

# ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ (CLASSIFICATION)



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS



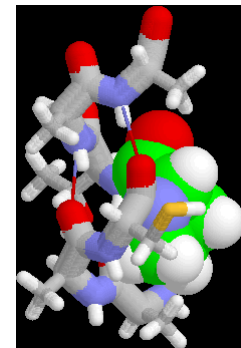
*Γεράσιμος Ραζής (razis@uth.gr)*

# Τι είναι η Κατηγοριοποίηση (classification);

Το γενικό πρόβλημα της ανάθεσης ενός αντικειμένου σε μία ή περισσότερες προκαθορισμένες κατηγορίες (κλάσεις)

## Παραδείγματα

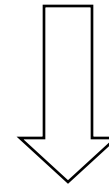
- Εντοπισμός spam emails
- Πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη
- Κατηγοριοποίηση συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης
- Κατηγοριοποίηση δευτερευόντων δομών πρωτεΐνης ως alpha-helix, beta-sheet, ή random coil
- Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού, κλπ



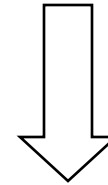
# Τι είναι η Κατηγοριοποίηση (classification);

**Κατηγοριοποίηση** είναι η διαδικασία εκμάθησης μιας συνάρτησης στόχου (target function)  $f$  (μοντέλο) που απεικονίζει κάθε σύνολο γνωρισμάτων  $x$  σε μια από τις προκαθορισμένες ετικέτες κλάσεις  $y$

Σύνολο εγγραφών  
( $x$ )



Μοντέλο  
Κατηγοριοποίησης



Ετικέτα κλάσης  
( $y$ )

# Τι είναι η Κατηγοριοποίηση (classification);

- **Είσοδος (x):** συλλογή από εγγραφές
- Κάθε εγγραφή περιέχει ένα σύνολο από γνωρίσματα/χαρακτηριστικά (attributes)
- Ένα από τα γνωρίσματα είναι η κλάση/κατηγορία (class)
- **Έξοδος (y):** ένα μοντέλο για το γνώρισμα «κλάση» ως μια συνάρτηση των τιμών των άλλων γνωρισμάτων

κατηγορικό      κατηγορικό      συνεχές      κλάση

<i>Tid</i>	Επιστροφή	Οικογενειακή Κατάσταση	Φορολογητέο Εισόδημα	Απάτη
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# Εισαγωγή

## Είδη μοντέλων κατηγοριοποίησης

- **Περιγραφικό μοντέλο (descriptive modeling)**: ως επεξηγηματικό εργαλείο
  - ποια χαρακτηριστικά κάνουν ένα ζώο να χαρακτηριστεί ως θηλαστικό
- **Μοντέλο πρόβλεψης (predictive modeling)**: για την πρόβλεψη της κλάσης άγνωστων εγγραφών
  - δοσμένων των χαρακτηριστικών κάποιου ζώου να προβλέψουμε αν είναι θηλαστικό, πτηνό, ερπετό ή αμφίβιο

# Εισαγωγή

Συνήθως το σύνολο δεδομένων εισόδου χωρίζεται σε

- ένα σύνολο εκπαίδευσης (training set), και
- ένα σύνολο ελέγχου (test set)

Το *σύνολο εκπαίδευσης* χρησιμοποιείται για να **κατασκευαστεί** το μοντέλο, ενώ το *σύνολο ελέγχου* για να την **επικύρωση** του μοντέλου

# Εισαγωγή

- Θεωρούμε ότι τιμή (ετικέτα) της κλάσης (γνώρισμα  $y$ ) είναι διακριτή τιμή
- Αν όχι, regression (παλινδρόμηση) όπου το γνώρισμα  $y$  παίρνει *συνεχείς τιμές*
- Κατάλληλη κυρίως για
  - δυαδικές κατηγορίες ή κατηγορίες για τις οποίες δεν υπάρχει διάταξη [διακριτές (nominal) vs διατεταγμένες (ordinal)]
  - για μη ιεραρχικές κατηγορίες

# Βήματα Κατηγοριοποίησης

## 1. Κατασκευή Μοντέλου

- Χρησιμοποιώντας το σύνολο εκπαίδευσης (στις εγγραφές του το γνώρισμα της κλάσης είναι προκαθορισμένο)

## 2. Εφαρμογή Μοντέλου για την κατηγοριοποίηση μελλοντικών ή άγνωστων αντικειμένων

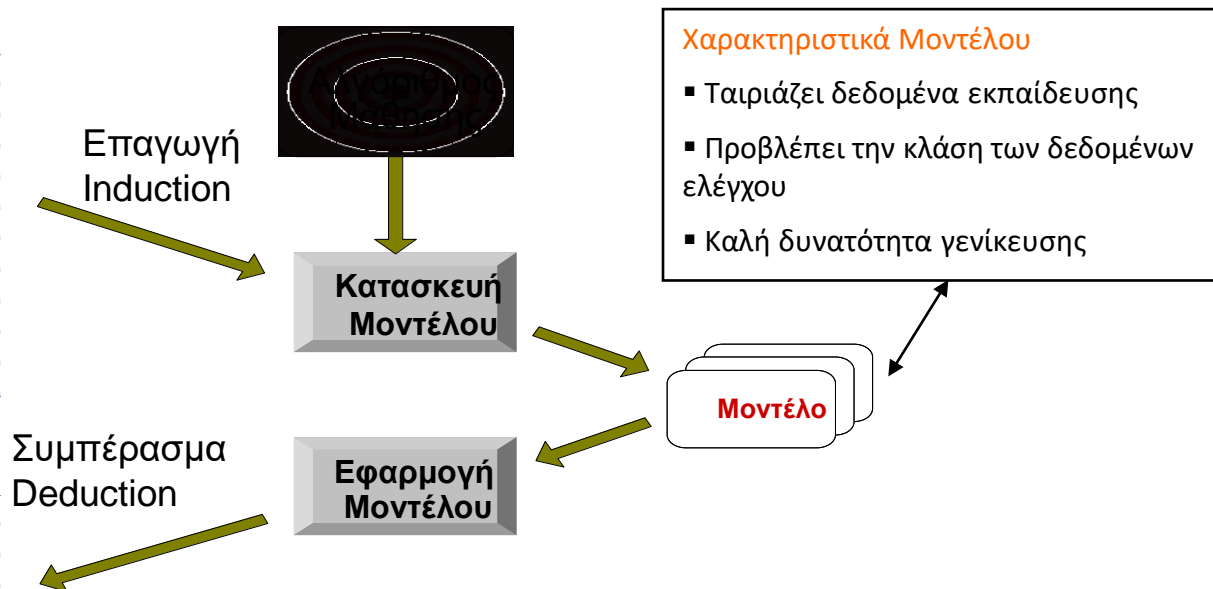
- Εκτίμηση της ακρίβειας του μοντέλου με χρήση συνόλου ελέγχου
- Ρυθμός ακρίβειας: το ποσοστό των εγγραφών του συνόλου ελέγχου που ταξινομούνται σωστά από το μοντέλο

Tid	Attrih1	Attrih2	Attrih3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σύνολο Εκπαίδευσης

Tid	Attrih1	Attrih2	Attrih3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σύνολο Ελέγχου





# Προεπεξεργασία

## 1. Καθαρισμός Δεδομένων (Data cleansing)

- Προεπεξεργασία δεδομένων και χειρισμός τιμών που λείπουν (π.χ. τις αγνοούμε ή τις αντικαθιστούμε με ειδικές τιμές)

## 2. Ανάλυση Σχετικότητας (Relevance analysis) (επιλογή χαρακτηριστικών (γνωρισμάτων) - feature selection)

- Απομάκρυνση των μη σχετικών ή περιττών γνωρισμάτων

## 3. Μετασχηματισμοί Δεδομένων (Data transformation)

- Κανονικοποίηση ή/και Γενίκευση
  - Πιθανών αριθμητικών γνωρισμάτων → κατηγορικά {low, medium, high}
  - Κανονικοποίηση αριθμητικών δεδομένων στο [0,1)

# Πόσο καλός είναι ο Κατηγοριοποιητής;

- Προβλεπόμενη ακρίβεια (Predictive accuracy)
- Ταχύτητα (speed)
  - Χρόνος κατασκευής (εκπαίδευσης) του μοντέλου
  - Χρόνος χρήσης/εφαρμογής του μοντέλου
- Ευρωστία (Robustness)
  - Χειρισμός θορύβου και τιμών που λείπουν
- Κλιμακωσιμότητα (Scalability)
  - Αποδοτικότητα σε βάσεις δεδομένων αποθηκευμένες στο δίσκο
- Ερμηνευσιμότητα (Interpretability)
  - Πόσο κατανοητό είναι το μοντέλο και τι νέα πληροφορία προσφέρει
- Ποιότητα (quality)

# Τεχνικές Κατηγοριοποίησης

Τεχνικές κατηγοριοποίησης βασισμένες σε

- Δέντρα Απόφασης (Decision Trees)
- Κανόνες (Rule-based Methods)
- Αλγόριθμοι Κοντινότερου Γείτονα (k Nearest Neighbor)
- Memory-based reasoning
- Νευρωνικά Δίκτυα (Neural Networks)
- Naïve Bayes και Bayesian Belief Δίκτυα
- Support Vector Machines



# Decision Trees (Δέντρα Απόφασης)

Εργαλείο υποστήριξης αποφάσεων που χρησιμοποιεί ένα μοντέλο αποφάσεων (μοιάζει με δέντρο) και τις πιθανές συνέπειές τους

# Δέντρα Απόφασης

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

# Παράδειγμα

Δεδομένα Εκπαίδευσης

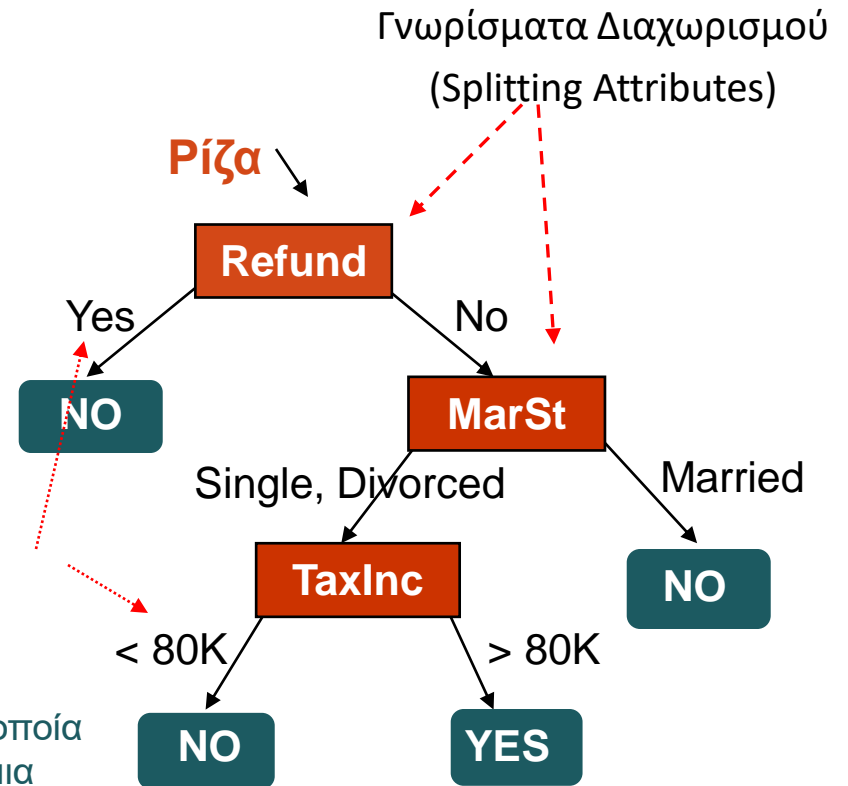
κατηγορικό  
κατηγορικό  
συνεχές  
κλάση

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Ετικέτα =  
συνθήκη

Φύλλα στα οποία  
αντιστοιχεί μια  
(ετικέτα) κλάση



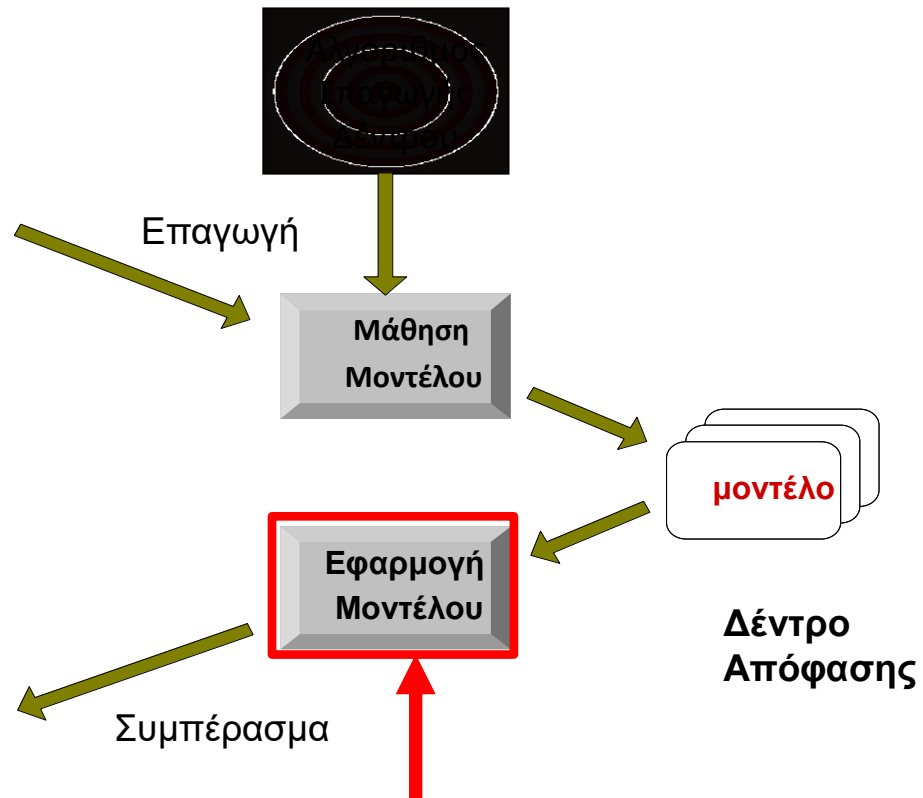
# Βήματα

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σύνολο Εκπαίδευσης

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σύνολο Ελέγχου



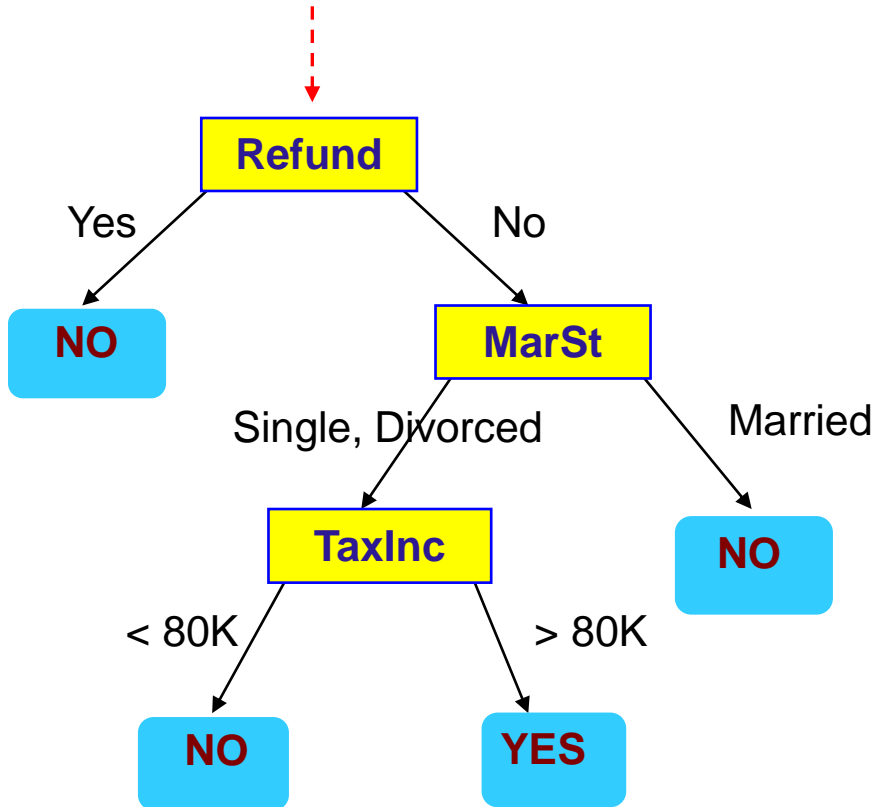
Αφού κατασκευαστεί το δέντρο, η εφαρμογή (χρήση) του στην κατηγοριοποίηση νέων εγγραφών είναι απλή → Διάσχιση από τη ρίζα στα φύλλα του

# Εφαρμογή Μοντέλου

Ξεκίνα από τη ρίζα του δέντρου

Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

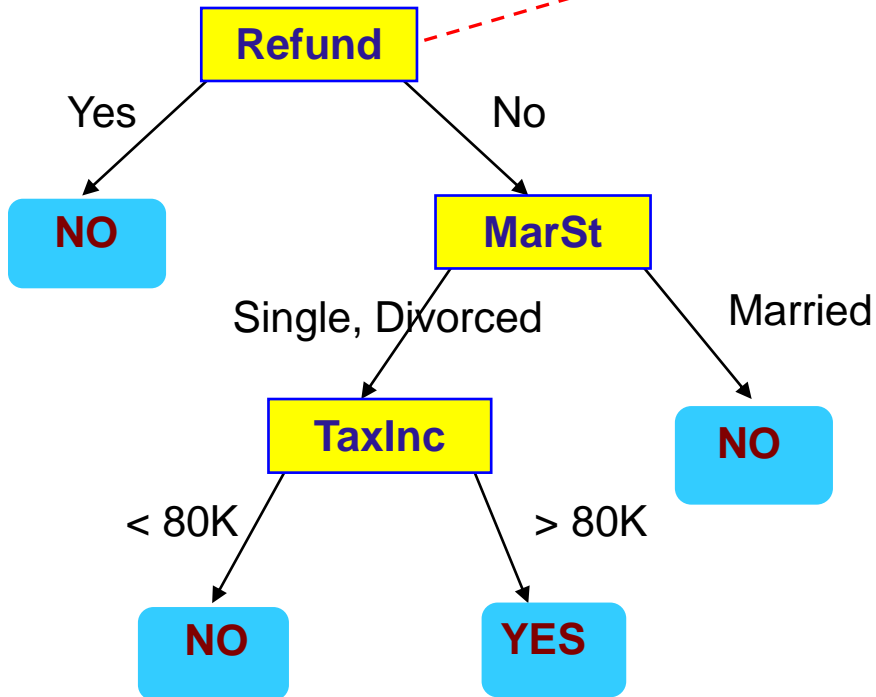




# Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

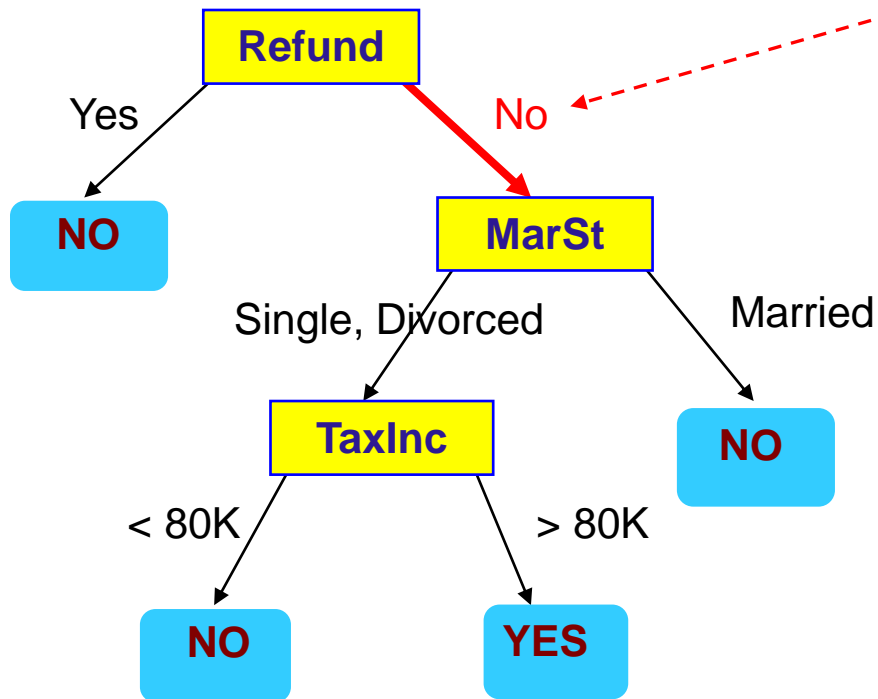
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



# Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

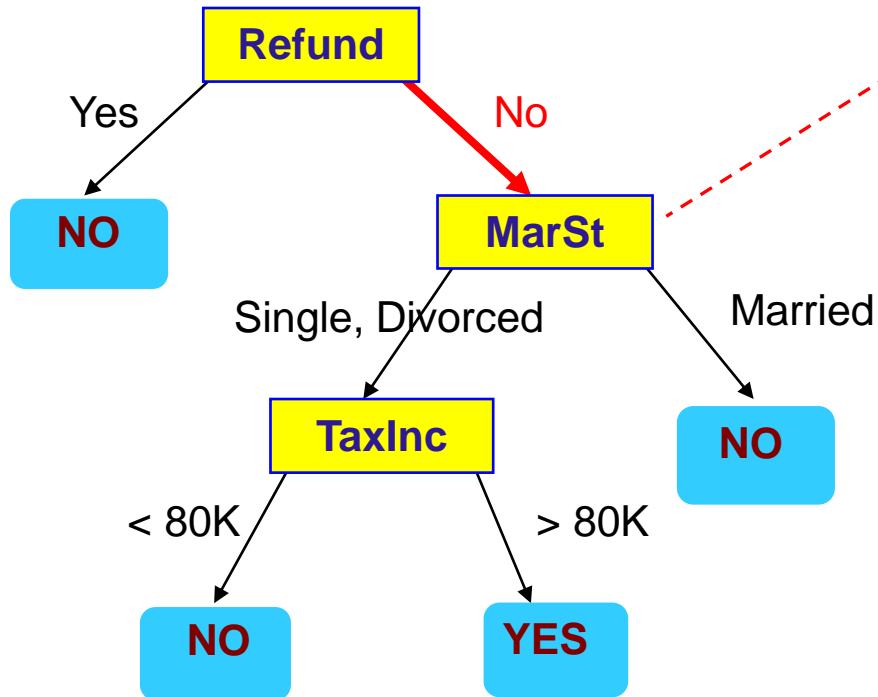
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



# Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

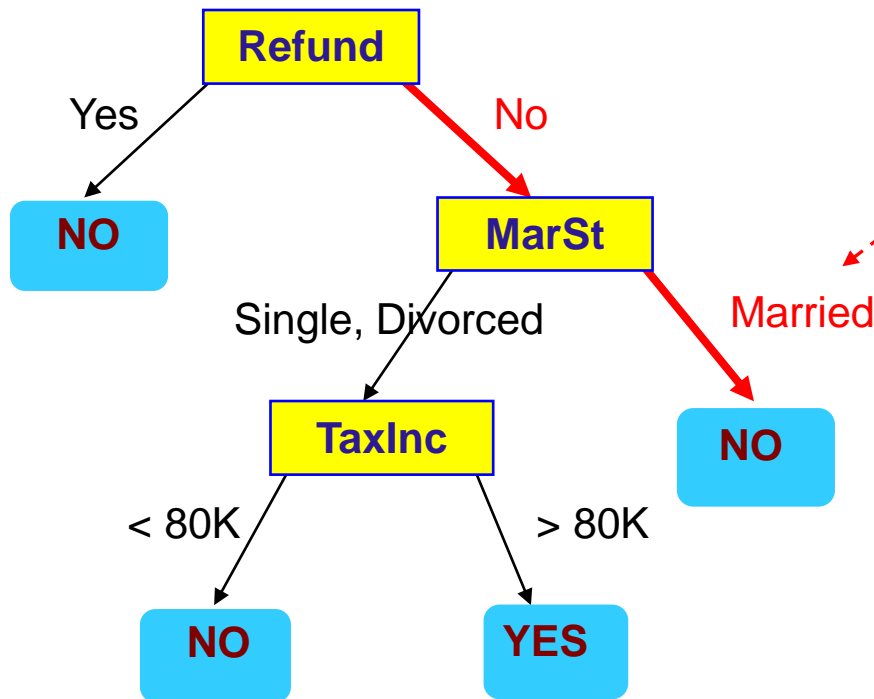
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



# Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

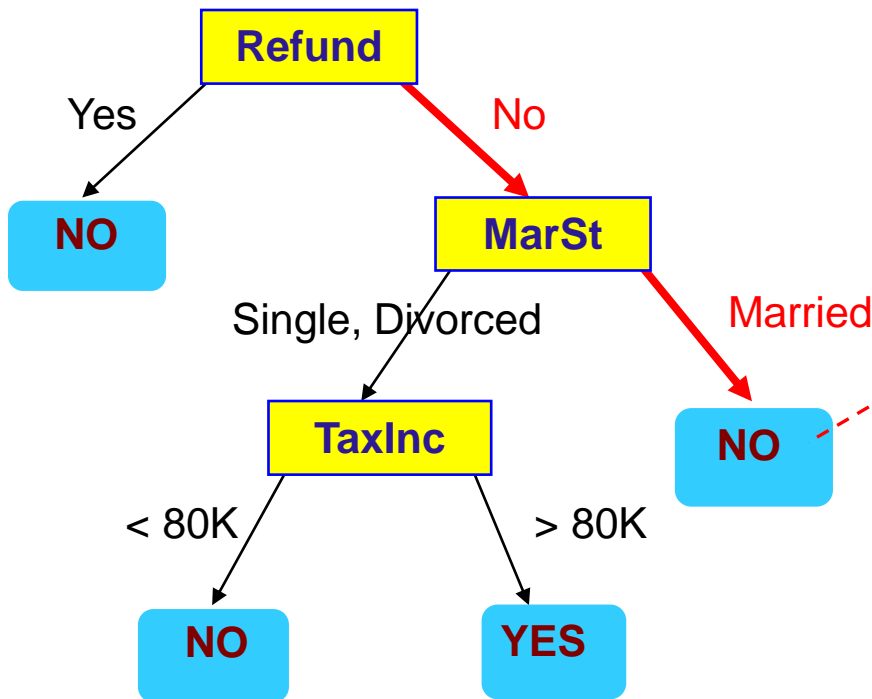
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



# Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Ανάθεση στο Cheat "No"

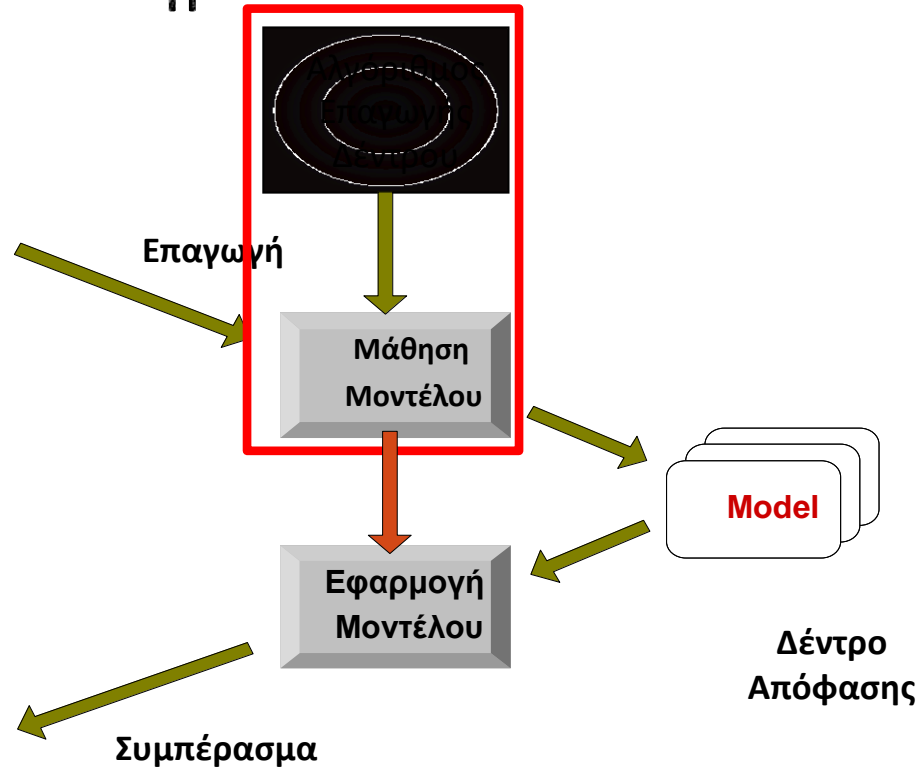
# Βήματα

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Σύνολο Εκπαίδευσης

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Σύνολο Ελέγχου



Στη συνέχεια, θα δούμε αλγορίθμους για την κατασκευή του δέντρου

# Δέντρα Απόφασης - Κατασκευή

Κατασκευή του δέντρου (συνοπτικά):

1. Ξεκίνα με έναν κόμβο που περιέχει όλες τις εγγραφές
2. **Διάσπαση** του κόμβου (μοίρασμα των εγγραφών) με βάση μια συνθήκη διαχωρισμού σε κάποιο από τα γνωρίσματα
3. Αναδρομική κλήση του Βήματος 2 σε κάθε κόμβο
4. Αφού κατασκευαστεί το δέντρο, κάποιες βελτιστοποιήσεις (tree pruning)

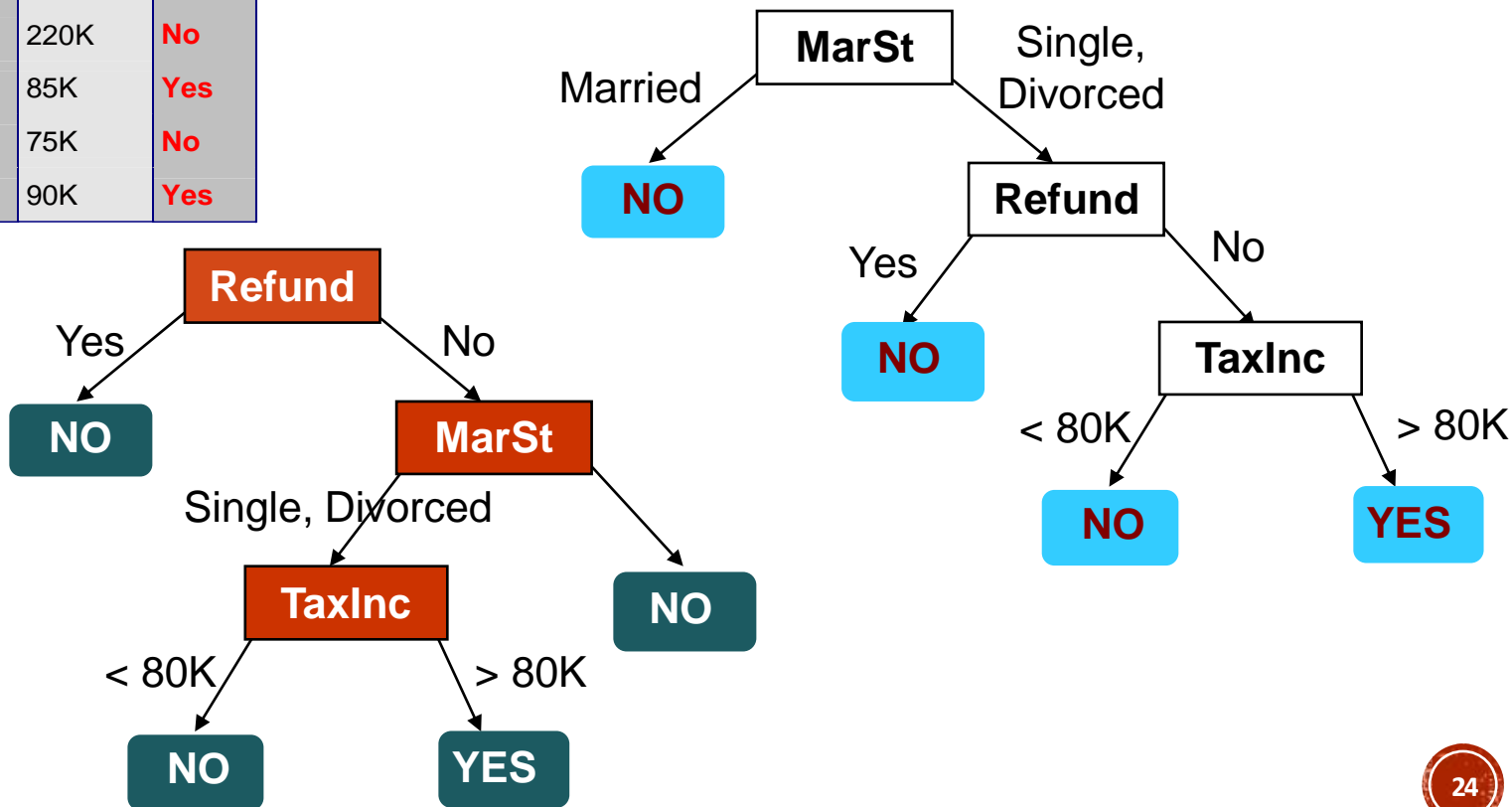
Το βασικό θέμα είναι:

*Ποιο γνώρισμα-συνθήκη διαχωρισμού να χρησιμοποιήσουμε για τη διάσπαση των εγγραφών κάθε κόμβου;*

# Παράδειγμα

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Για το ίδιο σύνολο εκπαίδευσης υπάρχουν διαφορετικά δέντρα





# Διάσπαση Δεδομένων (Data Fragmentation)

- Ο αριθμός των εγγραφών μειώνεται όσο κατεβαίνουμε στο δέντρο
- Ο αριθμός των εγγραφών στα φύλλα μπορεί να είναι πολύ μικρός για να πάρουμε οποιαδήποτε στατιστικά σημαντική απόφαση
- Μπορούμε να αποτρέψουμε την περαιτέρω διάσπαση όταν ο αριθμός των εγγραφών πέσει κάτω από ένα όριο

# Πλεονεκτήματα Δέντρων Απόφασης

- Μη παραμετρική προσέγγιση: Δε στηρίζεται σε υπόθεση εκ των προτέρων γνώσης σχετικά με τον τύπο της κατανομής πιθανότητας που ικανοποιεί η κλάση ή τα άλλα γνωρίσματα
- Η κατασκευή του βέλτιστου δέντρου απόφασης είναι ένα NP-complete πρόβλημα
- Ευρετικοί: Αποδοτική κατασκευή ακόμα και στην περίπτωση πολύ μεγάλου συνόλου δεδομένων
- Αφού το δέντρο κατασκευαστεί, η κατηγοριοποίηση νέων εγγραφών είναι πολύ γρήγορη  $O(h)$ , όπου  $h$  το μέγιστο ύψος του δέντρου
- Εύκολα στην κατανόηση (ιδιαίτερα τα μικρά δέντρα)
- Η ακρίβεια τους συγκρίσιμη με άλλες τεχνικές για μικρά σύνολα δεδομένων
- Καλή συμπεριφορά στο θόρυβο
- Η ύπαρξη πλεοναζόντων γνωρισμάτων (γνωρίσματα των οποίων η τιμή εξαρτάται από κάποιο άλλο) δεν είναι καταστροφική για την κατασκευή
  - Χρησιμοποιείται ένα από τα δύο
  - Αν πάρα πολλά, μπορεί να οδηγήσουν σε δέντρα πιο μεγάλα από ότι χρειάζεται

# Εκφραστικότητα

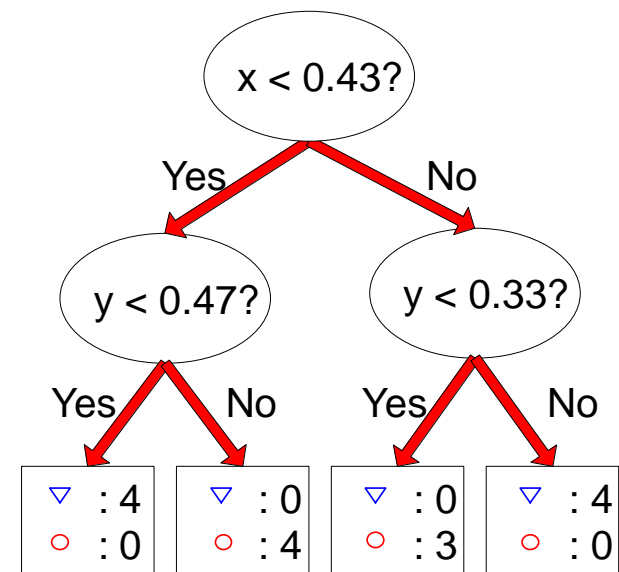
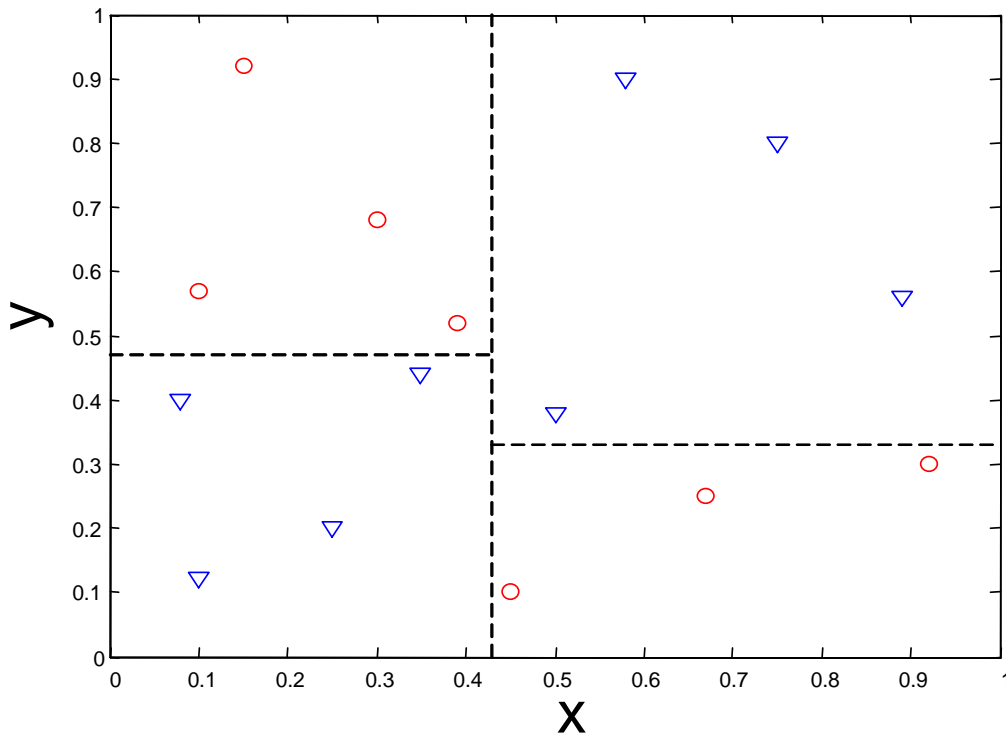
- Δυνατότητα αναπαράστασης για συναρτήσεις διακριτών τιμών, αλλά δε δουλεύουν σε κάποια είδη δυαδικών προβλημάτων
  - Π.χ. ισοτιμία (parity) -  $O(1)$  αν υπάρχει μονός (ζυγός) αριθμός από δυαδικά γνωρίσματα  $2^d$  κόμβοι για  $d$  γνωρίσματα
- Όχι καλή συμπεριφορά για συνεχείς μεταβλητές
  - Ιδιαίτερα όταν η συνθήκη ελέγχου αφορά ένα γνώρισμα τη φορά

# Σύνоро Απόφασης (Decision Boundary)

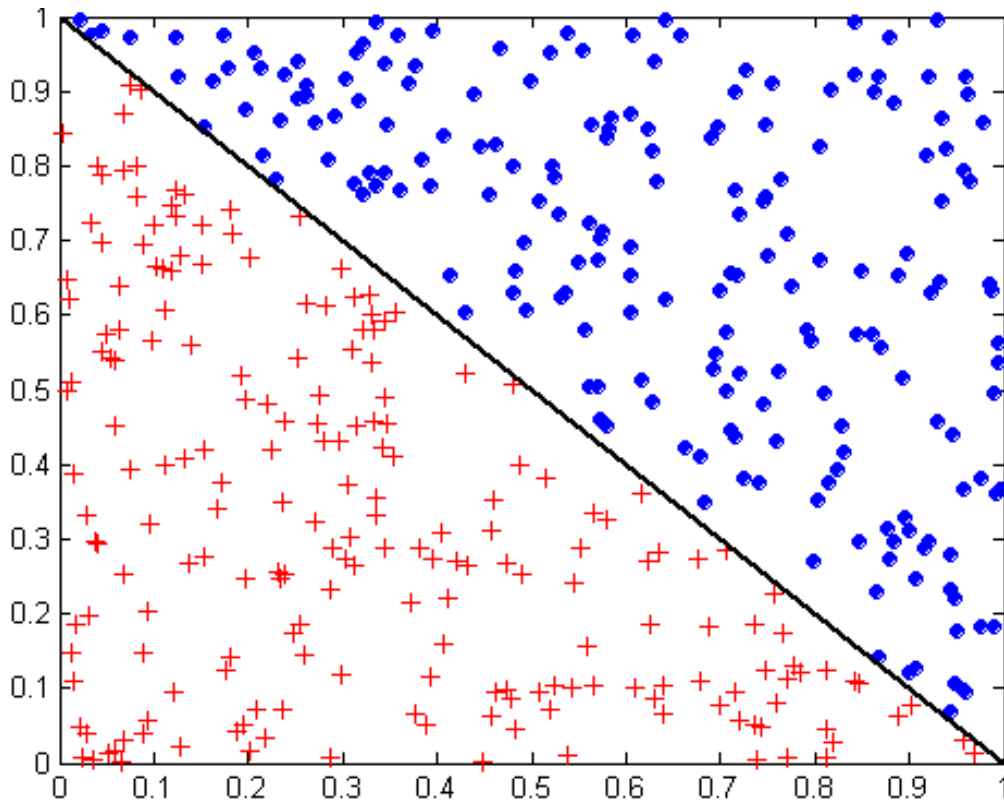
- Μπορούμε να δούμε τη διαδικασία ως τη διαδικασία *διαμερισμού του χώρου* των γνωρισμάτων σε ξένες περιοχές μέχρι κάθε περιοχή να περιέχει εγγραφές που να ανήκουν στην ίδια κλάση
- Η οριακή γραμμή (Border line) μεταξύ δυο γειτονικών περιοχών που ανήκουν σε διαφορετικές κλάσεις ονομάζεται και **σύνоро απόφασης (decision boundary)**

# Σύνορο Απόφασης (Decision Boundary)

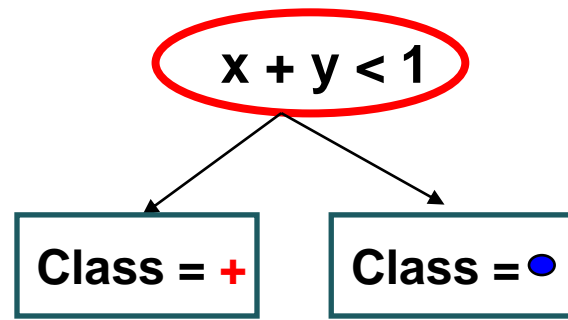
Όταν η συνθήκη ελέγχου περιλαμβάνει μόνο ένα γνώρισμα τη φορά τότε το Decision boundary είναι παράλληλη στους άξονες (τα σύνολα είναι ορθογώνια παραλληλόγραμμα)



# Σύνορο Απόφασης (Decision Boundary)



Oblique (πλάγιο) Δέντρο  
Απόφασης



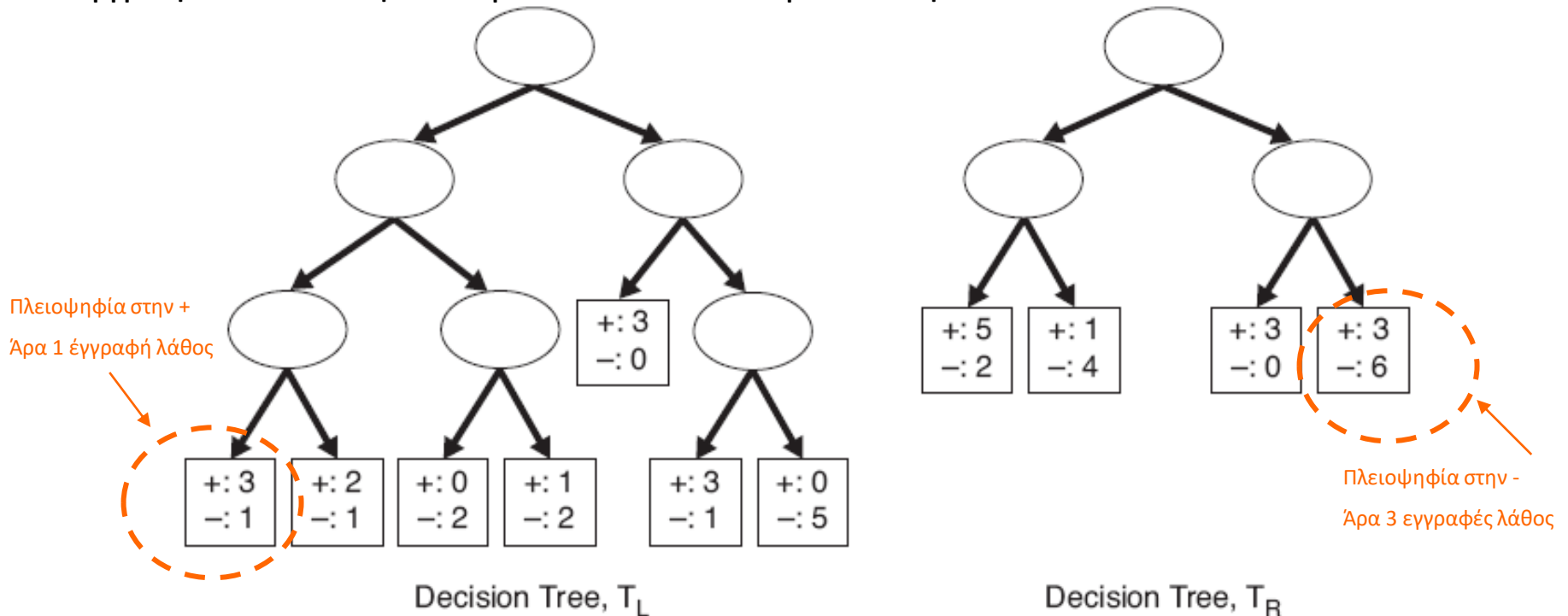
- Οι συνθήκες ελέγχου μπορούν να περιλαμβάνουν περισσότερα από ένα γνωρίσματα
- Μεγαλύτερη εκφραστικότητα
- Η εύρεση βέλτιστων συνθηκών ελέγχου είναι υπολογιστικά ακριβή

# Εκτίμηση του Λάθους

- Ως λάθος (σφάλμα) μετράμε τις εγγραφές που το μοντέλο τοποθετεί σε λάθος κλάση
- 2 ειδών λάθη
  - *Εκπαίδευσης* (training, re-substitution, apparent): λάθη κατηγοριοποίησης στα δεδομένα του συνόλου εκπαίδευσης (ποσοστό δεδομένων εκπαίδευσης που κατηγοριοποιούνται σε λάθος κλάση)
  - *Γενίκευσης* (generalization): τα αναμενόμενα λάθη κατηγοριοποίησης του μοντέλου σε δεδομένα που δεν έχει δει

# Εκτίμηση του Λάθους

Λάθη στα δεδομένα εκπαίδευσης, επειδή χρησιμοποιούμε την πλειοψηφία των εγγραφών σε ένα φύλλο για να αποδώσουμε κλάση



Παράδειγμα δύο δέντρων για τα ίδια δεδομένα εκπαίδευσης

Με βάση το λάθος εκπαίδευσης

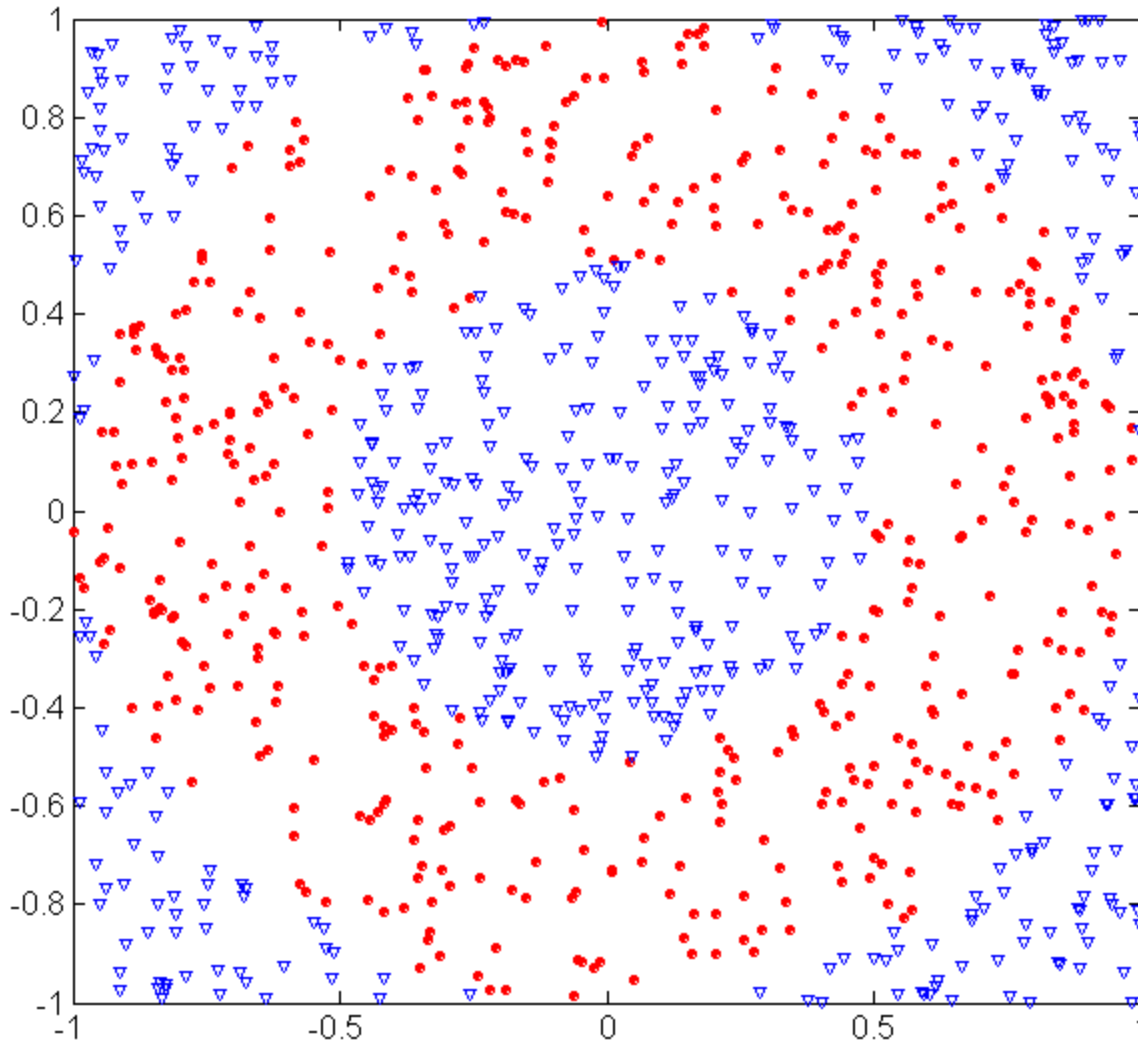
- Αριστερό  $4/24 = 0.167$
- Δεξί:  $6/24 = 0.25$



# Υπερπροσαρμογή (Overfitting)

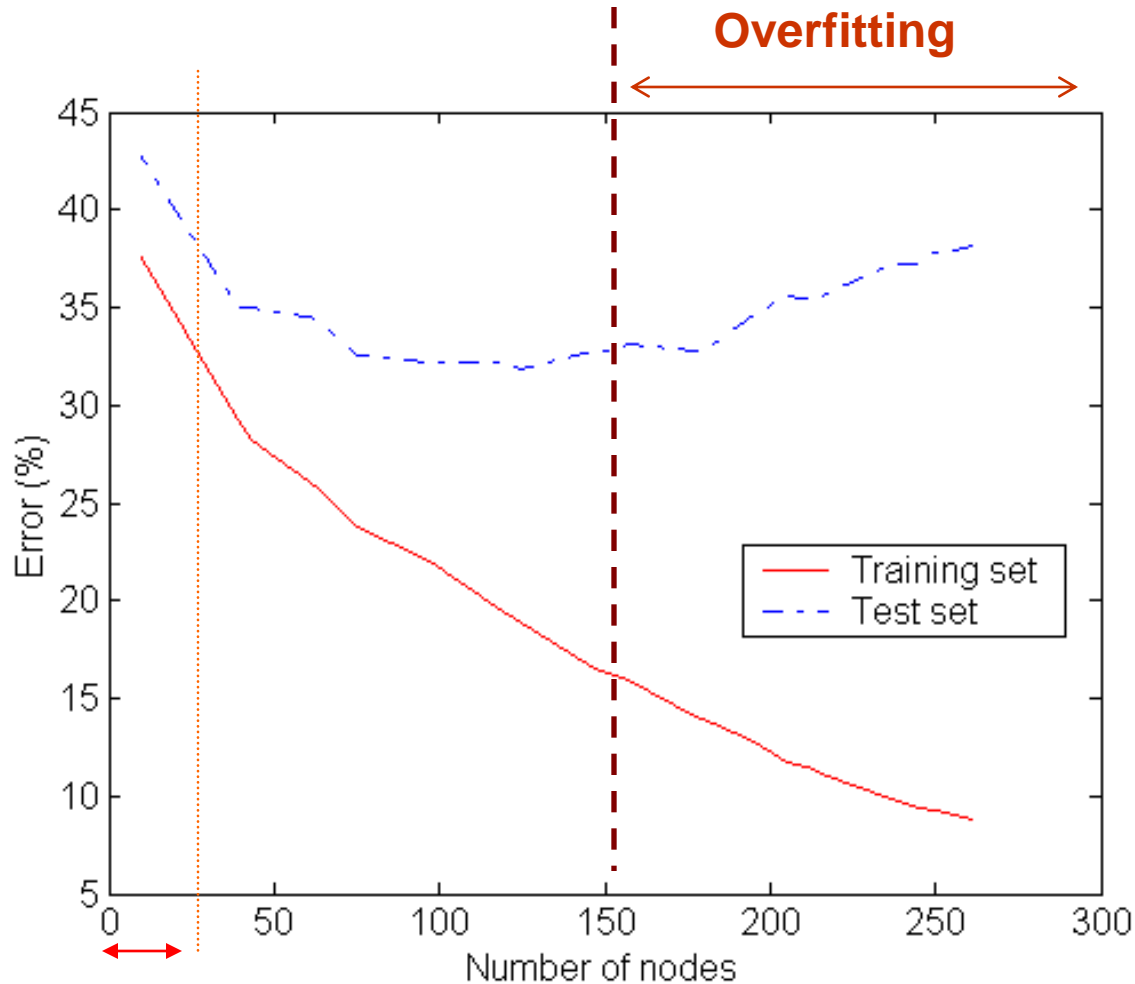
Μπορεί ένα μοντέλο που ταιριάζει πολύ καλά με τα δεδομένα εκπαίδευσης να έχει μεγαλύτερο λάθος γενίκευσης από ένα μοντέλο που ταιριάζει λιγότερο καλά στα δεδομένα εκπαίδευσης

# Υπερπροσαρμογή (Overfitting)



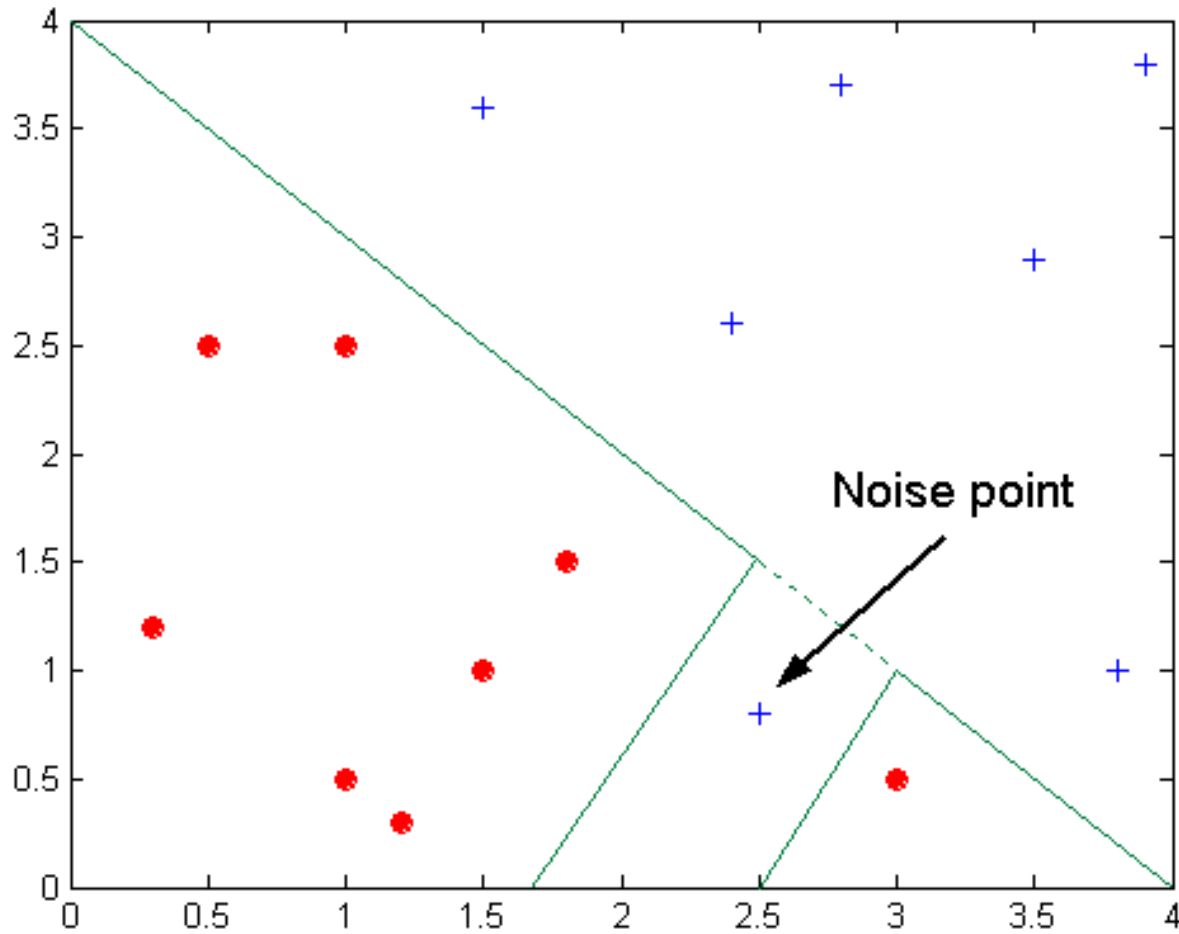
- Δύο κλάσεις:
  - κλάση 1 (500 κυκλικά σημεία)
  - κλάση 2 (500 τριγωνικά σημεία)
- Για τα σημεία της κλάσης 1 (κυκλικά σημεία):
  - $0.5 \leq \text{sqrt}(x_1^2 + x_2^2) \leq 1$
- Για τα σημεία της κλάσης 2 (τριγωνικά σημεία):
  - $\text{sqrt}(x_1^2 + x_2^2) > 0.5$  ή
  - $\text{sqrt}(x_1^2 + x_2^2) < 1$

# Υπερπροσαρμογή (Overfitting)



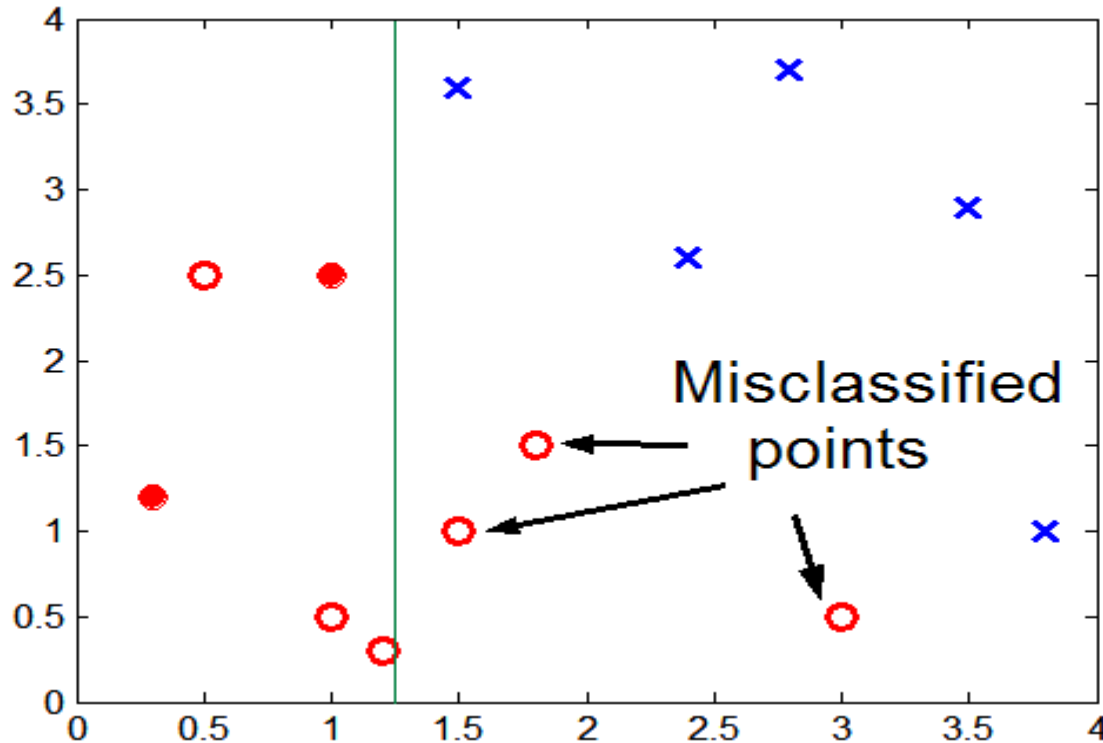
**Underfitting (υποπροσαρμογή):** όταν το μοντέλο είναι πολύ απλό και τα λάθη εκπαίδευσης και ελέγχου είναι μεγάλα

# Overfitting εξαιτίας Θορύβου



Το σημείο θορύβου επηρεάζει τη συνθήκη

# Overfitting εξαιτίας μη Επαρκών Δειγμάτων



- Κόκκινοι κύκλοι ανήκουν στην ίδια κλάση
- Οι γεμάτοι είναι στο σύνολο εκπαίδευσης, οι άδειοι στο σύνολο ελέγχου

- Η έλλειψη κόκκινων σημείων στο κάτω μισό του διαγράμματος κάνει δύσκολη την πρόβλεψη των κλάσεων σε αυτήν την περιοχή
- Μη επαρκής αριθμός εγγραφών εκπαίδευσης έχει ως αποτέλεσμα το δέντρο απόφασης να κάνει πρόβλεψη για τα σημεία αυτής της περιοχής χρησιμοποιώντας εγγραφές εκπαίδευσης μη σχετικές με το έργο της ταξινόμησης

# Υπερπροσαρμογή (Overfitting)

- Πρόβλημα λόγω πολλαπλών επιλογών
  - Επειδή σε κάθε βήμα εξετάζουμε πάρα πολλές διαφορετικές διασπάσεις κάποια διάσπαση βελτιώνει το δέντρο *κατά τύχη*
  - Το πρόβλημα χειροτερεύει όταν αυξάνει ο αριθμός των επιλογών και μειώνεται ο αριθμός των δειγμάτων (σημείων εκπαίδευσης)
- Το overfitting έχει ως αποτέλεσμα μοντέλα που είναι πιο περίπλοκα από ό,τι χρειάζεται
- Τα λάθη εκπαίδευσης δεν αποτελούν πια μια καλή εκτίμηση για τη συμπεριφορά του δέντρου σε εγγραφές που δεν έχει δει ξανά
- Νέοι μέθοδοι για την εκτίμηση του λάθους;

# Πολυπλοκότητα Μοντέλου

- Το Ξυράφι του Όκαμ (Occam's Razor)
  - «Κανείς δεν θα πρέπει να προβαίνει σε περισσότερες εικασίες από όσες είναι απαραίτητες»
- Δοθέντων δυο μοντέλων με παρόμοια λάθη γενίκευσης, πρέπει να προτιμάται το απλούστερο από το πιο περίπλοκο
  - Ένα πολύπλοκο μοντέλο είναι πιο πιθανό να έχει ταιριαστεί (fitted) τυχαία λόγω λαθών στα δεδομένα
- Για αυτό η πολυπλοκότητα του μοντέλου θα πρέπει να αποτελεί έναν από τους παράγοντες της αξιολόγησής του



# Εκτίμηση του Λάθους

- Κλάδεμα με Μειωμένο Σφάλμα (Reduced Error Pruning - REP)
  - Από τις απλούστερες μορφές κλαδέματος
  - Πλεονέκτημα της απλότητας και της ταχύτητας
  - Ξεκινώντας από τα φύλλα, κάθε κόμβος αντικαθίσταται με την πιο δημοφιλή κατηγορία του
    - Εάν η ακρίβεια της πρόβλεψης δεν επηρεάζεται, τότε η αλλαγή διατηρείται
- Χρήση ενός συνόλου επαλήθευσης για την εκτίμηση του λάθους γενίκευσης
  - Χώρισε τα δεδομένα εκπαίδευσης
    - $2/3$  εκπαίδευση
    - $1/3$  σύνολο επαλήθευσης (validation set) για υπολογισμό λάθους
- Χρήση για εύρεση του κατάλληλου μοντέλου





# Αποτίμηση Μοντέλου

- Precision (ακρίβεια)
- Recall (ανάκληση)
- F1 measure

# Μέτρα Εκτίμησης (κόστος)

- Όχι όλα τα σφάλματα το ίδιο σημαντικά  $\rightarrow$  «βάρη»
- Εισάγουμε την έννοια του Πίνακα Κόστους

Πίνακας Κόστους

	PREDICTED CLASS		
	$C(i j)$	Class = +	Class = -
ACTUAL CLASS	Class = +	$C(+, +)$	$C(+, -)$
	Class = -	$C(-, +)$	$C(-, -)$

- $C(i|j)$ : κόστος λανθασμένης κατηγοριοποίησης ενός παραδείγματος της κλάσης  $i$  ως κλάση  $j \rightarrow$  βάρος
  - Για  $C(+, +)$ ,  $C(-, -) \rightarrow$  αρνητικό
  - Για  $C(+, -)$ ,  $C(-, +) \rightarrow$  θετικό
- Αρνητική τιμή κόστους σημαίνει επιπρόσθετη «επιβράβευση» σωστής πρόβλεψης

# Μέτρα Εκτίμησης (κόστος) Παράδειγμα

Πίνακας κόστους	Predicted		
	C(i j)	+	-
Actual	+	-1	<b>10</b>
	-	<b>3</b>	0

# Πίνακας Σύγχυσης (Confusion Matrix)

		Εκτιμώμενη κλάση (PREDICTED CLASS)	
		Class=Yes	Class=No
Πραγματική κλάση (ACTUAL CLASS)	Class=Yes	TP	FN
	Class=No	FP	TN

**TP = True Positive**  
**FP = False Positive**  
**TN = True Negative**  
**FN = False Negative**

Ιδανικά = 0

Το πιο συνηθισμένο μέτρο

$$\text{Ακρίβεια - Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \rightarrow \text{Σωστές προβλέψεις}$$

$$\text{Ρυθμός σφάλματος - Error rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

$$\text{ErrorRate}(C) = 1 - \text{Accuracy}(C)$$

Τι συμβαίνει αν μια κατηγορία είναι σπάνια;

# Πίνακας Σύγχυσης (Confusion Matrix)

$f_{ij}$ : αριθμός των εγγραφών της κλάσης  $i$  που προβλέπονται ως κλάση  $j$

		πρόβλεψη PREDICTED CLASS	
		Class=Yes	Class=No
πραγματική ACTUAL CLASS	Class=Yes	$f_{11}$ <b>TP</b>	$f_{10}$ <b>FN</b>
	Class=No	$f_{01}$ <b>FP</b>	$f_{00}$ <b>TN</b>

**TP (true positive)  $f_{11}$**

**FN (false negative)  $f_{10}$**

**FP (false positive)  $f_{01}$**

**TN (true negative)  $f_{00}$**

$$\text{Accuracy} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{01} + f_{10}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error rate} = \frac{f_{01} + f_{10}}{f_{11} + f_{00} + f_{01} + f_{10}} = \frac{FP + FN}{TP + TN + FP + FN}$$

$$\text{ErrorRate}(C) = 1 - \text{Accuracy}(C)$$

# Μέτρα Εκτίμησης (κόστος)

Πίνακας Κόστους

	PREDICTED CLASS		
	C(i j)	Class = +	Class = -
ACTUAL CLASS	Class = +	C(+, +)	C(+, -)
	Class = -	C(-, +)	C(-, -)

Πίνακας Σύγχυσης

	PREDICTED CLASS		
	C(i j)	Class = +	Class = -
ACTUAL CLASS	Class = +	<b>TP</b> F <sub>11</sub>	<b>FN</b> F <sub>10</sub>
	Class = -	<b>FP</b> F <sub>01</sub>	<b>TN</b> F <sub>00</sub>

$$C(M) = \mathbf{TP} \times C(+, +) + \mathbf{FN} \times C(+, -) + \mathbf{FP} \times C(-, +) + \mathbf{TN} \times C(-, -)$$

- Στα προηγούμενα, είχαμε
- $C(+, +) = C(-, -) = 0$  -> όχι επιβράβευση
- $C(+, -) = C(-, +) = 1$  -> κάθε λάθος μετρά 1

# Παράδειγμα: Υπολογισμός του Κόστους της Κατηγοριοποίησης

$C(i|j)$ : κόστος λανθασμένης ταξινόμησης ενός παραδείγματος της κλάσης  $i$  ως κλάση  $j$

Cost Matrix	PREDICTED CLASS		
	$C(i j)$	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model $M_1$	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

210                      290                      400

Accuracy = 80%

Cost = 3910

Χάνει κάποια θετικά

Model $M_2$	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

255                      245                      400

Accuracy = 90%

Cost = 4255

# Κατηγοριοποίηση που λαμβάνει υπό όψιν της το Κόστος

## Κατασκευή Δέντρου Απόφασης

- Επιλογή γνωρίσματος στο οποίο θα γίνει η διάσπαση
- Στην απόφαση αν θα ψαλιδιστεί κάποιο υποδέντρο
- Στον καθορισμό της κλάσης του φύλλου



# Καθορισμός κλάσης

- Κανονικά, ως ετικέτα ενός φύλλου την πλειοψηφούσα κλάση
  - Έστω  $p(j)$  τον ποσοστό των εγγραφών του κόμβου που ανήκουν στην κλάση  $j$
  - Τότε: Leaf-label =  $\max p(j)$ , το ποσοστό των εγγραφών της κλάσης  $j$  που έχουν ανατεθεί στον κόμβο
- Τώρα, δίνουμε την κλάση  $i$  στον κόμβο που έχει το ελάχιστο

$$\sum_j p(j)C(j,i)$$

για όλες τις κλάσεις



# Κόστος vs Accuracy

- Αφού κατασκευαστεί ένα μοντέλο, θα θέλαμε να αξιολογήσουμε / εκτιμήσουμε την ποιότητα του / την ακρίβεια της κατηγοριοποίησης που πετυχαίνει
- Έμφαση στην *ικανότητα πρόβλεψης* του μοντέλου παρά στην αποδοτικότητα του (πόσο γρήγορα κατασκευάζει το μοντέλο ή ταξινομεί μια εγγραφή, κλιμάκωση κλπ)

# Κόστος vs Accuracy

- Μέτρα (metrics) για την εκτίμηση της απόδοσης του μοντέλου
  - Πώς να εκτιμήσουμε την απόδοση ενός μοντέλου
  - Τι θα μετρήσουμε (π.χ., είδαμε τα σφάλματα, ακρίβεια)
- Μέθοδοι για την εκτίμηση της απόδοσης
  - Πώς μπορούμε να πάρουμε αξιόπιστες εκτιμήσεις
  - Πώς θα το μετρήσουμε (π.χ. δεδομένα εκπαίδευσης, ελέγχου)
- Μέθοδοι για την σύγκριση μοντέλων
  - Πώς να συγκρίνουμε τη σχετική απόδοση δύο ανταγωνιστικών μοντέλων

Ισχύουν για όλα τα μοντέλα κατηγοριοποίησης (όχι μόνο για τα δέντρα απόφασης)

# Κόστος vs Accuracy

Cost	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	p	q
	Class=No	q	p

Η ακρίβεια (accuracy) είναι ανάλογη του κόστους αν:

1.  $C(\text{Yes} | \text{No}) = C(\text{No} | \text{Yes}) = q$
2.  $C(\text{Yes} | \text{Yes}) = C(\text{No} | \text{No}) = p$

Count	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

$$N = a + b + c + d$$

$$\text{Accuracy} = (a + d) / N$$

$$\text{Cost} = p(a + d) + q(b + c)$$

$$= p(a + d) + q(N - a - d)$$

$$= qN - (q - p)(a + d)$$

$$= N [q - (q - p) \times \text{Accuracy}]$$

# Μειονεκτήματα της Ακρίβειας (Accuracy)

- Θεωρείστε ένα πρόβλημα με 2 κλάσεις
  - Πλήθος οντοτήτων κλάσης «0»: 9990
  - Πλήθος οντοτήτων κλάσης «1»: 10
- Αν ένα μοντέλο προβλέπει οτιδήποτε ως κλάση «0», τότε
  - $accuracy = 9990/10000 = 99,9\%$
- Η accuracy είναι παραπλανητική γιατί το μοντέλο δεν προβλέπει κανένα παράδειγμα της κλάσης «1»

# Άλλες μετρήσεις με βάση τον Πίνακα Σύγκρισης

	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

**True Positive Rate** or **sensitivity** (ευαισθησία):

Το ποσοστό των θετικών οντοτήτων που κατηγοριοποιούνται σωστά

$$TPR = \frac{TP}{TP + FN}$$

**True Negative Rate** or **specificity** (ιδιαιτερότητα):

Το ποσοστό των αρνητικών οντοτήτων που κατηγοριοποιούνται σωστά

$$TNR = \frac{TN}{TN + FP}$$

# Άλλες μετρήσεις με βάση τον Πίνακα Σύγκυσης

	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	TP	FN
	Class=No	FP	TN

**False Positive Rate:** Το ποσοστό των αρνητικών οντοτήτων που κατηγοριοποιούνται λάθος

$$FPR = \frac{FP}{TN + FP}$$

**False Negative Rate:** Το ποσοστό των θετικών οντοτήτων που κατηγοριοποιούνται λάθος

$$FNR = \frac{FN}{TP + FN}$$



# Ακρίβεια (Precision) και Ανάκληση (Recall)

	PREDICTED CLASS		
	Class=Yes	Class=No	
ACTUAL CLASS	Class=Yes	<b>TP</b>	<b>FN</b>
	Class=No	<b>FP</b>	<b>TN</b>

Precision  $p = \frac{TP}{TP + FP}$

Πόσες από τις οντότητες που το μοντέλο έχει κατηγοριοποιήσει ως θετικές είναι πραγματικά θετικές. Όσο πιο μεγάλη η ακρίβεια, τόσο μικρότερος ο αριθμός των FP.

Recall  $r = \frac{TP}{TP + FN}$

Πόσες από τις θετικές οντότητες κατάφερε ο κατηγοριοποιητής να βρει. Όσο πιο μεγάλη η ανάκληση, τόσο λιγότερες θετικές οντότητες έχουν κατηγοριοποιηθεί λάθος (=TPR)

# Ακρίβεια (Precision) και Ανάκληση (Recall)

- Συχνά το ένα καλό είναι και το άλλο όχι
  - Π.χ. ένας κατηγοριοποιητής που όλα τα ταξινομεί ως θετικά, έχει την καλύτερη ανάκληση με τη χειρότερη ακρίβεια
- Πώς μπορούμε να συνδυάσουμε τις μετρικές;

# F<sub>1</sub> score

F<sub>1</sub> score (ή F<sub>1</sub> measure)

- Ο Αρμονικός μέσος (Harmonic mean) της Ακρίβειας και Ανάκλησης
- Τείνει να είναι πιο κοντά στο μικρότερο από τα δύο
- Υψηλή τιμή σημαίνει ότι και τα δύο είναι ικανοποιητικά μεγάλα
- Χρήσιμο ως μέσο ρυθμών (rate)

$$F_1 = \frac{2rp}{r+p} = \frac{2TP}{2TP + FP + FN} \quad \text{ή} \quad F_1 = \frac{2}{1/r + 1/p}$$

# F<sub>1</sub> score

$$\text{Precision (p)} = \frac{TP}{TP + FP}$$

$$\text{Recall (r)} = \frac{TP}{TP + FN}$$

$$\text{F - measure (F)} = \frac{2rp}{r + p} = \frac{2TP}{2TP + FN + FP}$$

$$\text{Weighted Accuracy} = \frac{w_1TP + w_4TN}{w_1TP + w_2FP + w_3FN + w_4TN}$$

	Ex_1	Ex_2
Recall	0,458	1
Precision	0,692	0,994
F1	0,551	0,997
Accuracy	0,88	0,999

# Micro και Macro $F_1$ score

- Τα Micro και Macro  $F_1$  score υπολογίζουν ελαφρώς διαφορετικά πράγματα
  - Επομένως η ερμηνεία τους διαφέρει
- Το Macro  $F_1$  score υπολογίζει τη μετρική ανεξάρτητα για κάθε κλάση και στη συνέχεια θα λάβει το μέσο όρο τους
  - Αντιμετωπίζει όλες τις κατηγορίες εξίσου
- Το Micro  $F_1$  score αξιοποιεί τις συνεισφορές όλων των κλάσεων για τον υπολογισμό της μετρικής

## Παράδειγμα

- Class A: 1 TP, 1 FP
- Class B: 10 TP, 90 FP
- Class C: 1 TP, 1 FP
- Class D: 1 TP, 1 FP
  
- $p_A = p_C = p_D = 0,5$  και  $p_B = 0,1$
- Macro  $F_1$  score:  $p = \frac{0,5+0,1+0,5+0,5}{4} = 0,4$
- Micro  $F_1$  score:  $p = \frac{1+10+1+1}{2+100+2+2} = 0,123$

# ROC - Receiver Operating Characteristic Curve (Χαρακτηριστική Καμπύλη Λειτουργίας Δέκτη)

- Αναπτύχθηκε στη δεκαετία 1950 για την ανάλυση θορύβου στα σήματα
- Χαρακτηρίζει το trade-off μεταξύ positive hits και false alarms
- Η καμπύλη ROC δείχνει τα TPR (στον άξονα των  $y$ ) προς τα FPR (στον άξονα των  $x$ )
- Η απόδοση κάθε μοντέλου αναπαρίσταται ως ένα σημείο στην καμπύλη ROC

True Positive Rate

Πόσα από τα θετικά βρίσκει

[πόσα από τα θετικά ταξινομεί σωστά]

$$\text{TPR} = \frac{TP}{TP + FN}$$

False Positive Rate

Πόσα από τα αρνητικά θεωρεί θετικά

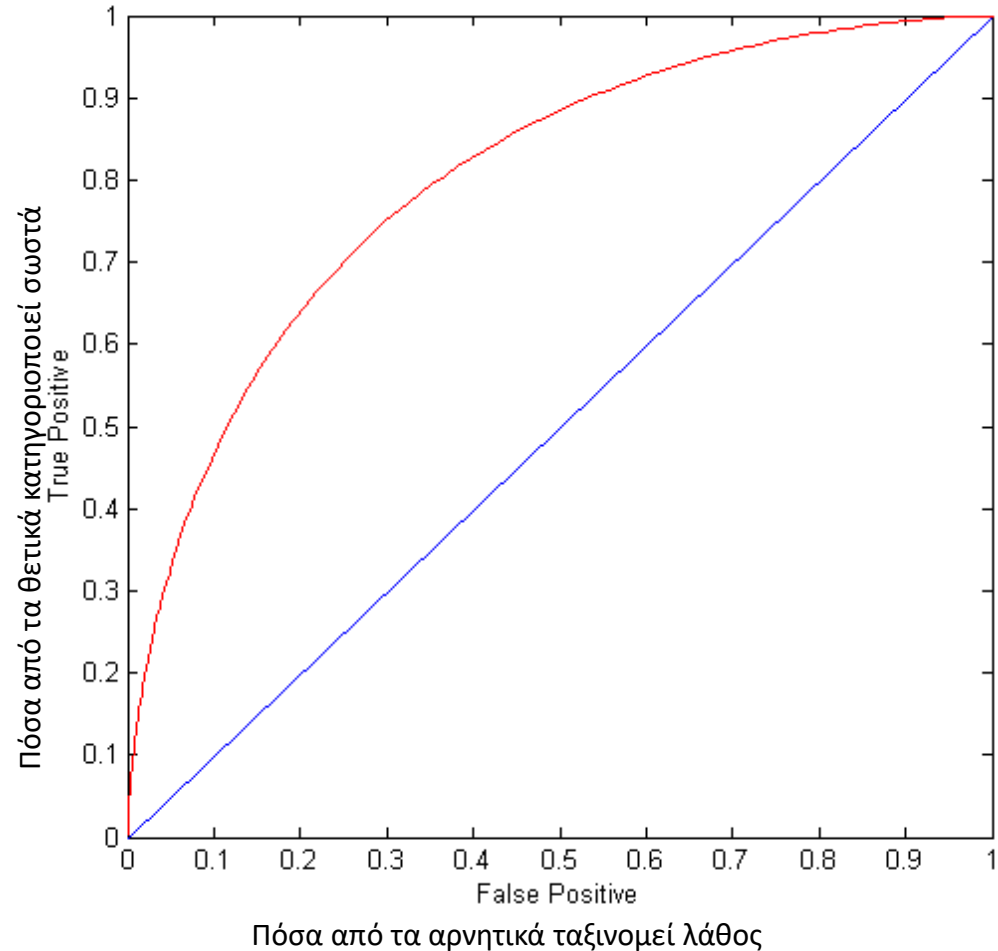
[πόσα από τα αρνητικά κατηγοριοποιεί λάθος]

$$\text{FPR} = \frac{FP}{TN + FP}$$

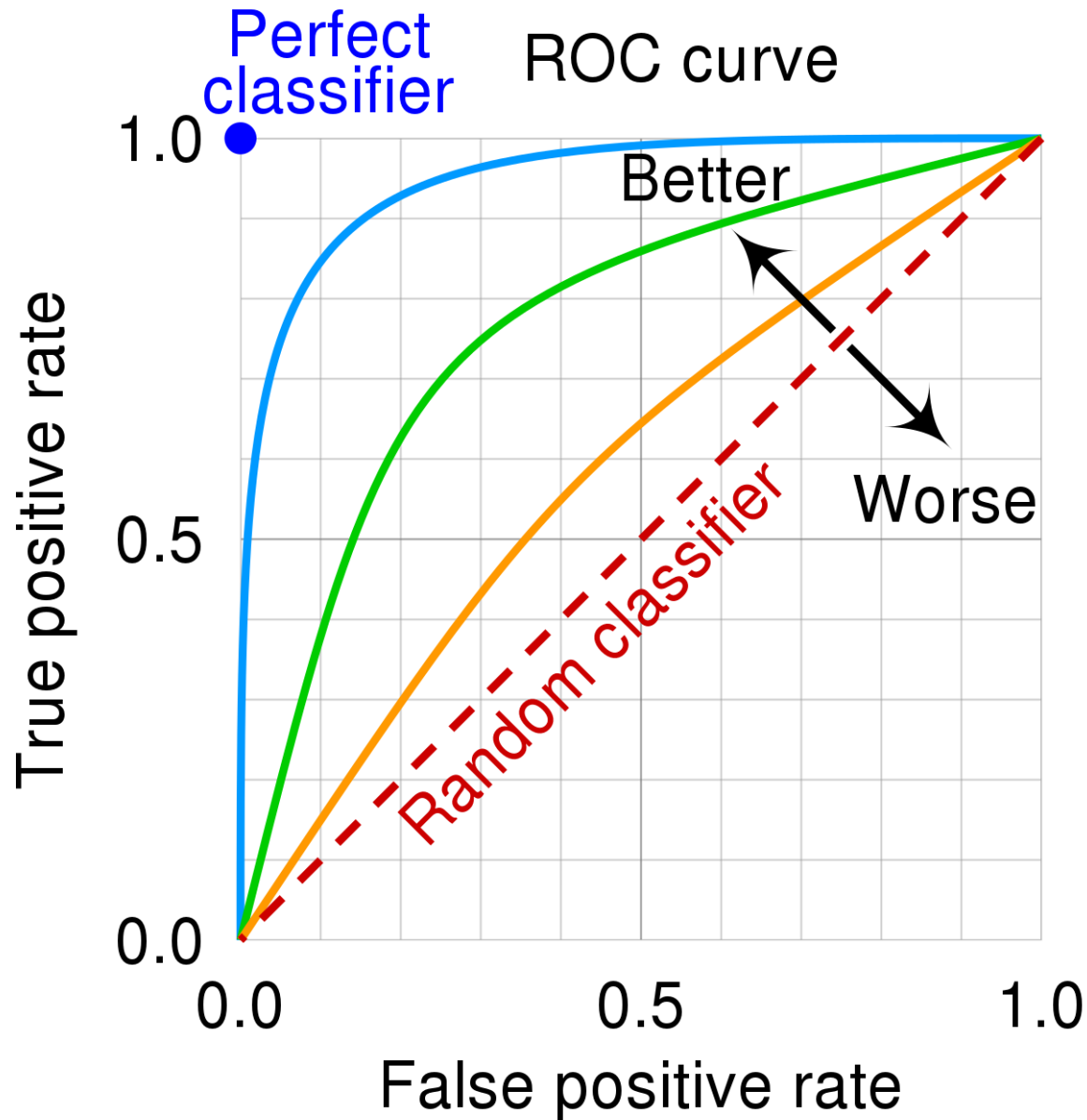
# ROC Curve

Κάθε σημείο αντιστοιχεί στα μοντέλα που παράγει κάθε κατηγοριοποιητής

- (0,0): το μοντέλο προβλέπει τα πάντα ως αρνητική κατηγορία
- (1,1): το μοντέλο προβλέπει τα πάντα ως θετική κατηγορία
- (0,1): ιδανικό (στην άνω αριστερή γωνία)
- Διαγώνια γραμμή: Random guessing
- Μια εγγραφή θεωρείται θετική με καθορισμένη πιθανότητα  $p$  ανεξάρτητα από τις τιμές των γνωρισμάτων της

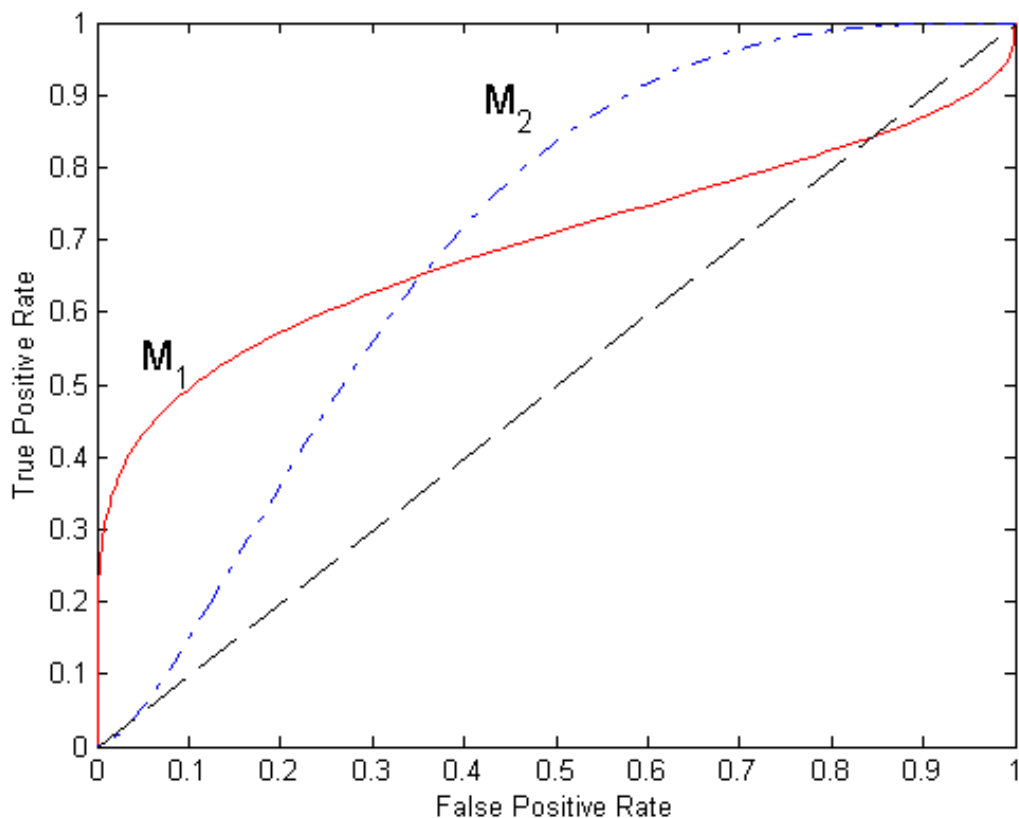


# ROC Curve





# ROC Curve - Σύγκριση Μοντέλων



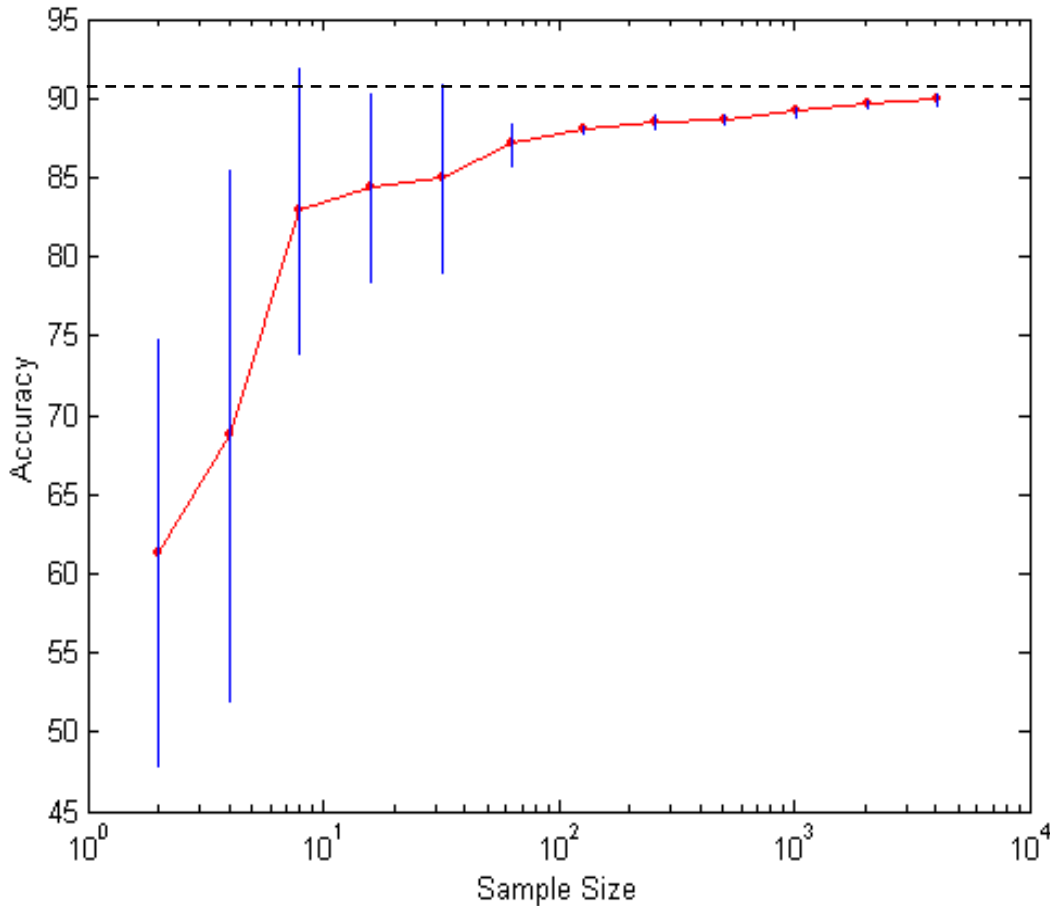
- Κανένα μοντέλο δεν είναι πάντα καλύτερο του άλλου
  - $M_1$  καλύτερο για μικρό FPR
  - $M_2$  καλύτερο για μεγάλο FPR
- Η περιοχή κάτω από την καμπύλη ROC
  - Ιδανικό μοντέλο:
    - Περιοχή = 1
  - Τυχαία πρόβλεψη:
    - Περιοχή = 0.5

Ένα μοντέλο είναι αυστηρά καλύτερο αν έχει μεγαλύτερη περιοχή κάτω από την καμπύλη του

# Μέθοδοι Αποτίμησης Μοντέλου

- Πώς μπορούμε να πάρουμε αξιόπιστες εκτιμήσεις της απόδοσης;
- Η απόδοση ενός μοντέλου μπορεί να εξαρτάται από πολλούς παράγοντες, εκτός του αλγορίθμου μάθησης:
  - Κατανομή των κλάσεων
  - Το κόστος της λανθασμένης κατηγοριοποίησης
  - Το μέγεθος του συνόλου εκπαίδευσης και του συνόλου ελέγχου

# Καμπύλη Μάθησης (Learning Curve)

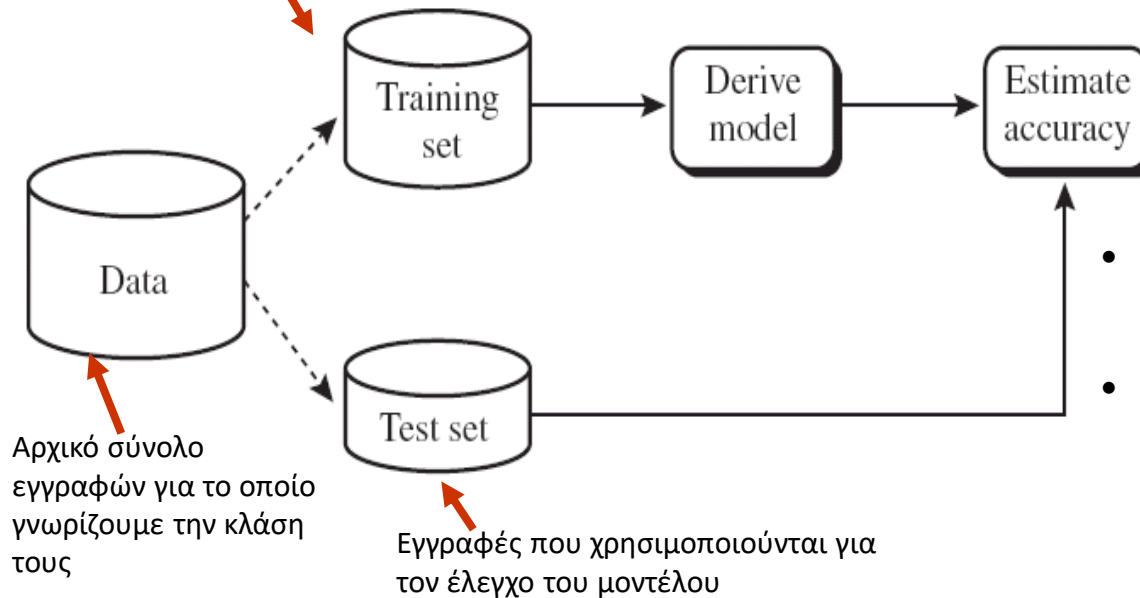


- Η καμπύλη μάθησης δείχνει πως μεταβάλλεται η ακρίβεια (accuracy) με την αύξηση του μεγέθους του δείγματος
- Επίδραση δείγματος μικρού μεγέθους:
  - Bias in the estimate
  - Variance of estimate

# Μέθοδος Holdout

- Διαμέριση του αρχικού συνόλου σε δύο ξένα σύνολα
  - Σύνολο εκπαίδευσης
  - Σύνολο Ελέγχου

Εγγραφές που χρησιμοποιούνται για την κατασκευή του μοντέλου



- Κατασκευή μοντέλου με βάση το σύνολο εκπαίδευσης
- Αποτίμηση μοντέλου με βάση το σύνολο ελέγχου

# Μειονεκτήματα

- Λιγότερες εγγραφές για εκπαίδευση
  - Πιθανόν όχι τόσο καλό μοντέλο, όσο αν χρησιμοποιούνταν όλες
- Το μοντέλο εξαρτάται από τη σύνθεση των συνόλων εκπαίδευσης και ελέγχου
  - Όσο μικρότερο το σύνολο εκπαίδευσης, τόσο μεγαλύτερη η variance του μοντέλου
  - Όσο μεγαλύτερο το σύνολο εκπαίδευσης, τόσο λιγότερο αξιόπιστη η πιστότητα του μοντέλου που υπολογίζεται με το σύνολο ελέγχου – wide confidence interval (διακύμανση)
- Τα σύνολα ελέγχου και εκπαίδευσης δεν είναι ανεξάρτητα μεταξύ τους (υποσύνολα του ίδιου συνόλου)

# Τυχαία Λήψη Δειγμάτων (Random Subsampling)

- Επανάληψη της μεθόδου για τη βελτίωσή της
  - Έστω  $k$  επαναλήψεις, παίρνουμε το μέσο όρο της ακρίβειας

$$acc_{sub} = \frac{1}{k} \sum_{i=1}^k acc_i$$

## Μειονεκτήματα

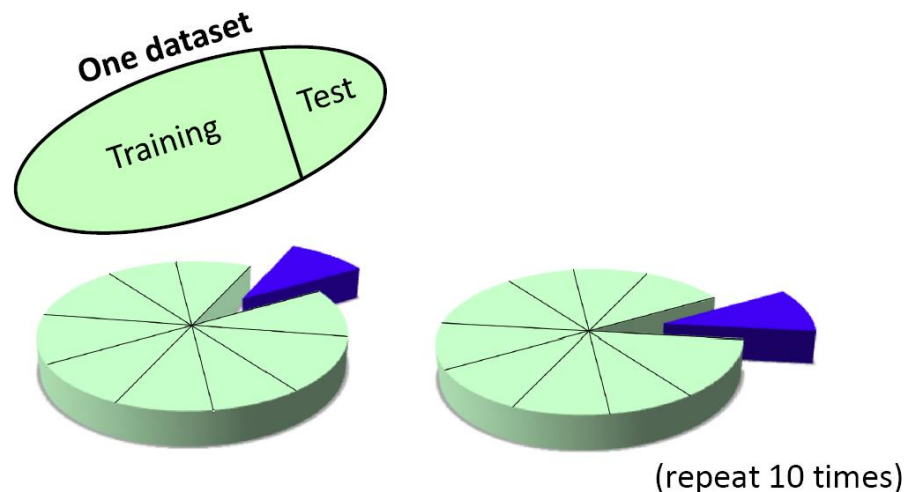
- Πάλι αφαιρούμε δεδομένα από το σύνολο εκπαίδευσης
- Μία εγγραφή μπορεί να επιλέγεται ως εγγραφή εκπαίδευσης πιο συχνά από κάποια άλλη

# Διασταυρωμένη Επικύρωση (Cross Validation)

- Κάθε εγγραφή χρησιμοποιείται τον ίδιο αριθμό φορές στην εκπαίδευση και ακριβώς μια φορά για έλεγχο
  - Διαμοίραση των δεδομένων σε  $k$  ίσα διαστήματα
  - Κατασκευή του μοντέλου αφήνοντας κάθε φορά ένα διάστημα ως σύνολο ελέγχου και χρησιμοποιώντας όλα τα υπόλοιπα ως σύνολα εκπαίδευσης
  - Επανάληψη  $k$  φορές
- **2-fold** (δύο ίσα υποσύνολα, το ένα μια φορά για έλεγχο, το άλλο για εκπαίδευση και μετά ανάποδα)
- Αν  $k = N$ , ( $N$  ο αριθμός των εγγραφών)  $\rightarrow$  *leave-one-out*
  - μεγαλύτερο δυνατό σύνολο εκπαίδευσης
  - σύνολα ελέγχου αμοιβαία αποκλειόμενα (καλύπτουν όλο το σύνολο)
  - υπολογιστικά ακριβή
  - υψηλή διακύμανση του μέτρου (μόνο μια τιμή)

# Διασταυρωμένη Επικύρωση (Cross Validation)

- Τα δεδομένα πρέπει να χωρίζονται σε training και test έτσι ώστε να εκτιμηθεί το overfitting
- 10-fold cross validation
  - Τα δεδομένα χωρίζονται σε 10 τμήματα (folds)
  - Με την σειρά επιλέγεται 1-fold
  - Εκτελείται ο αλγόριθμος
  - Υπολογίζεται ο μέσος όρος όλων των αποτελεσμάτων







# k-Nearest Neighbors algorithm (Κατηγοριοποιητής Κοντινότερου Γείτονα)

Χρησιμοποιείται για ταξινόμηση και παλινδρόμηση

# Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

Μέχρι στιγμής είδαμε: Κατηγοριοποίηση βασισμένη σε δύο βήματα

- Βήμα 1: Induction Step → Κατασκευή Μοντέλου
- Βήμα 2: Deduction Step → Εφαρμογή του μοντέλου για έλεγχο παραδειγμάτων

**Eager Learners** vs **Lazy Learners** → Μην κατασκευάζεις μοντέλο αν δε χρειαστεί!

# Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

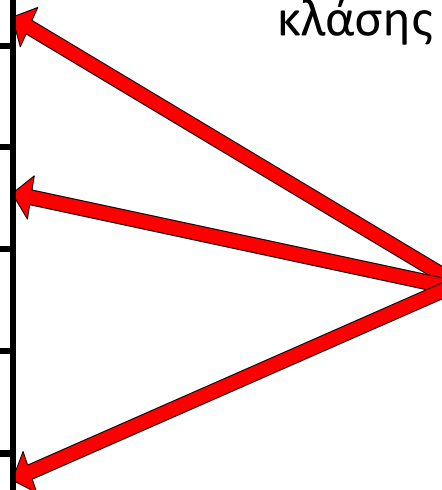
Σύνολο Αποθηκευμένων Περιπτώσεων

Atr1	.....	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- Αποθήκευσε τις εγγραφές του συνόλου εκπαίδευσης
- Χρησιμοποίησε τις αποθηκευμένες εγγραφές για την εκτίμηση της κλάσης των νέων περιπτώσεων

Unseen Case

Atr1	.....	AtrN



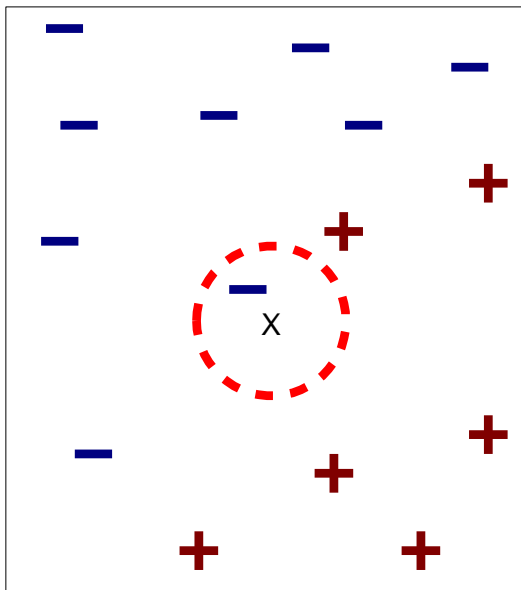
# Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

## Παραδείγματα

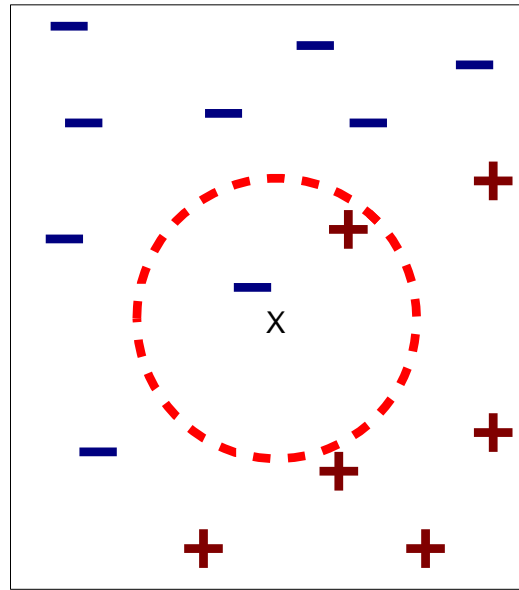
- **Rote-learner (Καθοδηγούμενος)**
  - Κρατά (Memorizes) όλο το σύνολο των δεδομένων εκπαίδευσης και ταξινομεί μια εγγραφή αν ταιριάζει πλήρως με κάποιο από τα δεδομένα εκπαίδευσης
- **Nearest neighbor (Κοντινότερος Γείτονας)**
  - Χρήση των  $k$  κοντινότερων σημείων (nearest neighbors) για την κατηγοριοποίηση

# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

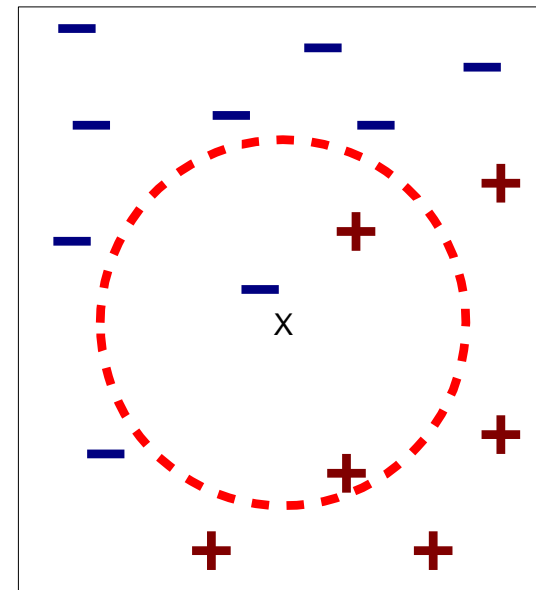
$k$ -κοντινότεροι γείτονες μιας εγγραφής  $x$  είναι τα σημεία που έχουν την  $k$ -οστή μικρότερη απόσταση από το  $x$



(a) 1-nearest neighbor



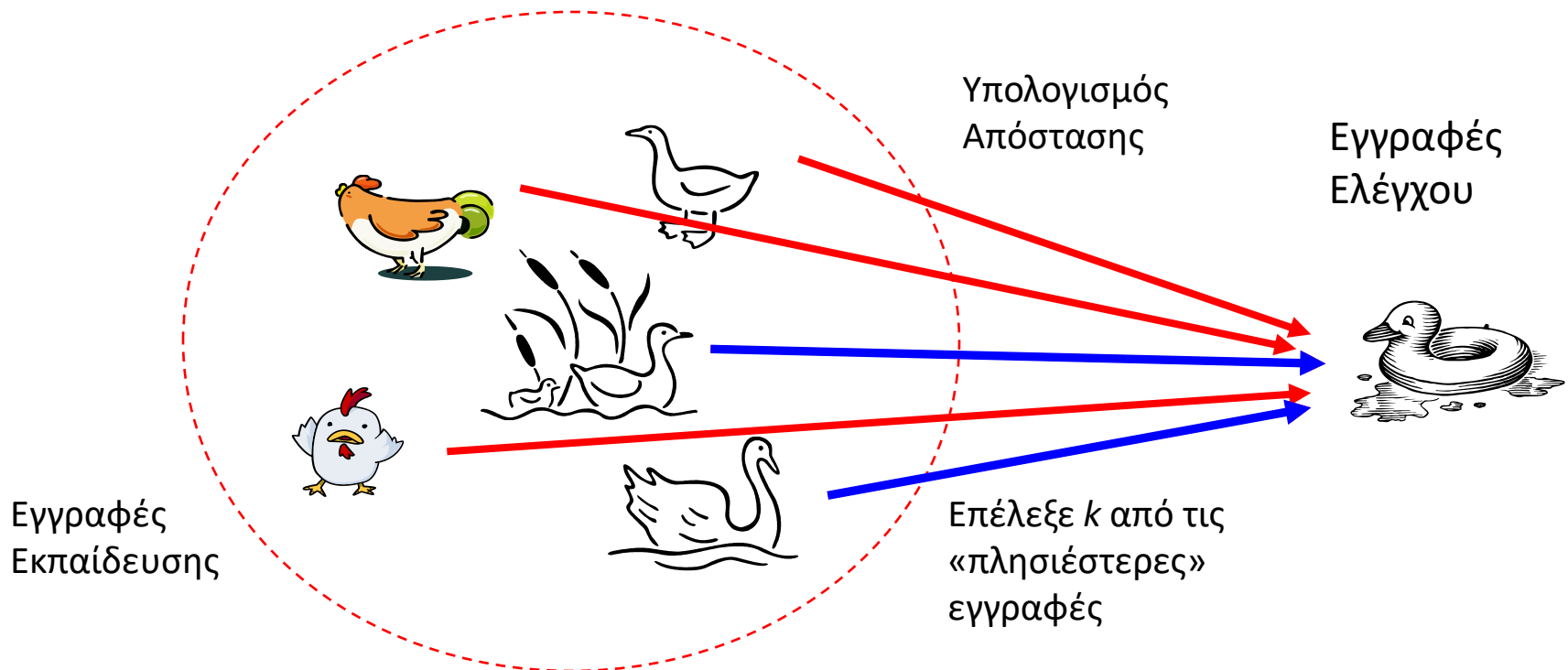
(b) 2-nearest neighbor



(c) 3-nearest neighbor

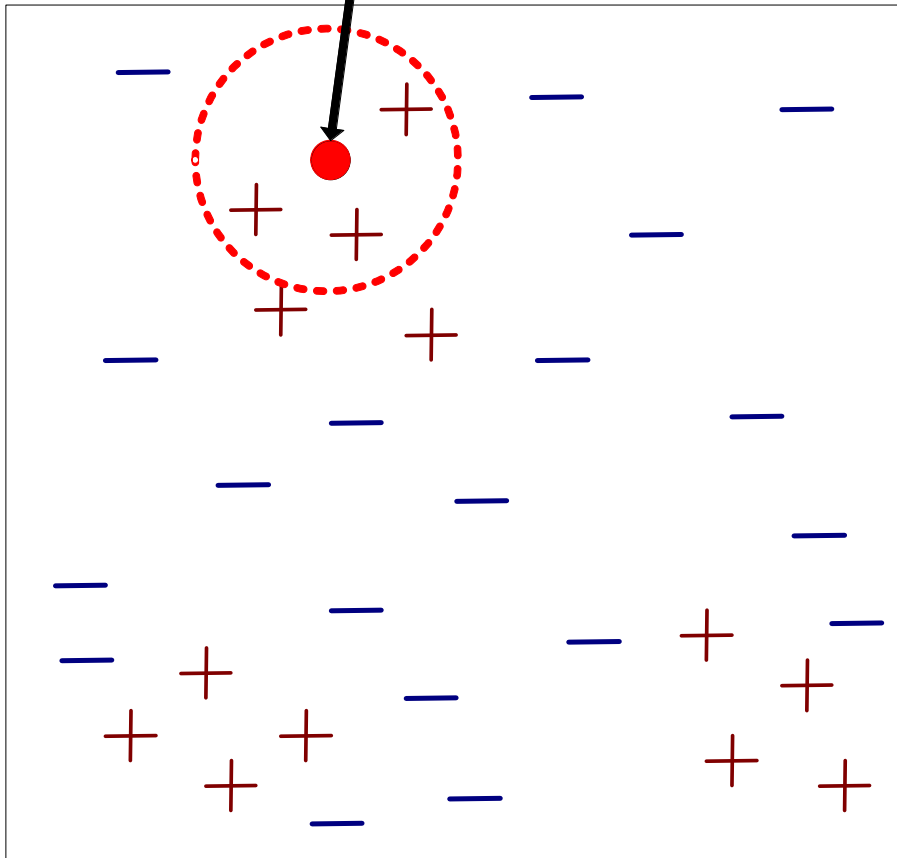
# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

Βασική ιδέα: Αν περπατάει σαν πάπια και ακούγεται σαν πάπια, τότε μάλλον είναι πάπια



# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

Άγνωστη Εγγραφή

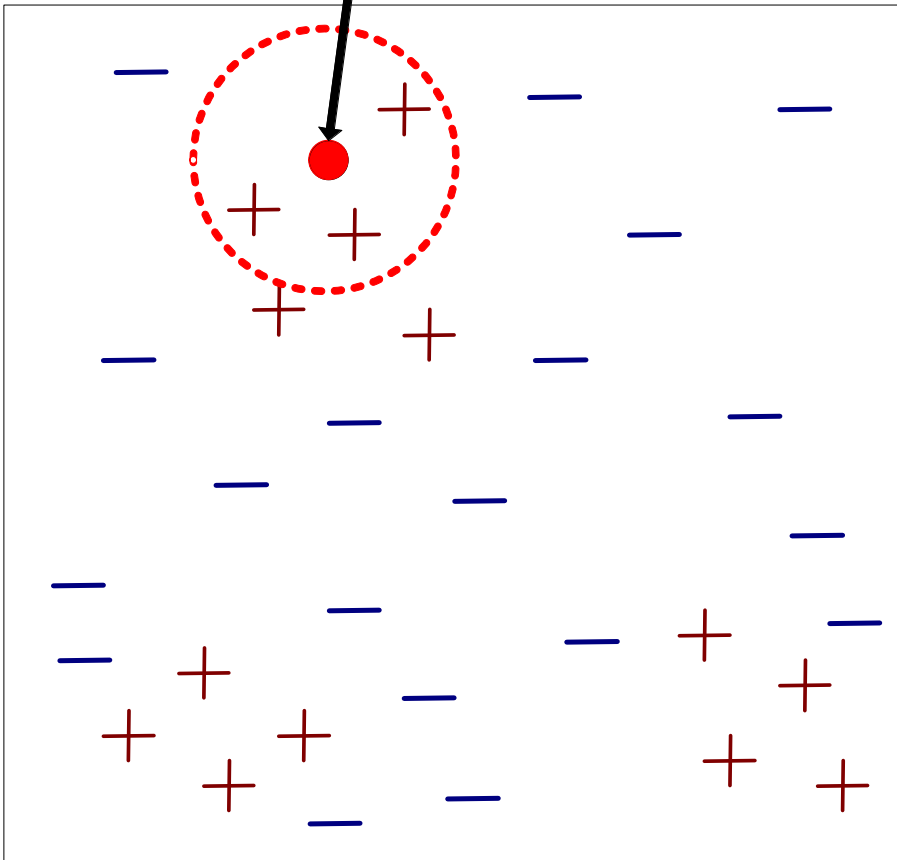


Για να κατηγοριοποιηθεί μια άγνωστη εγγραφή:

- 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 Κοντινότερου Γείτονα (kNN)

Άγνωστη Εγγραφή



Χρειάζεται

1. Το σύνολο των αποθηκευμένων εγγραφών
2. Μετρική απόστασης για να υπολογίσουμε την απόσταση μεταξύ εγγραφών
3. Την τιμή του  $k$ , δηλαδή τον αριθμό των κοντινότερων γειτόνων που πρέπει να ανακληθούν

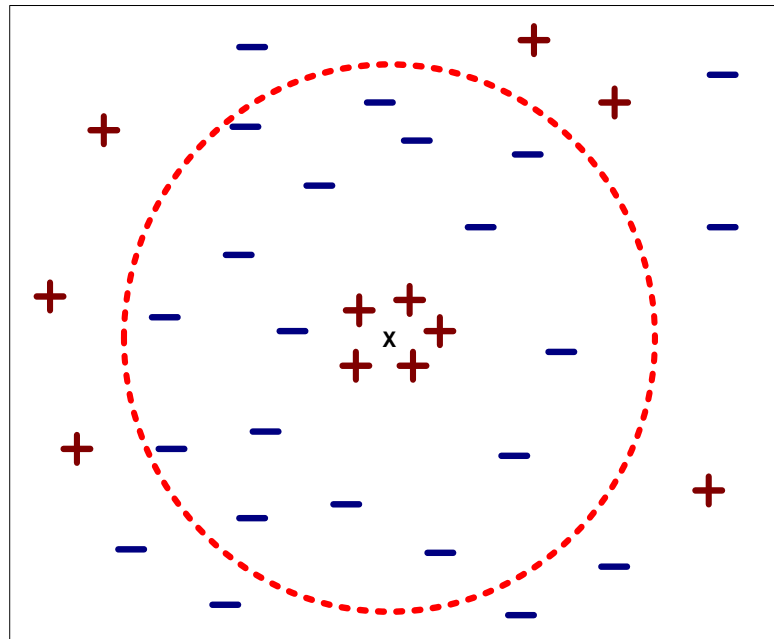


# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

- Απόσταση μεταξύ εγγραφών
  - Π.χ. Ευκλείδεια απόσταση
- Καθορισμός κλάσης
  - Απλά τη πλειοψηφική κλάση
    - Βάρος σε κάθε ψήφο με βάση την απόσταση
      - weight factor,  $w = 1/d^2$

# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

- Επιλογή της τιμής του  $k$ 
  - $k$  πολύ μικρό: ευαισθησία στα σημεία θορύβου
  - $k$  πολύ μεγάλο: η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις
  - συχνά  $k = \text{sqr}(n)$ , όπου  $n$  το μέγεθος του συνόλου εκπαίδευσης
  - default:  $k=10$



# Κατηγοριοποιητής Κοντινότερου Γείτονα (kNN)

## Θέματα Κλιμάκωσης

- Τα γνωρίσματα ίσως πρέπει να κλιμακωθούν ώστε οι αποστάσεις να μην κυριαρχηθούν από κάποιο γνώρισμα
- Παράδειγμα:
  - το ύψος μπορεί να κυμαίνεται από 1.5m σε 1.8m
  - το βάρος μπορεί να κυμαίνεται από 90lb σε 300lb
  - το εισόδημα μπορεί να κυμαίνεται από \$10K σε \$1M
- Δεν κατασκευάζεται μοντέλο, μεγάλο κόστος για την εφαρμογή της κατηγοριοποίησης
- Πολλές διαστάσεις (Η κατάρα των πολλών διαστάσεων)
- Θόρυβος (ελάττωση μέσω k-γειτόνων)



# Naive Bayes Classifier

Οικογένεια απλών «πιθανολογικών ταξινομητών» βασιζόμενοι στο θεωρήματος του Bayes με ισχυρές (αφελείς) υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών

# Κατηγοριοποιητής Bayes

- $X, Y$  τυχαίες μεταβλητές
- Δεσμευμένη πιθανότητα (Conditional probability):  $P(Y=y | X=x)$ 
  - Ποια είναι η πιθανότητα να συμβεί το  $Y$  δεδομένου ότι συνέβη το  $X$ ;

**Το θεώρημα του Bayes:** 
$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

- Από κοινού πιθανότητα:  $P(X=x, Y=y)$ 
  - Ποια η πιθανότητα το  $X$  και  $Y$  να συμβούν μαζί και στο ίδιο χρονικό σημείο;
- Σχέση μεταξύ από κοινού (joint) και δεσμευμένης (conditional) πιθανότητας

$$P(Y | X) = \frac{P(X \cap Y)}{P(X)} \quad P(X | Y) = \frac{P(Y \cap X)}{P(Y)}$$

# Θεώρημα του Bayes - Παράδειγμα 1

Δοθέντων ότι

1. Αν κάποιος έχει περάσει το μάθημα «Προγραμματισμός σε C», περνάει το μάθημα «Δομές Δεδομένων» με πιθανότητα  $4/5$
  2. Η εκ των προτέρων πιθανότητα κάποιος να περάσει το μάθημα «Προγραμματισμός σε C» είναι  $1/3$
  3. Η εκ των προτέρων πιθανότητα κάποιος να περάσει το μάθημα «Δομές Δεδομένων» είναι  $2/3$
- Πόσοι περνούν και τα δύο μαθήματα;
  - Αν ξέρουμε ότι ένας φοιτητής έχει περάσει το μάθημα «Δομές Δεδομένων» ποια είναι η πιθανότητα να έχει περάσει το μάθημα «Προγραμματισμός σε C»;

$$P(C = 1 | \Delta = 1) = \frac{P(\Delta = 1 | C = 1)P(C = 1)}{P(\Delta = 1)} = \frac{4/5 \times 1/3}{2/3} = 0.8$$

# Θεώρημα του Bayes - Παράδειγμα 2

Έστω 2 ομάδες, η Ομάδα Α και η Ομάδα Β

1. Η Ομάδα Α νικά στο 65% των μεταξύ τους αγώνων
2. Από τα παιχνίδια στα οποία νίκησε η Ομάδα Β, μόνο το 30% έγιναν στην έδρα της Ομάδας Β
3. 75% των νικών της Ομάδας Β γίνονται στην έδρα της

Αν η Ομάδα Β αναμένεται να φιλοξενήσει την Ομάδα Α στον επόμενο αγώνα, ποια ομάδα εμφανίζεται ως πιθανότερη νικήτρια;

# Κατηγοριοποιητής Bayes

Πώς μπορούμε να χρησιμοποιήσουμε αυτό το θεώρημα για το πρόβλημα της κατηγοριοποίησης;



# Κατηγοριοποιητής Bayes

- $X$ : σύνολο των γνωρισμάτων
- $Y$ : η μεταβλητή της κλάσης (κατηγορίας)
- $Y$  εξαρτάται από το  $X$  με μη ντετερμινιστικό τρόπο (*non-deterministic*)
- $P(Y | X)$ : Posterior probability (εκ των υστέρων πιθανότητα)
- $P(Y)$ : Prior probability (εκ των προτέρων πιθανότητα)

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training set for predicting borrowers who will default on loan payments.

$X' = (\text{Homeowner} = \text{No}, \text{Marital Status} = \text{Married}, \text{Annual Income} = 120\text{K})$

Υπολόγισε:  $P(\text{Yes} | X')$ ,  $P(\text{No} | X')$ , επέλεξε No ή Yes, ανάλογα με ποιο έχει τη μεγαλύτερη πιθανότητα

**Πώς θα υπολογίσουμε αυτές τις πιθανότητες;**

# Κατηγοριοποιητής Bayes

## Φάση Εκπαίδευσης:

Εκμάθηση των εκ των υστέρων πιθανοτήτων  $P(Y | X)$  για κάθε συνδυασμό των  $X$  και  $Y$  βασισμένη στα δεδομένα εκπαίδευσης

## Φάση Εφαρμογής:

Για κάθε εγγραφή ελέγχου  $X'$ , υπολόγισε την κλάση  $Y'$  που μεγιστοποιεί την εκ των υστέρων πιθανότητα  $P(Y' | X')$  δηλαδή, την πιο πιθανή κλάση με βάση τα δεδομένα ελέγχου

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)}$$

- $P(X)$  είναι σταθερή και μπορούμε να την αγνοήσουμε (μαρτυρία - evidence)
- $P(Y)$ : εκτιμάται εύκολα από τα δεδομένα εισόδου, είναι το ποσοστό των δεδομένων εκπαίδευσης που ανήκουν στην κλάση  $Y$  (εκ των προτέρων πιθανότητα)
- $P(X|Y)$ ?

# Κατηγοριοποιητής Bayes

*Υπολογισμός της εξαρτώμενης από τη κατηγορία πιθανότητας  $P(X | Y)$*

Υπάρχουν δύο βασικές μέθοδοι:

1. Απλοϊκός (Naive) Κατηγοριοποιητής
2. Δίκτυο πεποίθησης

Θα δούμε την πρώτη μέθοδο

# Παράδειγμα με 1 γνώρισμα

Οικογενειακή Κατάσταση	Αγοραστής
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Έγγαμος	ΟΧΙ
Άγαμος	ΝΑΙ
Άγαμος	ΝΑΙ
Έγγαμος	ΟΧΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Άγαμος	ΟΧΙ

*Αν κάποιος είναι άγαμος, είναι αγοραστής ή όχι?*

$$P(\text{Ναι} \mid \text{Άγαμος}) \rightarrow P(\text{Άγαμος} \mid \text{Ναι}) \quad P(\text{Ναι}) = 2/7 * 7/10 = 0.2$$

$$P(\text{Όχι} \mid \text{Άγαμος}) \rightarrow P(\text{Άγαμος} \mid \text{Όχι}) \quad P(\text{Όχι}) = 1/3 * 3/10 = 0.1$$

# Κατηγορική ιδιότητα (Categorical attribute)

Κατηγορική ιδιότητα (Categorical attribute)  $X_i$

$P(X_i=x_i | Y=y)$ : Ποσοστό των δεδομένων εκπαίδευσης της κλάσης  $y$  που έχουν τιμή  $x_i$  στο  $i$ -οστό γνώρισμα

$$P(\text{HomeOwner}=\text{Yes} | \text{No}) = 3/7$$

$$P(\text{MaritalStatus}=\text{Single} | \text{Yes}) = 2/3$$

- Τι γίνεται όταν έχουμε παραπάνω από ένα γνωρίσματα;
- Τι γίνεται όταν τα γνωρίσματα παίρνουν συνεχής τιμές;

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

# Παράδειγμα Κατηγοριοποιητή Bayes

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

**A:** γνωρίσματα

**M:** mammals

**N:** non-mammals

$$P(A | M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A | N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A | M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A | N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

**P(A | M) \* P(M) > P(A | N) \* P(N)**

**→ Mammal Class**

# Παράδειγμα με πολλαπλά γνωρίσματα

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

*Αν κάποιος είναι άγαμος και 35 χρονών, είναι αγοραστής ή όχι?*

Πρέπει να υπολογιστούν τα:  $P(\text{Ναι} \mid \text{Άγαμος}, 35)$  και  $P(\text{Όχι} \mid \text{Άγαμος}, 35)$   
με βάση τα:  $P(\text{Άγαμος}, 35 \mid \text{Ναι})$  και  $P(\text{Άγαμος}, 35 \mid \text{Όχι})$

# Κατηγοριοποίηση με πολλαπλά γνωρίσματα

- Σύνολο  $X = \{X_1, \dots, X_d\}$  από  $d$  γνωρίσματα
- Υπό συνθήκη ανεξαρτησία (Conditional independence):
  - Το  $X$  είναι υπό συνθήκη ανεξάρτητο του  $Y$ , δοθέντος του  $Z$ , αν:  
 $P(X | Y, Z) = P(X | Z) * P(Y | Z)$

$$P(X | Y = y) = \prod_{i=1}^d P(X_i | Y = y)$$

$$P(Y | X) = \frac{P(Y) \prod_{i=1}^d P(X_i | Y)}{P(X)}$$



# Παράδειγμα με πολλαπλά γνωρίσματα

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

*Αν κάποιος είναι άγαμος και 35 χρονών, είναι αγοραστής ή όχι?*

Πρέπει να υπολογιστούν τα:  $P(\text{Ναι} \mid \text{Άγαμος}, 35)$  και  $P(\text{Όχι} \mid \text{Άγαμος}, 35)$

## Κατηγοριοποιητής Bayes

$$P(\text{Ναι}|\text{Άγαμος, 35}) \rightarrow P(\text{Άγαμος, 35}|\text{Ναι}) * P(\text{Ναι}) = ;$$

$$P(\text{Όχι}|\text{Άγαμος, 35}) \rightarrow P(\text{Άγαμος, 35}|\text{Όχι}) * P(\text{Όχι}) = ;$$

Υπόθεση: Ανεξαρτησία οικογενειακής κατάστασης και ηλικίας

$$P(\text{Ναι}|\text{Άγαμος, 35}) \rightarrow P(\text{Άγαμος}|\text{Ναι}) * P(35|\text{Ναι}) * P(\text{Ναι}) = ;$$

$$P(\text{Όχι}|\text{Άγαμος, 35}) \rightarrow P(\text{Άγαμος}|\text{Όχι}) * P(35|\text{Όχι}) * P(\text{Όχι}) = ;$$

*Από το παράδειγμα μιας ιδιότητας, έχω ήδη υπολογίσει:*

$$P(\text{Άγαμος}|\text{Ναι}) * P(\text{Ναι}) = 0.2$$

$$P(\text{Άγαμος}|\text{Όχι}) * P(\text{Όχι}) = 0.1$$

# Εκτίμηση των Υπό Συνθήκη Πιθανοτήτων για Συνεχή Γνωρίσματα

## Διακριτοποίηση (discretization)

- Χωρίζουμε σε διαστήματα και η εκτίμηση γίνεται με βάση την αναλογία των εγγραφών εκπαίδευσης στο αντίστοιχο διάστημα
  - Πολλά διαστήματα → λίγες εγγραφές εκπαίδευσης
  - Λίγα διαστήματα → πιθανόν να συναθροίζουν εγγραφές που ανήκουν σε διαφορετικές κατηγορίες

# Εκτίμηση των Υπό Συνθήκη Πιθανοτήτων για Συνεχή Γνωρίσματα

- Χρήση κάποιας κατανομής
- Υποθέτουμε μια συγκεκριμένη μορφή κατανομής πιθανοτήτων
- Συνήθως Gauss (κανονική) κατανομή
- Χαρακτηρίζεται από δύο παραμέτρους
  - Μέσο ( $\mu$ )
  - Διακύμανση ( $\sigma^2$ )

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- Το  $\mu_{ij}$  είναι το μέσο για όλα τα δεδομένα εκπαίδευσης της κατηγορίας (κλάσης)  $y_i$
- Όμοια εκτιμάται και η διακύμανση

# Παράδειγμα Κατηγοριοποιητή Bayes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Κανονική κατανομή
- Income, Class=No
  - Μέσος = 110K
  - Διακύμανση= 2975

$$P(\text{Income} = 120 \mid \text{No}) = \frac{1}{\sqrt{2\pi(54.54)}} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

# Κατηγοριοποιητής Bayes

Επιστροφή στο παράδειγμα των αγοραστών...

... με χρήση κανονικής κατανομής!

$$\bar{x}_N = (20+30+30+\dots 40)/7 = 30.71 \quad s_N = 2.7$$

$$\bar{x}_O = (25 + 20 + 20)/3 = 21.67 \quad s_O = 1.7$$

$$P(35 \leq \text{Ηλικία} \leq 35+\varepsilon | \text{'Ναι'}) = \int_{35}^{35+\varepsilon} \frac{1}{\sqrt{2\pi}s_N} e^{-\frac{(35-\bar{x}_N)^2}{2s_N^2}} \simeq \varepsilon \frac{1}{\sqrt{2\pi}s_N} e^{-\frac{(35-\bar{x}_N)^2}{2s_N^2}} = 0.11\varepsilon$$

$$P(35 \leq \text{Ηλικία} \leq 35+\varepsilon | \text{'Όχι'}) = \int_{35}^{35+\varepsilon} \frac{1}{\sqrt{2\pi}s_O} e^{-\frac{(35-\bar{x}_O)^2}{2s_O^2}} \simeq \varepsilon \frac{1}{\sqrt{2\pi}s_O} e^{-\frac{(35-\bar{x}_O)^2}{2s_O^2}} = 10^{-14}\varepsilon$$

το  $\varepsilon$  κανονικοποιείται οπότε μπορούμε να χρησιμοποιήσουμε την προηγούμενη εξίσωση

# Κατηγοριοποιητής Bayes

$P(\text{Ναι} | \text{Άγαμος, 35}) \rightarrow$

$$P(\text{Άγαμος} | \text{Ναι}) P(35 | \text{Ναι}) * P(\text{Ναι}) = 0.2 * 0.11\epsilon = 0.022 \epsilon$$

$P(\text{Όχι} | \text{Άγαμος, 35}) \rightarrow$

$$P(\text{Άγαμος} | \text{Όχι}) P(35 | \text{Όχι}) * P(\text{Όχι}) = 0.1 * 10^{-14} \epsilon = 10^{-15} \epsilon$$

Άρα, αγοραστής

# Κατηγοριοποιητής Bayes

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $X' = (\text{HomeOwner} = \text{No}, \text{MaritalStatus} = \text{Married}, \text{Income} = 120\text{K})$
- Πρέπει να υπολογιστεί το  $P(Y | X')$ , δηλαδή το  $P(Y) * P(X' | Y)$
- Έχουμε δύο τιμές για το  $Y$ 
  - $Y = \text{No}$ :
    - $P(\text{HomeOwner} = \text{No} | \text{No}) * P(\text{MaritalStatus} = \text{Married} | \text{No}) * P(\text{Income} = 120\text{K} | \text{No}) = 4/7 * 4/7 * 0.0072 = 0.0024$
  - $Y = \text{Yes}$ :
    - $P(\text{HomeOwner} = \text{No} | \text{Yes}) * P(\text{MaritalStatus} = \text{Married} | \text{Yes}) * P(\text{Income} = 120\text{K} | \text{Yes}) = 3/3 * 0 * 1.2 = 0$



# Διαδικασία Διόρθωσης

- $X' = (\text{HomeOwner} = \text{No}, \text{MaritalStatus} = \text{Married}, \text{Income} = 120\text{K})$ 
  - $P(X' | Y = \text{Yes})$  είναι 0!
- Επειδή τα δείγματα εκπαίδευσης μπορεί να μην καλύπτουν όλες τις κατηγορίες -> Διαδικασία Διόρθωσης

$$\Pr(X_i = x_i | Y = y_j) = \frac{n_c + mp}{n + m}$$

- $n_c$ : ο αριθμός των εγγραφών εκπαίδευσης της κλάσης  $y_j$  που παίρνουν την τιμή  $x_i$
- $n$ : συνολικός αριθμός εγγραφών της κλάσης  $y_j$
- $m$ : μια παράμετρος που καλείται ισοδύναμο μέγεθος δείγματος (equivalent sample size) (ισορροπεί την εκ των υστέρων ( $n_c/n$ ) και την εκ των προτέρων ( $p$ ) πιθανότητα)
- $p$ : μια παράμετρος που καθορίζει ο χρήστης (η εκ των προτέρων πιθανότητα εμφάνισης της τιμής  $x_i$  για το γνώρισμα  $X_i$  μεταξύ των εγγραφών της κλάσης  $y_j$ )

# Κατηγοριοποιητής Bayes

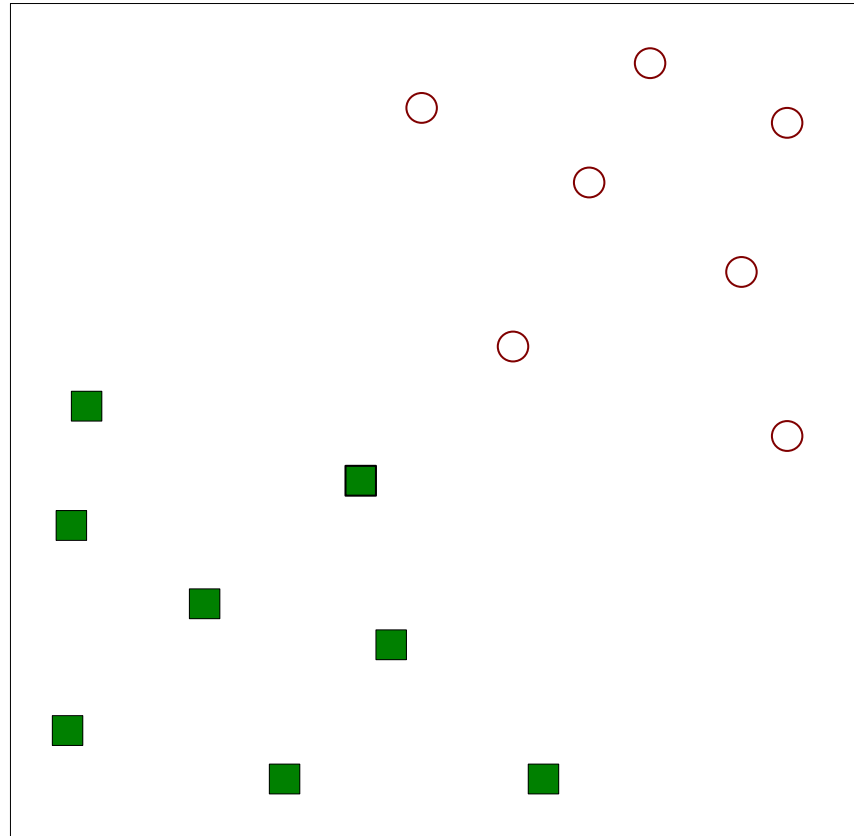
- Ανοχή σε μη σχετικά γνωρίσματα - Αν το  $X_i$  δεν είναι σχετικό (irrelevant),  $P(X_i | Y)$  είναι σχεδόν ομοιόμορφο
- Πρόβλημα όταν υπάρχουν εξαρτήσεις μεταξύ των γνωρισμάτων (μεταβλητών) (correlated attributes)
- Καλή κλιμάκωση σε μεγάλο όγκο δεδομένων, μια απλή ανάγνωση των δεδομένων εκπαίδευσης
- Καλή ανοχή στο θόρυβο, γιατί τα σημεία θορύβου εξομαλύνονται
- Δεν επηρεάζονται από τιμές που λείπουν γιατί αυτές μπορούμε να τις αγνοήσουμε



# Support Vector Machines (Μηχανές Διανυσμάτων Υποστήριξης)

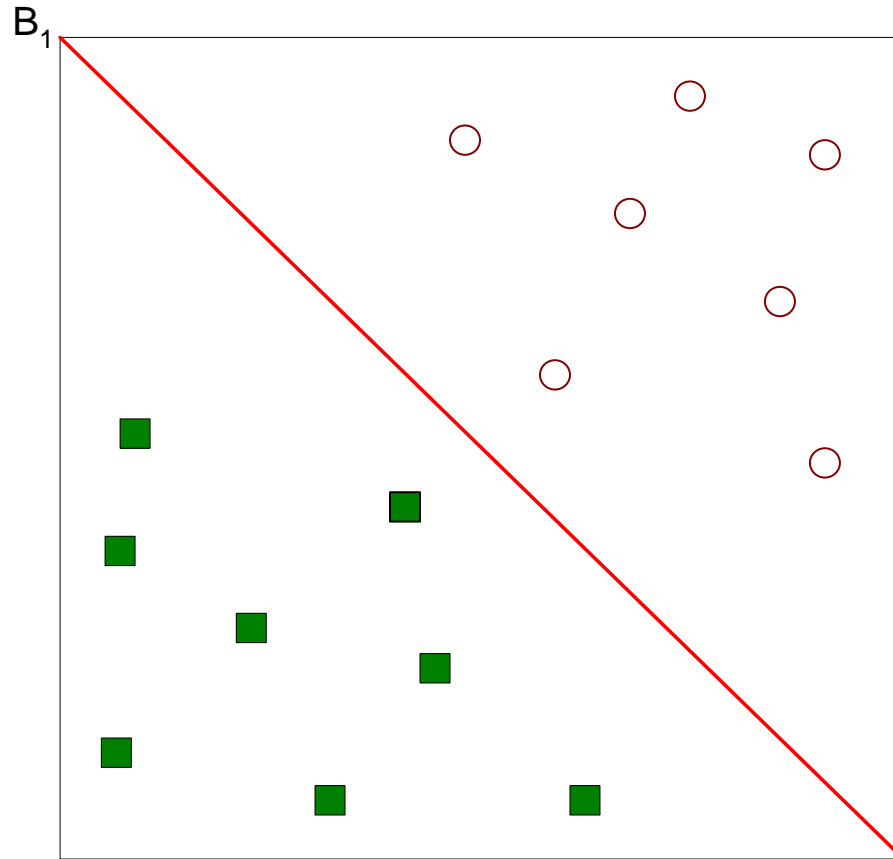
Εποπτευόμενα μοντέλα μάθησης με σχετικούς αλγόριθμους που αναλύουν δεδομένα για ανάλυση ταξινόμησης και παλινδρόμησης

# Κατηγοριοποιητής SVM



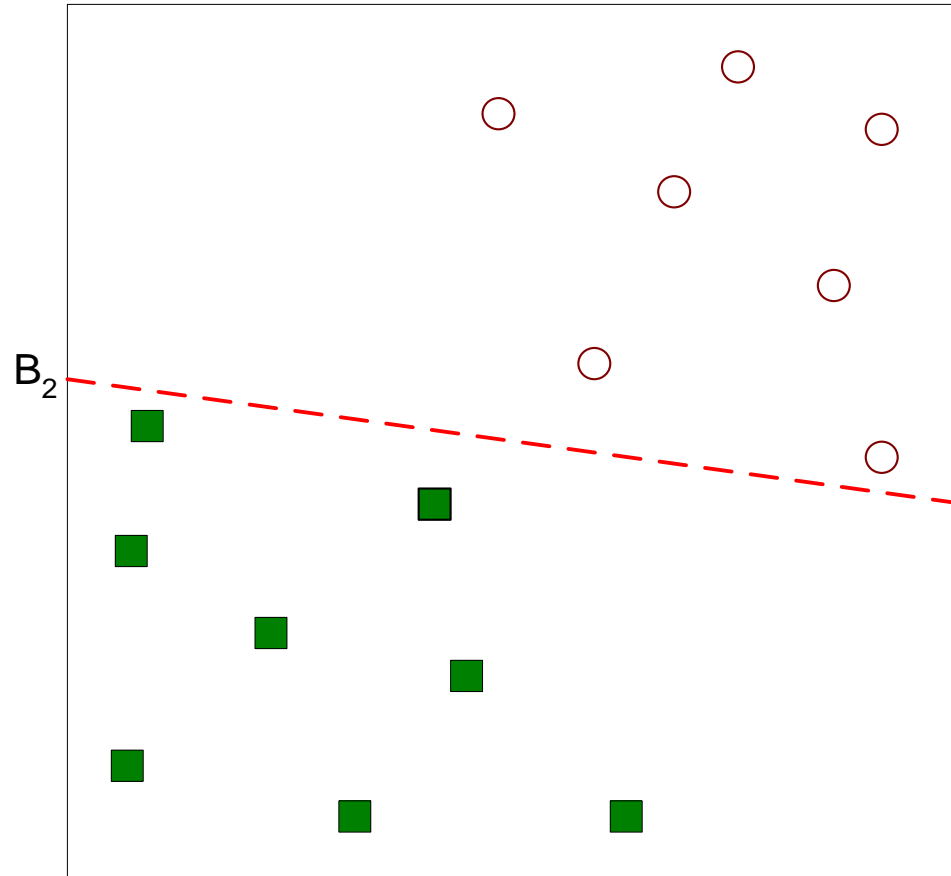
Βρες ένα γραμμικό υπερ-επίπεδο (όριο απόφασης) που να διαχωρίζει τα δεδομένα

# Κατηγοριοποιητής SVM



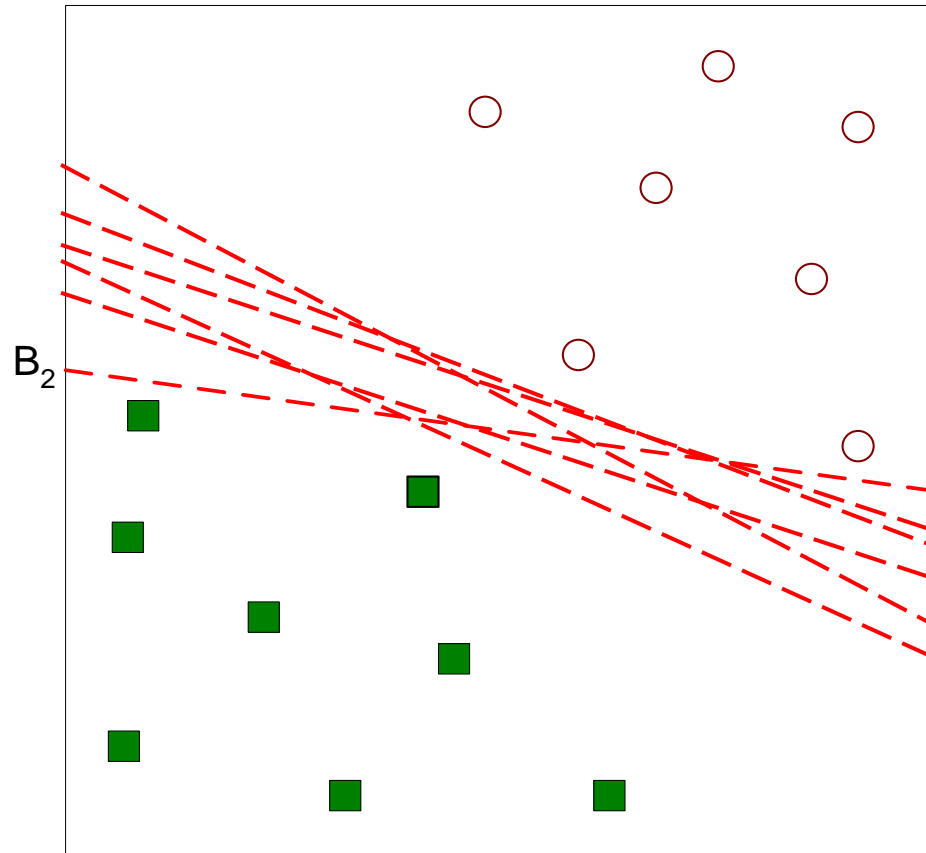
Μία πιθανή λύση

# Κατηγοριοποιητής SVM



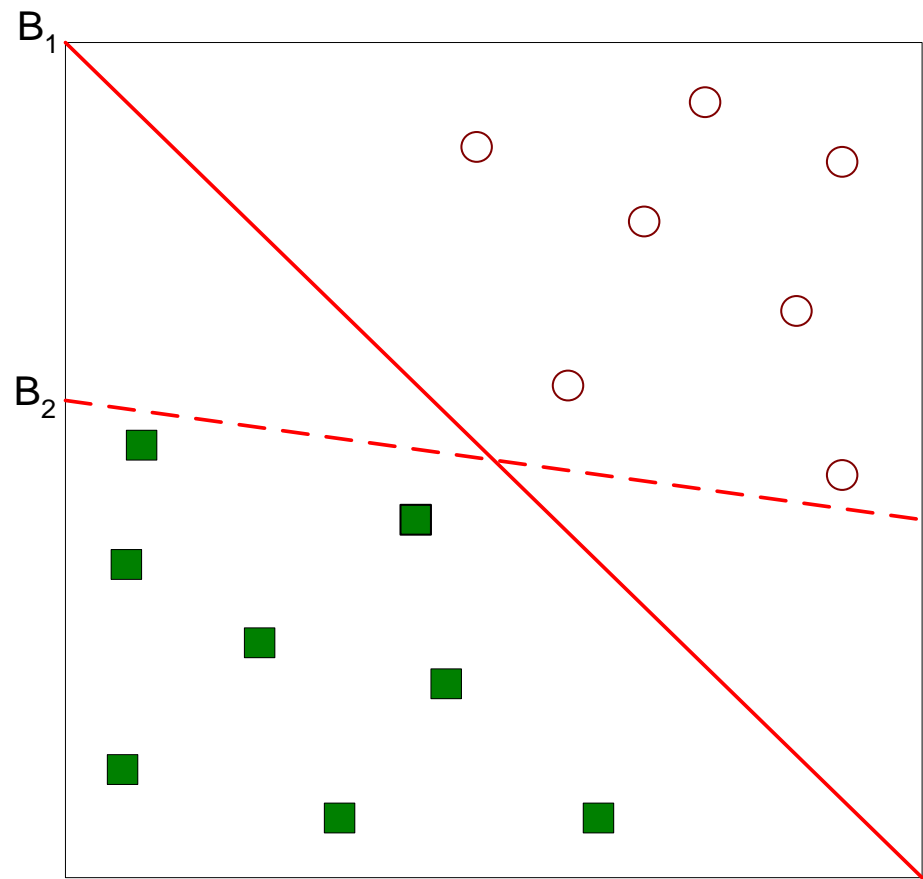
Μία ακόμη πιθανή λύση...

# Κατηγοριοποιητής SVM



Πολλές πιθανές λύσεις!!

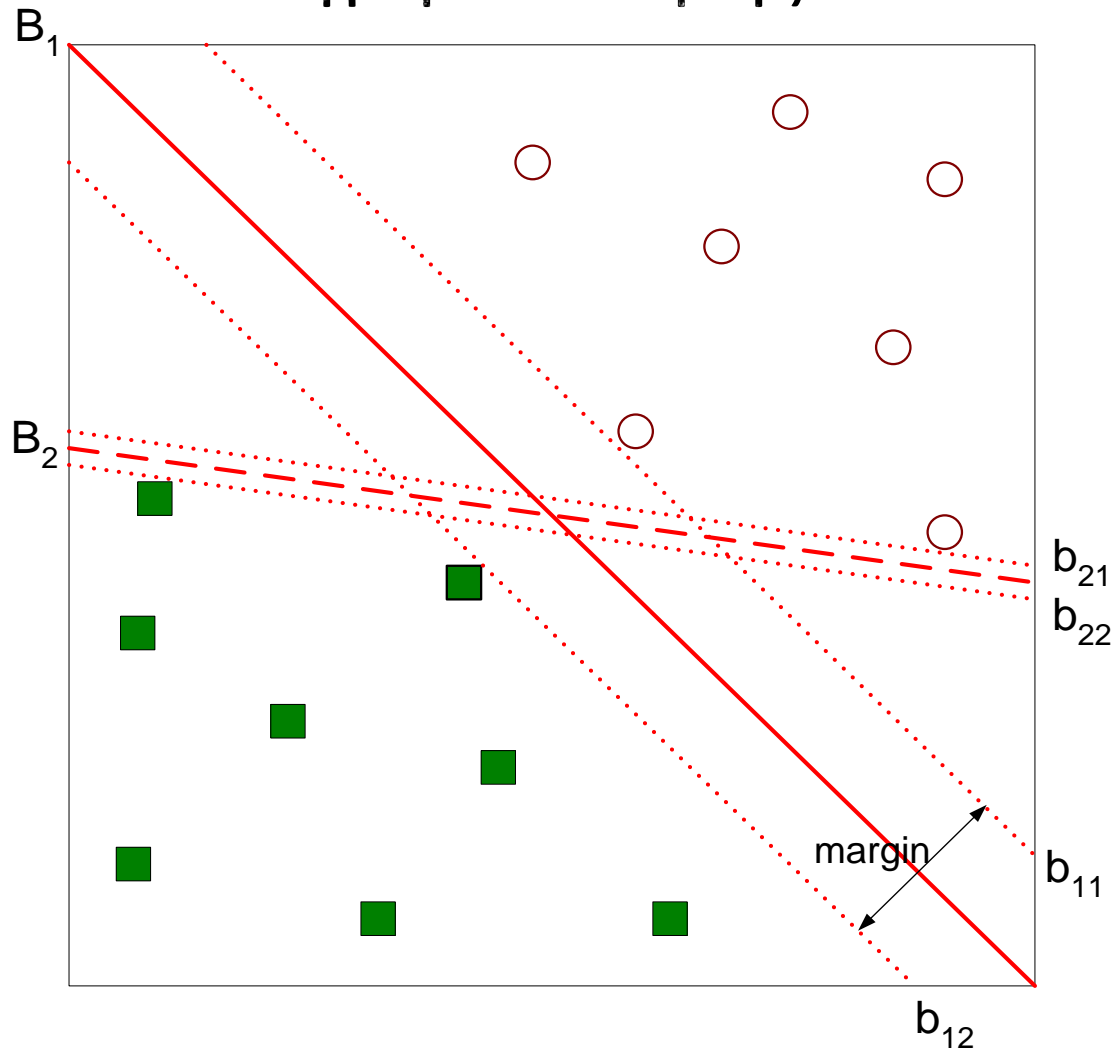
# Κατηγοριοποιητής SVM



- Ποια είναι καλύτερη η  $B_1$  ή η  $B_2$ ?
- Πώς ορίζεται το καλύτερη; Με ποιο κριτήριο;

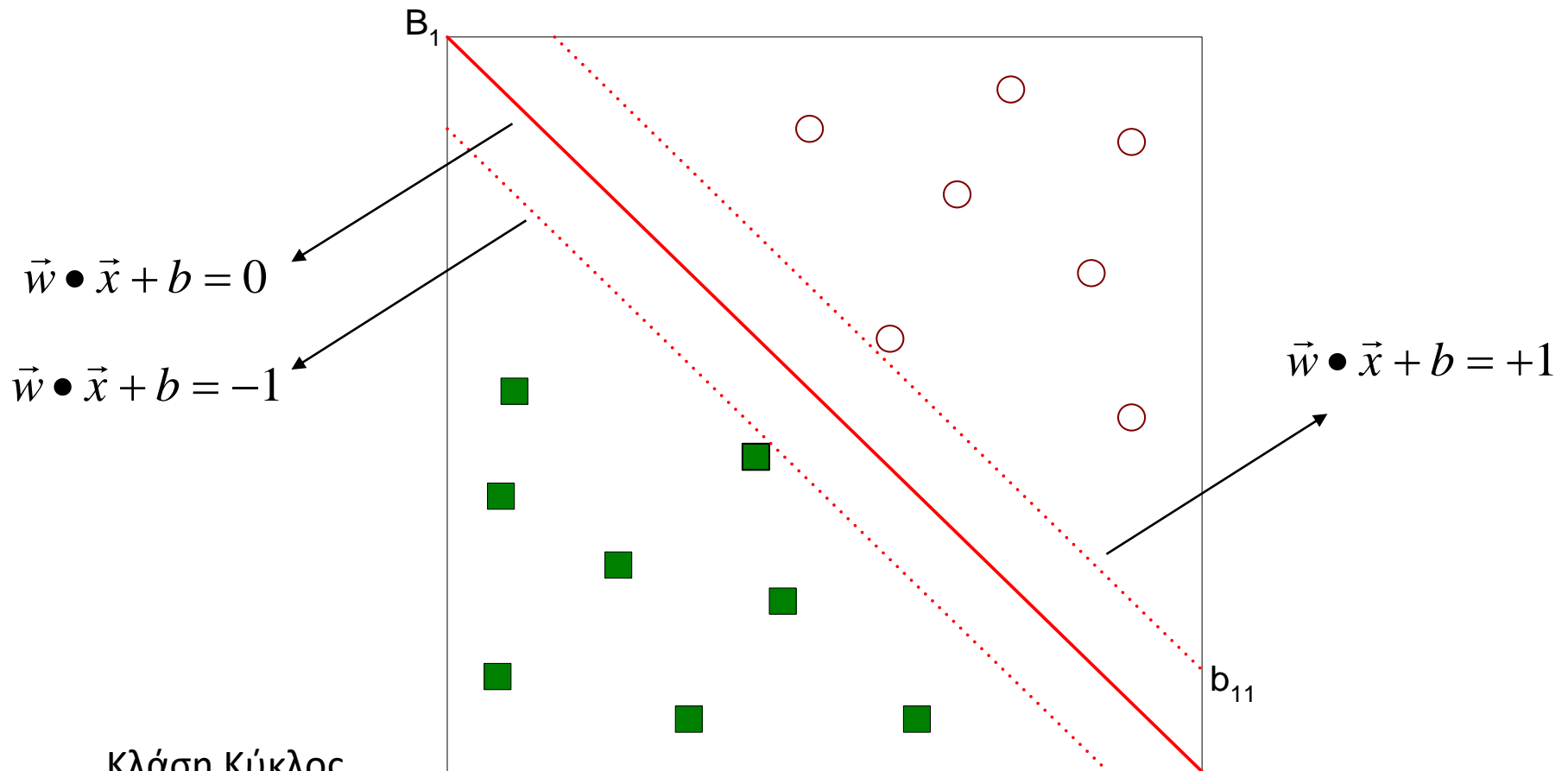


# Κατηγοριοποιητής SVM



- Προτιμότερο είναι το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο (margin)
  - το  $B_1$  είναι καλύτερο από το  $B_2$  (χωρητικότητα)

# Γραμμικό SVM



Κλάση Κύκλος

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

Κλάση Τετράγωνο

$$\text{Margin} = \frac{2}{\|\vec{w}\|^2}$$

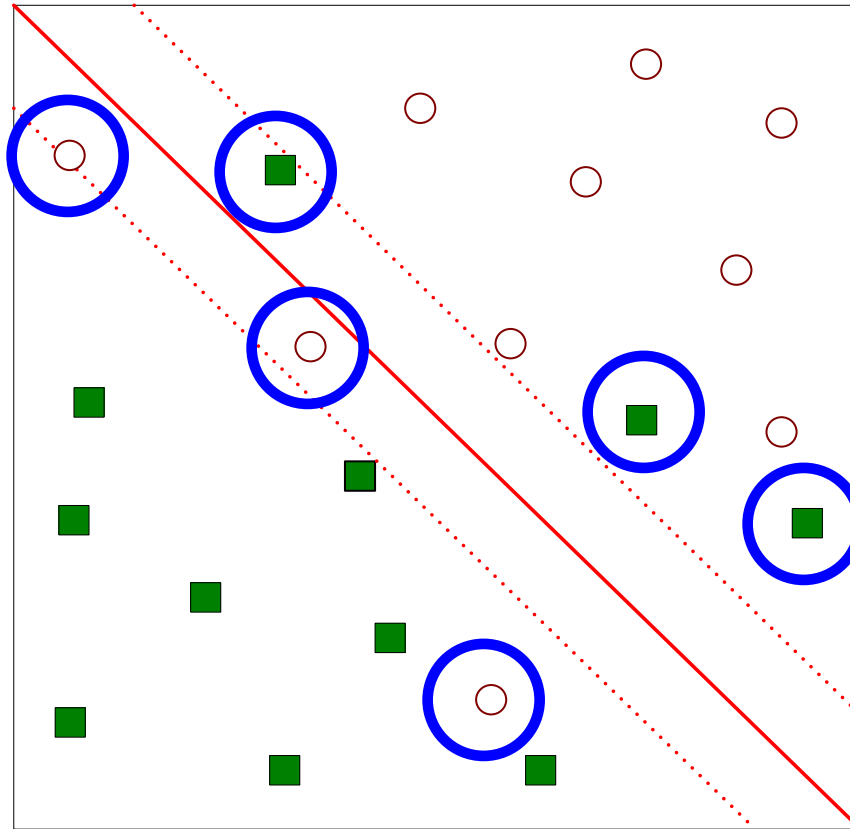
# Κατηγοριοποιητής SVM

- Θέλουμε να μεγιστοποιήσουμε:  $\text{Margin} = \frac{2}{\|\vec{w}\|^2}$
- Το οποίο είναι ισοδύναμο με το να ελαχιστοποιήσουμε:  $L(w) = \frac{\|\vec{w}\|^2}{2}$
- Με βάση τους παρακάτω περιορισμούς (constraints):

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

- Ένα πρόβλημα βελτιστοποίησης περιορισμών (constrained optimization problem)
  - Αριθμητικές μέθοδοι για την επίλυση του

# Μη Γραμμικώς διαχωρίσιμο πρόβλημα



Τι συμβαίνει αν το πρόβλημα δεν είναι γραμμικώς διαχωρίσιμο;

# Μη Γραμμικώς διαχωρίσιμο πρόβλημα

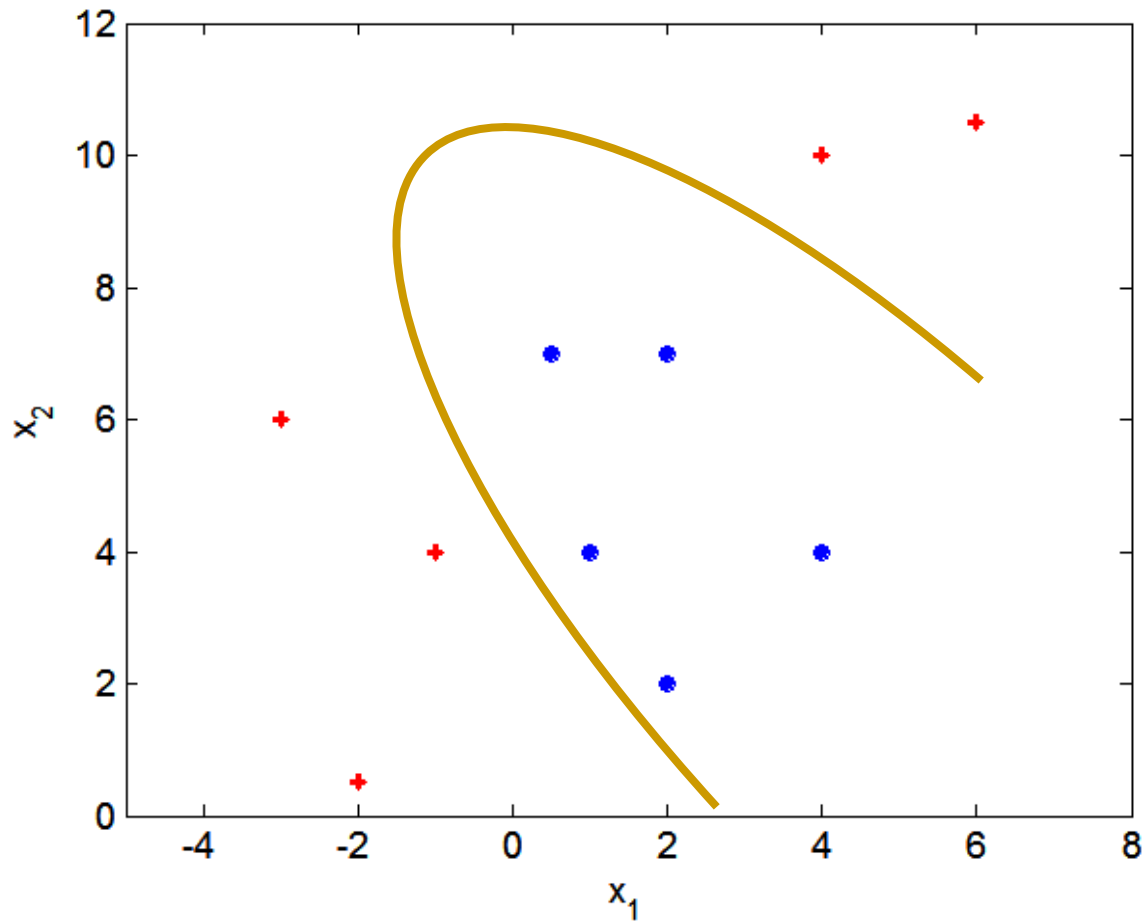
- Εισαγωγή χαλαρών μεταβλητών (slack variables)

- Ελαχιστοποίηση:
$$L(\mathbf{w}) = \frac{\|\vec{\mathbf{w}}\|^2}{2} + C \left( \sum_{i=1}^N \xi_i^k \right)$$

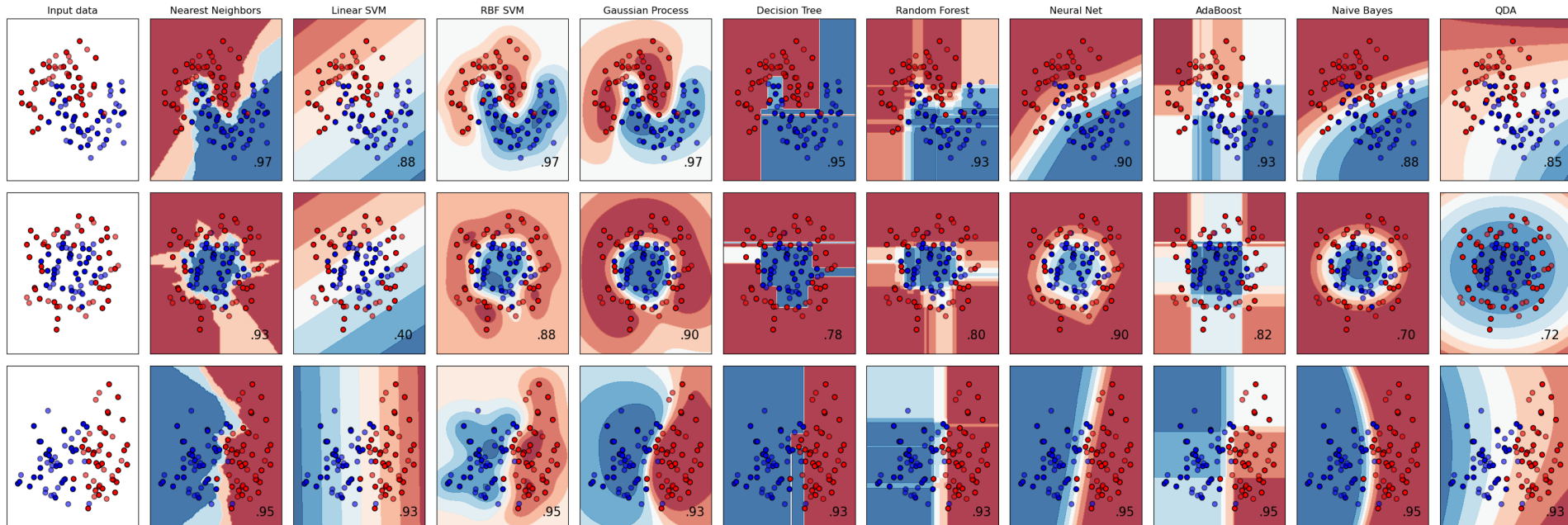
- Με τους περιορισμούς:

$$f(\vec{\mathbf{x}}_i) = \begin{cases} 1 & \text{if } \vec{\mathbf{w}} \bullet \vec{\mathbf{x}}_i + \mathbf{b} \geq 1 - \xi_i \\ -1 & \text{if } \vec{\mathbf{w}} \bullet \vec{\mathbf{x}}_i + \mathbf{b} \leq -1 + \xi_i \end{cases}$$

# Μη Γραμμικώς διαχωρίσιμο πρόβλημα

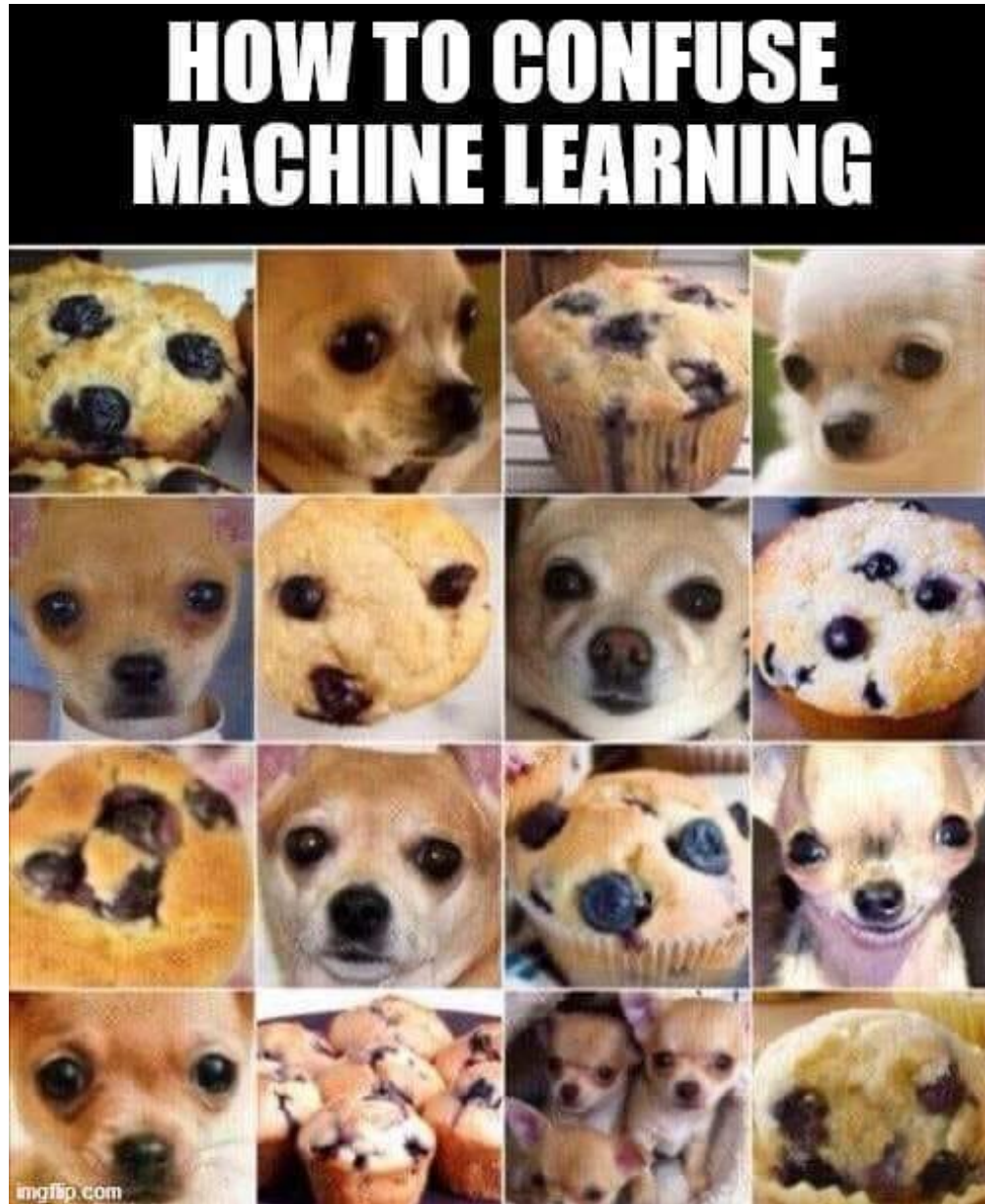


# Σύγκριση Ταξινομητών



[https://scikit-learn.org/auto\\_examples/classification/plot\\_classifier\\_comparison.html](https://scikit-learn.org/auto_examples/classification/plot_classifier_comparison.html)

# “How to confuse Machine Learning” meme





# “How to confuse Machine Learning” meme



“How to confuse Machine Learning” meme

Ποιο το πρόβλημα;

Ποια η Λύση;;

# GPT-4 ως Ταξινομητής

GPT-4 ⓘ

G



Please tell me what is it in each cell in a tabular format aligned with this picture grid.



Certainly! Here's a breakdown of the content of each cell in the format of a table:



Row	Column 1	Column 2	Column 3	Column 4
1	Muffin with blueberries	Chihuahua (side profile)	Blueberry Muffin	Chihuahua (profile looking up)
2	Chihuahua (frontal face)	Cookie with chocolate chips	Chihuahua (face close-up)	Blueberry Muffin with sugar on top
3	Blueberry Muffin with a smiley face	Chihuahua (frontal face close-up)	Blueberry Muffin with blueberries on top	Chihuahua (slightly showing teeth)
4	Chihuahua (puppy face close-up)	Group of Cookies with chocolate chips	Group of Chihuahua puppies	Muffin with blueberries

# Αναφορές

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, «Introduction to Data Mining», 2<sup>nd</sup> edition, Pearson, 2018, ISBN-13: 978-0133128901
- Jure Leskovec, Anand Rajaraman, Jeff Ullman, «Mining of Massive Datasets», Cambridge University Press, 2019, <http://www.mmds.org/>
- <https://www.cs.uoi.gr/~pitoura/courses/dm/>
- Ευστάθιος Κύρκος, «Επιχειρηματική ευφυΐα και εξόρυξη δεδομένων», 2015, ISBN: 978-960-603-109-0