Τμήμα Πληροφορικής με Εφαρμογές στη Βιοϊατρική
Σχολή Θετικών Επιστημών
Πανεπιστήμιο Θεσσαλίας

ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Κατηγοριοποίηση

Αριστείδης Γ. Βραχάτης, Dipl-Ing, M.Sc, PhD

Κατηγοριοποιητής K πλησιέστερων γειτόνων

- Ας μελετήσουμε μια μη παραμετρική μέθοδο για την εκτίμηση πιθανοφάνειας η οποία χρησιμοποιεί την εκτίμηση πυκνότητας με βάση τους πλησιέστερους γείτονες.

- Έστω ότι $\mathbf{D}$ είναι ένα σύνολο δεδομένων εκπαίδευσης το οποίο αποτελείται από n σημεία $\mathbf{x}_i \in \mathbb{R}^d$, και έστω ότι συμβολίζουμε με $\mathbf{D}_i$ το υποσύνολο των σημείων του $\mathbf{D}$ που έχουν ως ετικέτα την κατηγορία c_i, με $n_i = |\mathbf{D}_i|$.

- Ας υποθέσουμε ότι δίνεται ένα σημείο δοκιμής $\mathbf{x} \in \mathbb{R}^d$, καθώς και το K (το πλήθος των γειτόνων που θα ληφθούν υπόψη)· έστω ότι η απόσταση του $\mathbf{x}$ από τον K-οστό πλησιέστερο γείτονά του στο $\mathbf{D}$ συμβολίζεται με r.

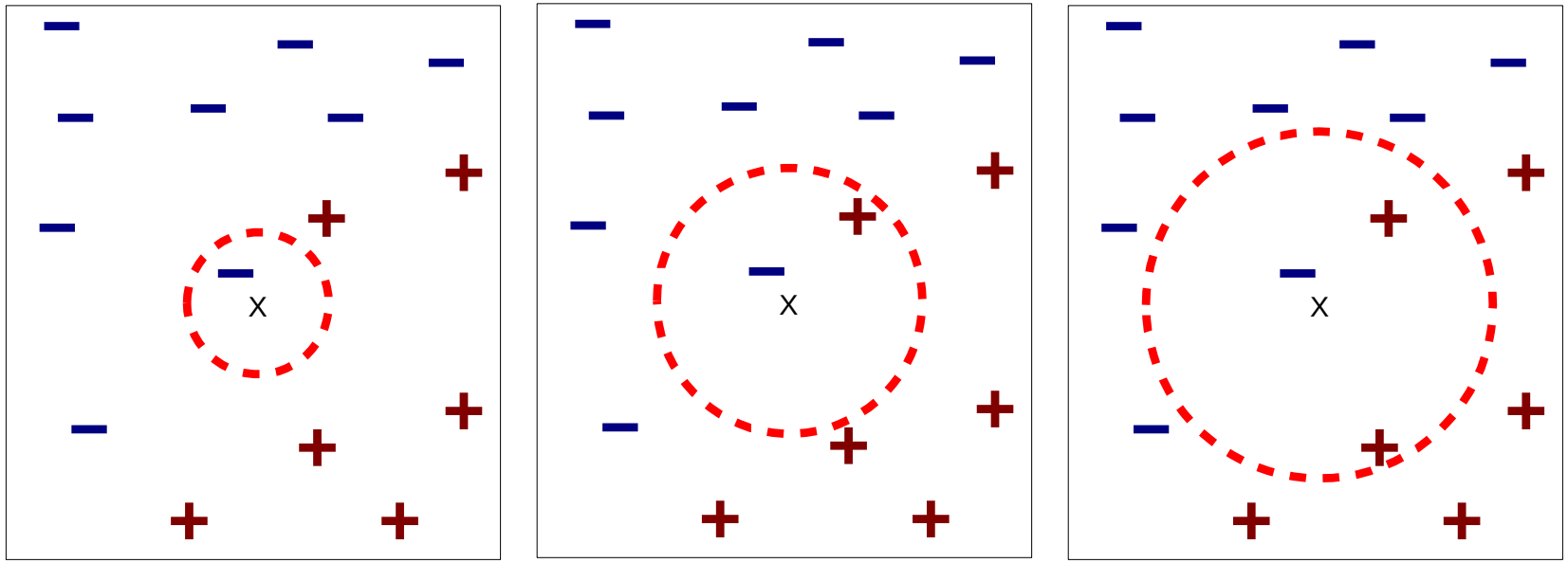
- Θεωρήστε την d-διάστατη «υπερμπάλα» ακτίνας r γύρω από το σημείο δοκιμής $\mathbf{x}$, η οποία ορίζεται ως

$$B_d(\mathbf{x}, r) = \left\{ \mathbf{x}_i \in \mathbf{D} | \ \delta(\mathbf{x}, \mathbf{x}_i) \leq r \right\}$$

- Εδώ, το $\delta(\mathbf{x}, \mathbf{x}_i)$ είναι η απόσταση των $\mathbf{x}$ και $\mathbf{x}_i$· συνήθως υποθέτουμε ότι πρόκειται για την Ευκλείδεια απόσταση, δηλαδή $\delta(\mathbf{x}, \mathbf{x}_i) = ||\mathbf{x} - \mathbf{x}_i||_2$. Επίσης, υποθέτουμε ότι $|B_d(\mathbf{x}, r)| = K$.

Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

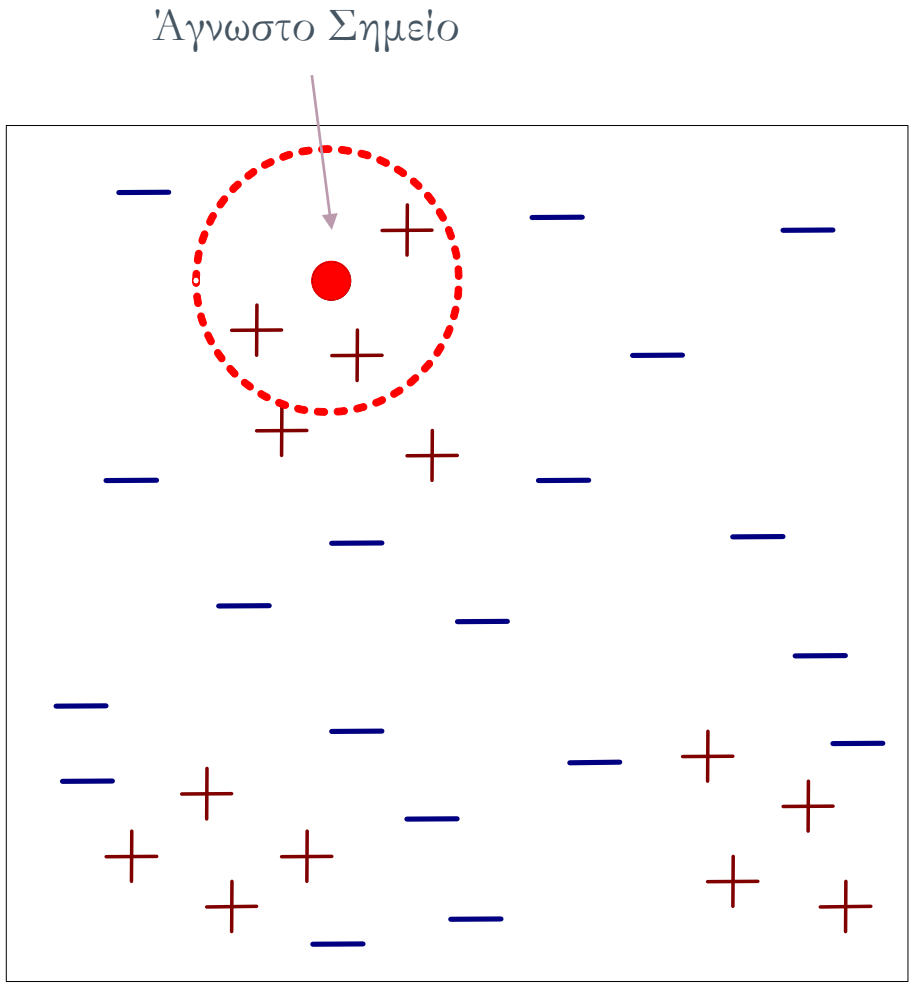
– k-κοντινότεροι γείτονες μιας εγγραφής x είναι τα σημεία που έχουν την k-οστή μικρότερη απόσταση από το x



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

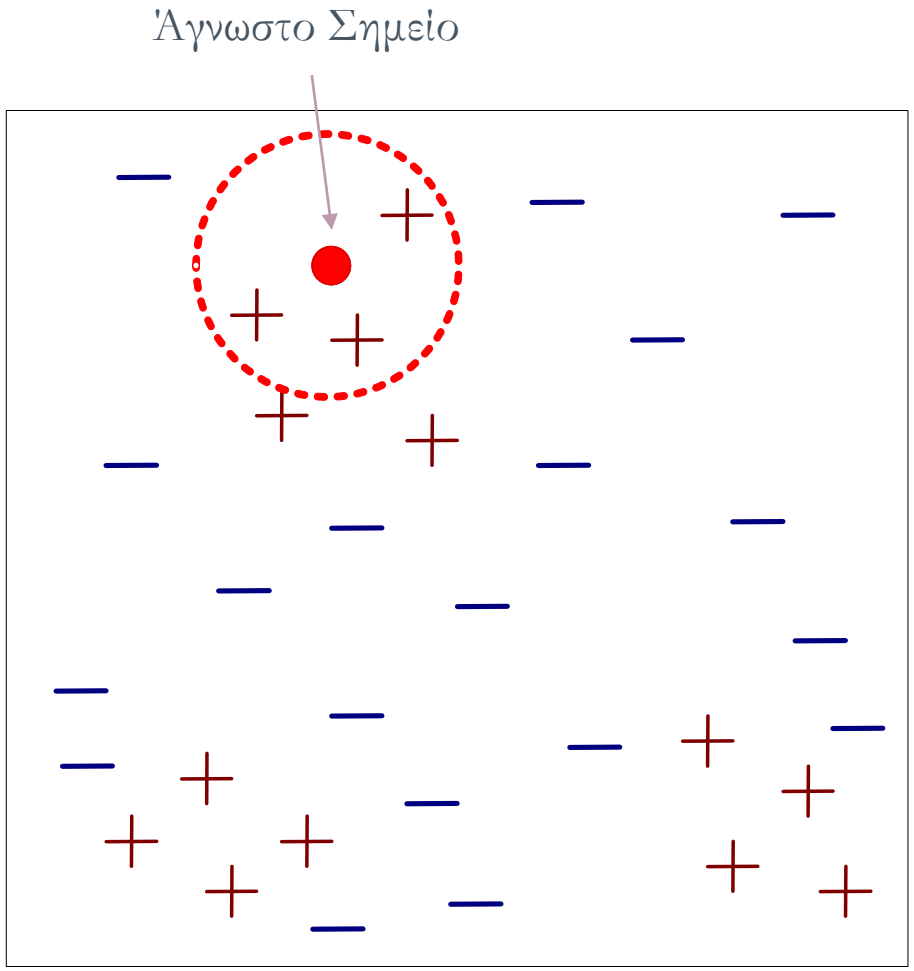
Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

- Για να κατηγοριοποιηθεί μια άγνωστη εγγραφή:
  - Υπολογισμός της απόστασης από τις εγγραφές του συνόλου
  - Εύρεση των k κοντινότερων γειτόνων
  - Χρήση των κλάσεων των κοντινότερων γειτόνων για τον καθορισμό της κλάσης της άγνωστης εγγραφής - π.χ., με βάση την πλειοψηφία (majority vote)



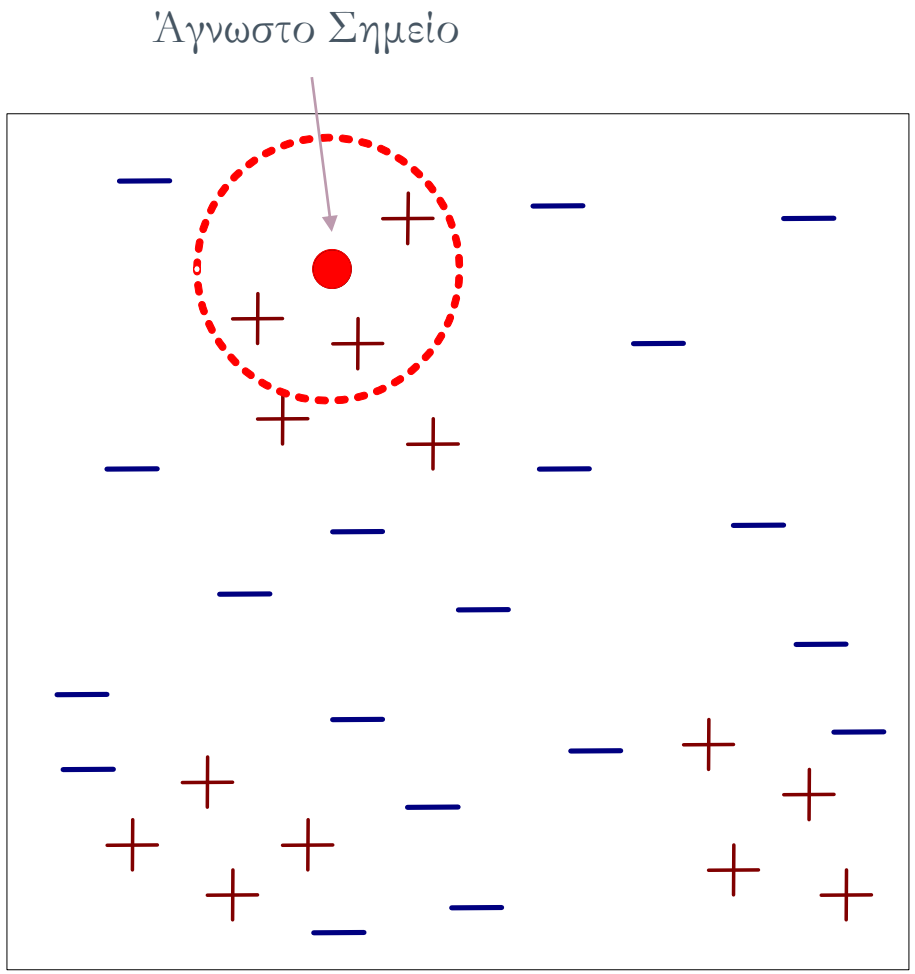
Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

- Χρειάζεται:
  - Το σύνολο των αποθηκευμένων εγγραφών
  - Distance Metric Μετρική απόστασης για να υπολογίσουμε την απόσταση μεταξύ εγγραφών
  - Την τιμή του k, δηλαδή τον αριθμό των κοντινότερων γειτόνων που πρέπει να ανακληθούν



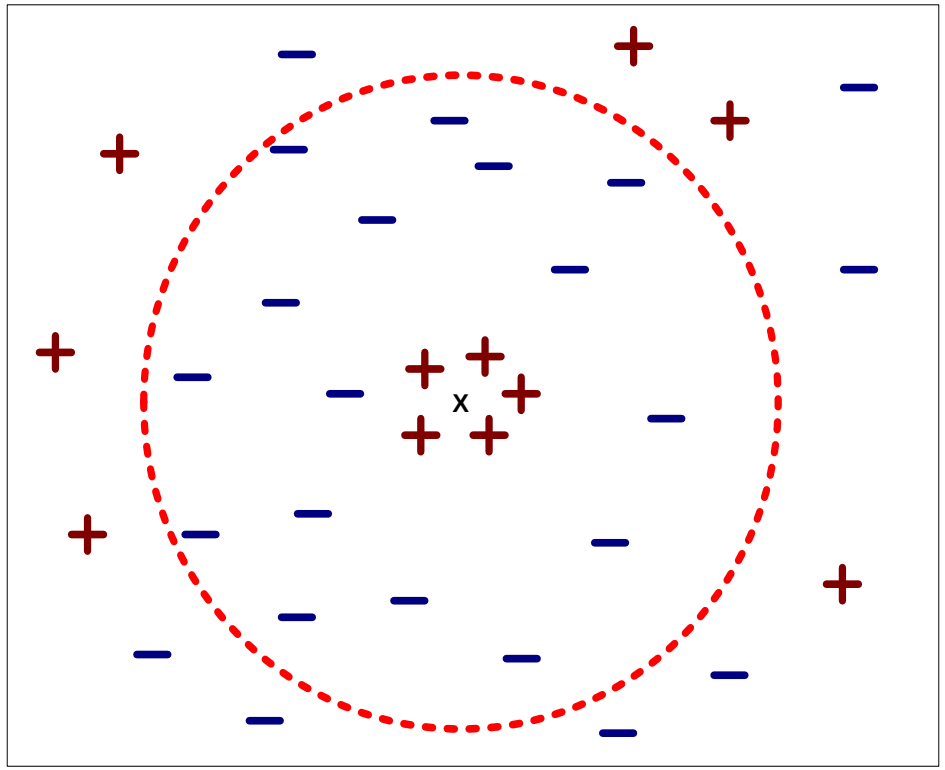
Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

- Απόσταση μεταξύ εγγραφών:
  - Πχ ευκλείδεια απόσταση
- Καθορισμός τάξης:
  - Η πλειοψηφική κλάση
  - Βάρος σε κάθε ψήφο με βάση την απόσταση
  - …



Κατηγοριοποιητής Κ πλησιέστερων γειτόνων

- Επιλογή της τιμής του k:
  - k πολύ μικρό, ευαισθησία στα σημεία θορύβου
  - k πολύ μεγάλο, η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις



Δέντρα Αποφάσεων

- Συχνά είναι απαραίτητο να κάνουμε μια σειρά ερωτήσεων πριν καταλήξουμε σε μια απόφαση για ένα πρόβλημα.

- Οι απαντήσεις σε μια ερώτηση
  - μπορεί να οδηγήσουν σε άλλη ερώτηση ή
  - μπορεί να οδηγήσουν σε μια απόφαση για να επιτευχθεί η λύση του προβλήματος

- Μοντέλο = Δέντρο Απόφασης
  - Εσωτερικοί κόμβοι αντιστοιχούν σε κάποιο γνώρισμα
  - Διαχωρισμός (split) ενός κόμβου σε παιδιά
    - η ετικέτα στην ακμή = συνθήκη/έλεγχος
  - Φύλλα αντιστοιχούν σε κλάσεις

Δέντρα Αποφάσεων

- Το πρόβλημα του εστιατορίου

- Χαρακτηριστικά (attributes) του προβλήματος:
  - Εναλλακτικό: Ναι, Όχι.
  - Μπαρ: Ναι, Όχι.
  - Π/Σ: Ναι, Όχι.
  - Πεινασμένος: Ναι, Όχι.
  - Πελάτες: Κανένας, Μερικοί, και Πλήρες.
  - Τιμή: $, $$, $$$.
  - Βρέχει: Ναι, Όχι.
  - Κράτηση: Ναι, Όχι.
  - Τύπος: Γαλλικό, Ιταλικό, Ταϋλανδέζικο, ή ταχυφαγείο.
  - ΕκτίμησηΑναμονής: 0'–10', 10'–30', 30'–60', >60'.

- Απόφαση για το αν ο πελάτης θα περιμένει: ΝΑΙ ή ΟΧΙ

Δέντρα Αποφάσεων

– Σύνολο εκπαίδευσης

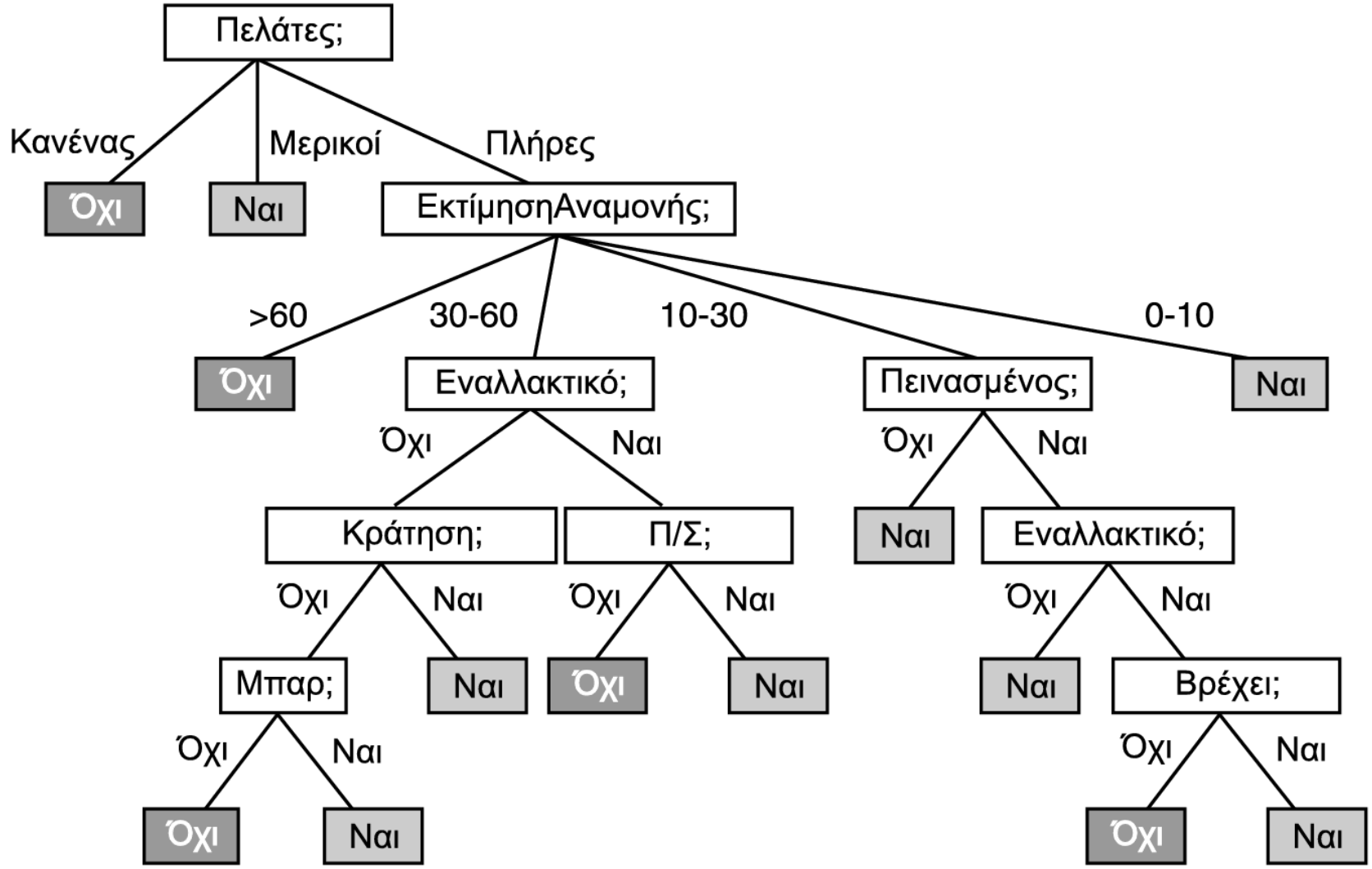
#	Εναλ	Μπαρ	Π/Σ	Πεινασμ	Πελατες	Τιμή	Βρέχει	Κράτηση	Τύπος	Εκτιμ	ΘαΠεριμένει
X_1	Ναι	Όχι	Όχι	Ναι	Μερικοί	$$$	Όχι	Ναι	Γαλλικό	0-10	**Ναι**
X_2	Ναι	Όχι	Όχι	Ναι	Πλήρες	$	Όχι	Όχι	Ταϋλ	30-60	**Όχι**
X_3	Όχι	Ναι	Όχι	Όχι	Μερικοί	$	Όχι	Όχι	Ταχυφ.	0-10	**Ναι**
X_4	Ναι	Όχι	Ναι	Ναι	Πλήρες	$	Ναι	Όχι	Ταϋλ	10-30	**Ναι**
X_5	Ναι	Όχι	Ναι	Όχι	Πλήρες	$$$	Όχι	Ναι	Γαλλικό	>60	**Όχι**
X_6	Όχι	Ναι	Όχι	Ναι	Μερικοί	$$	Ναι	Ναι	Ιταλικό	0-10	**Ναι**
X_7	Όχι	Ναι	Όχι	Όχι	Κανένας	$	Ναι	Όχι	Ταχυφ.	0-10	**Όχι**
X_8	Όχι	Όχι	Όχι	Ναι	Μερικοί	$$	Ναι	Ναι	Ταϋλ	0-10	**Ναι**
X_9	Όχι	Ναι	Ναι	Όχι	Πλήρες	$	Ναι	Όχι	Ταχυφ.	>60	**Όχι**
X_{10}	Ναι	Ναι	Ναι	Ναι	Πλήρες	$$$	Όχι	Ναι	Ιταλικό	10-30	**Όχι**
X_{11}	Όχι	Όχι	Όχι	Όχι	Κανένας	$	Όχι	Όχι	Ταϋλ	0-10	**Όχι**
X_{12}	Ναι	Ναι	Ναι	Ναι	Πλήρες	$	Όχι	Όχι	Ταχυφ.	30-60	**Ναι**

Δέντρα Αποφάσεων

– Σύνολο εκπαίδευσης

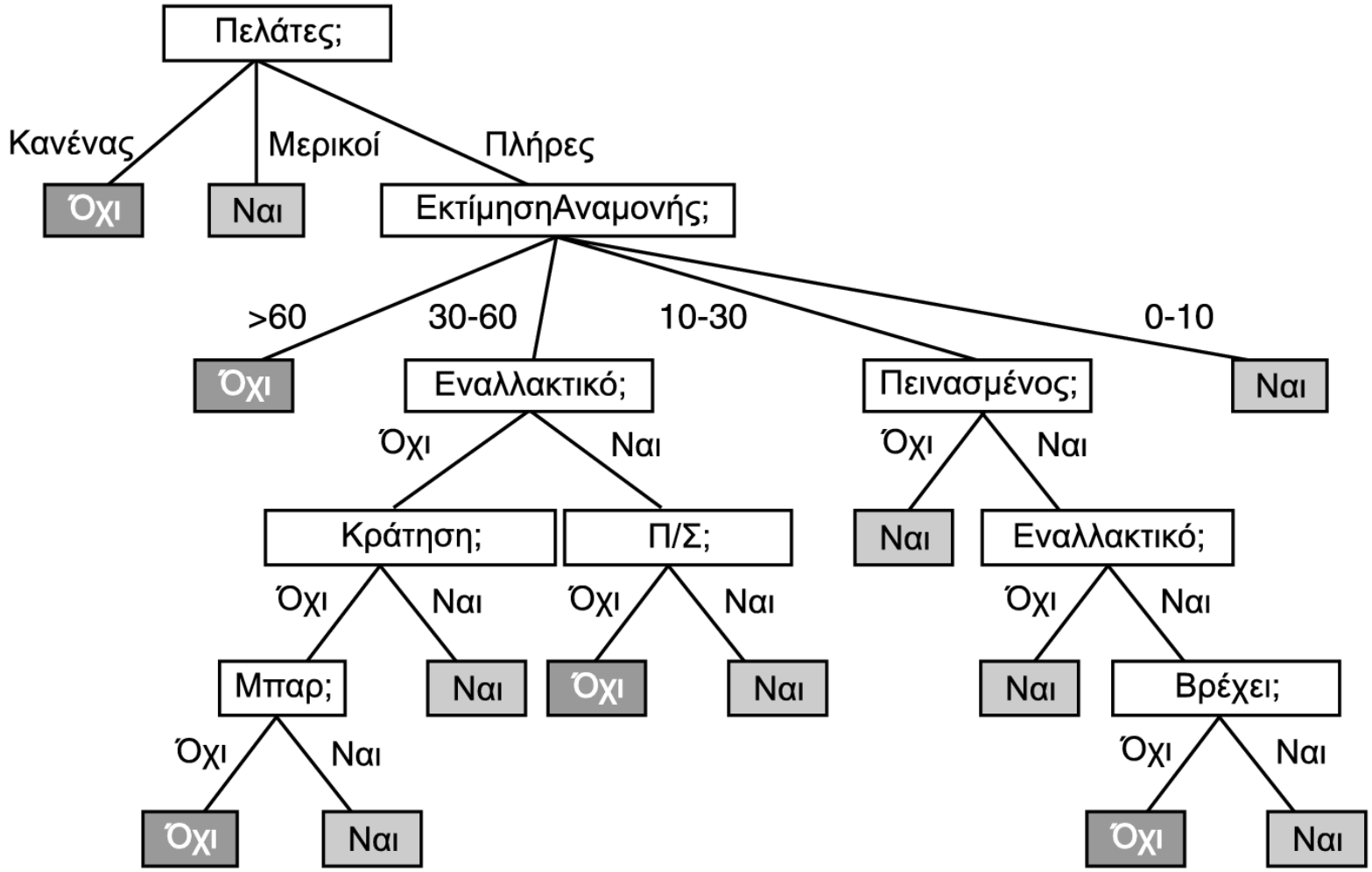
#	Εναλ	Μπαρ	Π/Σ	Πεινασμ	Πελατες	Τιμή	Βρέχει	Κράτηση	Τύπος	Εκτιμ	ΘαΠεριμένει
X_1	Ναι	Όχι	Όχι	Ναι	Μερικοί	$$$	Όχι	Ναι	Γαλλικό	0-10	**Ναι**
X_2	Ναι	Όχι	Όχι	Ναι	Πλήρες	$	Όχι	Όχι	Ταϋλ	30-60	**Όχι**
X_3	Όχι	Ναι	Όχι	Όχι	Μερικοί	$	Όχι	Όχι	Ταχυφ.	0-10	**Ναι**
X_4	Ναι	Όχι	Ναι	Ναι	Πλήρες	$	Ναι	Όχι	Ταϋλ	10-30	**Ναι**
X_5	Ναι	Όχι	Ναι	Όχι	Πλήρες	$$$	Όχι	Ναι	Γαλλικό	>60	**Όχι**
X_6	Όχι	Ναι	Όχι	Ναι	Μερικοί	$$	Ναι	Ναι	Ιταλικό	0-10	**Ναι**
X_7	Όχι	Ναι	Όχι	Όχι	Κανένας	$	Ναι	Όχι	Ταχυφ.	0-10	**Όχι**
X_8	Όχι	Όχι	Όχι	Ναι	Μερικοί	$$	Ναι	Ναι	Ταϋλ	0-10	**Ναι**
X_9	Όχι	Ναι	Ναι	Όχι	Πλήρες	$	Ναι	Όχι	Ταχυφ.	>60	**Όχι**
X_{10}	Ναι	Ναι	Ναι	Ναι	Πλήρες	$$$	Όχι	Ναι	Ιταλικό	10-30	**Όχι**
X_{11}	Όχι	Όχι	Όχι	Όχι	Κανένας	$	Όχι	Όχι	Ταϋλ	0-10	**Όχι**
X_{12}	Ναι	Ναι	Ναι	Ναι	Πλήρες	$	Όχι	Όχι	Ταχυφ.	30-60	**Ναι**

Δέντρα Αποφάσεων



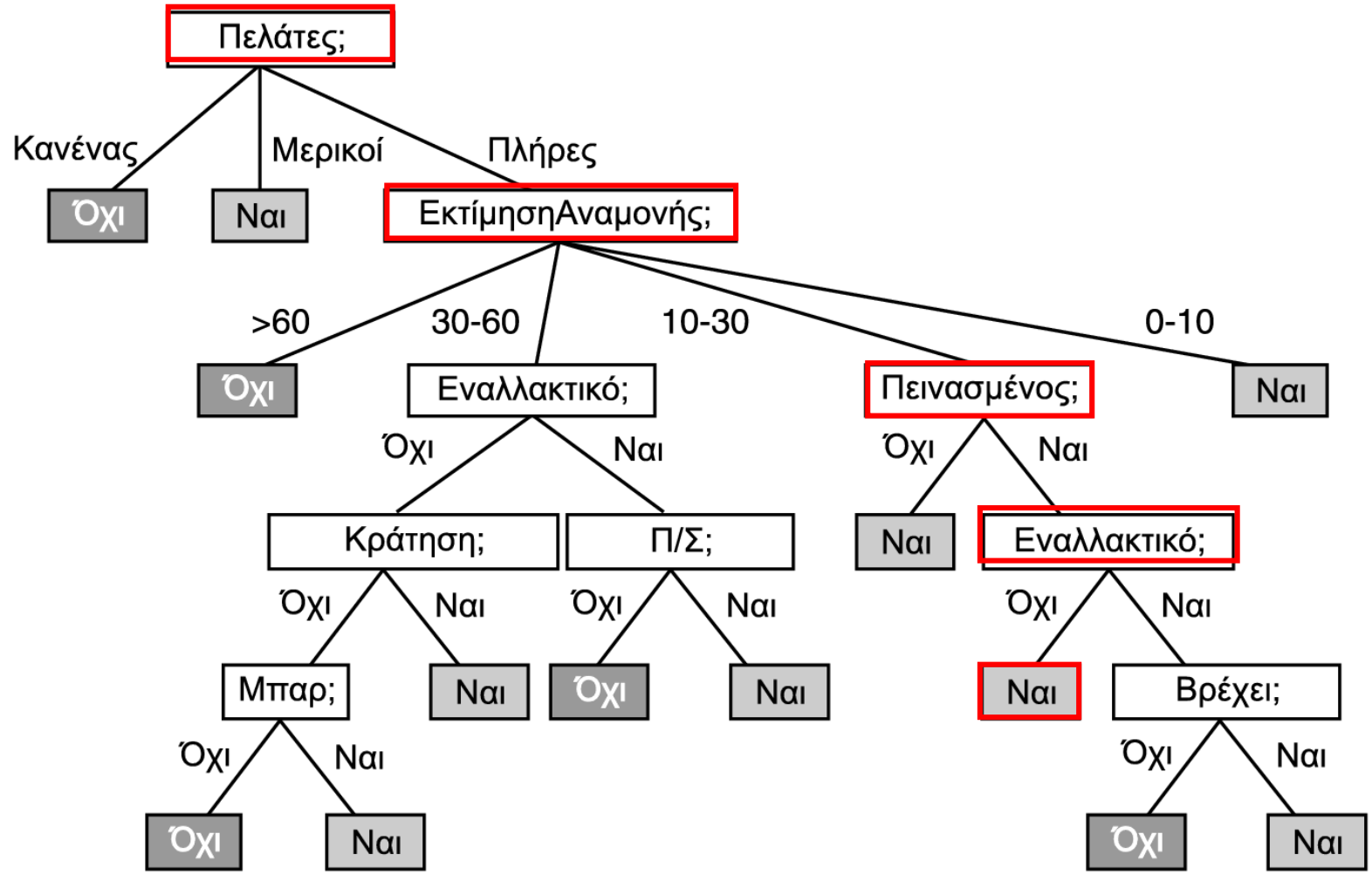
Δέντρα Αποφάσεων

Νέος Πελάτης: Πλήρες, Εκτίμηση Αναμονής = 15 min, Πεινασμένος = ΝΑΙ, Ενναλακτικό = ΟΧΙ



Δέντρα Αποφάσεων

Νέος Πελάτης: Πλήρες, Εκτίμηση Αναμονής = 15 min, Πεινασμένος = ΝΑΙ, Ενναλακτικό = ΟΧΙ



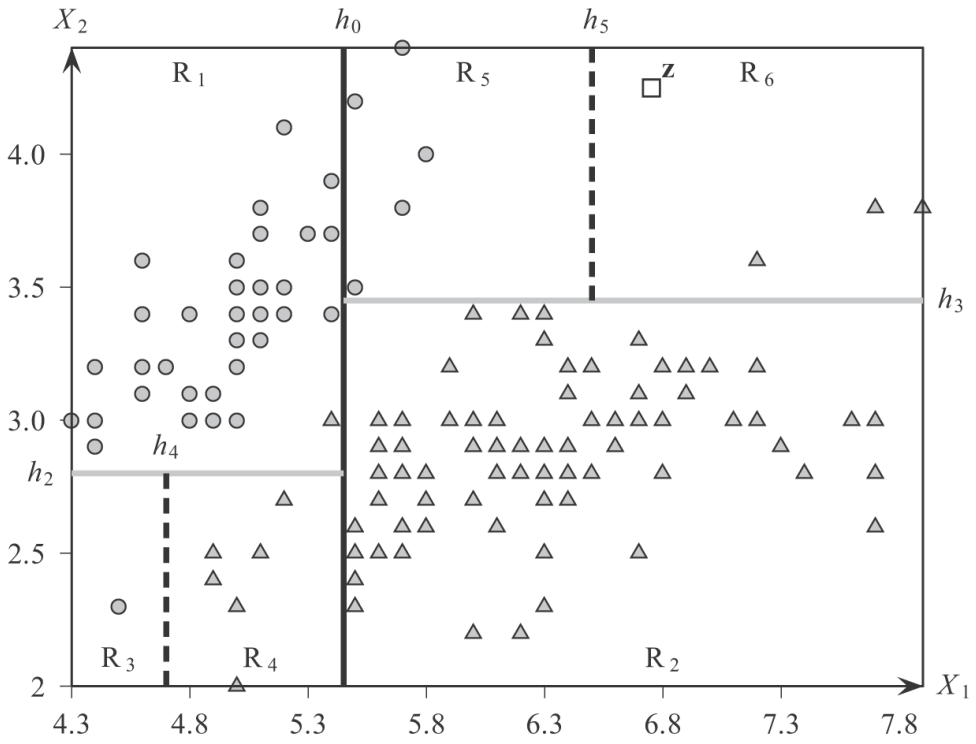
Δέντρα Αποφάσεων

- Εσωτερικοί κόμβοι:
  - έλεγχος και απόφαση με βάση την τιμή κάποιου χαρακτηριστικού (attribute test)

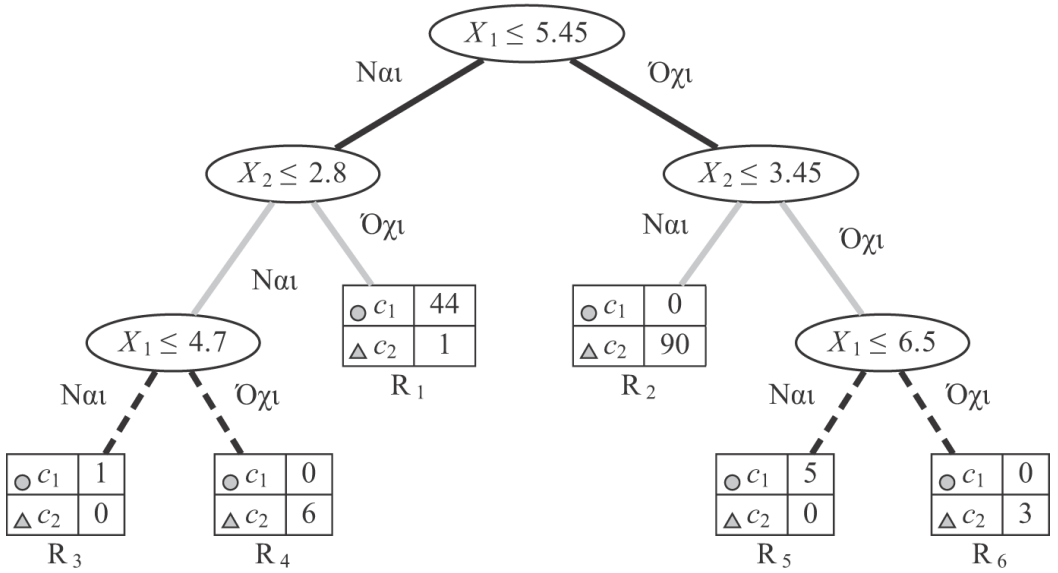
- Φύλλα: απόφαση ταξινόμησης σε κάποια κατηγορία

- Πως επιτυγχάνεται ο διαχωρισμός σε δέντρα αποφάσεων.
  - Υπάρχουν περισσότεροι από δυο τρόποι διαχωρισμού μιας απόφασης.
  - Ωστόσο κάθε φορά θα πρέπει να επιλέγεται η μέθοδος με όποια δεν χάνεται ο έλεγχος – πληροφορία μιας παραμέτρου που μπορεί να επηρεάσει μια απόφαση.

Δέντρα Αποφάσεων – Iris Data

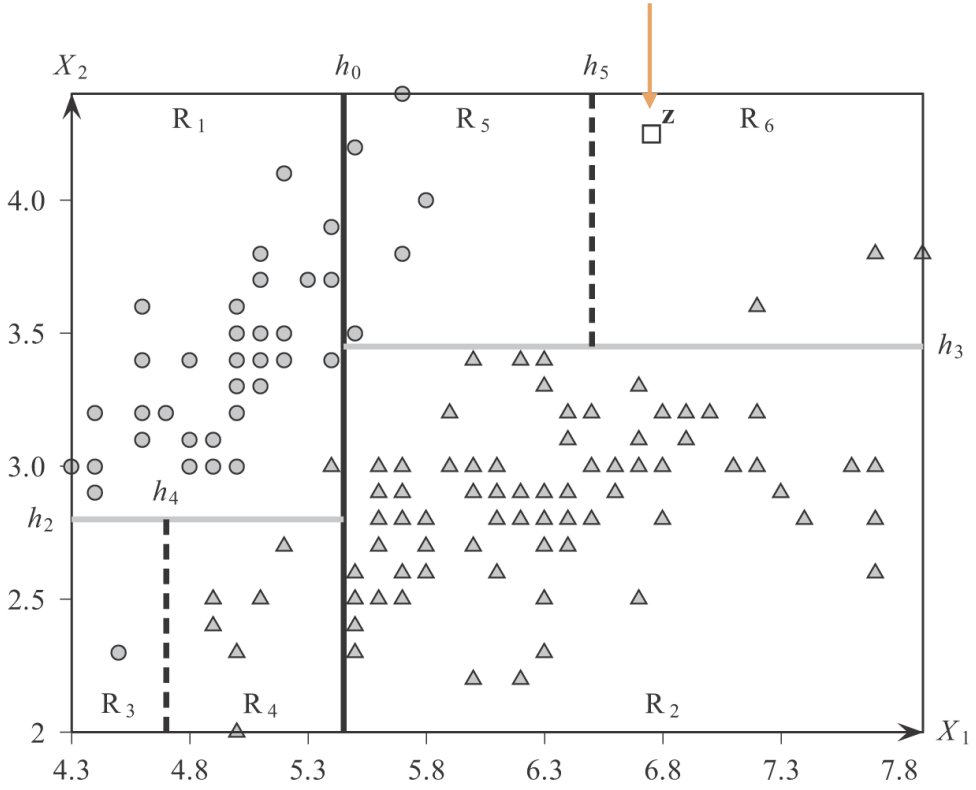


(α) Αναδρομικοί διαμερισμοί

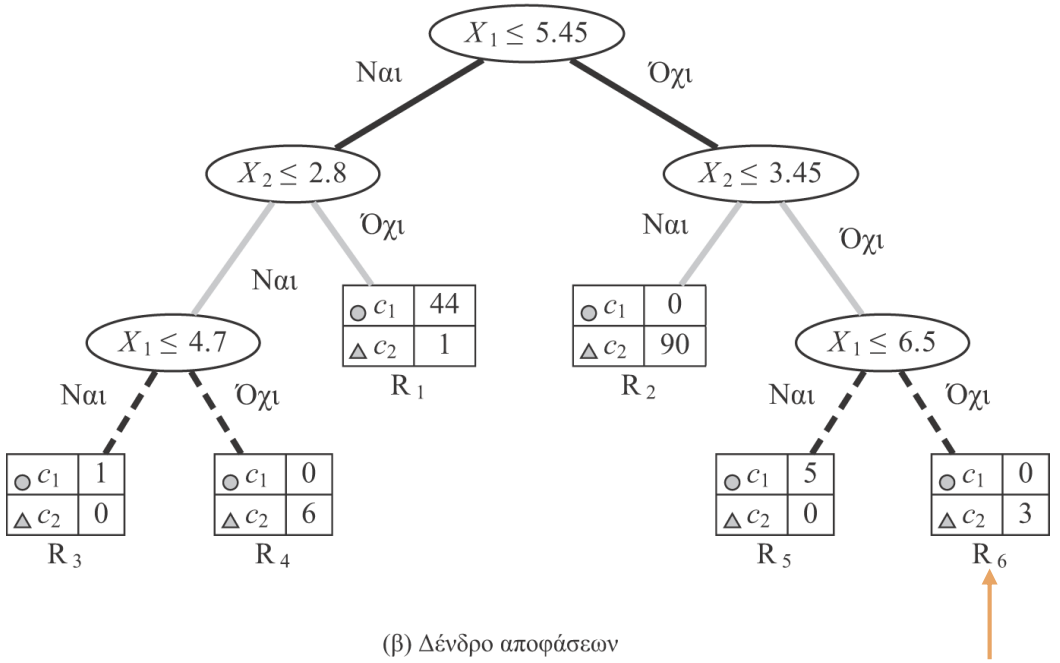


(β) Δένδρο αποφάσεων

Δέντρα Αποφάσεων – Iris Data

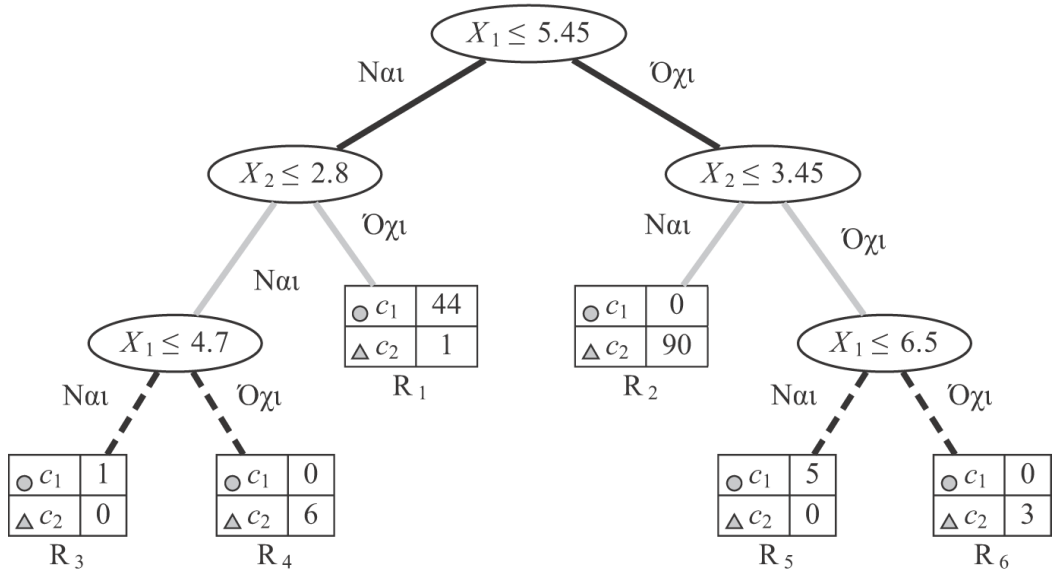


(α) Αναδρομικοί διαμερισμοί



(β) Δένδρο αποφάσεων

Κανόνες από δένδρα αποφάσεων



(β) Δένδρο αποφάσεων

- Ένα δένδρο είναι ένα σύνολο κανόνων αποφάσεων· κάθε κανόνας περιλαμβάνει τις αποφάσεις κατά μήκος της διαδρομής προς ένα φύλλο:

  R3 : Αν $X_1 \le 5.45$ και $X_2 \le 2.8$ και $X_1 \le 4.7$, τότε η κατηγορία είναι η c_1, ή
  R4: Αν $X_1 \le 5.45$ και $X_2 \le 2.8$ και $X_1 > 4.7$, τότε η κατηγορία είναι η c_2, ή
  R1: Αν $X_1 \le 5.45$ και $X_2 > 2.8$, τότε η κατηγορία είναι η c_1, ή
  R2: Αν $X_1 > 5.45$ και $X_2 \le 3.45$, τότε η κατηγορία είναι η c_2, ή
  R5: Αν $X_1 > 5.45$ και $X_2 > 3.45$ και $X_1 \le 6.5$, τότε η κατηγορία είναι η c_1, ή
  R6: Αν $X_1 > 5.45$ και $X_2 > 3.45$ και $X_1 > 6.5$, τότε η κατηγορία είναι η c_2.

Μέτρα αποτίμησης σημείων διαμερισμού: Εντροπία

- Η πρώτη προφανής επιλογή μας είναι ένα σημείο διαμερισμού το οποίο παράγει τον καλύτερο διαχωρισμό ή διάκριση των ετικετών για τις διαφορετικές κατηγορίες.

- Η εντροπία μετρά την ποσότητα της αταξίας ή αβεβαιότητας σε ένα σύστημα.

- Μια διαμέριση έχει χαμηλότερη εντροπία (ή αταξία) αν είναι σχετικά «καθαρή», δηλαδή αν τα περισσότερα από τα σημεία έχουν την ίδια ετικέτα.

- Από την άλλη πλευρά, μια διαμέριση έχει υψηλότερη εντροπία (ή αταξία) αν οι ετικέτες των κατηγορίες είναι ανάμεικτες, με αποτέλεσμα να μην προκύπτει πλειοψηφική κατηγορία.

- Αν μια περιοχή είναι «καθαρή», δηλαδή περιλαμβάνει σημεία από την ίδια κατηγορία, τότε η εντροπία της είναι μηδενική.

$$Entropy(t) = -\sum_{j=1}^{c} p(j \mid t) \log_2 p(j \mid t)$$

- ✓ p(j | t) σχετική συχνότητα της κλάσης j στον κόμβο t
- ✓ c αριθμός κλάσεων

C1	**0**
C2	**6**
Entropy=0.000	

C1	**1**
C2	**5**
Entropy=0.650	

C1	**3**
C2	**3**
Entropy = 1.000	

Κέρδος Πληροφορίας (Gain)

- Και σε αυτήν την περίπτωση, όταν ένας κόμβος p διασπάται σε k σύνολα (παιδιά), η ποιότητα του διαχωρισμού υπολογίζεται ως:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

- ✓ ni = αριθμός εγγραφών του παιδιού i,
- ✓ n = αριθμός εγγραφών του κόμβου p.

– Όσο υψηλότερη είναι η τιμή του κέρδους πληροφορίας, τόσο μεγαλύτερη είναι η μείωση της εντροπίας, και άρα τόσο καλύτερο είναι το σημείο διαμερισμού.

– Μπορούμε να βαθμολογήσουμε κάθε σημείο διαμερισμού και να επιλέξουμε εκείνο το οποίο παράγει το υψηλότερο πληροφοριακό κέρδος.

Μέτρα αποτίμησης σημείων διαμερισμού: Δείκτης Gini

Δείκτης Gini:

– Ο δείκτης Gini ορίζεται ως $$Gini(t) = 1 - \sum_{j=1}^{c} p(j \mid t)^2$$

- p(j | t) σχετική συχνότητα της κλάσης j στον κόμβο t (ποσοστό εγγραφών της κλάσης j στον κόμβο t)
- c αριθμός κλάσεων

– Αν η διαμέριση είναι «καθαρή», τότε ο δείκτης Gini θα είναι 0.

Παραδείγματα:

C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

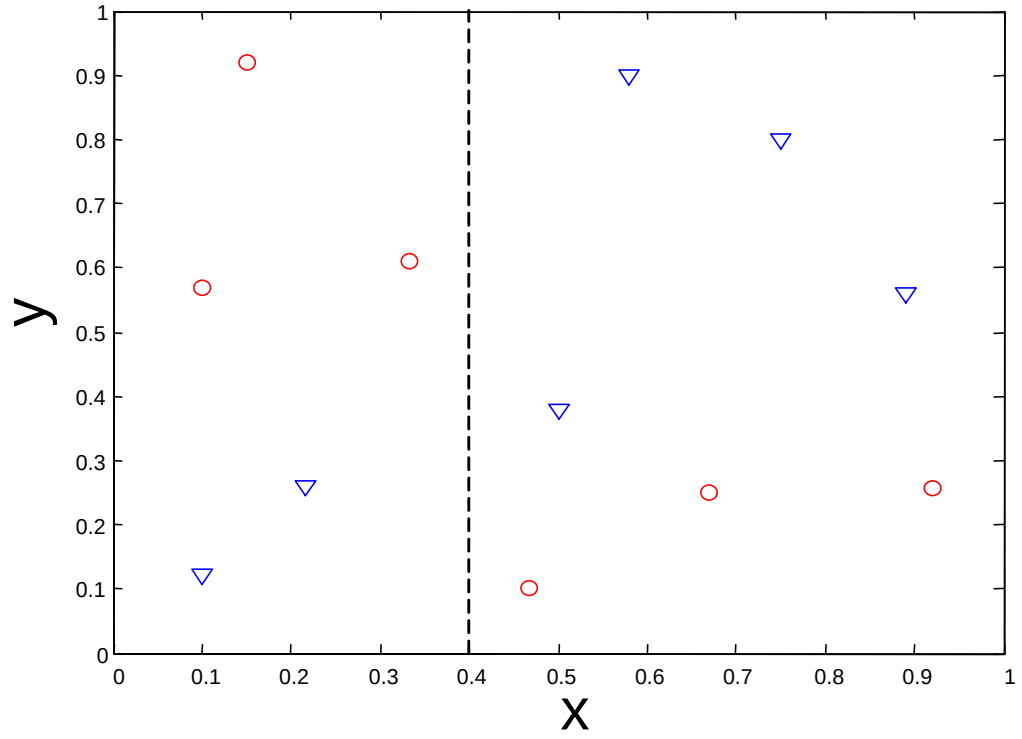
C1	**3**
C2	**3**
Gini=0.500	

Μέτρα αποτίμησης σημείων διαμερισμού: Δείκτης Gini

– Όταν ένας κόμβος p διασπάται σε k κόμβους (παιδιά), (που σημαίνει ότι το σύνολο των εγγραφών του κόμβου χωρίζεται σε k υποσύνολα), η ποιότητα του διαχωρισμού υπολογίζεται ως:

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i)$$

- όπου, ni = αριθμός εγγραφών του παιδιού i,
- n = αριθμός εγγραφών του κόμβου p.



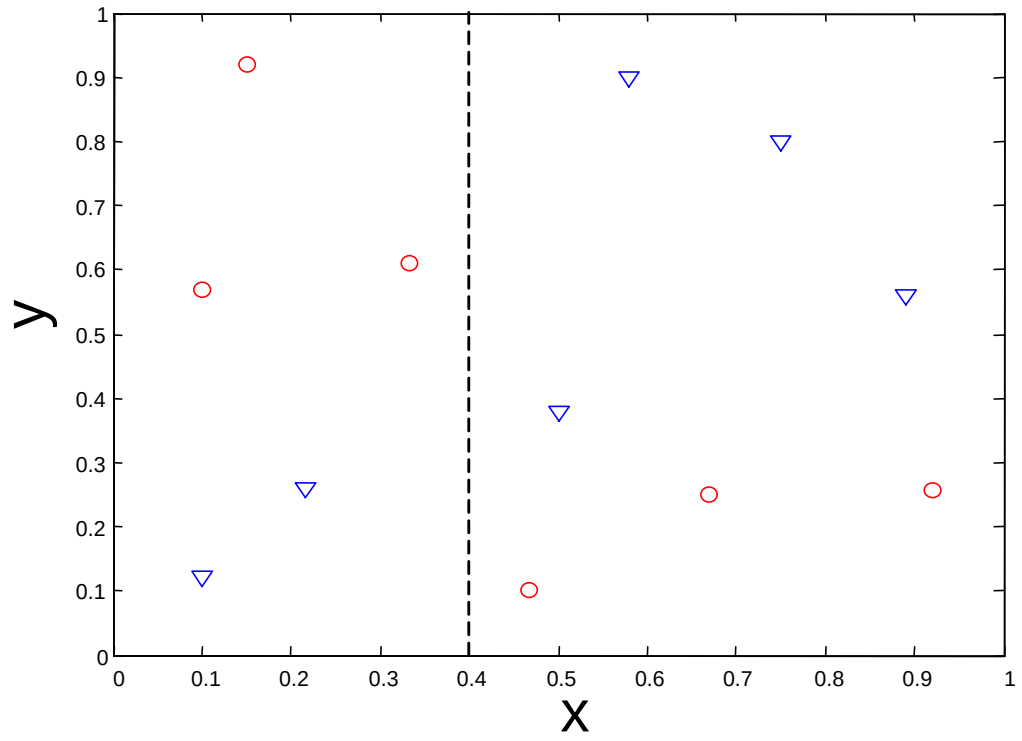
Δείκτης Gini για το διαχωρισμό στο σημείο X = 0.4 ???

Μέτρα αποτίμησης σημείων διαμερισμού: Δείκτης Gini

– Όταν ένας κόμβος p διασπάται σε k κόμβους (παιδιά), (που σημαίνει ότι το σύνολο των εγγραφών του κόμβου χωρίζεται σε k υποσύνολα), η ποιότητα του διαχωρισμού υπολογίζεται ως:

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i)$$

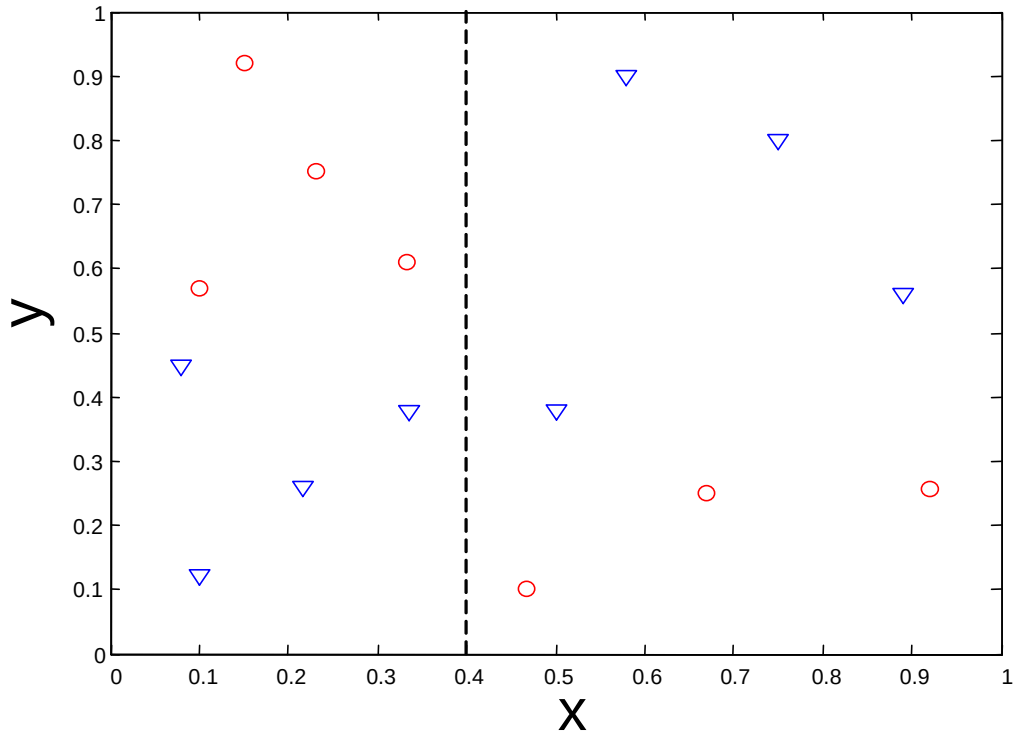
- όπου, ni = αριθμός εγγραφών του παιδιού i,
- n = αριθμός εγγραφών του κόμβου p.



$$Gini_{split} = 0.486$$

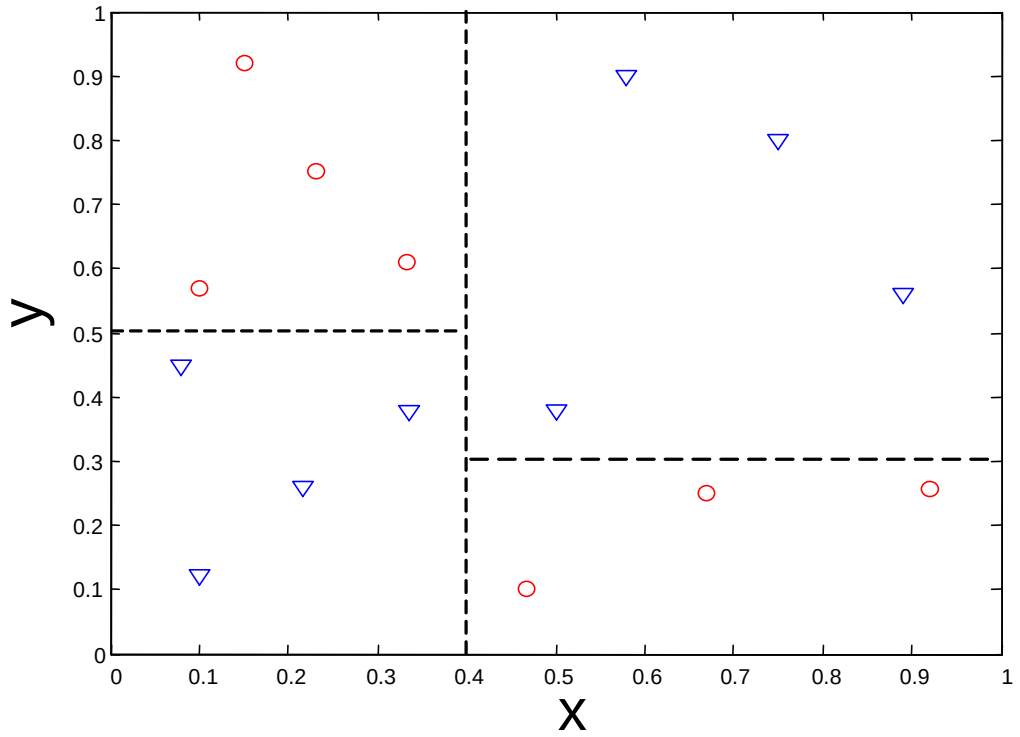
Άσκηση

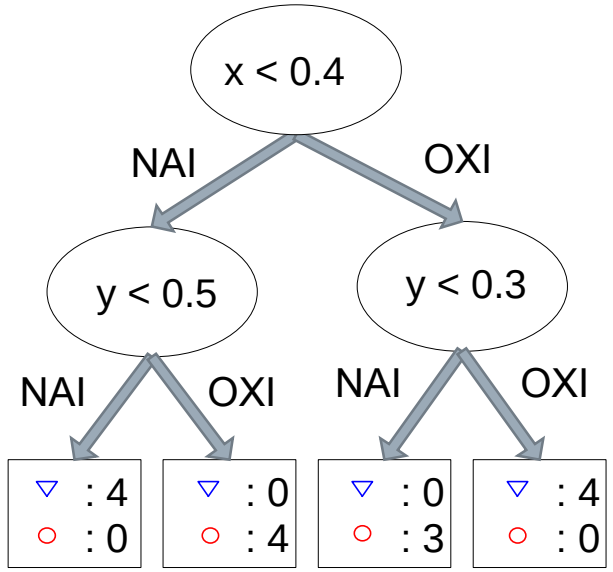
- Έστω η ρίζα του δέντρου στο X = 0.4.
- Κατασκευάστε ένα δένδρο αποφάσεων χρησιμοποιώντας ως κατώφλι καθαρότητας τη τιμή 100%.



Άσκηση

- Έστω η ρίζα του δέντρου στο X = 0.4.
- Κατασκευάστε ένα δένδρο αποφάσεων χρησιμοποιώντας ως κατώφλι καθαρότητας τη τιμή 100%.





Άσκηση

- Κατασκευάστε ένα δένδρο αποφάσεων χρησιμοποιώντας ως κατώφλι καθαρότητας τη τιμή 100%. Χρησιμοποιείστε στο κέρδος πληροφορίας ως μέτρο αποτίμησης των σημείων διαμερισμού.
- Κατηγοριοποιείστε το σημείο (Age = 27, Car = Vintage)

Point	Age	Car	**Risk**
$\mathbf{x}_1$	25	Sports	*L*
$\mathbf{x}_2$	20	Vintage	*H*
$\mathbf{x}_3$	25	Sports	*L*
$\mathbf{x}_4$	45	SUV	*H*
$\mathbf{x}_5$	20	Sports	*H*
$\mathbf{x}_6$	25	SUV	*H*