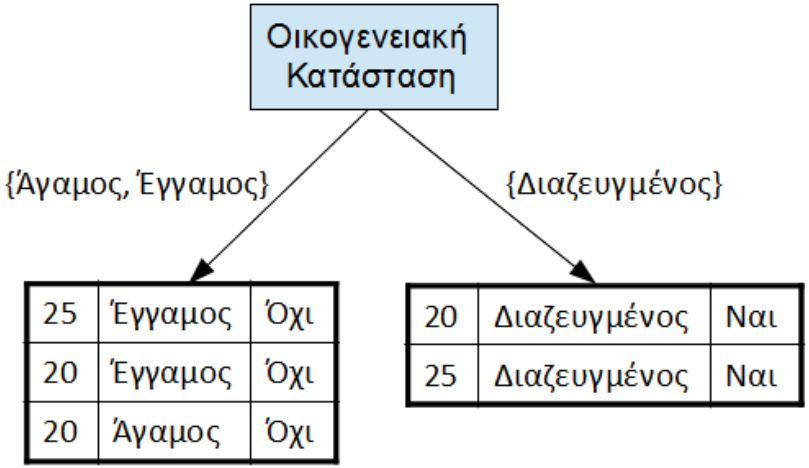
+ εύκολη διάσπαση
- περίπλοκο δένδρο
- αριθμητικές ιδιότητες? (μόνο με κβάντωση)

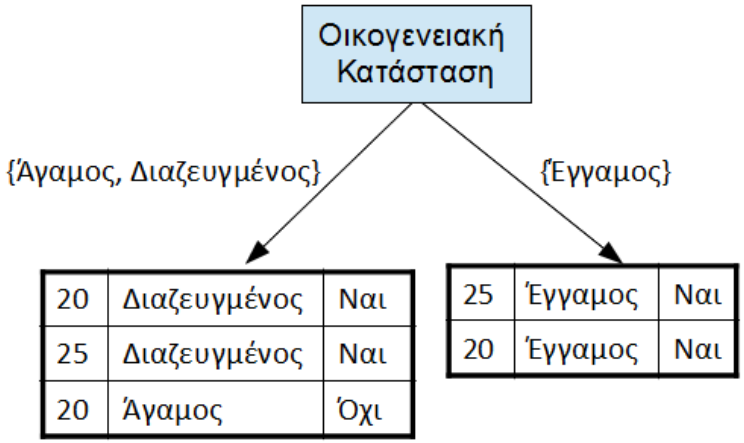
2 παιδιά

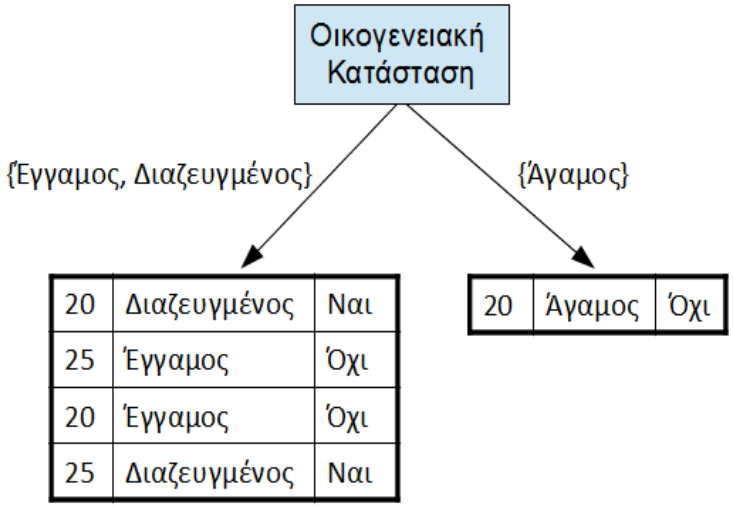
- δύσκολη διάσπαση
+ από/ευνόητο δένδρο
+ χειρισμός αριθμητικών ιδιοτήτων
+ καλύτερη ακρίβεια!

- Επιλογή: Δυαδικά Δένδρα (2 παιδιά)

Διάσπαση σε δυαδικά δένδρα







δυνατές επιλογές: $2^{n-1} - 1$

Παράδειγμα

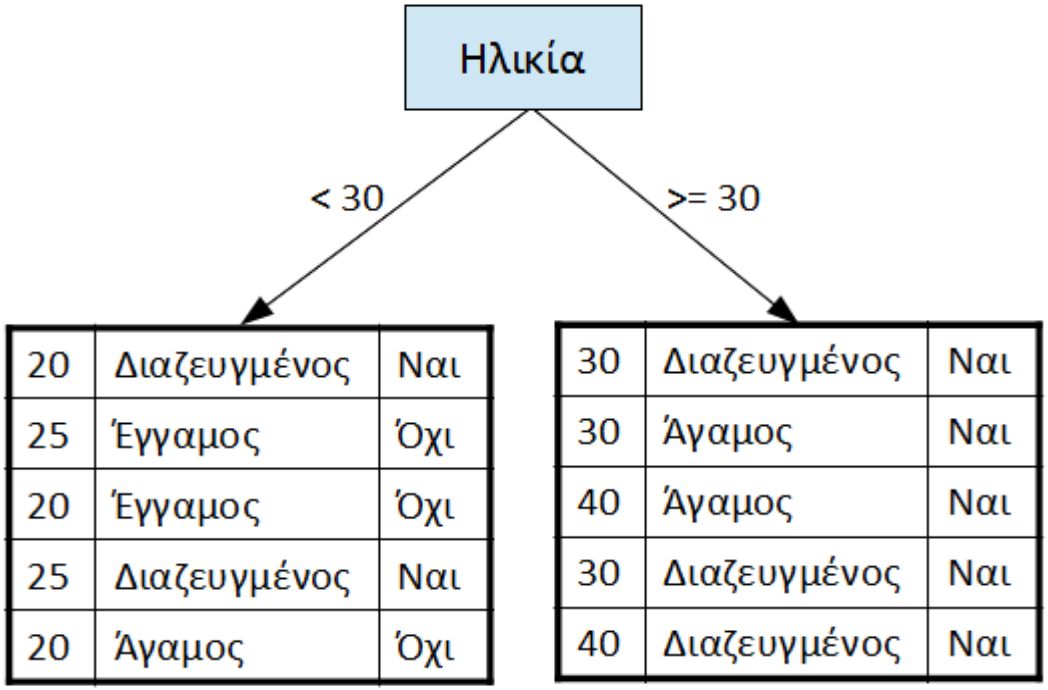
- a
- b
- c
- a, b
- a, c
- a, b, c
- -

- b, c
- a, c
- a, b
- c
- b
- a
- -
- a, b, c

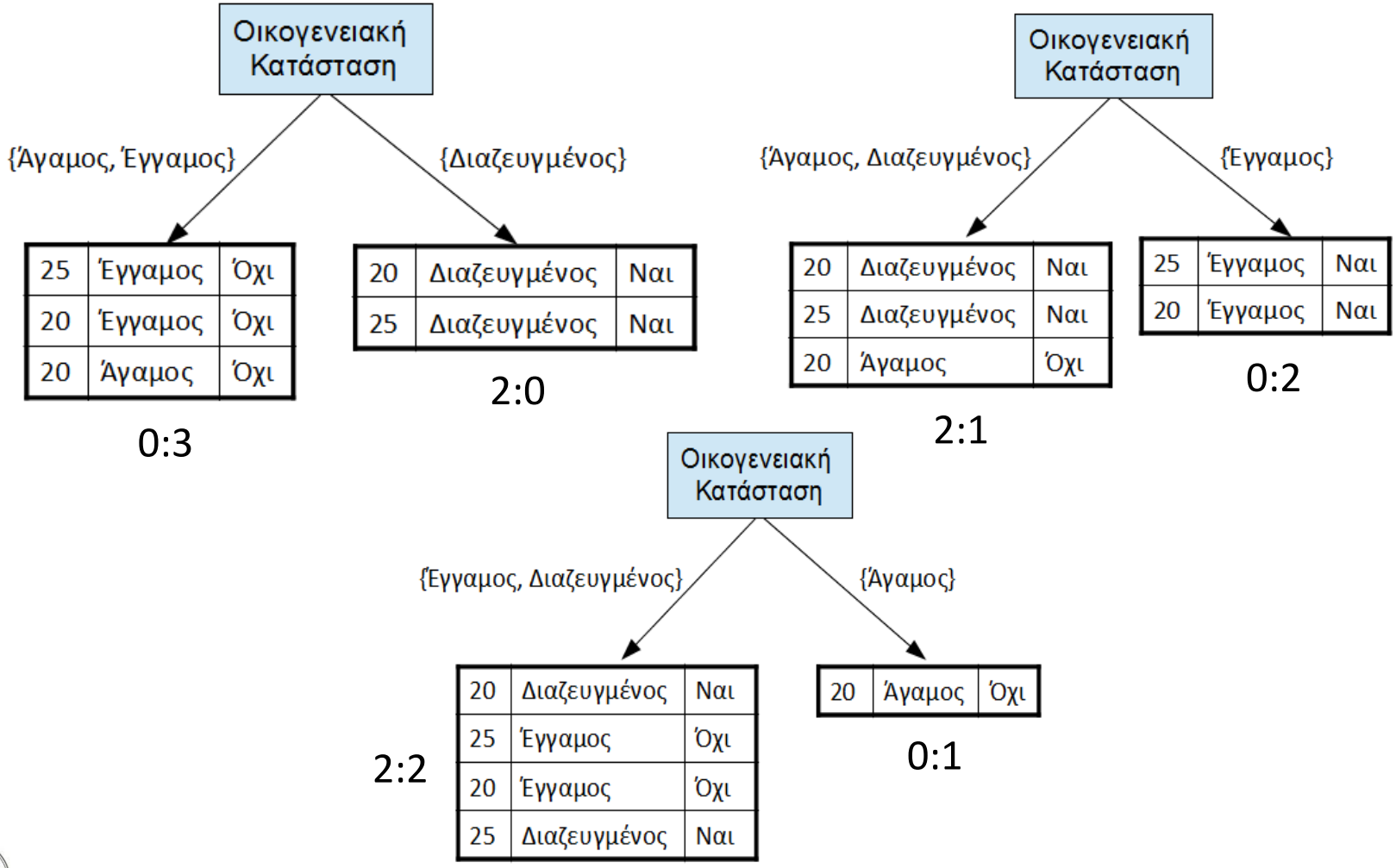
$(2^n-2) / 2 = 2^{n-1}-1$

Διάσπαση αριθμητικών ιδιοτήτων

- Ταξινόμηση.
- Εξέταση n-1 διαχωριστικών θέσεων.



Πως επιλέγουμε τη διαχωριστική ιδιότητα



Μέτρο ανομοιογένειας

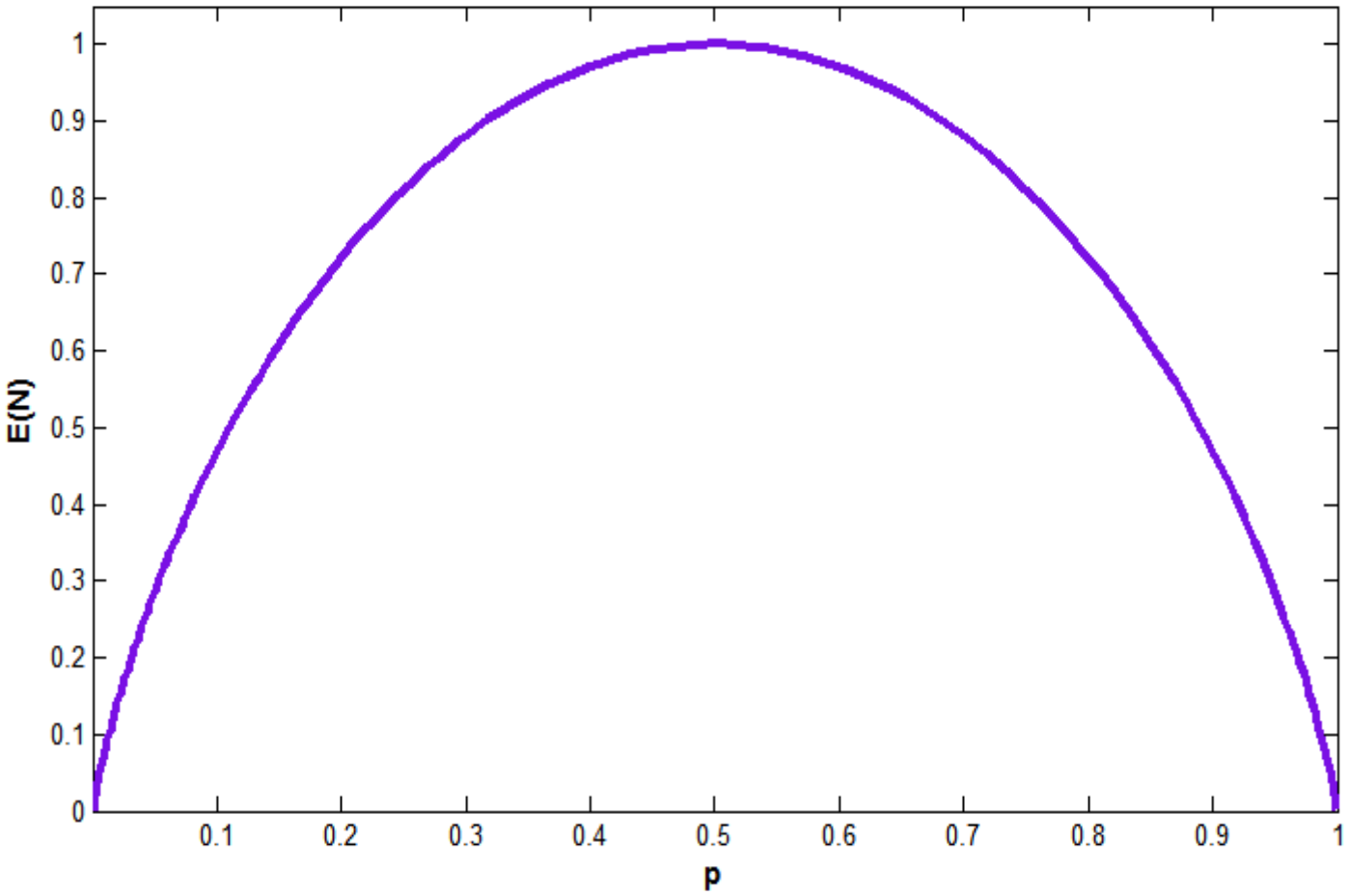
- Εντροπία για c κλάσεις:

( ) ,

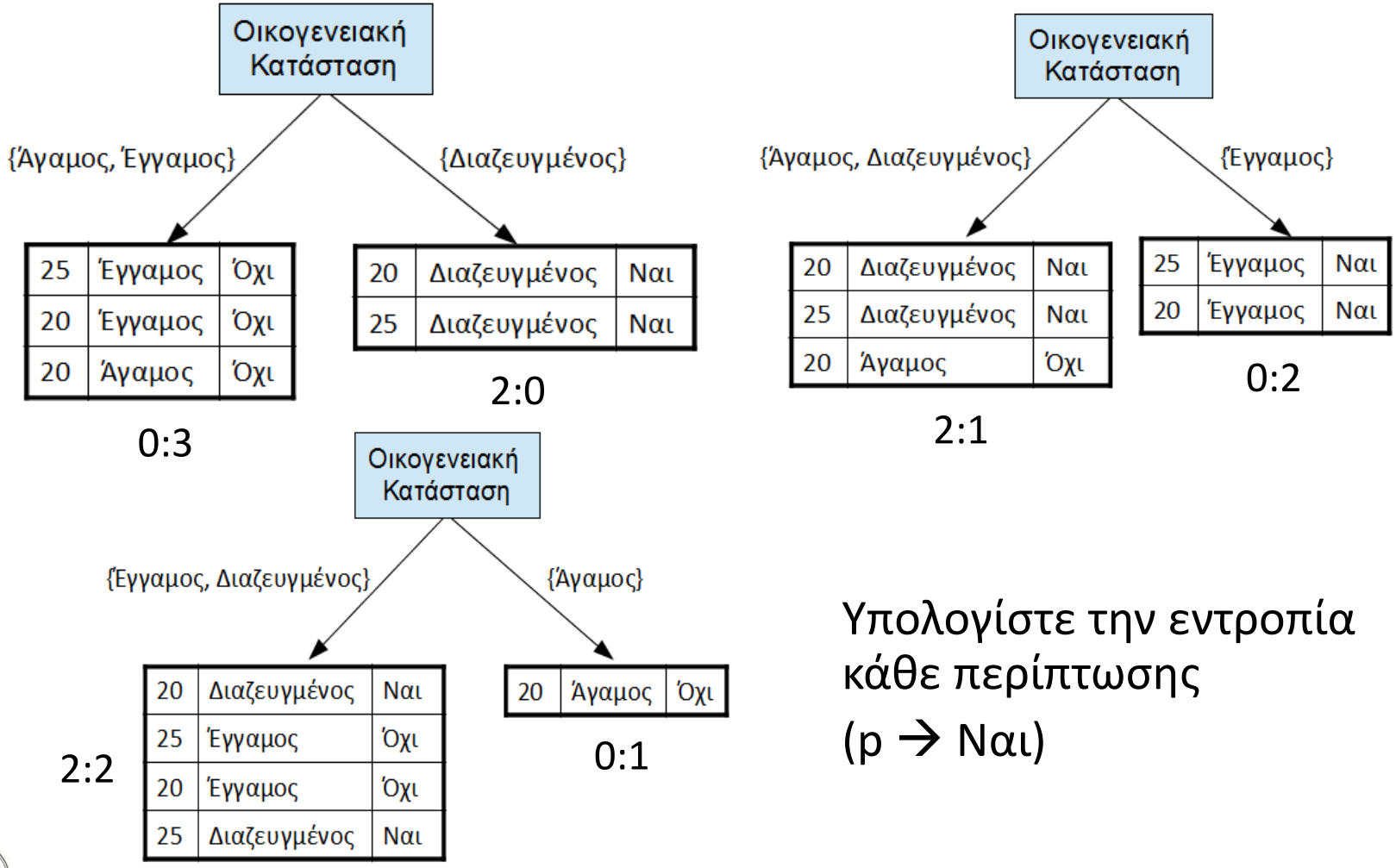
- Για c = 2:

( ) ( ) ( )

Εντροπία για c = 2



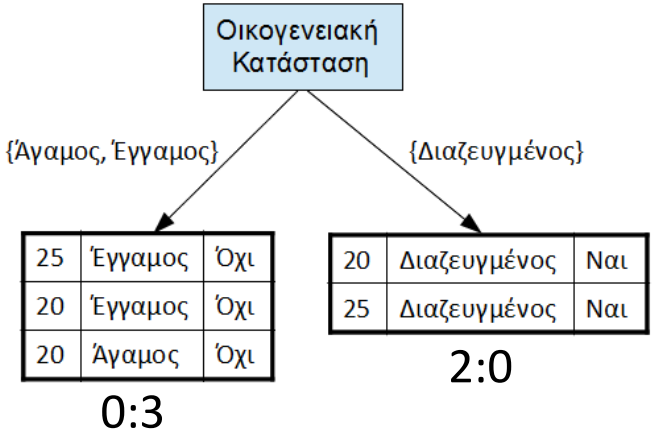
Παράδειγμα (1/2)

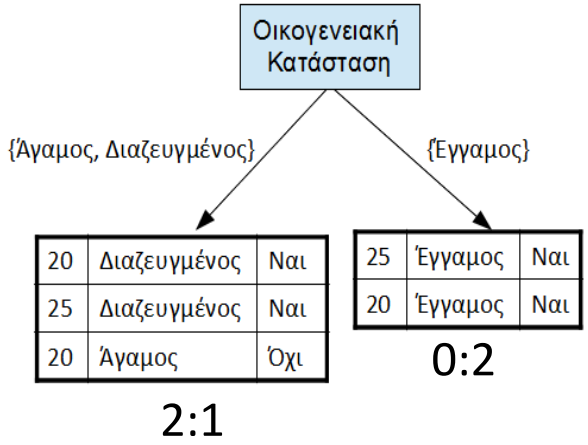


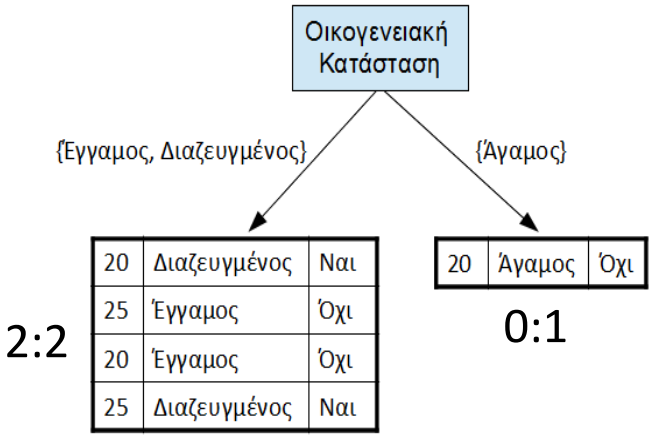
Υπολογίστε την εντροπία κάθε περίπτωσης

(p → Ναι)

Παράδειγμα (2/2)







1. A: -0log0-1log1=0
   Δ: -1log1-0log0=0

2. A: -2/3 log (2/3) - 1/3 log (1/3) = 0.9183
   Δ: -0log0 -1log1=0

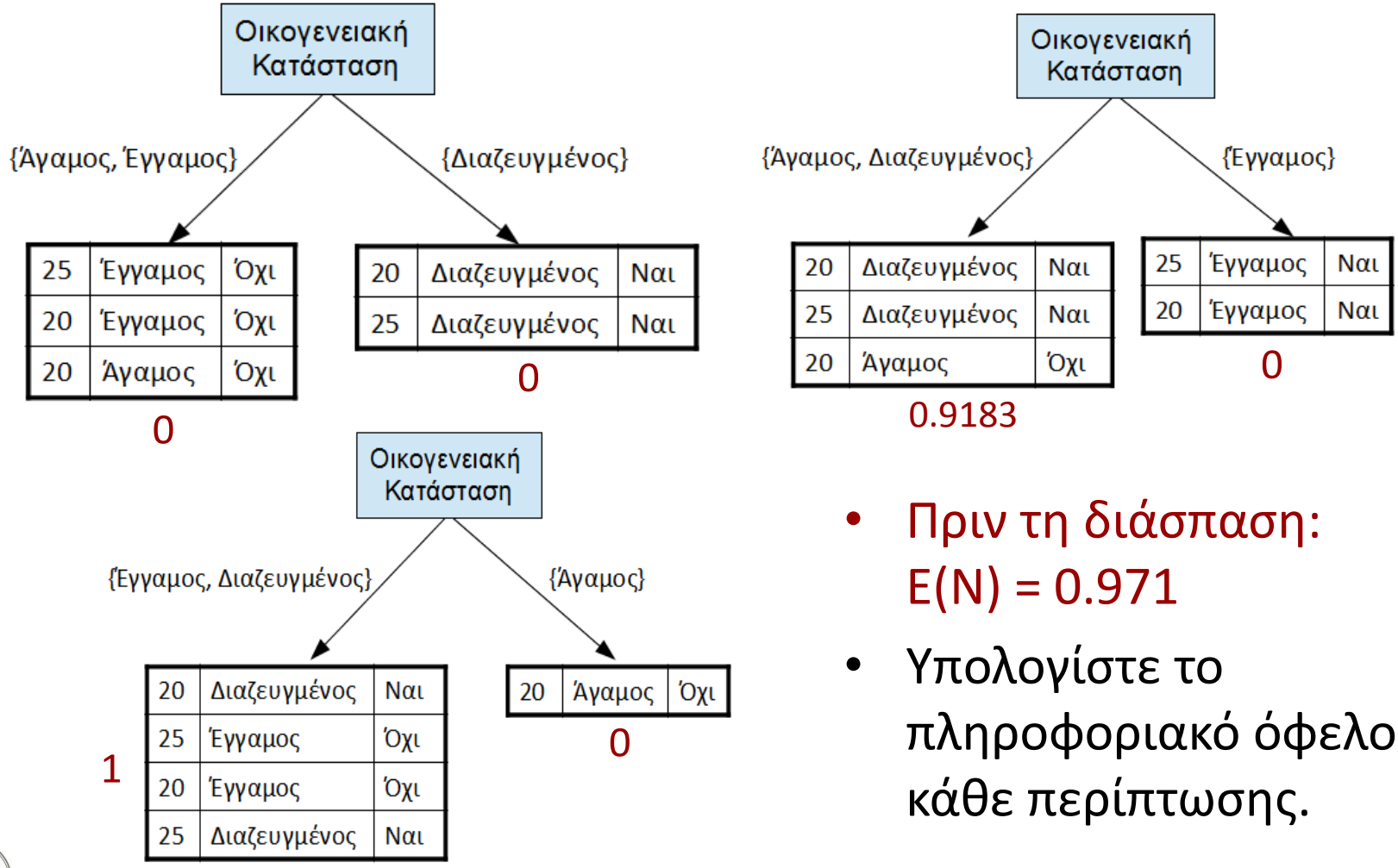
3. A: - 2/4 log (2/4) - 2/4 log (2/4) = 1
   Δ: -0log0 -1log1=0

Αξιολόγηση διαχωριστικής ιδιότητας

- Πληροφοριακό όφελος για μία ιδιότητα:

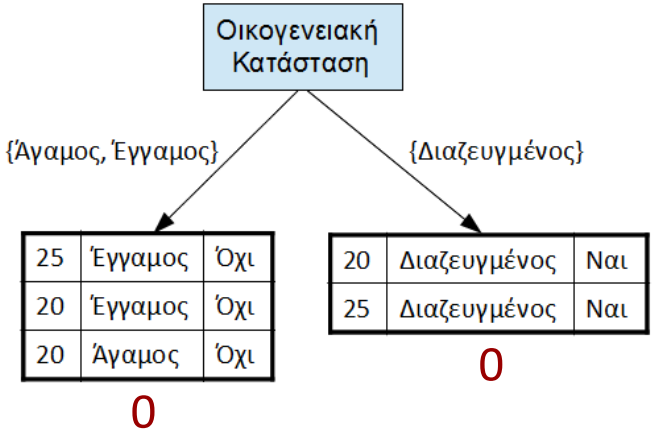
$$(\quad) \quad - \quad (\quad) \quad -$$

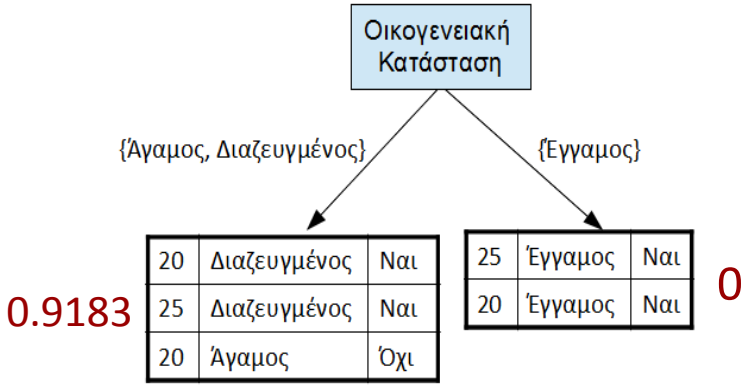
Παράδειγμα πληροφοριακού οφέλους (1/2)

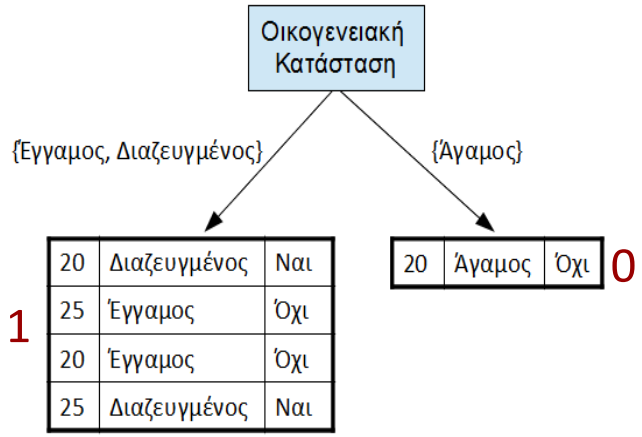


- Πριν τη διάσπαση: E(N) = 0.971
- Υπολογίστε το πληροφοριακό όφελος κάθε περίπτωσης.

Παράδειγμα πληροφοριακού οφέλους (2/2)







Πριν τη διάσπαση: E(N) = 0.971

1. 0.971 – 3/5 *0 -2/5*0 = 0.971

2. 0.971 – 3/5*0.9183 – 2/5 * 0 = 0.42

3. 0.971 – 4/5 *1 – 1/5 * 0 = 0.171

Επιλογή διαχωριστικής ιδιότητας

- Εξετάζουμε για κάθε ιδιότητα όλους τους δυνατούς διαχωρισμούς.
  - επιλέγουμε το διαχωρισμό με το μέγιστο όφελος.
- Επιλέγουμε την ιδιότητα με το μεγαλύτερο πληροφοριακό όφελος.
  - εφαρμόζουμε το διαχωρισμό μεγίστου οφέλους.

Κατασκευή δένδρου απόφασης (greedy)

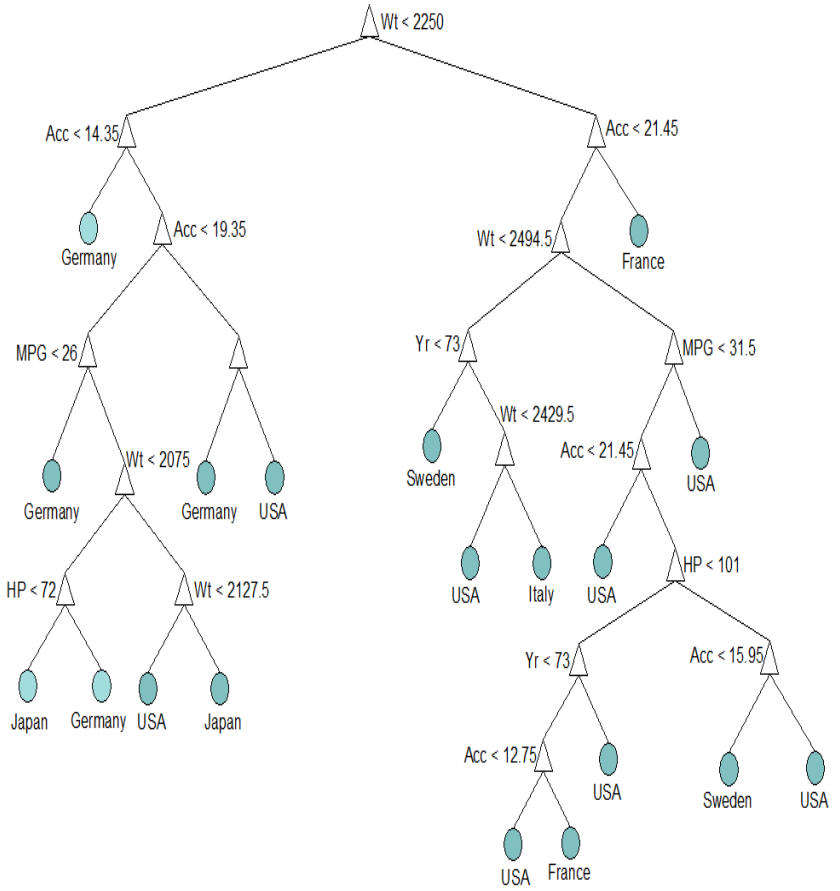
Κατασκευή (κόμβος N, σύνολο D) **ΜΗ ΠΡΑΚΤΙΚΟ ΚΡΙΤΗΡΙΟ**

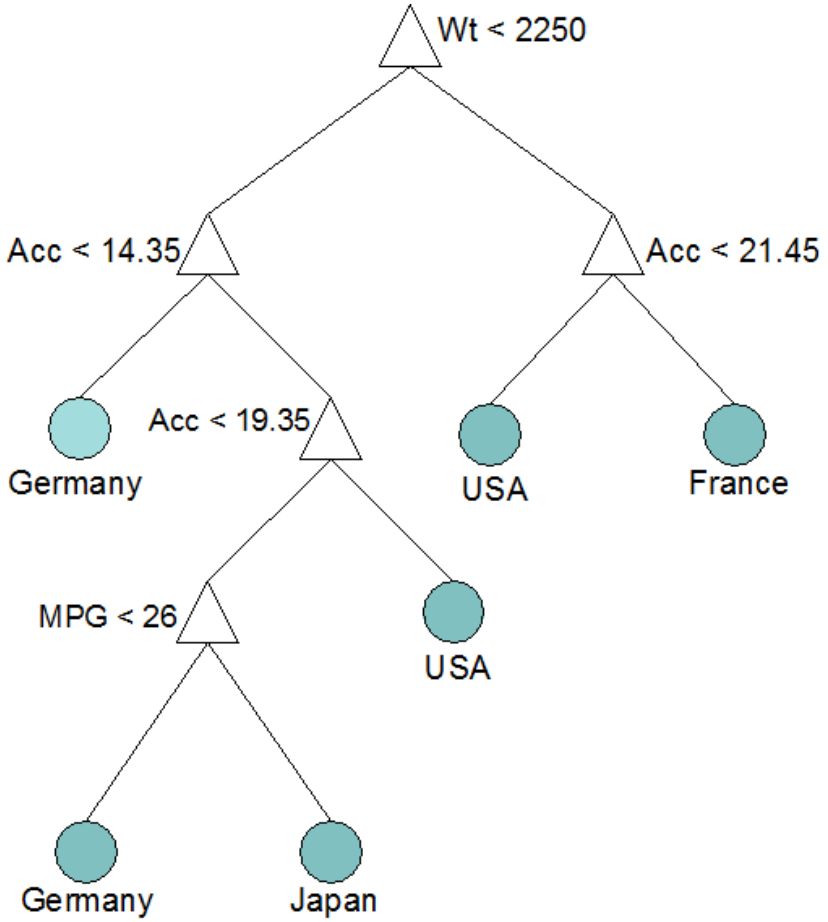
1. **If** όλα τα αντικείμενα στο D ανήκουν στην ίδια κλάση C
2. Κάνε τον N φύλλο και ανάθεσέ του την κλάση C.
3. **Else**
4. Επέλεξε μία διαχωριστική ιδιότητα για τον κόμβο N.
5. Καθόρισε k συνθήκες ελέγχου για τη διαχωριστική ιδιότητα.
6. Δημιούργησε k κόμβους, n_1, ..., n_k, και θέσε τους ως παιδιά του N.
7. Διαμοίρασε με τη συνθήκη ελέγχου το D σε k ομάδες d_1, ..., d_k.
8. **For** $i = 1$ έως k
9. **Call** *Κατασκευή*(n_i, d_i)
10. **End**

Εναλλακτικά κριτήρια τερματισμού

- Ένα ποσοστό (π.χ., 80%) ανήκουν στην ίδια κλάση.
- Αν ο κόμβος περιέχει λιγότερα από, π.χ., 10, αντικείμενα:
  - η κλάση του φύλλου είναι η πλειοψηφούσα
- Μπορούμε να έχουμε και τα δύο κριτήρια.

Αποτέλεσμα εναλλακτικών κριτηρίων





Χαρακτηριστικά Δένδρων Απόφασης (1/2)

- Η κατασκευή του βέλτιστου δένδρου απόφασης απαιτεί αποτρεπτικό χρόνο (είναι NP-complete πρόβλημα).
  - Για το λόγο αυτό χρησιμοποιούνται ευρετικοί αλγόριθμοι, οι οποίοι είναι άπληστοι και δεν χρησιμοποιούν οπισθοδρόμηση.
  - Τα ευρετικά μειώνουν κατά πολύ το χρόνο κατασκευής.
  - Το αποτέλεσμα είναι ότι τα δένδρα απόφασης **κλιμακώνονται σε μεγάλους όγκους δεδομένων.**
- **Γρήγορη εφαρμογή.**

Χαρακτηριστικά Δένδρων Απόφασης (2/2)

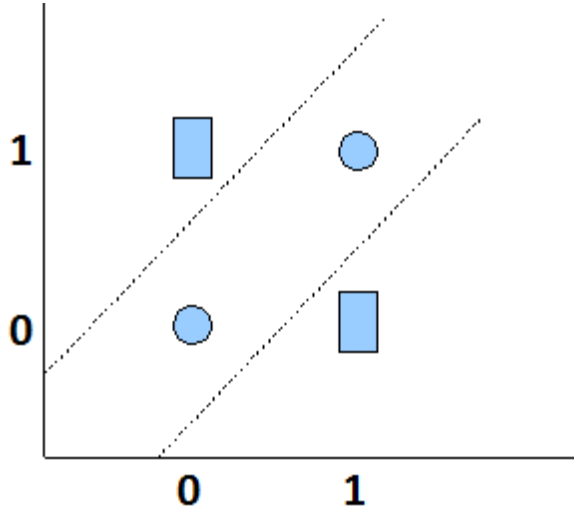
- Η **ακρίβεια πρόβλεψης** των δένδρων απόφασης είναι αποδεκτή για τις περισσότερες περιπτώσεις, συγκρίσιμη με την ακρίβεια άλλων κατηγοριοποιητών.
- Το μοντέλο που προκύπτει είναι πολύ **εύκολο στην κατανόηση**.
- Τα δένδρα απόφασης έχουν καλή **ανοχή στο θόρυβο:**
  - ειδικά όταν εφαρμόζεται ψαλιδισμός.

Επιπλέον

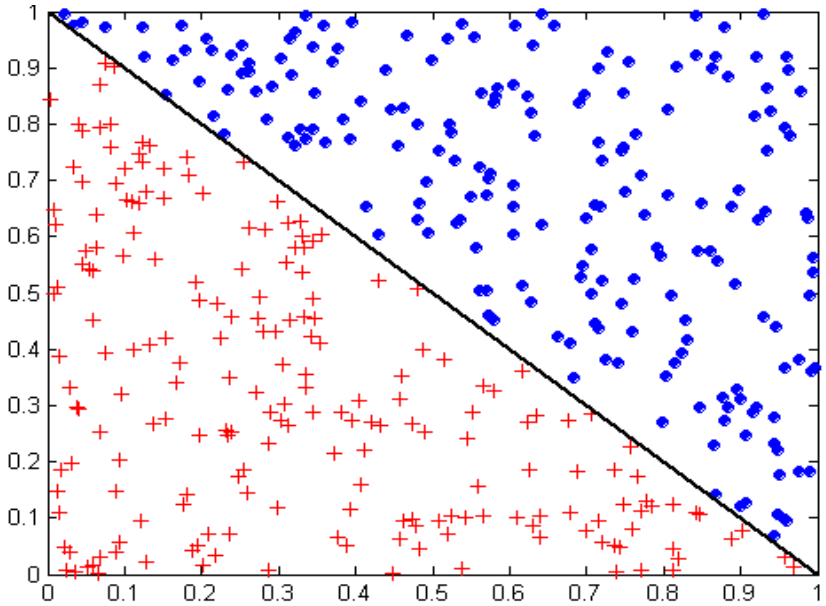
- Τα ΔΑ μπορούν να διαχειριστούν πολυδιάστατα δεδομένα.
  - 1 διάσταση τη φορά χρησιμοποιείται κατά την ανάπτυξη του μοντέλου.
- … και κάθε τύπο μεταβλητών.
  - Συμβολικές, αριθμητικές, κλπ.

Μειονεκτήματα

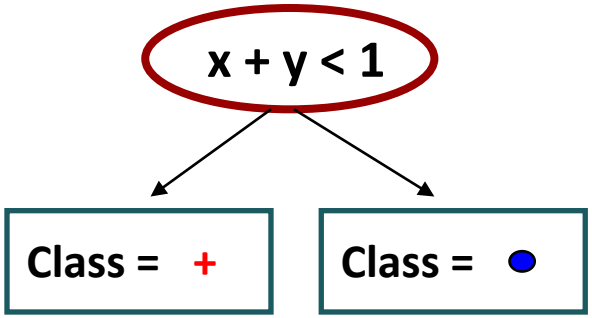
- Αγνοούν εξαρτήσεις μεταξύ των ιδιοτήτων.
- Προβλήματα όταν λείπουν πολλά δεδομένα.
- Διάσπαση ως προς μία ιδιότητα:
  - αντιστοίχιση με περιοχές, τα όρια των οποίων είναι παράλληλα με τους άξονες.



Πλάγια Δένδρα Απόφασης

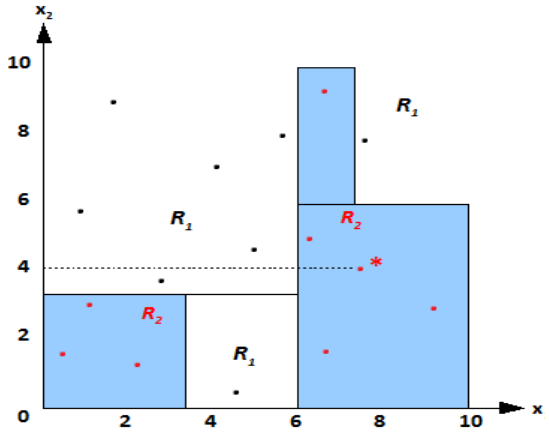


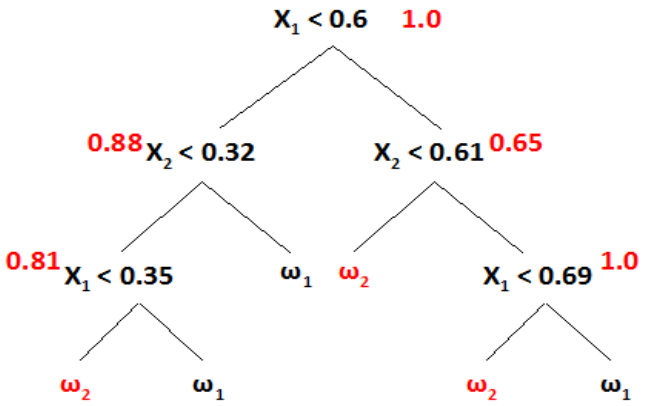
Oblique (πλάγιο) Δέντρο Απόφασης



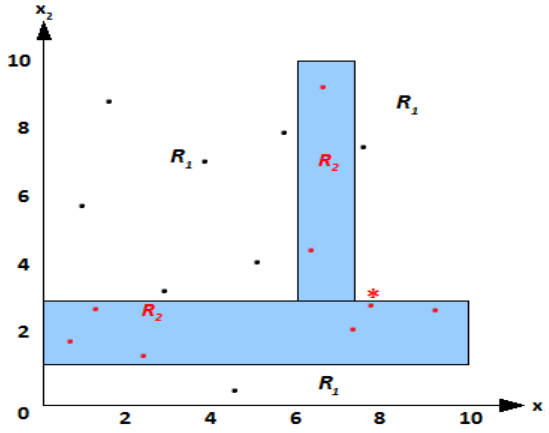
- Οι συνθήκες ελέγχου μπορούν να περιλαμβάνουν περισσότερα από ένα γνωρίσματα.
- Μεγαλύτερη εκφραστικότητα.
- Η εύρεση βέλτιστων συνθηκών ελέγχου είναι υπολογιστικά ακριβή.

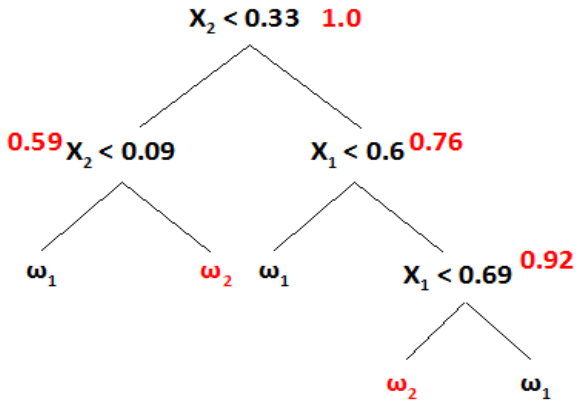
Πρόβλημα αστάθειας





- Μικρή μετακίνηση ενός μόνο δείγματος, οδηγεί σε πολύ διαφορετικά αποτελέσματα.

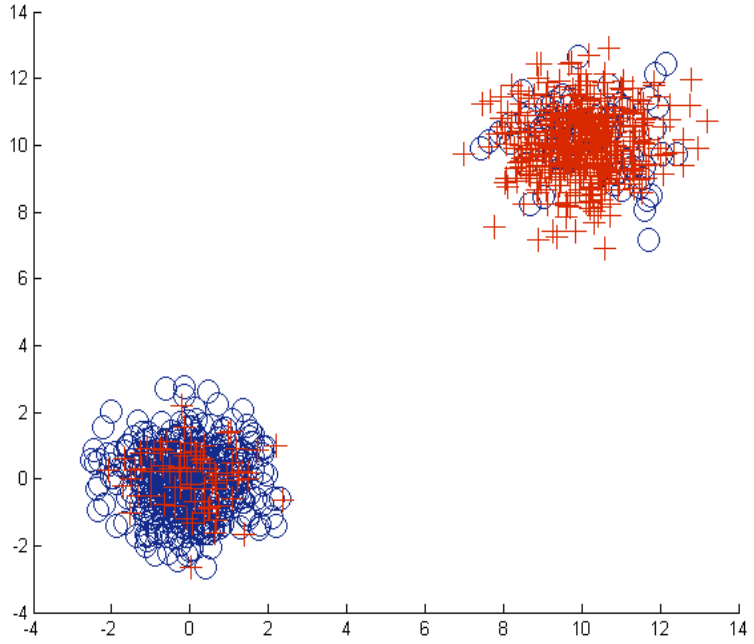




Βελτίωση της ακρίβειας

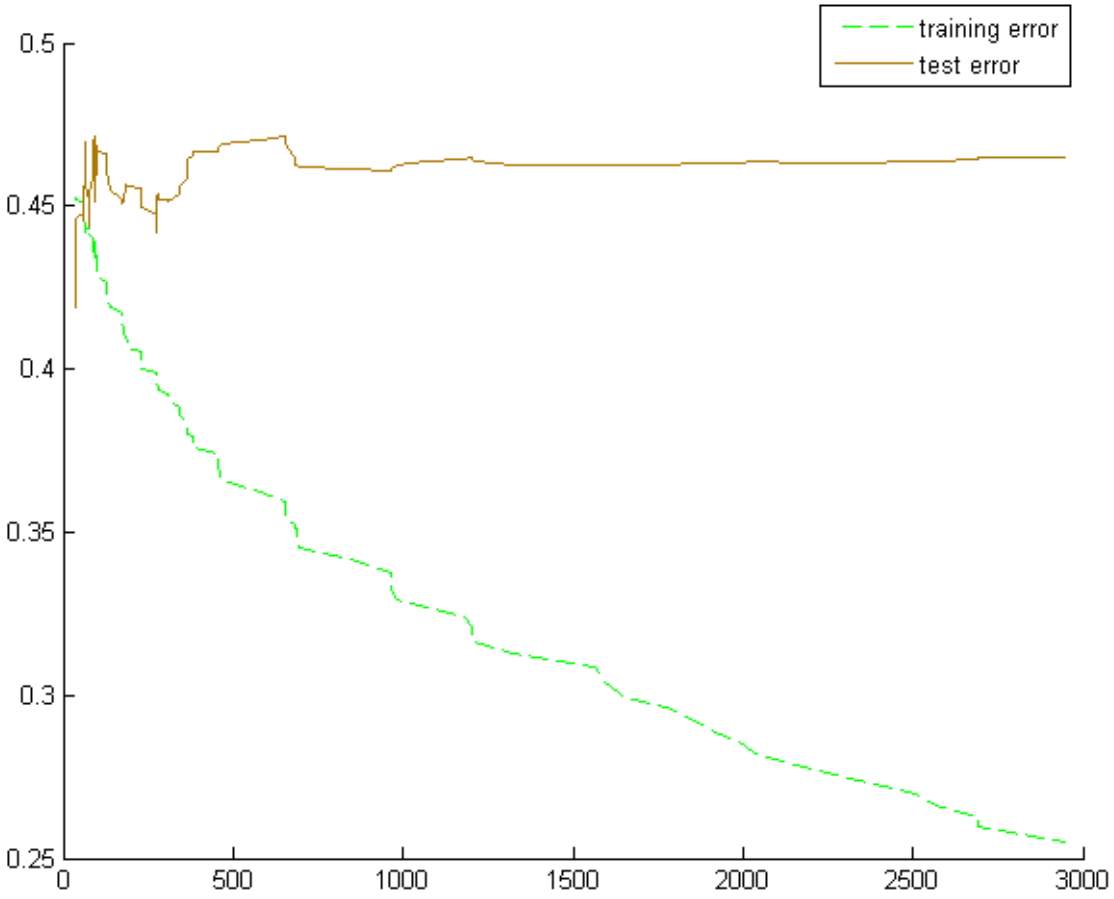
- Ξέρουμε πώς να εκτιμούμε σωστά την ακρίβεια.
- Μπορούμε να βελτιώσουμε την ακρίβεια χρησιμοποιώντας διαφορετικά τους γνωστούς μας κατηγοριοποιητές;
  - Κλάδεμα.
  - Σύνολα κατηγοριοποιητών.

Το φαινόμενο της υπερβολικής προσαρμογής (1/3)



- 2000 δισδιάστατα σημεία σε δύο κλάσεις (1000 σημεία ανά κλάση), που ακολουθούν κανονική κατανομή.
- Προσθέτουμε θόρυβο: ανταλλάζουμε την κλάση 150 σημείων.
- Διαχωρίζουμε 1000 σημεία στο σύνολο ελέγχου.
- Προσθέτουμε επιπλέον θόρυβο στο σύνολο εκμάθησης: ανταλλάζουμε την κλάση άλλων 200 τυχαίων σημείων.

Το φαινόμενο της υπερβολικής προσαρμογής (2/3)



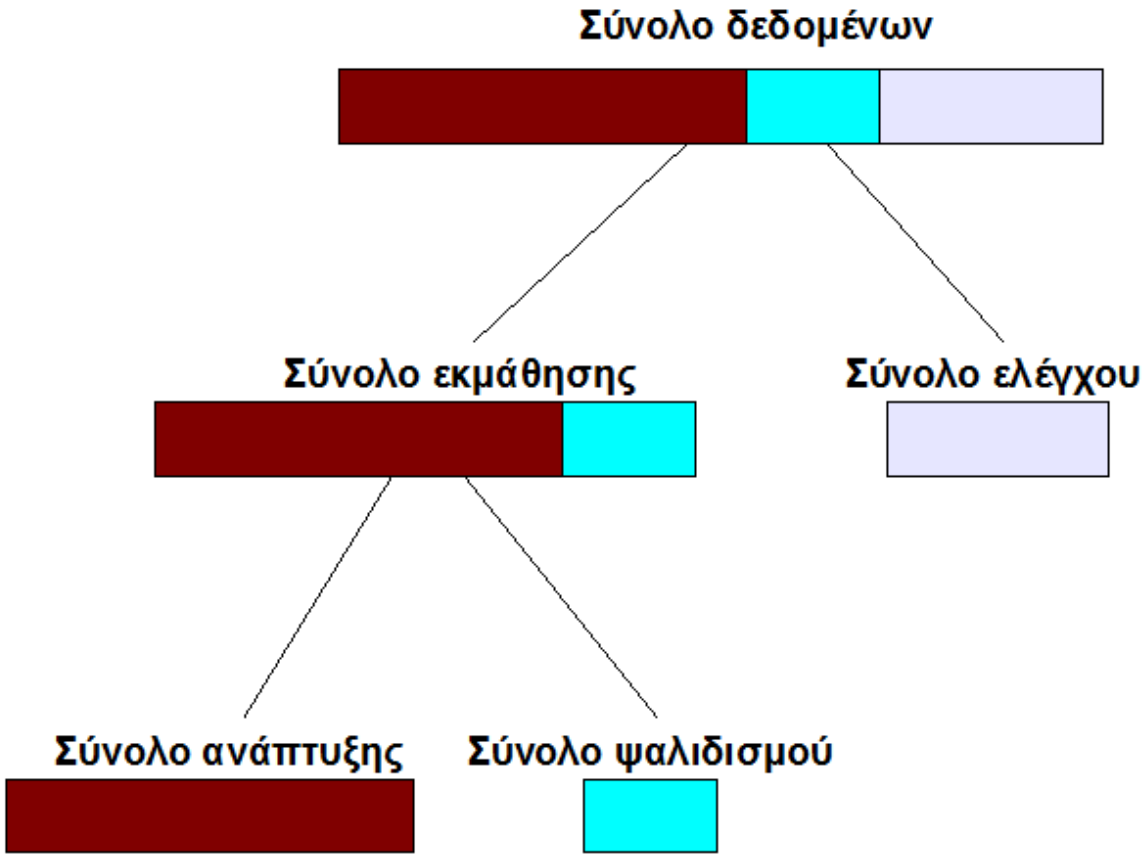
Το φαινόμενο της υπερβολικής προσαρμογής (3/3)

- Μικρός αριθμός κόμβων: ανεπαρκής προσαρμογή (underfiting).
- Στο δένδρο απόφασης προστίθενται ολοένα και περισσότεροι κόμβοι με σκοπό να καλυφθούν (δηλαδή, να γίνει εκμάθηση) τα σημεία θορύβου.
- Η υπερβολική προσαρμογή, που δηλώνεται από τη μείωση του λάθους εκμάθησης, μειώνει όμως τη δυνατότητα του δένδρου απόφασης να γενικεύει σε άγνωστα αντικείμενα, κάτι που δηλώνεται από την αύξηση του λάθους ελέγχου.
- Το πρόβλημα είναι γενικό σε όλους τους κατηγοριοποιητές, όχι μόνο στα ΔΑ.

Κλάδεμα (pruning)

- Προ-κλάδεμα: εναλλακτικά κριτήρια τερματισμού κατά τη δημιουργία (διαφ. 37).
  - Δύσκολη επιλογή κατωφλίου.
- Μετά-κλάδεμα:
  - Πρώτα δημιουργούμε το δένδρο.
  - Μετά, διαγράφουμε υποδένδρα ώστε να μειώσουμε την υπερβολική προσαρμογή.

Αλγόριθμος REP (1/3)



REP: Reduced Error Pruning

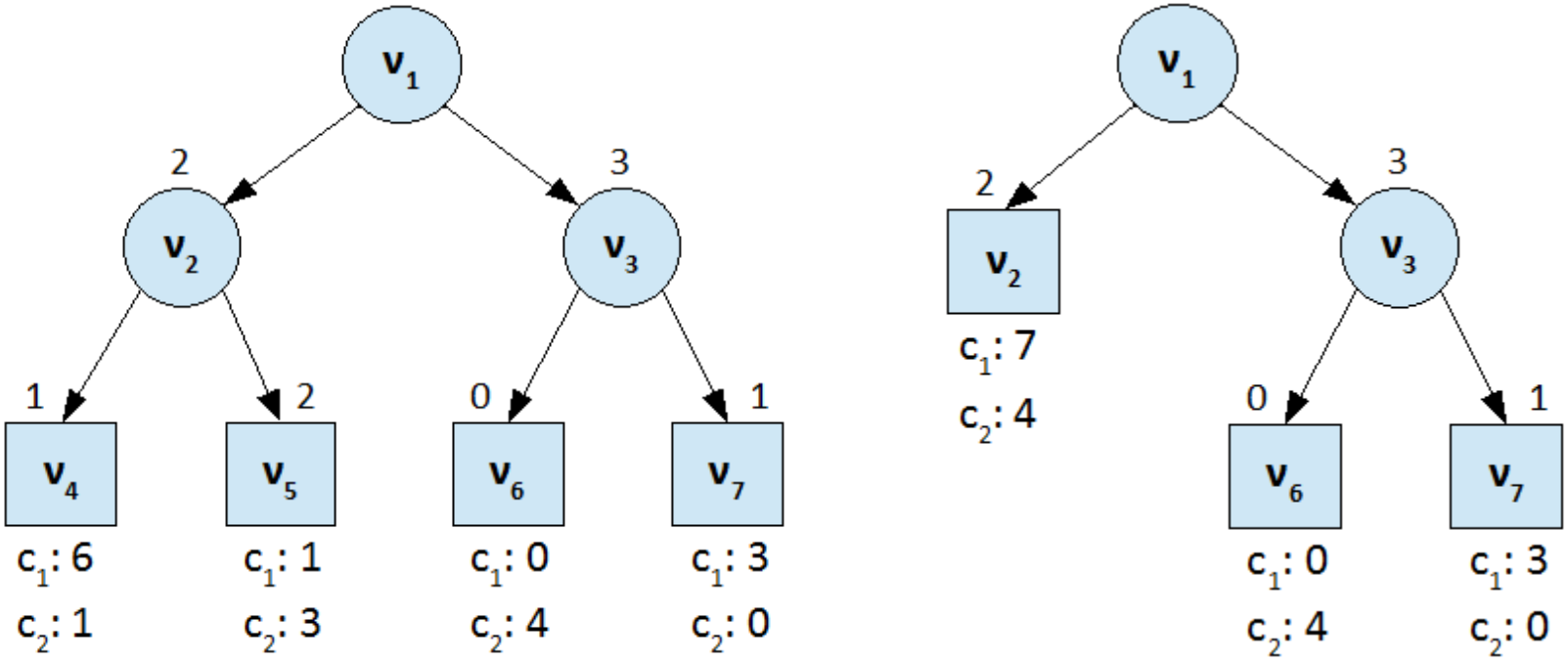
Αλγόριθμος REP (2/3)

1. Κατηγοριοποιούμε τα αντικείμενα του συνόλου ψαλιδισμού.
2. Εφαρμόζουμε μετα-διατεταγμένη διάσχιση στο δένδρο.
3. Εξετάζουμε κάθε εσωτερικό κόμβο *v* που είναι πατέρας φύλλου:
   - E(T_v): αντικείμενα του συνόλου ψαλιδισμού που κατατάσσονται λανθασμένα σε όλα τα φύλλα του *v*.
   - E(v): αντικείμενα του συνόλου ψαλιδισμού που κατατάσσονται λανθασμένα αν θέσουμε το *v* ως φύλλο.
     - κλάση του νέου φύλλου *v* : η πιο συχνή κλάση των αντικειμένων των τωρινών φύλλων.

Αλγόριθμος REP (3/3)

4. Αν για κάθε κόμβο *v* ισχύει E(v)>E(T_v), τότε ο αλγόριθμος τερματίζει.
5. Διαφορετικά βρίσκουμε τον κόμβο *v* με τη μεγαλύτερη διαφορά E(T_v)-E(v).
6. Ψαλιδίζουμε το υποδένδρο με ρίζα το *v*, θέτοντας το *v* ως φύλλο και αναθέτοντάς του την κλάση που πλειοψηφεί μεταξύ των αντικειμένων του συνόλου ανάπτυξης.
7. Πίσω στο Βήμα 1.

Παράδειγμα


v1
2
3
v2
v3
1
2
0
1
v4
v5
v6
v7
c1: 6
c2: 1
c1: 1
c2: 3
c1: 0
c2: 4
c1: 3
c2: 0
v1
2
3
v2
v3
c1: 7
c2: 4
0
1
v6
v7
c1: 0
c2: 4
c1: 3
c2: 0

- ( ) ( ) ( ) ( )

Κλιμακούμενοι αλγόριθμοι

- Δένδρα απόφασης:
  - απαίτηση τα δεδομένα να είναι στην κύρια μνήμη,
  - αλλεπάλληλες ταξινομήσεις αριθμητικών δεδομένων σε κάθε κόμβο.
- Δεν κλιμακώνονται εύκολα.
- Κλιμακούμενες παραλλαγές: SLIQ, SPRINT.
  - **ΔΕΝ** μειώνουν την ακρίβεια.

SLIQ: Supervised Learning In Quest

- Αρχές του SLIQ:
  - Προ-ταξινόμηση.
  - Ανάπτυξη κατά πλάτος.

Προ-ταξινόμηση: Λίστες Ιδιοτήτων/Κλάσεων

RID	Ηλικία	Οικ. Κατάσταση	Αγοραστής
1	20	Διαζευγμένος	Ναι
2	30	Διαζευγμένος	Ναι
3	25	Έγγαμος	Όχι
4	30	Άγαμος	Ναι
5	40	Άγαμος	Ναι
6	20	Έγγαμος	Όχι
7	30	Διαζευγμένος	Ναι
8	25	Διαζευγμένος	Ναι
9	40	Διαζευγμένος	Ναι
10	20	Άγαμος	Όχι

Λίστα ιδιότητας

Ηλικία	RID
20	1
20	6
20	10
25	3
25	8
30	2
30	4
30	7
40	5
40	9

Λίστα ιδιότητας

Οικ. Κατάσταση	RID
Διαζευγμένος	1
Διαζευγμένος	2
Διαζευγμένος	7
Διαζευγμένος	8
Διαζευγμένος	9
Έγγαμος	6
Έγγαμος	3
Άγαμος	4
Άγαμος	5
Άγαμος	10

Λίστα κλάσεων

RID	Αγοραστής	Φύλλο
1	Ναι	NO
2	Ναι	NO
3	Όχι	NO
4	Ναι	NO
5	Ναι	NO
6	Όχι	NO
7	Ναι	NO
8	Ναι	NO
9	Ναι	NO
10	Όχι	NO

Αρχική ρίζα

Διάσπαση με gini index

- Έστω *c* κλάσεις και *n* αντικείμενα.
- P_i : σχετική συχνότητα της κλάσης *i* στο σύνολο *S*.
- Δείκτης gini: $gini(S) = 1 - \sum_{i=1}^{c} p_i^2$
- Αν ένα διαχωριστικό σημείο χωρίζει το σύνολο *S* σε δύο υποσύνολα, S_1 και S_2 , μεγέθους n_1 και n_2 αντίστοιχα:

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

Παράδειγμα SLIQ (1/3)

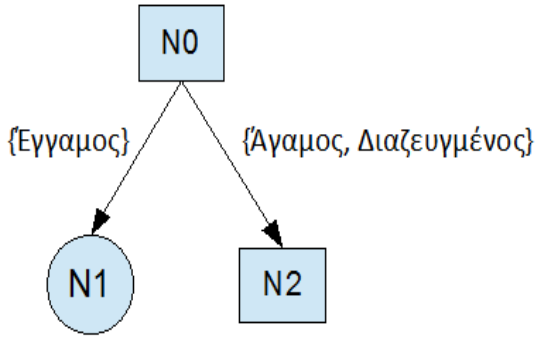
Ηλικία								
	20		25			30	40	
		22.5		27.5		35		
	>	<=	>	<=	>	<=	>	<=
Ναι	7	1	6	2	5	5	2	7
Όχι	3	2	1	3	0	3	0	3
Gini(i)		0.44	0.24	0.48	0	0.46	0	
Gini split		0.3		0.24		0.36		

RID	Αγοραστής	Φύλλο
1	Ναι	N2
2	Ναι	N2
3	Όχι	N1
4	Ναι	N2
5	Ναι	N2
6	Όχι	N1
7	Ναι	N2
8	Ναι	N2
9	Ναι	N2
10	Όχι	N2

	Ναι	Όχι	Gini(i)	Gini split
Άγαμος	2	1	0.44	0.559
Διαζευγμένος - Έγγαμος	5	2	0.61	

	Ναι	Όχι	Gini(i)	Gini split
Διαζευγμένος	5	0	0	0.24
Άγαμος-Έγγαμος	2	3	0.48	

	Ναι	Όχι	Gini(i)	Gini split
Έγγαμος	0	2	0	**0.168**
Άγαμος-Διαζευγμένος	7	1	0.21	

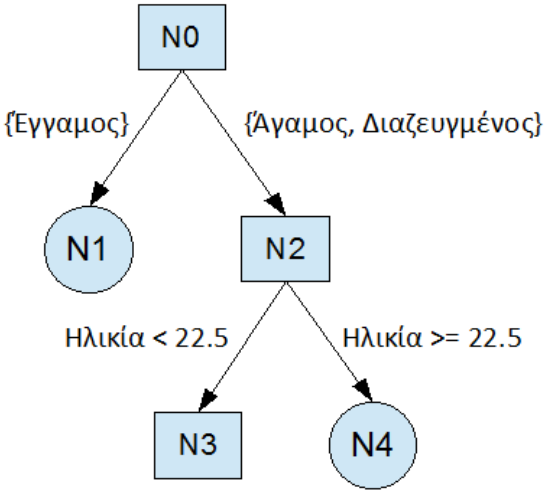


Παράδειγμα SLIQ (2/3)

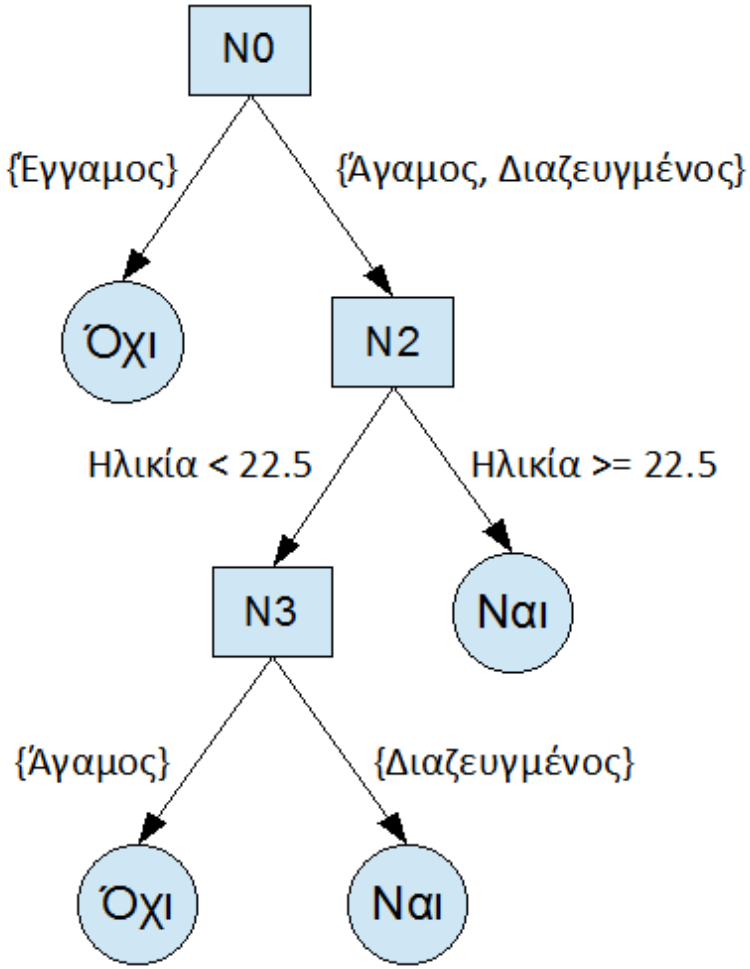
Ηλικία	RID
20	1
20	6
20	10
25	3
25	8
30	2
30	4
30	7
40	5
40	9

Οικ. Κατάσταση	RID
Διαζευγμένος	1
Διαζευγμένος	2
Διαζευγμένος	7
Διαζευγμένος	8
Διαζευγμένος	9
Έγγαμος	6
Έγγαμος	3
Άγαμος	4
Άγαμος	5
Άγαμος	10

RID	Αγοραστής	Φύλλο
1	Ναι	N3
2	Ναι	N4
3	Όχι	N1
4	Ναι	N4
5	Ναι	N4
6	Όχι	N1
7	Ναι	N4
8	Ναι	N4
9	Ναι	N4
10	Όχι	N3



Παράδειγμα SLIQ (3/3)



SPRINT (Scalable PaRallelizable INndution of decision Tree)

- Ο SLIQ απαιτεί η λίστα κλάσης να παραμένει στη μνήμη.
- Πρόβλημα αν αυτό δεν είναι εφικτό.
- Ο SPRINT καταργεί τη λίστα κλάσης.
  - προσθέτει την κλάση στις λίστες ιδιοτήτων.
- Κάθε κόμβος έχει τη δική του λίστα ιδιοτήτων.
  - Δυνατότητα για παραλληλισμό.

Λίστες ιδιοτήτων SPRINT

Ηλικία	Αγοραστής	RID
20	Ναι	1
20	Όχι	6
20	Όχι	10
25	Όχι	3
25	Ναι	8
30	Ναι	2
30	Ναι	4
30	Ναι	7
40	Ναι	5
40	Ναι	9

Οικ. Κατάσταση	Αγοραστής	RID
Διαζευγμένος	Ναι	1
Διαζευγμένος	Ναι	2
Διαζευγμένος	Όχι	7
Διαζευγμένος	Ναι	8
Διαζευγμένος	Ναι	9
Έγγαμος	Όχι	6
Έγγαμος	Ναι	3
Άγαμος	Ναι	4
Άγαμος	Ναι	5
Άγαμος	Όχι	10

Ενημέρωση λιστών

- Για την ιδιότητα διάσπασης:
  - απλώς διαχωρίζεται η αντίστοιχη λίστα σε 2 άλλες (δυαδική διάσπαση).
- Για τις υπόλοιπες ιδιότητες:
  - ο διαχωρισμός γίνεται βάσει RID.

SPRINT vs. SLIQ

SPRINT

- Δεν έχει περιορισμό μνήμης.
- Απαιτεί χρόνο ενημέρωσης και επαναποθήκευσης λιστών.
- Παραλληλίζεται εύκολα.

SLIQ

- Μειώνει τους περιορισμούς μνήμης αλλά δεν τους εξαλείφει (λίστα κλάσης).
- Ενημερώνει μόνο τη λίστα κλάσεων (στη μνήμη).
- Δεν παραλληλίζεται εύκολα (λόγω μίας κεντρικής λίστας κλάσης).

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.