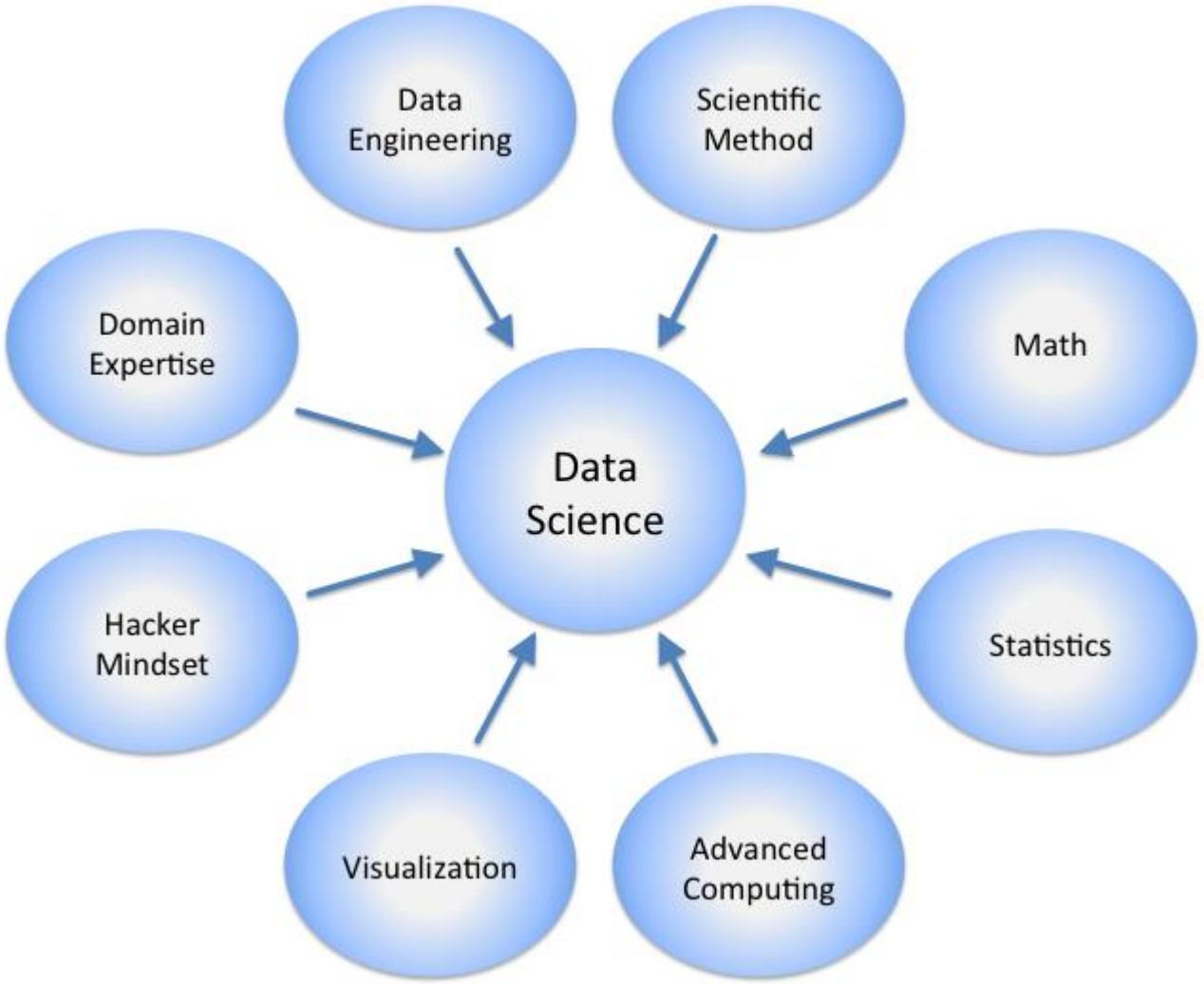
Τι είναι η Εξόρυξη Δεδομένων;

- Με απλά λόγια είναι:
  - Αποδοτικές τεχνικές για να αναλύσουμε πολύ μεγάλες συλλογές από δεδομένα και να εξάγουμε χρήσιμες πληροφορίες από αυτά

- Επιστήμη των Δεδομένων (Data Science)
  - Είναι ένας καινούριος όρος, ο οποίος ήρθε να αντικαταστήσει προγενέστερους όρους, όπως Ανακάλυψη Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Data-base) ή Εξόρυξη Δεδομένων (Data Mining).



Πηγή: https://en.wikipedia.org/wiki/Data_science

Παραδείγματα δεδομένων (1)

- Κυβερνητικά:
  - IRS (εφορία), δημογραφικά δεδομένα, «ΔΙΑΥΓΕΙΑ», …
- Αρχεία κειμένου (document data)
  - Web ως συλλογή κειμένων: δισεκατομμύρια σελίδες
  - Wikipedia: 4 εκατομμύρια λήμματα (που συνεχώς αυξάνονται)
  - Online συλλογές επιστημονικών άρθρων
- Μεγάλες εταιρίες
  - WALMART: 20M συναλλαγές την ημέρα
  - MOBIL: 100 TB γεωλογικά σύνολα δεδομένων
  - AT&T 300 M κλήσεις την ημέρα
  - Εταιρίες πιστωτικών καρτών

Παραδείγματα δεδομένων (2)

- Επιστημονικά
  - NASA, EOS project: 50 GB την ώρα
- Παράδειγμα: Διαδικτυακά δεδομένα
  - Web: 50 δισεκατομμύρια σελίδες διασυνδεδεμένες
  - Facebook: 400 εκατομμύρια χρήστες
  - MySpace: 300 εκατομμύρια χρήστες
  - Instant messenger: ~ 1 δισεκατομμύρια χρήστες
  - Blogs: 250 εκατομμύρια blogs

Εξόρυξη Δεδομένων (Ορισμός)

- Πολύ μεγάλα σύνολα δεδομένων (data sets)
- (1) η διαδικασία ανακάλυψης (discovery) προτύπων (patterns) που πριν δεν ήταν γνωστά, ισχύουν, είναι πιθανόν χρήσιμα και είναι κατανοητά
- (2) η ανάλυση τους για να βρούμε μη αναμενόμενες σχέσεις ανάμεσα στα δεδομένα καθώς και να τα συνοψίσουμε με νέους τρόπους που είναι κατανοητοί και χρήσιμοι στους χρήστες

Γιατί είναι χρήσιμη; Εμπορική πλευρά

- Πολλά δεδομένα συγκεντρώνονται και εισάγονται σε αποθήκες δεδομένων ή είναι διαθέσιμα στο διαδίκτυο
  - Αγορές σε πολύ-καταστήματα/αλυσίδες
  - Συναλλαγές με τράπεζες/πιστωτικές κάρτες
  - Web, web logs
  - Network traffic
  - Κοινωνικά δίκτυα (emails, συστήματα δικτύωσης)
- Σχεδιασμός καλύτερων συστημάτων
  - Αποφυγή spam, αποδοτικότητα
- Μεγάλος ανταγωνισμός
  - Παροχή καλύτερων, προσωπικών υπηρεσιών σε κάποιο πεδίο (fraud detection, target marketing)

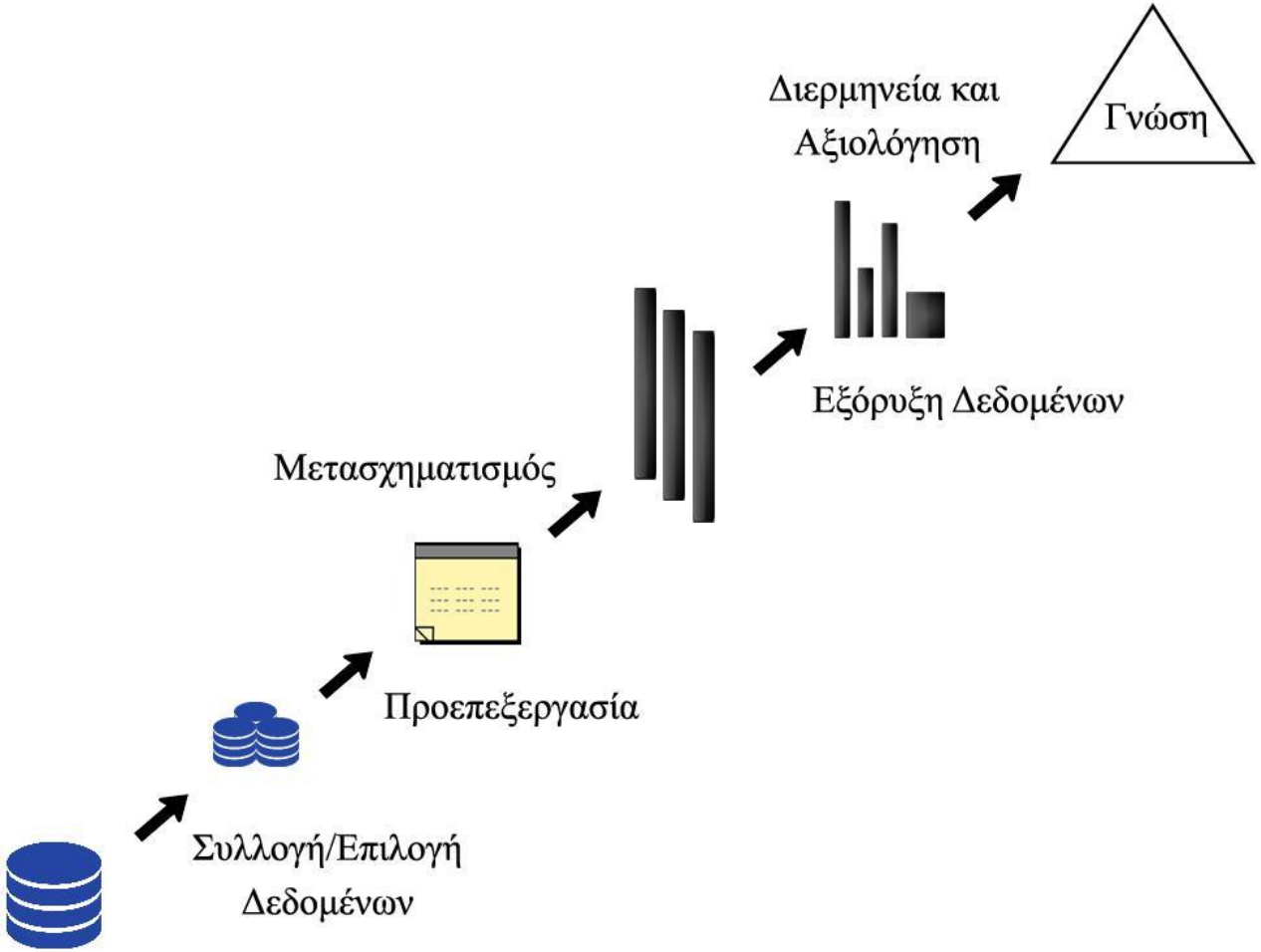
Γιατί είναι χρήσιμη; Επιστημονική πλευρά

- Τα δεδομένα συλλέγονται και αποθηκεύονται σε τρομερές ταχύτητες (GB/hour)
  - Απομακρυσμένοι αισθητήρες (remote sensors) σε δορυφόρους
  - Τηλεσκόπια στον ουρανό
  - Microarrays που παράγουν γονιδιακά δεδομένα
  - Επιστημονικές προσομοιώσεις που παράγουν terabytes δεδομένων
- Η εξόρυξη δεδομένων μπορεί να βοηθήσει τους επιστήμονες
  - Στην κατηγοριοποίηση και την τμηματοποίηση των δεδομένων
  - Στη διατύπωση υποθέσεων

Ανακάλυψη Γνώσης από Βάσεις Δεδομένων

- Συλλογή Δεδομένων (Data Collection)
- Προεπεξεργασία Δεδομένων (Preprocessing)
- Μετασχηματισμός Δεδομένων (Transformation)
- **Εξόρυξη Δεδομένων (Data Mining)**
- Διερμηνεία και Αξιολόγηση (Interpretation/Evaluation)

Βασικά στάδια Ανακάλυψης Γνώσης από Βάσεις Δεδομένων



Πηγή: http://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2966/1/02_chapter_01.pdf

Προεπεξεργασία

- Ένα απλό παράδειγμα

Ετήσιο Εισόδημα	Πιστοληπτική ικανότητα	Έγκριση δανείου
15000	Μέτρια	Ναι
12000	Κακή	Όχι
	Μέτρια	Όχι
50000	Καλή	Ναι
30000		Ναι
16000	Κακή	Όχι

  - Διαγραφή ολόκληρης της γραμμής;
  - Αναζήτηση και καταχώρηση της πραγματικής τιμής;
  - Χρήση μιας σταθερής τιμής για όλες τις χαμένες τιμές;
  - Αντικατάσταση της χαμένης τιμής με τη μέση τιμή της στήλης;

Γιατί χρειάζεται η προεπεξεργασία;

- Χαμένες Τιμές
- Θορυβώδη Δεδομένα
- Κανονικοποίηση
- Κατασκευή νέων πεδίων
- Μείωση Διαστάσεων (δηλ. στηλών) και Επιλογή Χαρακτηριστικών

Δείτε περισσότερες πληροφορίες:
http://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/1234/2/Kef._7.pdf

Είδη/Τεχνικές Εξόρυξης Δεδομένων (συνοπτικά)

- Ομαδοποίηση (συσταδοποίηση) – **clustering**
  - χωρίζουμε τα δεδομένα σε ομάδες από «όμοια» σύνολα
- Κανόνες συσχέτισης (**Association rule mining**)
  - βρίσκουμε συσχετίσεις ανάμεσα στα δεδομένα, π.χ. ποια δεδομένα εμφανίζονται συχνά μαζί σε συναλλαγές
- Κατηγοριοποίηση (**Classification**)
  - κατηγοριοποιούμε τα δεδομένα τοποθετώντας τα σε μια (ή περισσότερες) από έναν αριθμό από δοσμένες κατηγορίες

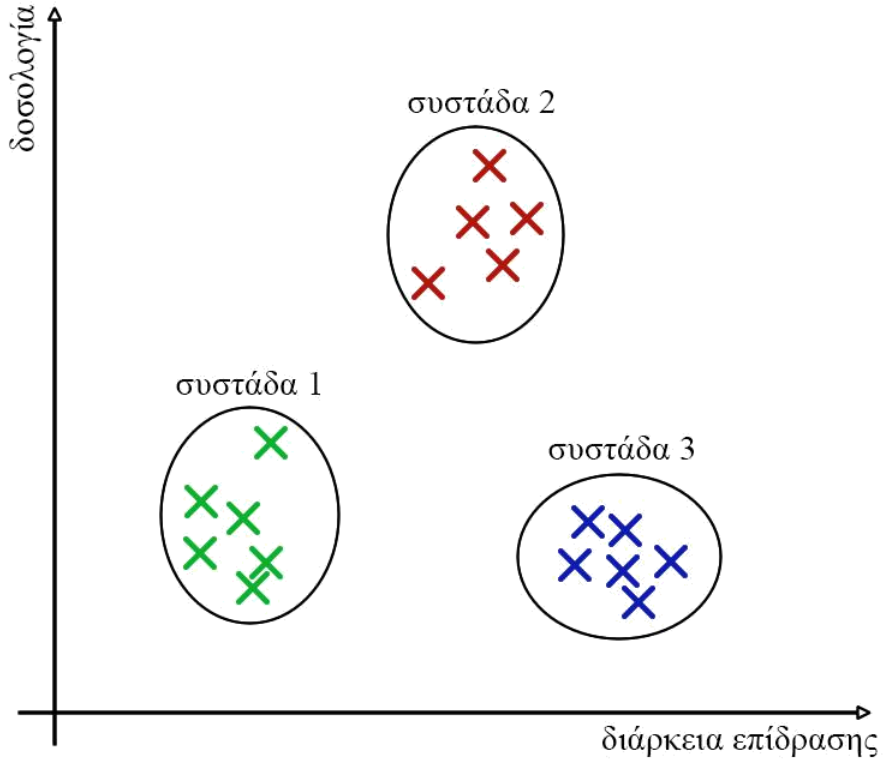
Κατηγοριοποίηση

- Πρόκειται για μια προγνωστική (προβλεπτική) μέθοδο.
  - Στόχος είναι η δημιουργία ενός μοντέλου – κατηγοριοποιητή (classifier) με βάση τα υπάρχοντα δεδομένα.
  - Ουσιαστικά, είναι η μάθηση μιας συνάρτησης, η οποία απεικονίζει ένα αντικείμενο σε μια κλάση (ή κατηγορία).

Συσταδοποίηση

- Η συσταδοποίηση (clustering) είναι μια περιγραφική μέθοδος.
  - Έχοντας ένα σύνολο δεδομένων, στόχος της συσταδοποίησης είναι η δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν όμοια ή παρεμφερή δείγματα.
  - Ουσιαστικά αναζητείται ένα πεπερασμένο σύνολο κατηγοριών ή συστάδων, για να περιγράψει τα δεδομένα. Οι κατηγορίες μπορεί να είναι αμοιβαία αποκλειόμενες και εξαντλητικές ή να έχουν μία πιο σύνθετη αναπαράσταση, όπως για παράδειγμα ιεραρχικές και επικαλυπτόμενες.

Παράδειγμα συσταδοποίησης



Πηγή: http://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2966/1/02_chapter_01.pdf

Παράδειγμα συσταδοποίησης

- Στόχος: Χωρισμός των καταναλωτών σε ομάδες έτσι ώστε τα μέλη κάθε ομάδας να είναι ο στόχος για μια συγκεκριμένη πολιτική marketing

- Προσέγγιση:
  - Συγκέντρωση διαφορετικών γνωρισμάτων για τους καταναλωτές
  - Ορισμός «ομοιότητας» ανάμεσα στους πελάτες
  - Δημιουργία ομάδων με όμοιους πελάτες
  - Μέτρηση της ποιότητας της ομαδοποίησης (πχ παρατηρώντας τις αγοραστικές συνήθειες στην ίδια ομάδα και ανάμεσα σε διαφορετικές ομάδες)

Γενική εικόνα της διαδικασίας εξόρυξης…

- Εκμάθηση του πεδίου εφαρμογής
  - Σχετική προηγούμενη γνώση και τους στόχους της εφαρμογής
- Δημιουργία του συνόλου δεδομένων: data selection
- Καθαρισμός και προ-επεξεργασία των δεδομένων: (έως και 60% της συνολικής προσπάθειας)
- Ελάττωση δεδομένων και μετασχηματισμοί
  - Χρήσιμα χαρακτηριστικά, ελάττωση διαστάσεων κλπ

- Επιλογή λειτουργίας εξόρυξης δεδομένων
  - πχ, συσταδοποίηση, ταξινόμηση, κλπ
- Επιλογή του αλγορίθμου εξόρυξης δεδομένων
- Εξόρυξη Δεδομένων: αναζήτηση προτύπων ενδιαφέροντος

- Εκτίμηση προτύπων και αναπαράσταση γνώσης
  - οπτικοποίηση, μετασχηματισμοί, απομάκρυνση περιττών προτύπων, κλπ

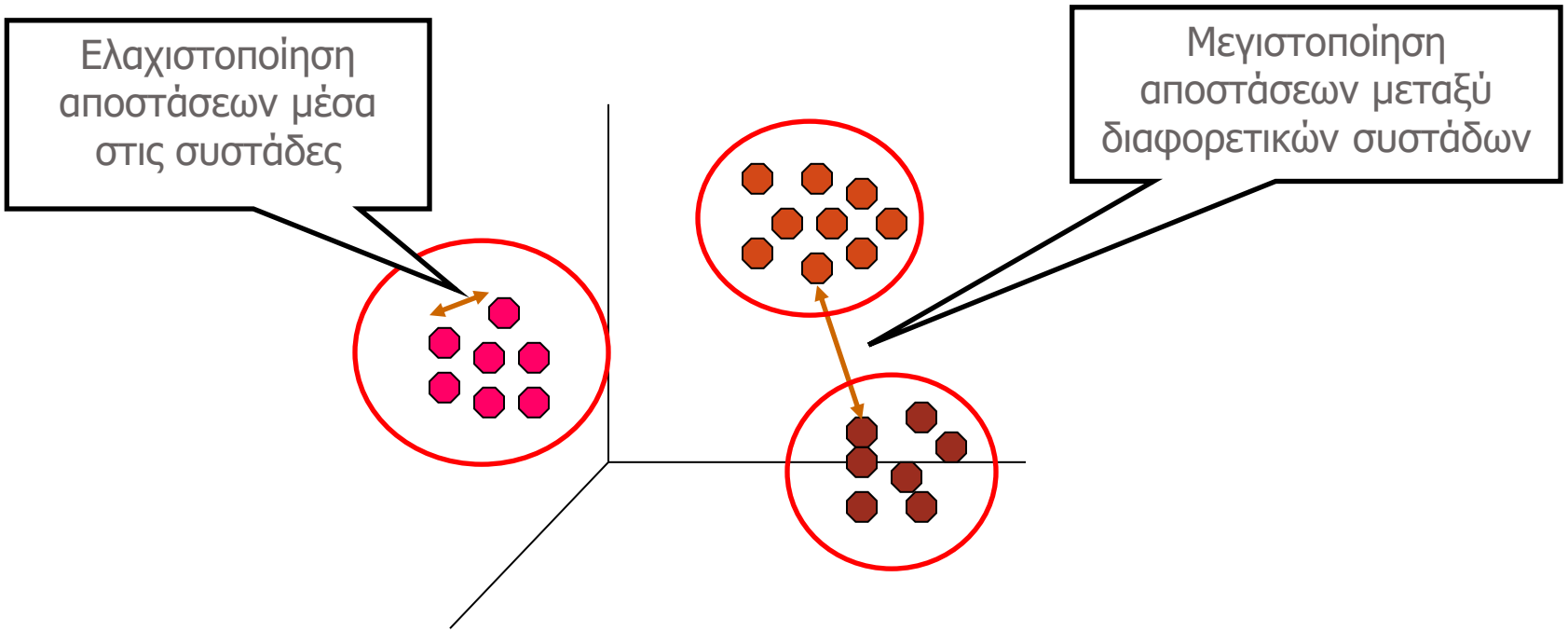
- Χρήση της γνώσης

Συσταδοποίηση (clustering)

- Στο πρόβλημα της συσταδοποίησης μας δίνεται ένα σύνολο δεδομένων, χωρίς τις αντίστοιχες κλάσεις ή ετικέτες και χρειαζόμαστε κάποιον αλγόριθμο, ο οποίος θα ομαδοποιήσει αυτόματα τα δεδομένα σε συστάδες.
- Οι συστάδες που δημιουργούνται θέλουμε να διαχωρίζουν ορθά τα δεδομένα.
  - Αυτό πρακτικά σημαίνει ότι μια συστάδα θέλουμε να απαρτίζεται από αντικείμενα, όπου κάθε αντικείμενο είναι πιο κοντά σε κάθε άλλο αντικείμενο της ίδιας συστάδας απ' ότι σε κάποιο άλλο αντικείμενο διαφορετικής συστάδας.

Βασικός στόχος

- Εύρεση συστάδων (ομάδων) αντικειμένων έτσι ώστε τα αντικείμενα σε κάθε συστάδα να είναι όμοια (ή να σχετίζονται) και διαφορετικά (ή μη σχετιζόμενα) από τα αντικείμενα των άλλων συστάδων



Εφαρμογές

- ομαδοποίηση γονιδίων και πρωτεϊνών που έχουν την ίδια λειτουργία,
- εικόνες,
- χαρακτηριστικά ασθενειών
- μετοχών με παρόμοια διακύμανση τιμών,
- ομαδοποίηση weblog για εύρεση παρόμοιων προτύπων προσπέλασης,
- ομαδοποίηση σχετιζόμενων αρχείων για browsing,
- ομαδοποίηση κειμένων
- πελάτες με παρόμοια συμπεριφορά

Συσταδοποίηση επιπέδου βροχής

Διφορούμενη η Έννοια Συστάδας

Αρχικά δεδομένα: Πόσες συστάδες υπάρχουν;

Έξι συστάδες

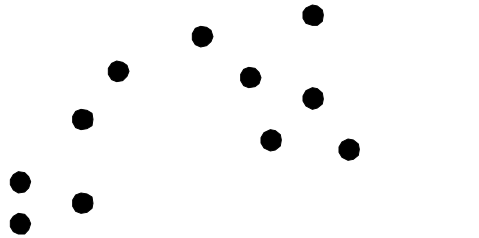
Δύο συστάδες

Τέσσερεις συστάδες

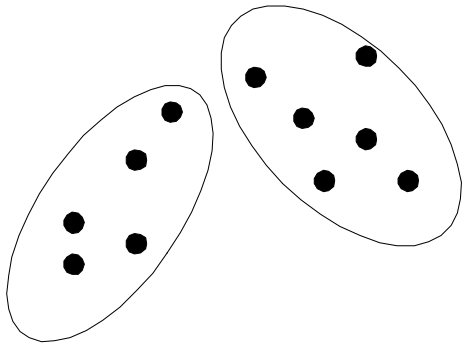
Τύποι Συσταδοποίησης

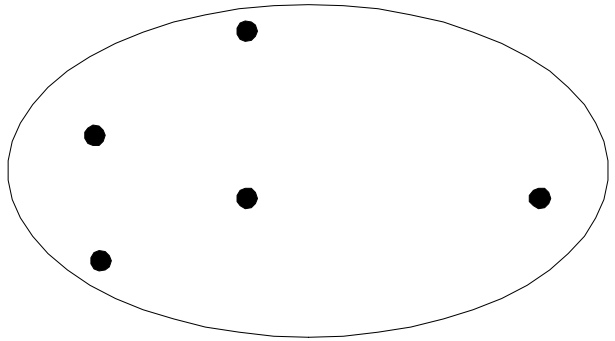
- Μία συσταδοποίηση είναι ένα σύνολο από συστάδες
- Σημαντική διάκριση: διαμεριστική και ιεραρχική
- Διαμεριστική:
  - μία διαίρεση των αντικειμένων σε μη επικαλυπτόμενα υποσύνολα
- Ιεραρχική:
  - ένα σύνολο ένθετων συστάδων οργανωμένες σε ένα ιεραρχικό δέντρο
  - Επιτρέπουμε σε μια συστάδα να έχει υπο-συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο

Διαμεριστική Συσταδοποίηση



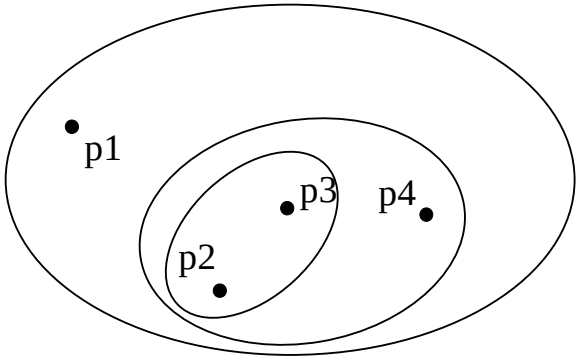
Αρχικά Σημεία



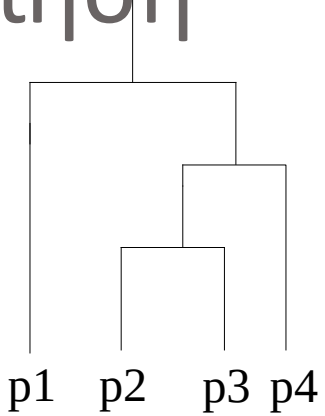


Διαμεριστική Συσταδοποίηση

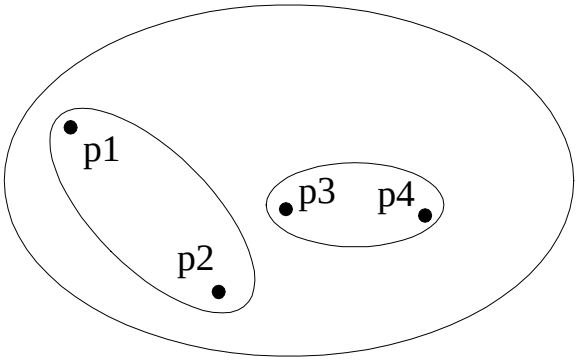
Ιεραρχική Συσταδοποίηση



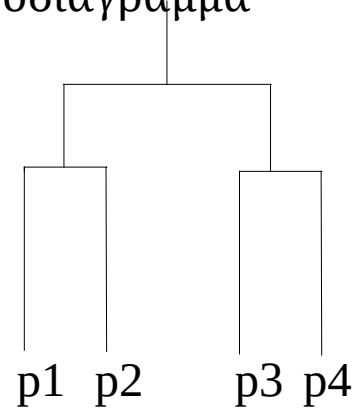
Παραδοσιακή Ιεραρχική Συσταδοποίηση



Παραδοσιακό Δενδροδιάγραμμα



Μη παραδοσιακή Ιεραρχική Συσταδοποίηση



Μη παραδοσιακό Δενδροδιάγραμμα

Τύποι Συσταδοποιήσεων

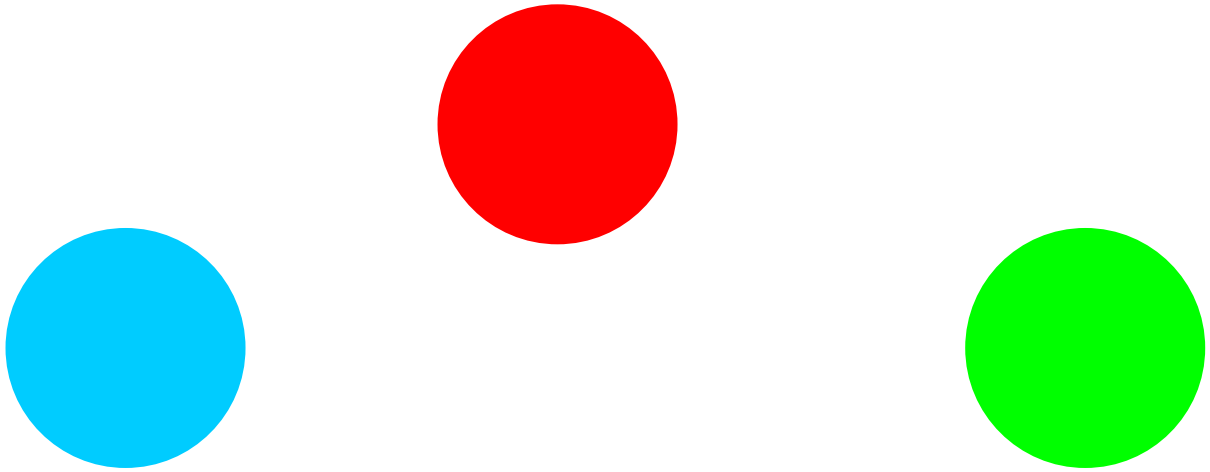
- Καλά διαχωρισμένες
- Βασισμένες σε πρότυπο
- Βασισμένες στη γειτνίαση
- Βασισμένες στην πυκνότητα
- Εννοιολογικές συστάδες

Ιδιότητες των Δεδομένων Εισόδου

- Μέτρο εγγύτητας/πυκνότητας
- Σποραδικότητα
- Τύποι γνωρισμάτων
- Τύποι δεδομένων
- Θόρυβος και ακραία σημεία

Τύποι συστάδων: Καλώς Διαχωρισμένες Συστάδες

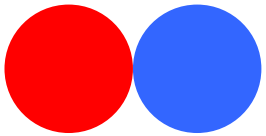
- Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας συστάδας είναι κοντινότερο σε (ή πιο όμοιο με) όλα τα άλλα σημεία της συστάδας από ότι σε οποιοδήποτε άλλο σημείο που δεν ανήκει στη συστάδα.



3 καλώς-διαχωρισμένες συστάδες

Τύποι συστάδων: Συστάδες βασισμένες σε κέντρο ή πρότυπο

- Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοιο ώστε ένα αντικείμενο στην συστάδα είναι κοντινότερο σε (ή πιο όμοιο με) το «κέντρο» ή πρότυπο της συστάδας από ότι από το κέντρο οποιασδήποτε άλλης συστάδας.
- Το κέντρο της ομάδας είναι συχνά
  - centroid (κεντροειδές), ο μέσος όρος των σημείων της συστάδας, ή
  - a medoid, το πιο «αντιπροσωπευτικό» σημείο της συστάδας

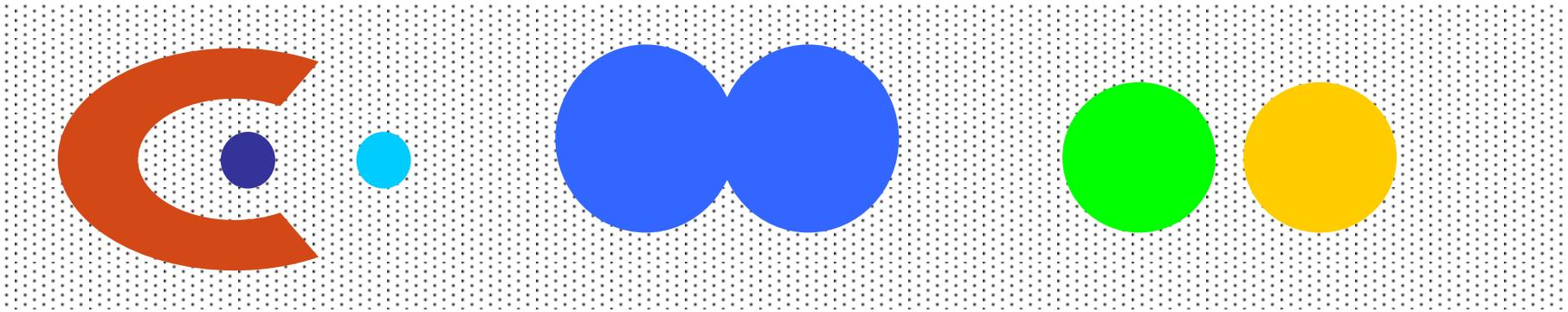


4 συστάδες βασισμένες σε κέντρο

Τείνουν στο να είναι κυκλικές

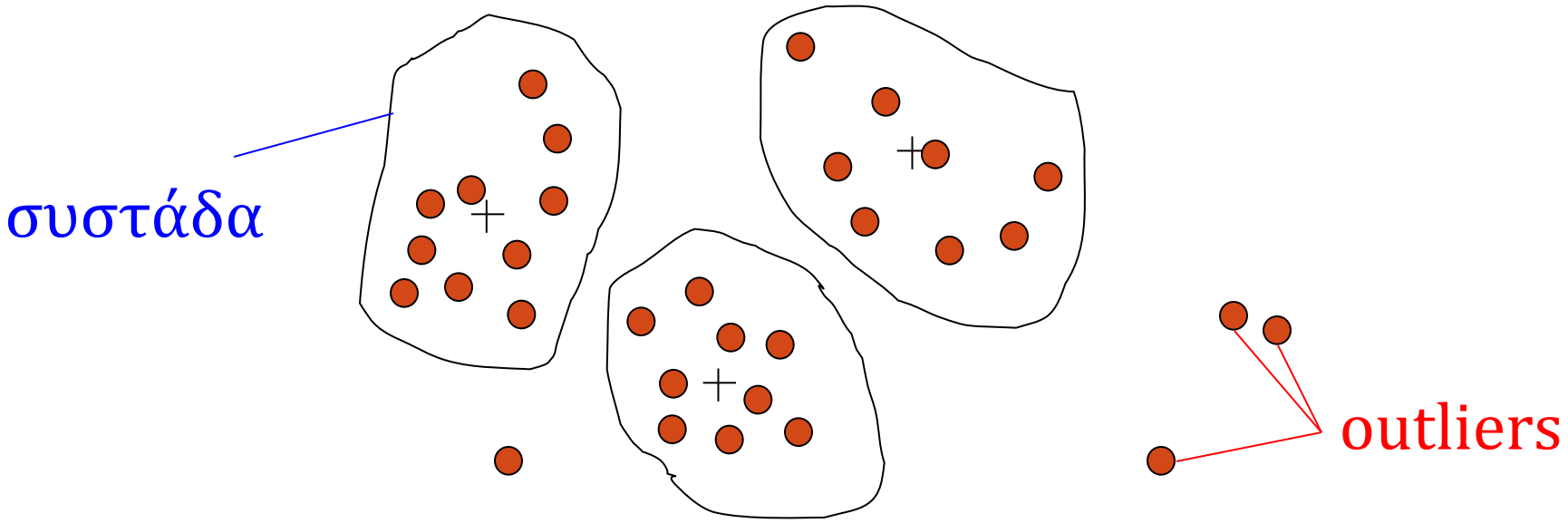
Τύποι συστάδων: Συστάδες βασισμένες στην πυκνότητα

- Μια συστάδα είναι μια πυκνή περιοχή από σημεία την οποία χωρίζουν από άλλες περιοχές μεγάλης πυκνότητας περιοχές χαμηλής πυκνότητας
- Συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν υπάρχει θόρυβος ή outliers



6 συστάδες βασισμένες στην πυκνότητα

Προβλήματα με την εύρεση συστάδων



Outlier (ακραίο σημείο) τιμές που είναι εξαιρέσεις ως προς τις συνηθισμένες ή αναμενόμενες τιμές

Γνωστοί μέθοδοι συσταδοποίησης

- Κ-μέσων και οι παραλλαγές του
- Ιεραρχική συσταδοποίηση
- Συσταδοποίηση βασισμένη στην πυκνότητα

Συσταδοποίηση Κ-Μέσων

- Διαμεριστική προσέγγιση
- Κάθε συστάδα σχετίζεται με ένα κέντρο βάρους (κεντρικό σημείο, centroid)
- Κάθε σημείο εκχωρείται στη συστάδα με το κοντινότερο κέντρο βάρους
- Το πλήθος των συστάδων, Κ, πρέπει να καθοριστεί, εξ αρχής
- Ο βασικός αλγόριθμος είναι απλός

Ο βασικός Αλγόριθμος των Κ-Μέσων

1: Επιλογή K σημείων ως τα αρχικά κεντρικά σημεία

2: **Repeat**

3: Ανάθεση όλων των αρχικών σημείων στο *κοντινότερο* τους από τα K κεντρικά σημεία

4: Επανα-υπολογισμός του *κεντρικού σημείου* κάθε συστάδας

5: **Until** τα κεντρικά σημεία να μην αλλάζουν

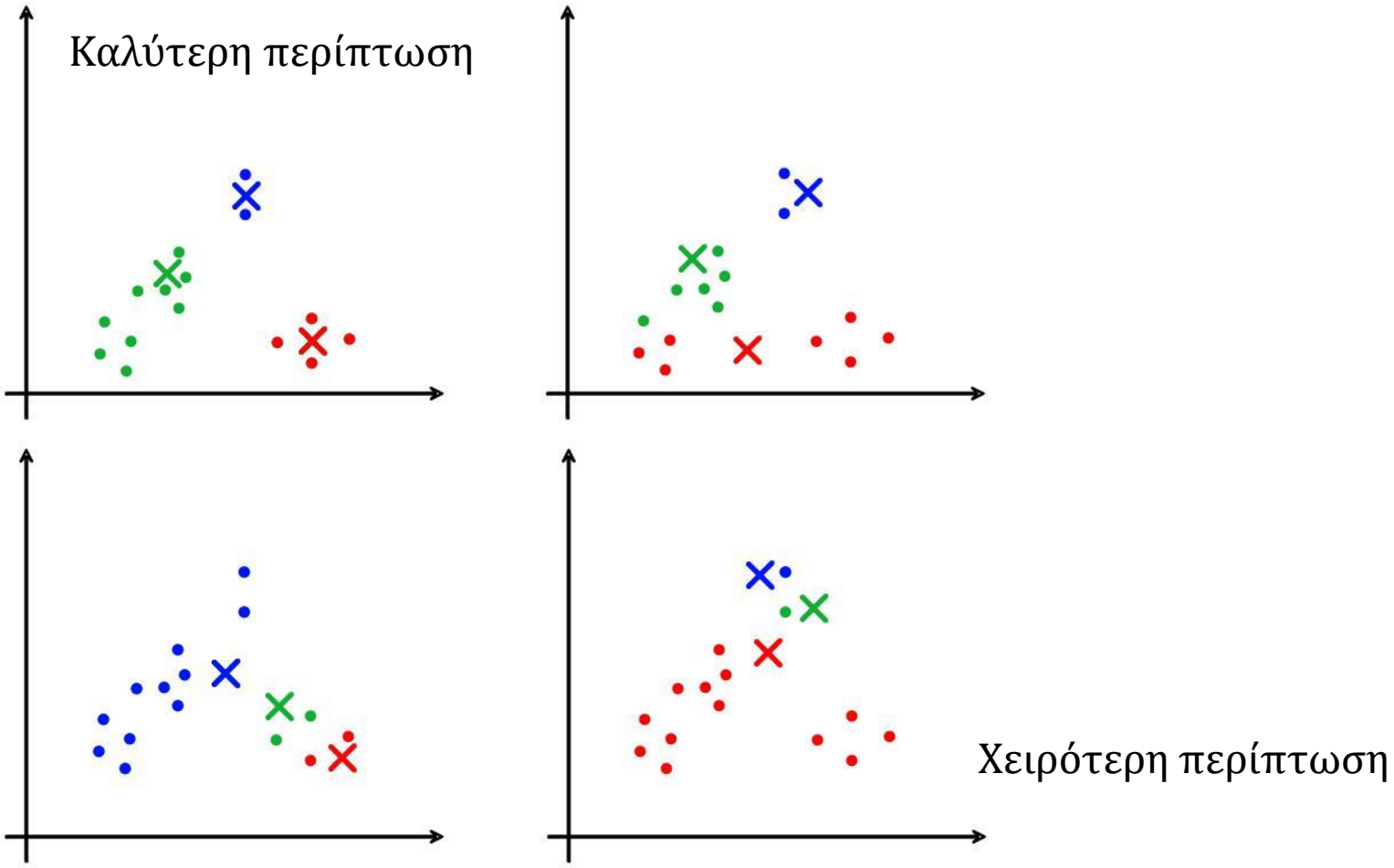
K-Μέσοι … Λεπτομέρειες

- Τα αρχικά κέντρα βάρους επιλέγονται τυχαία
- Το κέντρο βάρους είναι (τυπικά) ο μέσος των σημείων της συστάδας
- Η «εγγύτητα» μετριέται με την Ευκλείδεια Απόσταση, ομοιότητα συνημιτόνου, συσχέτιση…

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2 + \cdots + (y_n - x_n)^2} = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}.$$

- Η σύγκλιση γίνεται στις πρώτες επαναλήψεις
- Για την επιλογή του αριθμού των συστάδων δεν υπάρχει κάποιος γενικός κανόνας, ο οποίος να λειτουργεί εγγυημένα

Τυχαία Αρχικοποίηση Κεντροειδών



Ζητήματα γύρω από τον αλγόριθμο

- Ουσιαστικά, ο αλγόριθμος προσπαθεί επαναληπτικά να «μειώσει» την απόσταση όλων των σημείων από ένα σημείο της συστάδας
- Η πιο συνηθισμένη μέτρηση είναι το άθροισμα των τετράγωνων του λάθους (Sum of Squared Error (SSE))

- Για κάθε σημείο, το λάθος είναι η απόστασή του από την κοντινότερη συστάδα

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- Για να πάρουμε το SSE, παίρνουμε το τετράγωνο αυτών των λαθών και τα προσθέτουμε

- Όπου dist Ευκλείδεια απόσταση, x είναι ένα σημείο στη συστάδα C_i και m_i είναι ο αντιπρόσωπος (κεντρικό σημείο) της συστάδας C_i

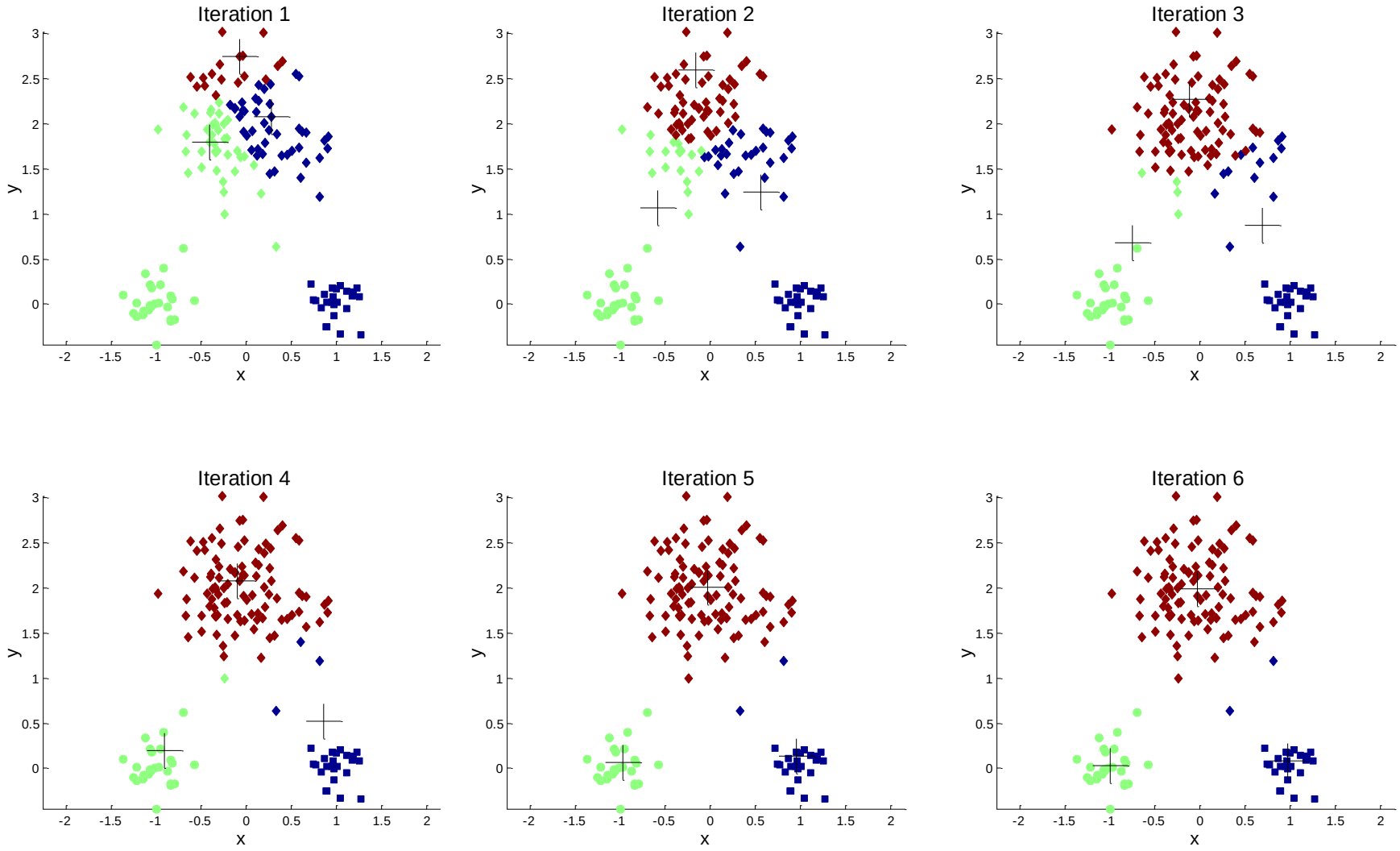
- Μπορούμε να δείξουμε ότι το σημείο που ελαχιστοποιεί το SSE για τη συστάδα είναι ο μέσος όρος $c_i = 1/m_i \Sigma_{x \in Ci} C_i x$

- Δοθέντων δύο συστάδων, μπορούμε να επιλέξουμε αυτήν με το μικρότερο λάθος

Εκτίμηση των Συστάδων

- Άθροισμα των τετραγωνικών λαθών (SSE)
- Για κάθε σημείο, το λάθος είναι η απόσταση από την κοντινότερη συστάδα
- *Ένας εύκολος τρόπος για την μείωση του SSE είναι η αύξηση του K, δηλαδή του πλήθους των συστάδων*

K-means: Επιλογή αρχικών σημείων



- Η πολυπλοκότητα του αλγορίθμου k-μέσων είναι O(I * n * K * d)

  n = αριθμός σημείων,

  K = αριθμός συστάδων,

  I = αριθμός επαναλήψεων,

  d = αριθμός γνωρισμάτων (διάσταση)

K-means: Επιλογή αρχικών σημείων

Αν υπάρχουν K «πραγματικές συστάδες» η πιθανότητα να επιλέξουμε ένα κέντρο από κάθε συστάδα είναι μικρή, συγκεκριμένα αν όλες οι συστάδες έχουν το ίδιο μέγεθος n, τότε:

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

Για παράδειγμα, αν K = 10, η πιθανότητα είναι = $10!/1010 = 0.00036$

Ένα από τα μειονεκτήματα του αλγορίθμου k-means είναι το γεγονός ότι δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του K, δηλαδή του αριθμού των συστάδων (δηλ. των αρχικών σημείων)

Κάποιες λύσεις για την επιλογή αρχικών σημείων

- Πολλαπλά τρεξίματα
- Βοηθά, αλλά πολλές περιπτώσεις

- Δειγματοληψία και χρήση κάποιας ιεραρχικής τεχνικής

- Επιλογή παραπάνω από k αρχικών σημείων και μετά επιλογή k από αυτά τα αρχικά κεντρικά σημεία (πχ τα πιο απομακρυσμένα μεταξύ τους)

- Σταδιακή επιλογή
- Επιλογή του πρώτου σημείου τυχαία ή ως το μέσο όλων των σημείων
- Για καθένα από τα υπόλοιπα αρχικά σημεία
- επέλεξε αυτό που είναι πιο μακριά από τα μέχρι τώρα επιλεγμένα αρχικά σημεία

- Μπορεί να οδηγήσει στην επιλογή outliers
- Ο υπολογισμός του πιο απομακρυσμένου σημείου είναι δαπανηρός
- Συχνά εφαρμόζεται σε δείγματα

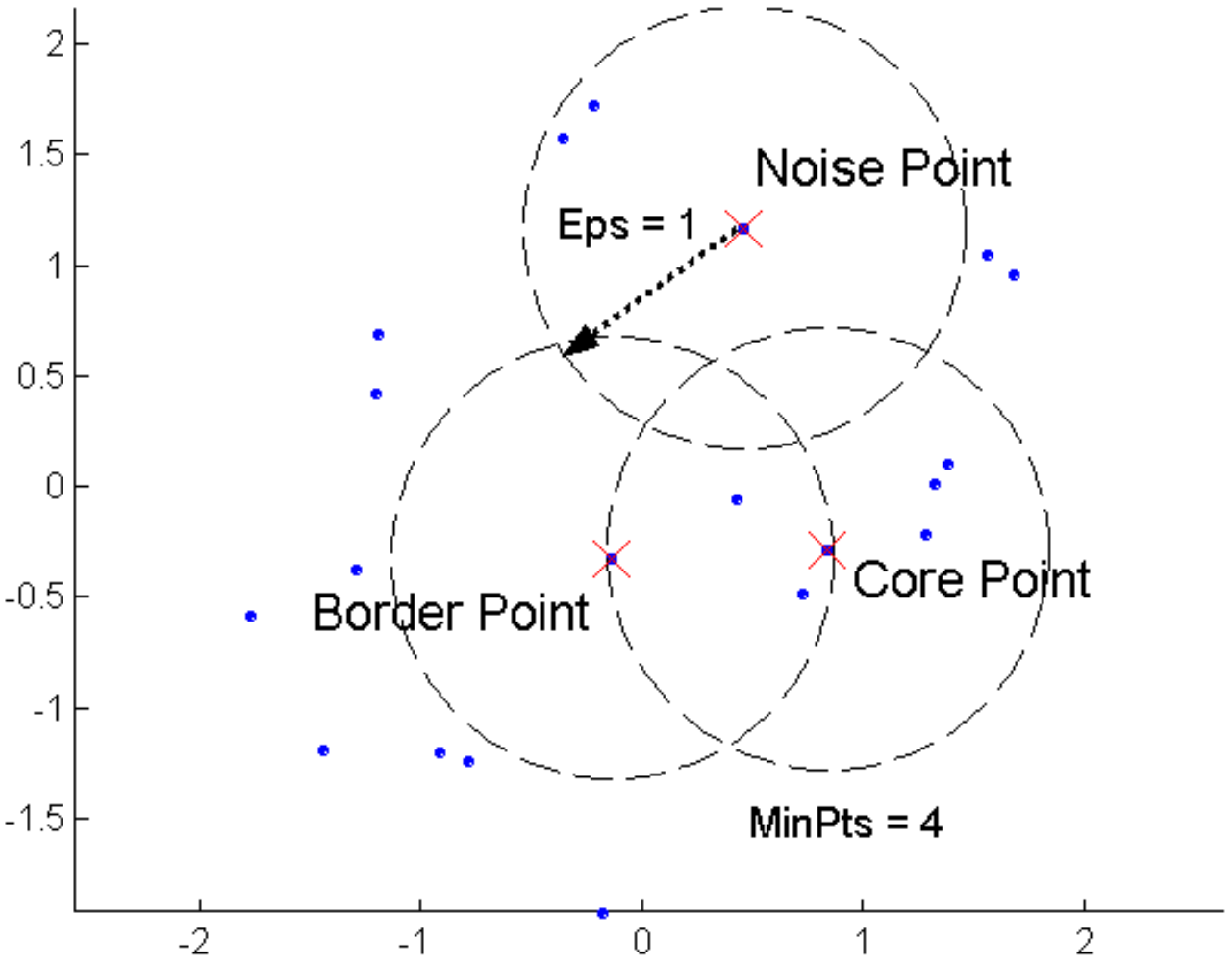
DBSCAN

- Είναι ένας αλγόριθμος συσταδοποίησης που βασίζεται στην πυκνότητα
- Η πυκνότητα ορίζεται ως το πλήθος των σημείων μέσα σε μία καθορισμένη ακτίνα (**Eps**)
- Ένα σημείο αποτελεί σημείο πυρήνα, εάν περιέχει περισσότερα από ένα καθορισμένο πλήθος σημείων (**MinPts**) μέσα σε ακτίνα Eps

DBSCAN

- Τα σημεία πυρήνα βρίσκονται στο εσωτερικό μίας συστάδας
- Ένα σημείο ορίου (περιθωρίου) έχει λιγότερα από MinPts μέσα σε ακτίνα Eps, αλλά βρίσκεται στην γειτονιά ενός σημείου πυρήνα
- Ένα σημείο θορύβου είναι οποιοδήποτε σημείο που δεν είναι ούτε σημείο πυρήνα ούτε σημείο ορίου.

DBSCAN: Σημεία Πυρήνα, Ορίου και Θορύβου

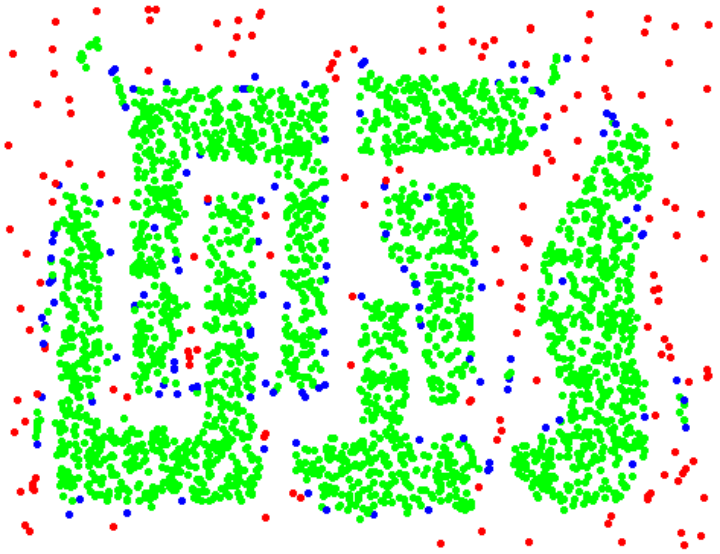


Αλγόριθμος DBSCAN

1. Όρισε όλα τα σημεία ως σημεία πυρήνα, ορίου, θορύβου
2. Εξάλειψε τα σημεία θορύβου
3. Τοποθέτησε μία ακμή ανάμεσα σε όλα τα σημεία πυρήνα που βρίσκονται εντός απόστασης *Eps* μεταξύ τους
4. Όρισε κάθε ομάδα συνδεδεμένων σημείων πυρήνα ως χωριστή συστάδα
5. Εκχώρησε κάθε σημείο ορίου σε μία από τις συστάδες των σημείων πυρήνα

Βήμα 1 & 2

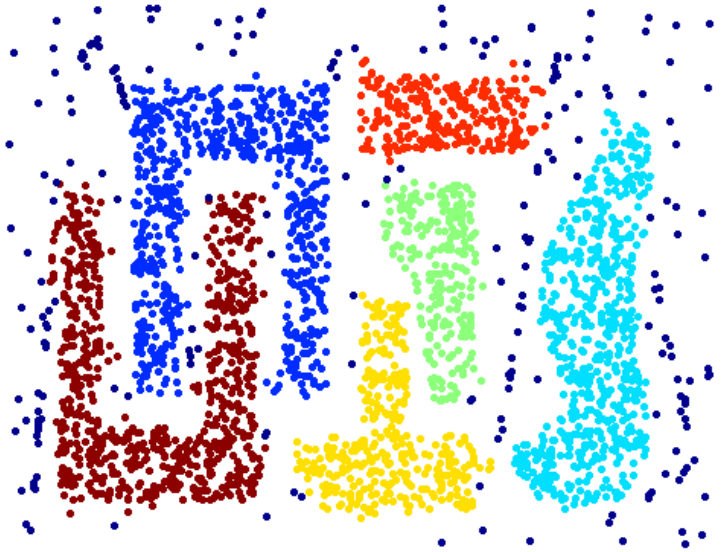
Αρχικά σημεία



Τύποι σημείων: core, border και noise

Eps = 10, MinPts = 4

Βήμα 3&4



Αρχικά Σημεία

Συστάδες

Πλεονεκτήματα του αλγορίθμου DBSCAN

- Σε αντίθεση με αλγορίθμους, όπως ο k-means, ο DBSCAN δεν απαιτεί τον εκ των προτέρων προσδιορισμό του αριθμού των συστάδων.
- Μπορεί να καταλήξει σε αυθαίρετα σχήματα συστάδων. Μπορεί να εντοπίσει ακόμα και μια συστάδα, η οποία βρίσκεται γύρω από κάποια άλλη. Αυτό συμβαίνει λόγω της παραμέτρου MinPts, η οποία ελαττώνει την εμφάνιση του φαινομένου της αλυσίδας συστάδων. Το φαινόμενο της αλυσίδας συστάδων συμβαίνει, όταν διαφορετικές συστάδες συνδέονται με μια λεπτή γραμμή σημείων-αντικειμένων.
- Έχει καλή ευαισθησία στον θόρυβο και δεν επηρεάζεται από ακραίες τιμές.
- Χρειάζεται μόνο δυο παραμέτρους και έχει μικρή ευαισθησία ως προς τη σειρά εμφάνισης των δεδομένων στη βάση.
- Εφόσον έχουν μελετηθεί τα δεδομένα και έχουν γίνει κατανοητά, ο προσδιορισμός των παραμέτρων *MinPts* και *Eps* δεν είναι δύσκολος.

Μειονεκτήματα του αλγορίθμου DBSCAN

- Δεν μπορεί να συσταδοποιήσει καλά σύνολα από δεδομένα με μεγάλες διαφορές πυκνότητας, καθώς δεν μπορεί να εντοπιστεί κάποιος συνδυασμός *MinPts-Eps*, που να είναι κατάλληλος για όλες τις συστάδες.
- Αν τα δεδομένα δεν έχουν γίνει κατανοητά, η επιλογή ενός κατωφλίου που να έχει νόημα μπορεί να είναι δύσκολη.

Δείτε

- http://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2972/1/02_chapter_06.pdf

- Κριτήριο αξιολόγησης 1 & Κριτήριο αξιολόγησης 3

Κατηγοριοποίηση (Classification)

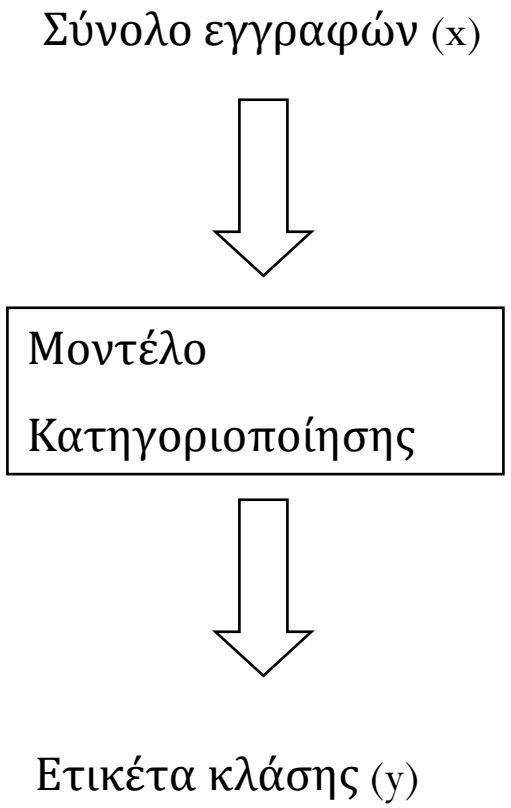
- Η κατηγοριοποίηση αποτελεί μια από τις βασικές εργασίες στο στάδιο της Εξόρυξης Δεδομένων.
- Βασίζεται στην εξέταση των χαρακτηριστικών ενός αντικειμένου, το οποίο, με βάση τα χαρακτηριστικά αυτά,αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων.

Παραδείγματα

- Το γενικό πρόβλημα της ανάθεσης ενός αντικειμένου σε μία ή περισσότερες προκαθορισμένες κατηγορίες (κλάσεις), π.χ.
  - Εντοπισμός spam emails, με βάση πχ την επικεφαλίδα τους ή το περιεχόμενό τους
  - Πρόβλεψη καρκινικών κυττάρων χαρακτηρίζοντας τα ως καλοήθη ή κακοήθη
  - Κατηγοριοποίηση συναλλαγών με πιστωτικές κάρτες ως νόμιμες ή προϊόν απάτης
  - Κατηγοριοποίηση δευτερευόντων δομών πρωτείνης ως alpha-helix, beta-sheet, ή random coil
  - Χαρακτηρισμός ειδήσεων ως οικονομικές, αθλητικές, πολιτιστικές, πρόβλεψης καιρού, κλπ

Ορισμός

- 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**.



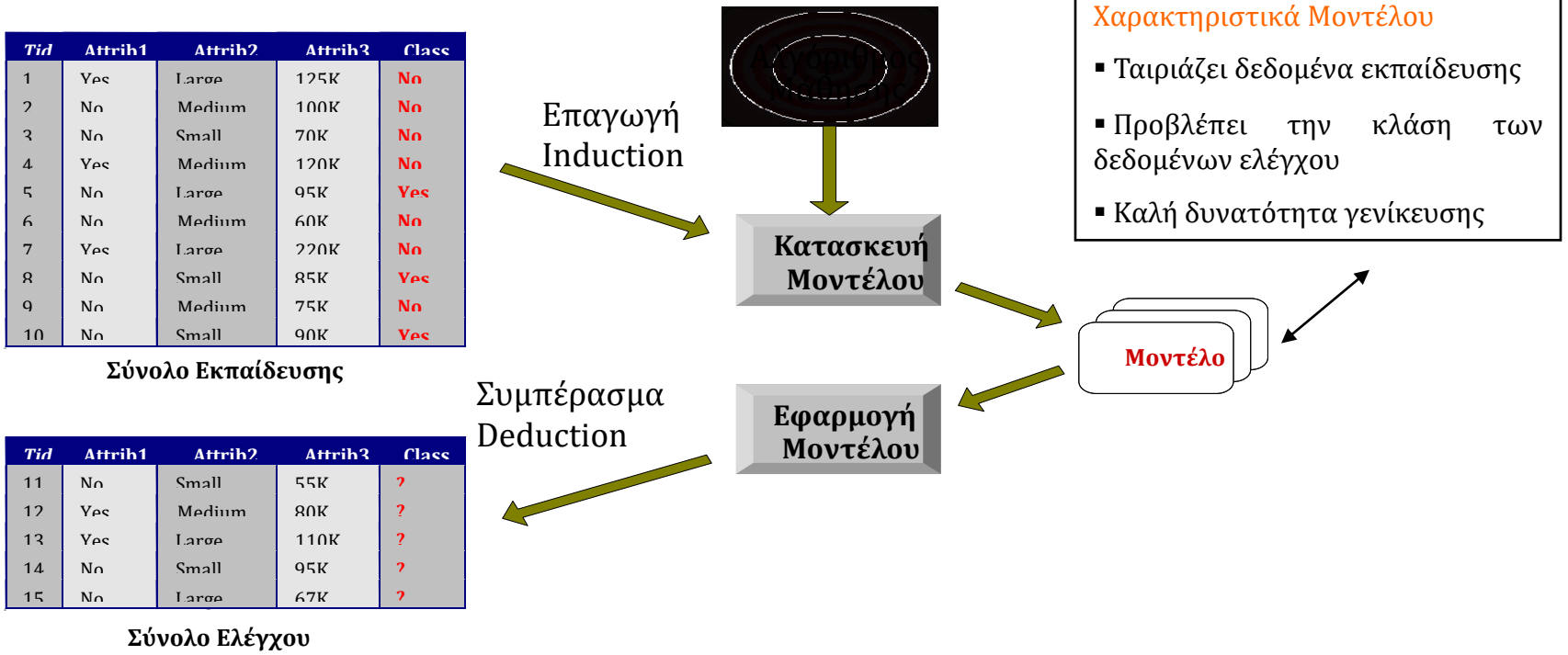
- Συνήθως το σύνολο δεδομένων εισόδου χωρίζεται σε:
  - ένα σύνολο εκπαίδευσης (training set) και
  - ένα σύνολο ελέγχου (test set)

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

Βήματα Κατηγοριοποίησης

1. Κατασκευή Μοντέλου
   - Χρησιμοποιώντας το σύνολο εκπαίδευσης (στις εγγραφές του το γνώρισμα της κλάσης είναι προκαθορισμένο)
   - Το μοντέλο μπορεί να είναι ένα δέντρο απόφασης, κανόνες, μαθηματικοί τύποι κλπ

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 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
Επαγωγή
Induction
Κατασκευή Μοντέλου
Χαρακτηριστικά Μοντέλου
▪ Ταιριάζει δεδομένα εκπαίδευσης
▪ Προβλέπει την κλάση των δεδομένων ελέγχου
▪ Καλή δυνατότητα γενίκευσης
Μοντέλο
Συμπέρασμα
Deduction
Εφαρμογή Μοντέλου
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

Προεπεξεργασία

1. Καθαρισμός Δεδομένων (data cleaning)
   - Προεπεξεργασία δεδομένων και χειρισμός τιμών που λείπουν (πχ 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

Εκτίμηση Μεθόδων Κατηγοριοποίησης

- Προβλεπόμενη ακρίβεια - Predictive accuracy
- Ταχύτητα (speed)
  - Χρόνος κατασκευής του μοντέλου
  - Χρόνος χρήσης/εφαρμογής του μοντέλου
- Robustness
  - Χειρισμός θορύβου και τιμών που λείπουν
- Κλιμάκωση - Scalability
  - Αποδοτικότητα σε βάσεις δεδομένων αποθηκευμένες στο δίσκο
- Ευκρίνεια - Interpretability
  - Πόσο κατανοητό είναι το μοντέλο και τι νέα πληροφορία προσφέρει
- Ποιότητα - Goodness of rules (quality)
  - π.χ. μέγεθος του δέντρου

Τεχνικές κατηγοριοποίησης

- Δέντρα Απόφασης (decision trees)
- Κανόνες (Rule-based Methods)
- Αλγόριθμοι Κοντινότερου Γείτονα
- Memory based reasoning
- Νευρωνικά Δίκτυα
- Naïve Bayes Δίκτυα
- Support Vector Machines
- …

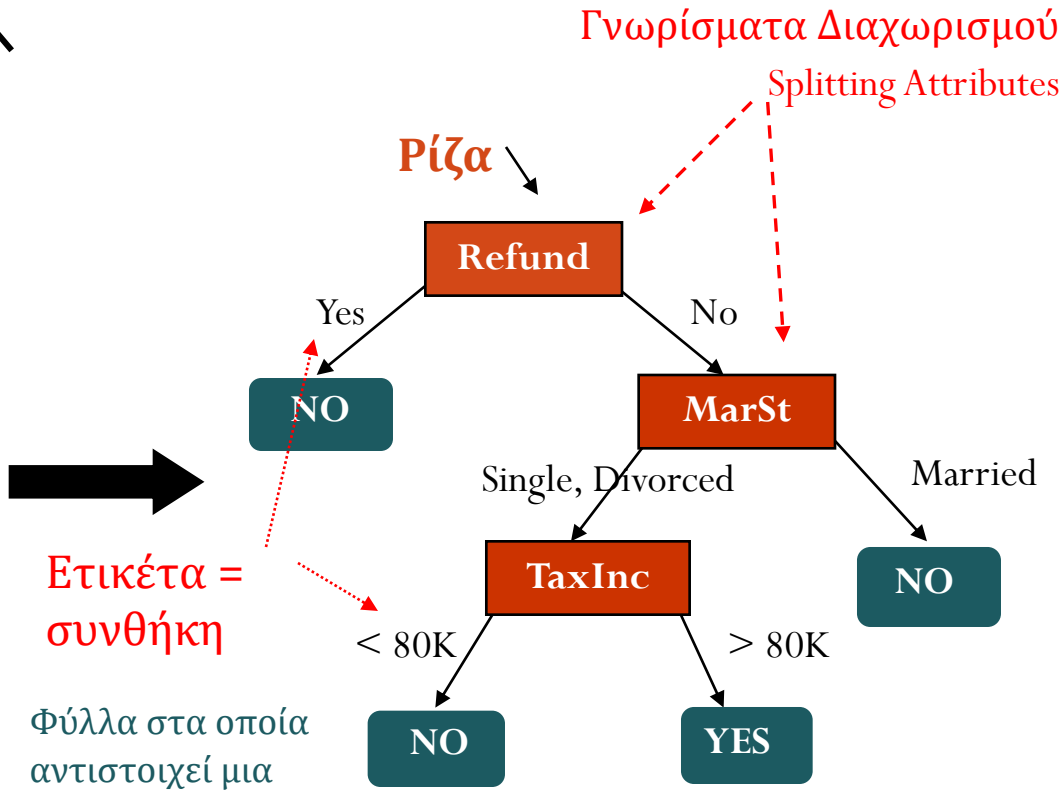
Μοντέλο = Δέντρο Απόφασης

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

Δέντρο Απόφασης: Παράδειγμα

Δεδομένα Εκπαίδευσης

Tid	Refund (κατηγορικό)	Marital Status (κατηγορικό)	Taxable Income (συνεχές)	Cheat (κλάση)
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



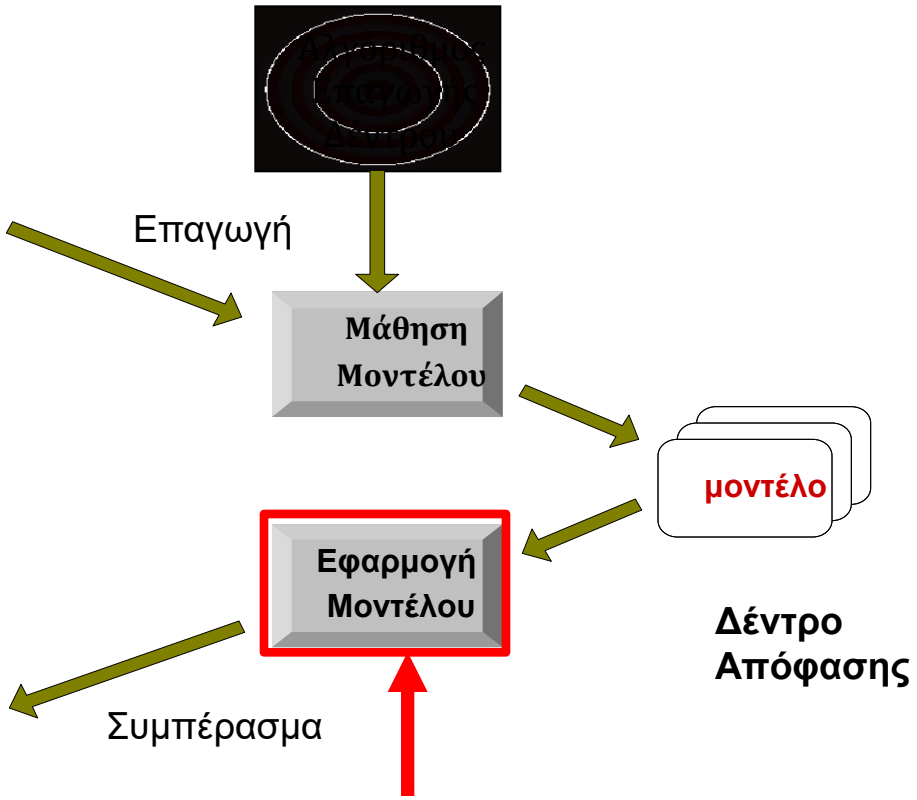
Δέντρο Απόφασης: Βήματα

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	?

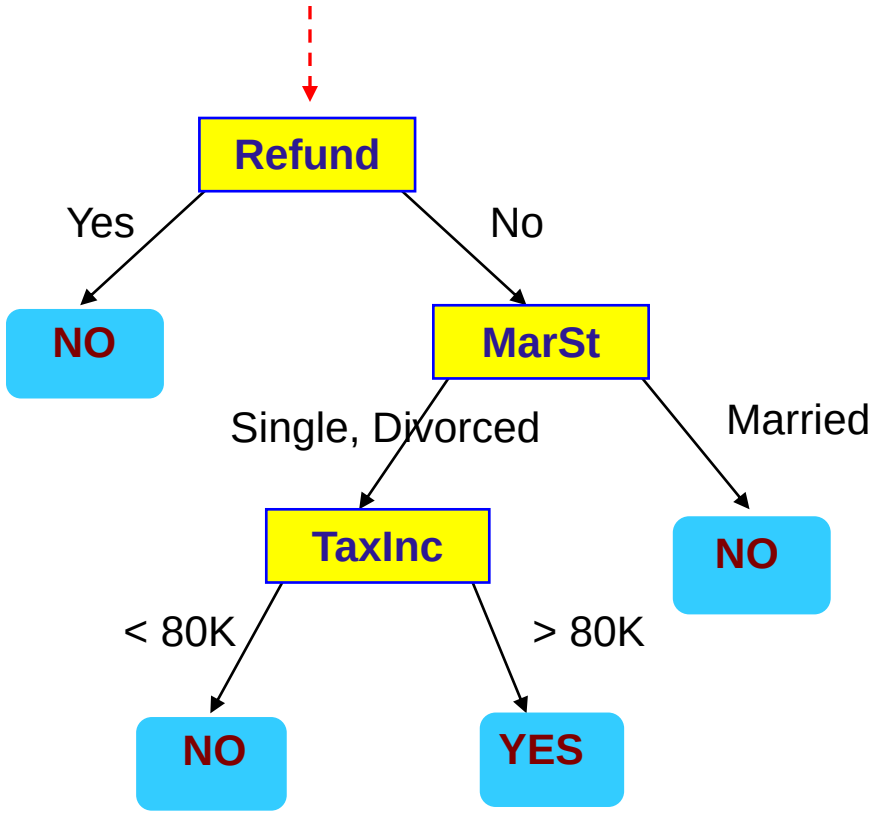
Σύνολο Ελέγχου



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

Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Ξεκίνα από τη ρίζα του δέντρου.



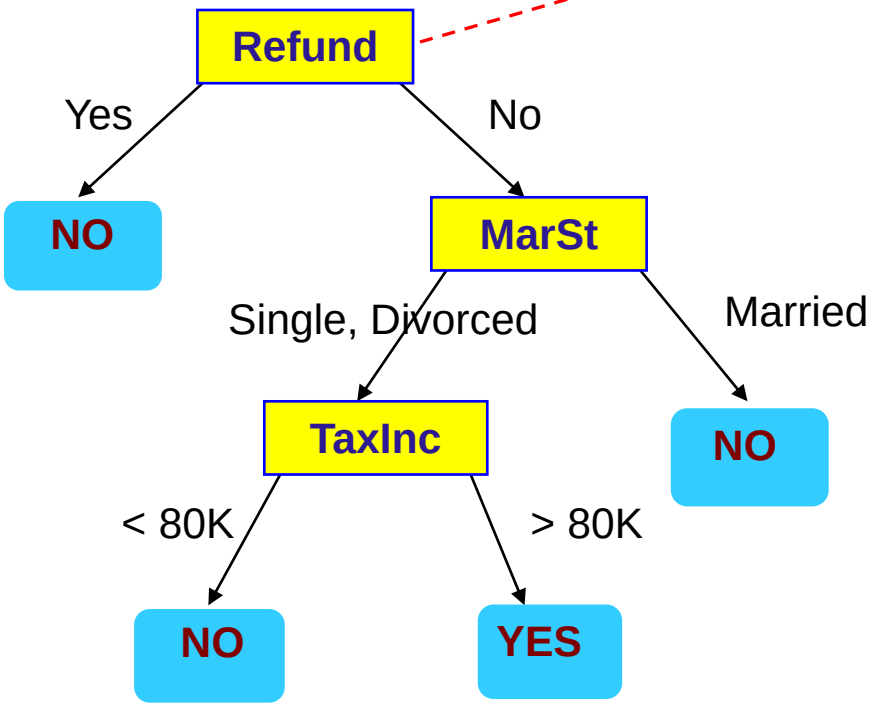
Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

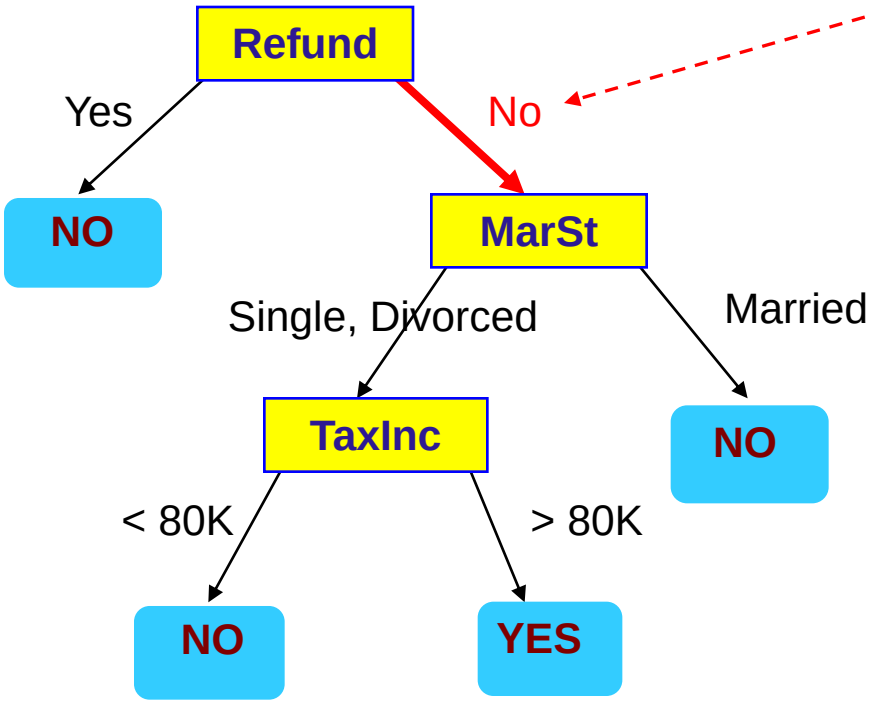
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

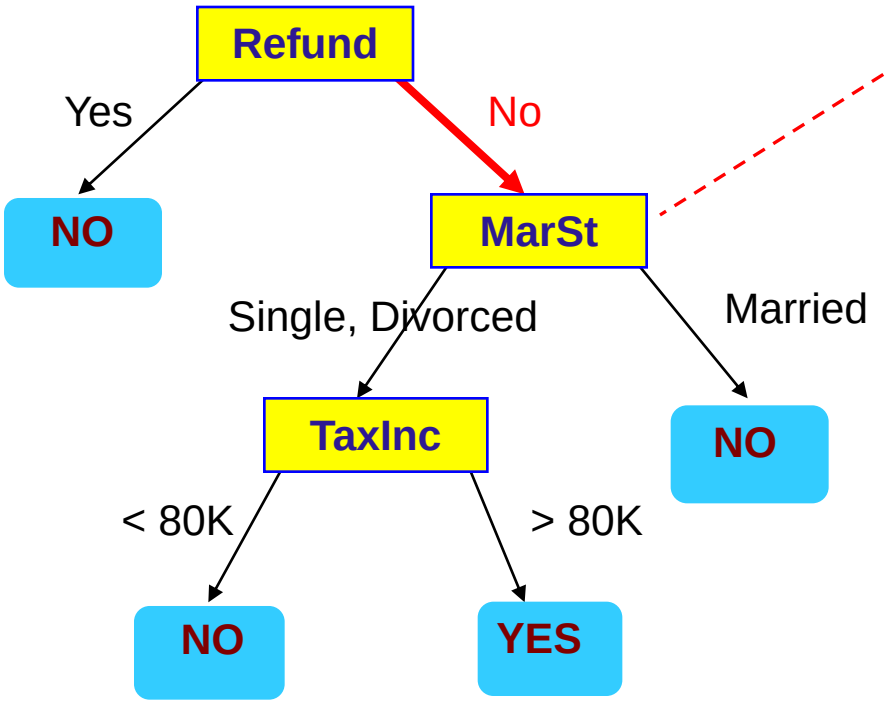
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

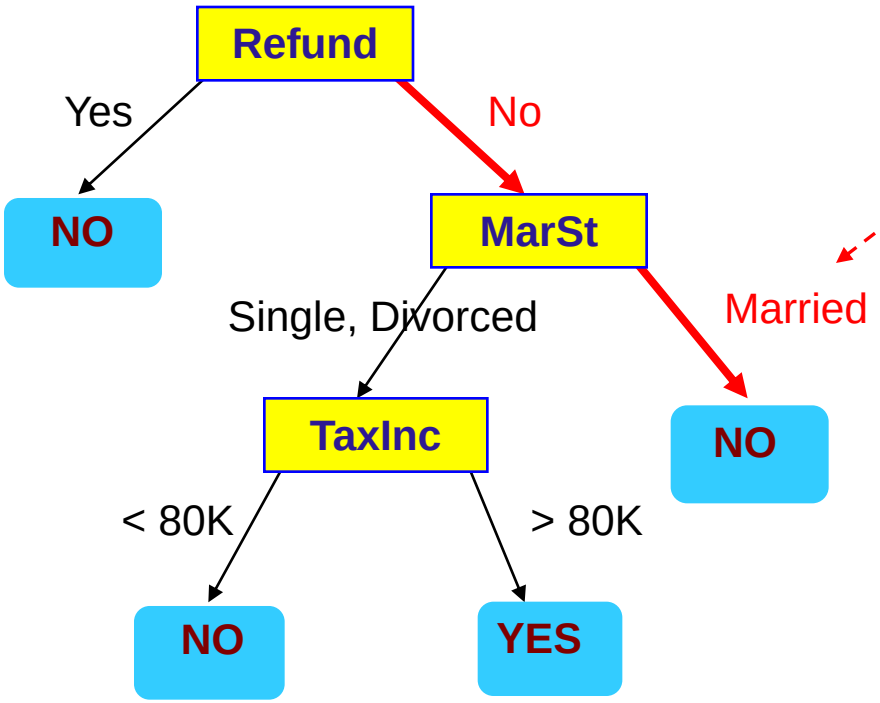
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

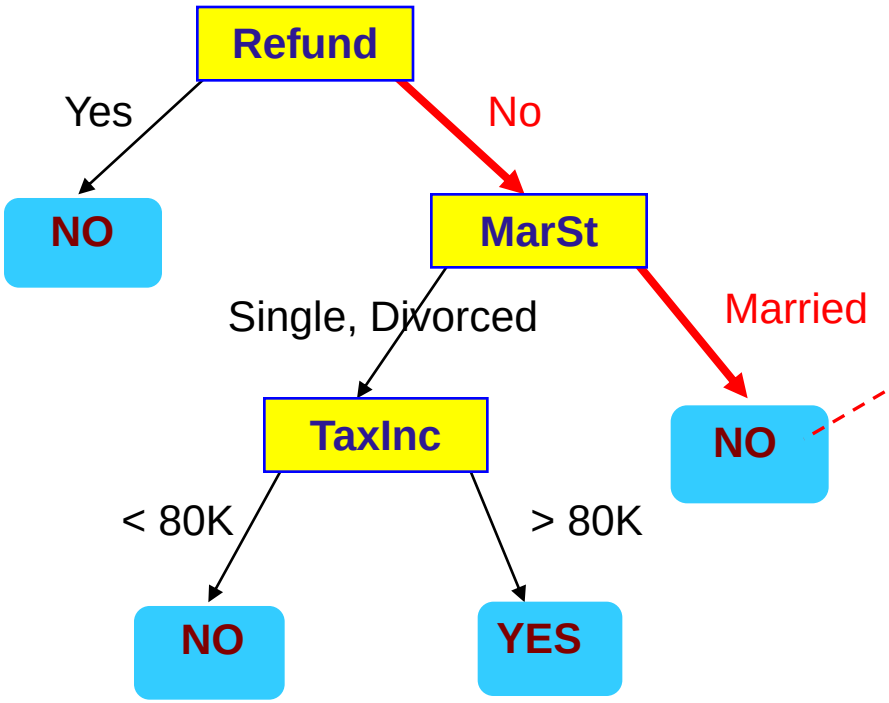
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Δέντρο Απόφασης: Εφαρμογή Μοντέλου

Δεδομένα Ελέγχου

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Ανάθεση στο Cheat “No”

Δέντρο Απόφασης

Κατασκευή του δέντρου (με λίγα λόγια):

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)

(top-down, recursive, divide-and-conquer προσέγγιση)

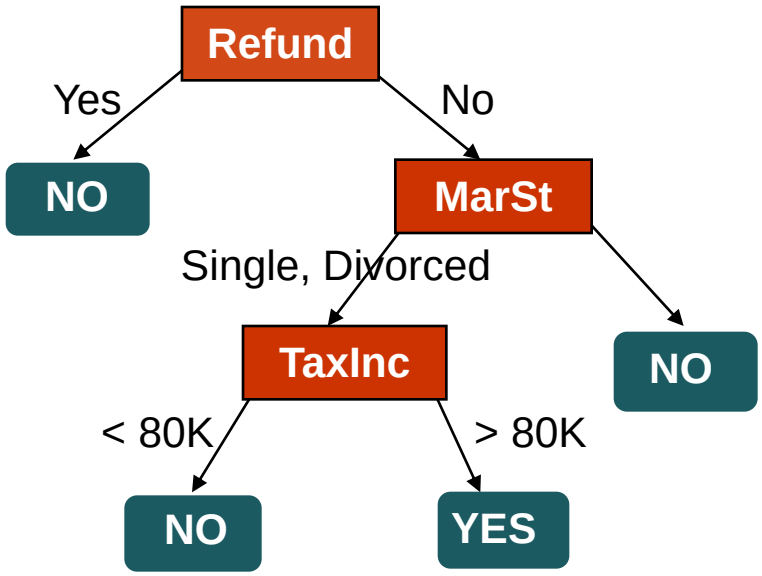
Το βασικό θέμα είναι

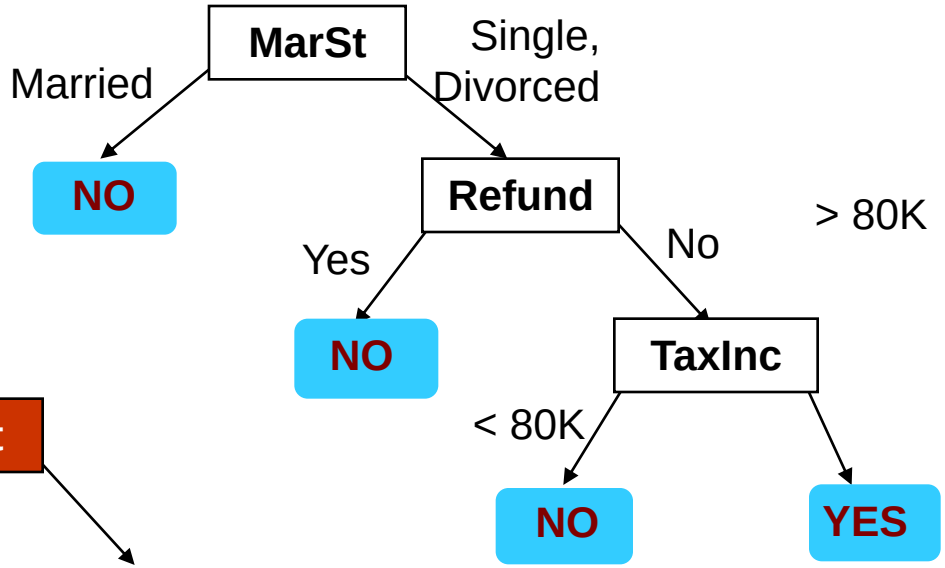
Ποιο γνώρισμα-συνθήκη διαχωρισμού να χρησιμοποιήσουμε για τη διάσπαση των εγγραφών κάθε κόμβου

Δέντρο Απόφασης: 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





Αλγόριθμοι για επαγωγή δένδρων απόφασης

- Ο αριθμός των πιθανών Δέντρων Απόφασης είναι εκθετικός.

- Πολλοί αλγόριθμοι για την επαγωγή (induction) του δέντρου οι οποίοι ακολουθούν μια greedy στρατηγική για να κτίσουν το δέντρο απόφασης παίρνοντας μια σειρά από τοπικά βέλτιστες αποφάσεις
  - Hunt's Algorithm (από τους πρώτους)
  - CART
  - ID3, C4.5
  - SLIQ, SPRINT

Κατηγοριοποιητής Κανόνων

- Κατηγοριοποίηση με κανόνες “if…then…”
- Κανόνας: (*Συνθήκη*) $\rightarrow y$
  - όπου
    - *Συνθήκη* είναι μία σύζευξη γνωρισμάτων
    - y είναι η ετικέτα της κατηγορίας
  - Παραδείγματα κανόνων κατηγοριοποίησης:
    - (Blood Type=Warm) $\wedge$ (Lay Eggs=Yes) $\rightarrow$ Birds
    - (Taxable Income < 50K) $\wedge$ (Refund=Yes) $\rightarrow$ Evade=No

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Εφαρμογή Κανόνων Κατ.

- Ένας κανόνας r *καλύπτει* ένα στιγμιότυπο x εάν τα γνωρίσματα του x ικανοποιούν την συνθήκη του κανόνα.

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds

R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes

R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals

R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles

R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Ο κανόνας R1 καλύπτει το παράδειγμα hawk => Bird

Ο κανόνας R3 καλύπτει το παράδειγμα grizzly bear => Mammal

Κάλυψη Κανόνα και Ακρίβεια

- Κάλυψη
  - Ποσοστό εγγραφών που ικανοποιούν την συνθήκη του κανόνα
- Ακρίβεια κανόνα
  - Ποσοστό εγγραφών που ικανοποιούν και την συνθήκη και το επάκολουθο

Tid	Refund	Marital Status	Taxable Income	Class
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**

(Status=Single) → No

Coverage = 40%, Accuracy = 50%

Πώς λειτουργεί ένας Κατ. Κανόνων;

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds

R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes

R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals

R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles

R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?
dogfish shark	cold	yes	no	yes	?

Το lemur πυροδοτεί τον κανόνα R3, και έτσι κατηγοριοποιείται ως mammal

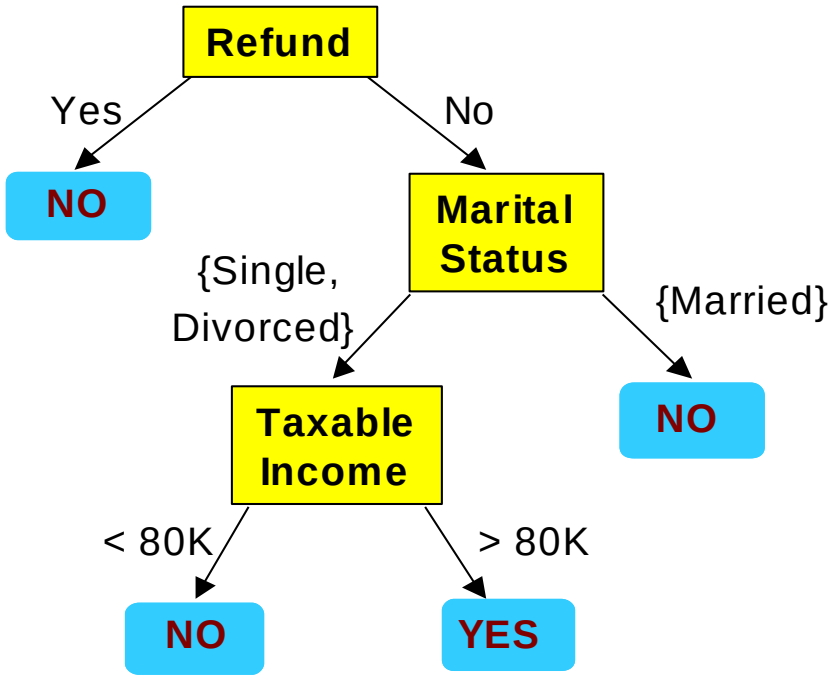
Η turtle και τον R4 και τον R5

Ένας dogfish shark δεν πυροδοτεί κανένα κανόνα

Χαρακτηριστικά Κατ. Κανόνων

- Αμοιβαία αποκλειόμενοι κανόνες
  - Ο κατ. περιέχει αμοιβαία αποκλειόμενους κανόνες εάν οι κανόνες είναι ανεξάρτητοι μεταξύ τους
  - Κάθε εγγραφή καλύπτεται από ένα κανόνα
- Πλήρεις κανόνες
  - Ένας κατ. έχει πλήρη κάλυψη εάν αντιπροσωπεύει κάθε συνδυασμό τιμών γνωρισμάτων
  - Κάθε εγγραφή καλύπτεται από ένα κανόνα τουλάχιστον

Από Δέντρα Απόφασης σε Κανόνες



Classification Rules

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Οι κανόνες είναι αμοιβαία αποκλειόμενοι και πλήρεις

Το σύνολο κανόνων περιέχει όση ακριβώς πληροφορία περιέχει και το δέντρο

Διατεταγμένο Σύνολο Κανόνων

- Οι κανόνες διατάσσονται με βάση την προτεραιότητά τους.
  - Ένα διατεταγμένο σύνολο κανόνων είναι γνωστό και ως λίστα απόφασης.
- Όταν μία εγγραφή δοκιμής δίνεται στον κατ.
  - Εκχωρείται στην κατηγορία του πιο υψηλά σε διάταξη κανόνα που έχει πυροδοτηθεί.
  - Εάν δεν έχει πυροδοτηθεί κανένας κανόνας, εκχωρείται στην προεπιλεγμένη κατηγορία.

Διατ. Σύνολο Κανόνων (Παράδειγμα)

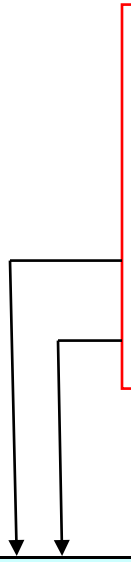
R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds

R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes

R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals

R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles

R5: (Live in Water = sometimes) → Amphibians



Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
turtle	cold	no	no	sometimes	?

Πλεονεκτήματα Κατ. Κανόνων

- Εκφραστικά ισοδύναμοι με τα δέντρα απόφασης
- Εύκολοι στην ερμηνεία
- Εύκολοι στη δημιουργία
- Γρήγορη κατηγοριοποίηση νέων παραδειγμάτων
- Απόδοση συγκρίσιμη με δέντρα απόφασης

Κατηγοριοποιητές Bayes

- Το θεώρημα Bayes ορίστηκε μαθηματικά ως η ακόλουθη εξίσωση:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)},$$

όπου A και B είναι γεγονότα.

$P(A)$ και $P(B)$ είναι οι πιθανότητες των A και B που είναι ανεξάρτητα μεταξύ τους.

$P(A \mid B)$, η υπό συνθήκη πιθανότητα, είναι η πιθανότητα του A δεδομένου του B να είναι αληθής.

$P(B \mid A)$, είναι η πιθανότητα του B δεδομένου του A να είναι αληθής.

Κατηγοριοποιητές Bayes

Το θεώρημα του Bayes: Παράδειγμα 1

- Ένας γιατρός γνωρίζει ότι η μηνιγγίτιδα προκαλεί αυχενική δυσκαμψία 50% των περιπτώσεων
- Εκ' των προτέρων πιθανότητα ενός ασθενή να έχει μηνιγγίτιδα είναι 1/50,000
- Εκ' των προτέρων πιθανότητα ενός ασθενή να έχει αυχενική δυσκαμψία είναι 1/20
- Εάν ο ασθενής έχει αυχενική δυσκαμψία (S) , ποια είναι η πιθανότητα ότι έχει μηνιγγίτιδα (M);

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Κατηγοριοποιητές Bayes

- Θεωρούμε κάθε γνώρισμα και ετικέτα κατηγορίας ως τυχαία μεταβλητή
- Με δεδομένη μία εγγραφή $(A_1, A_2, \ldots, A_n)$
  - στόχος είναι η πρόβλεψη της κατηγορίας C
  - Συγκεκριμένα, θέλουμε να βρούμε την τιμή του C που μεγιστοποιεί το $P(C \mid A_1, A_2, \ldots, A_n)$
- Μπορούμε να εκτιμήσουμε το $P(C \mid A_1, A_2, \ldots, A_n)$ απευθείας από τα δεδομένα;

Κατηγοριοποιητές Bayes

- Υπολογίζουμε την εκ' των υστέρων πιθανότητα $P(C| A_1,A_2, \ldots, A_n)$ για όλες τις τιμές του C χρησιμοποιώντας το θεώρημα του Bayes.

- Επιλέγουμε την τιμή του C που μεγιστοποιεί το $P(C| A_1,A_2, \ldots, A_n)$ ή το $P(A_1,A_2, \ldots, A_n|C)P(C)$

- Πώς υπολογίζουμε (εκτιμούμε) το $P(A_1,A_2, \ldots, A_n|C)$;

Απλοϊκός Κατ. Bayes

- Θεωρούμε ανεξαρτησία ανάμεσα στα γνωρίσματα A_i δεδομένης της κλάσης

- $P(A_1, A_2, \ldots, A_n \mid C_j) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$

- Μπορούμε να εκτιμήσουμε τα $P(A_i \mid C_j)$ για όλα τα A_i και C_j

- Ένα νέο σημείο κατηγοριοποιείται ως C_j εάν το $P(C_j) * \Pi\, P(A_i \mid C_j)$ είναι μέγιστο.

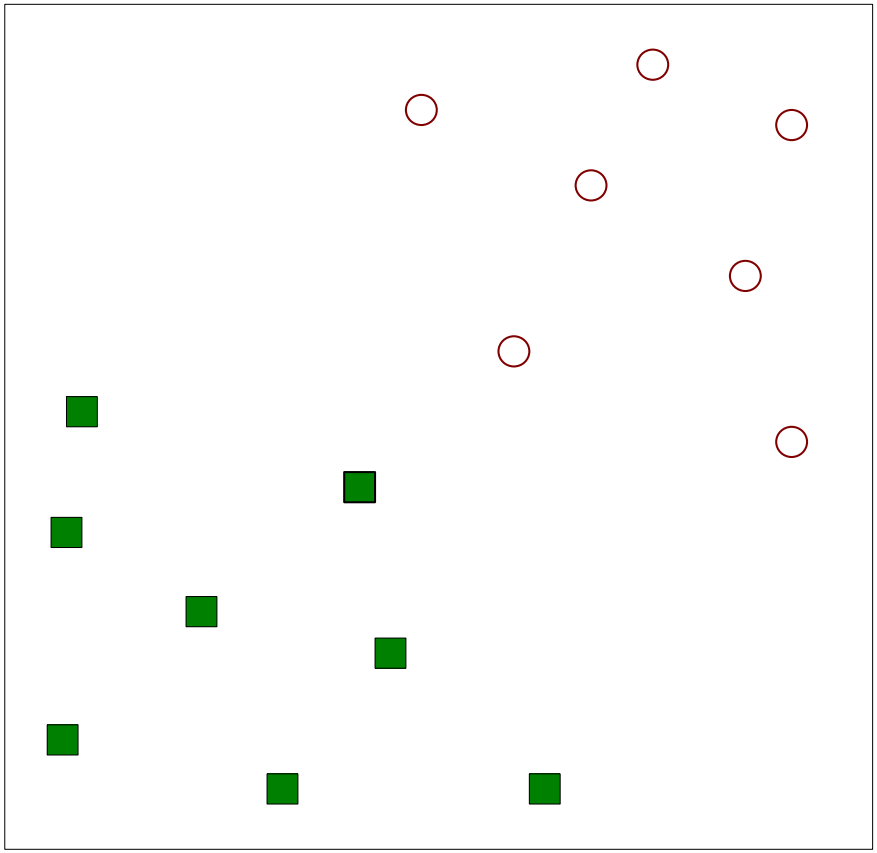
Πώς υπολογίζουμε πιθανότητες;

- Κατηγορία: $P(C)=N_c/N$
  - π.χ., $P(No)=7/10$ και $P(Yes)=3/10$
- Για διακριτά γνωρίσματα
  - $P(A_i \mid C_k) = |A_{ik}|/ N_c$
    - όπου $|A_{ik}|$ είναι το πλήθος των παραδειγμάτων που έχουν το γνώρισμα A_i και ανήκουν στην κατηγορία C_k
  - Παραδείγματα:
  - P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

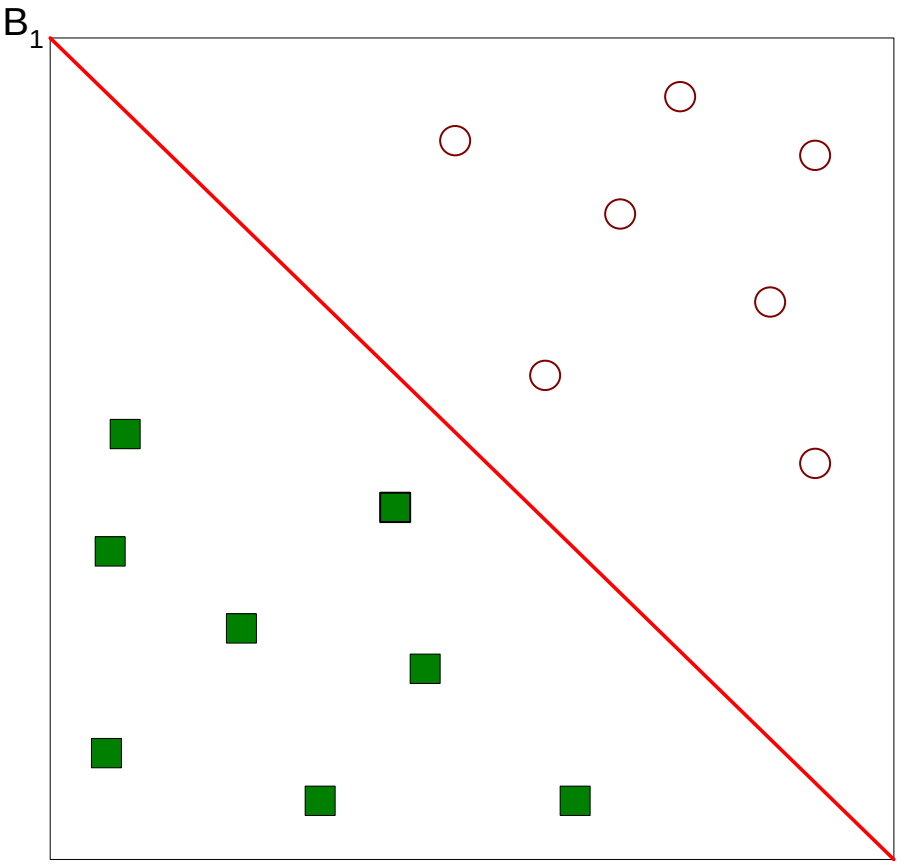
Πώς υπολογίζουμε τις πιθανότητες;

- Για συνεχή γνωρίσματα:
  - Διακριτοποιούμε το εύρος σε τμήματα
  - Μία κατηγορική μεταβλητή ανά τμήμα
- Διάσπαση δύο κατευθύνσεων
  - επιλέγουμε μία από τις δύο διασπάσεις σαν νέο γνώρισμα
- Εκτίμηση πυκνότητας πιθανότητας
  - Θεωρούμε ότι το γνώρισμα ακολουθεί κανονική κατανομή
  - χρησιμοποιούμε δεδομένα για την εκτίμηση των παραμέτρων (μέσος, τυπική απόκλιση)

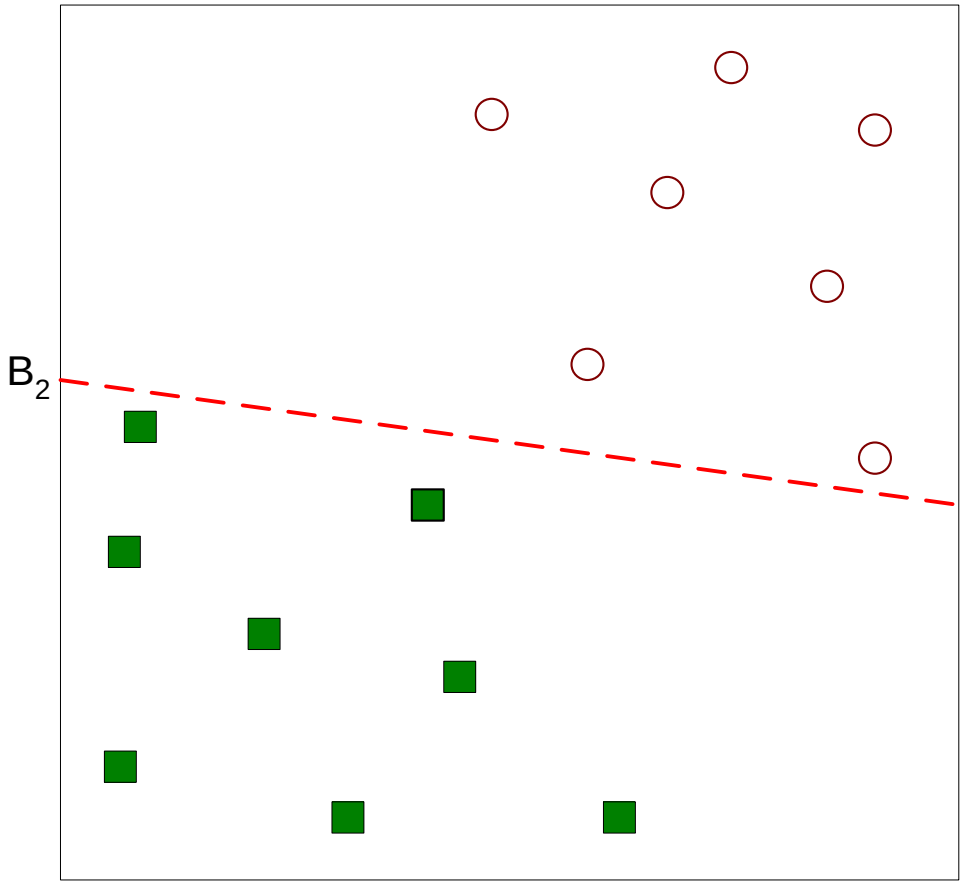
Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)



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

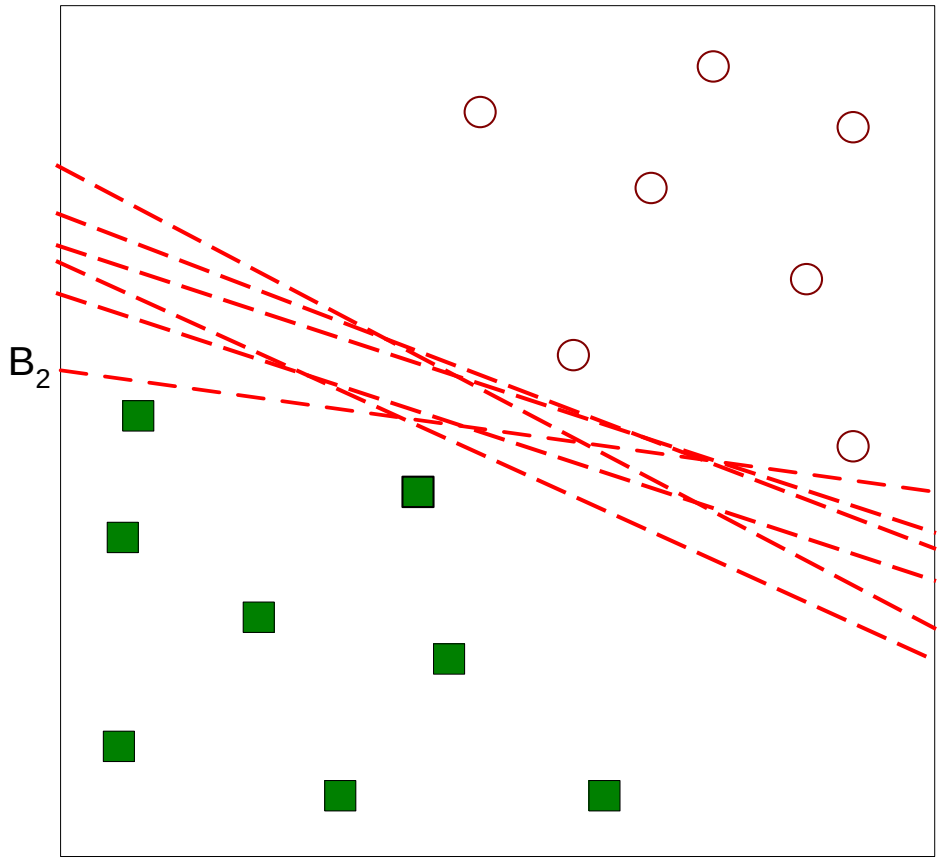


- Μία πιθανή λύση



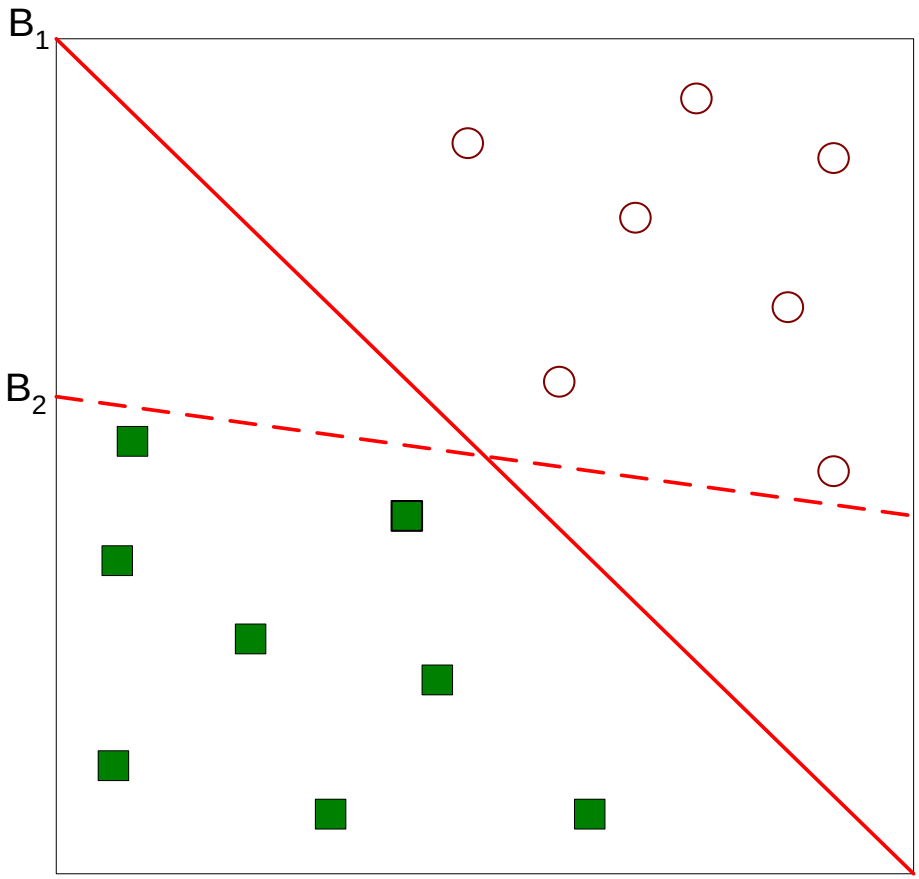
- Μια ακόμα πιθανή λύση

Κατηγοριοποιητές SVM



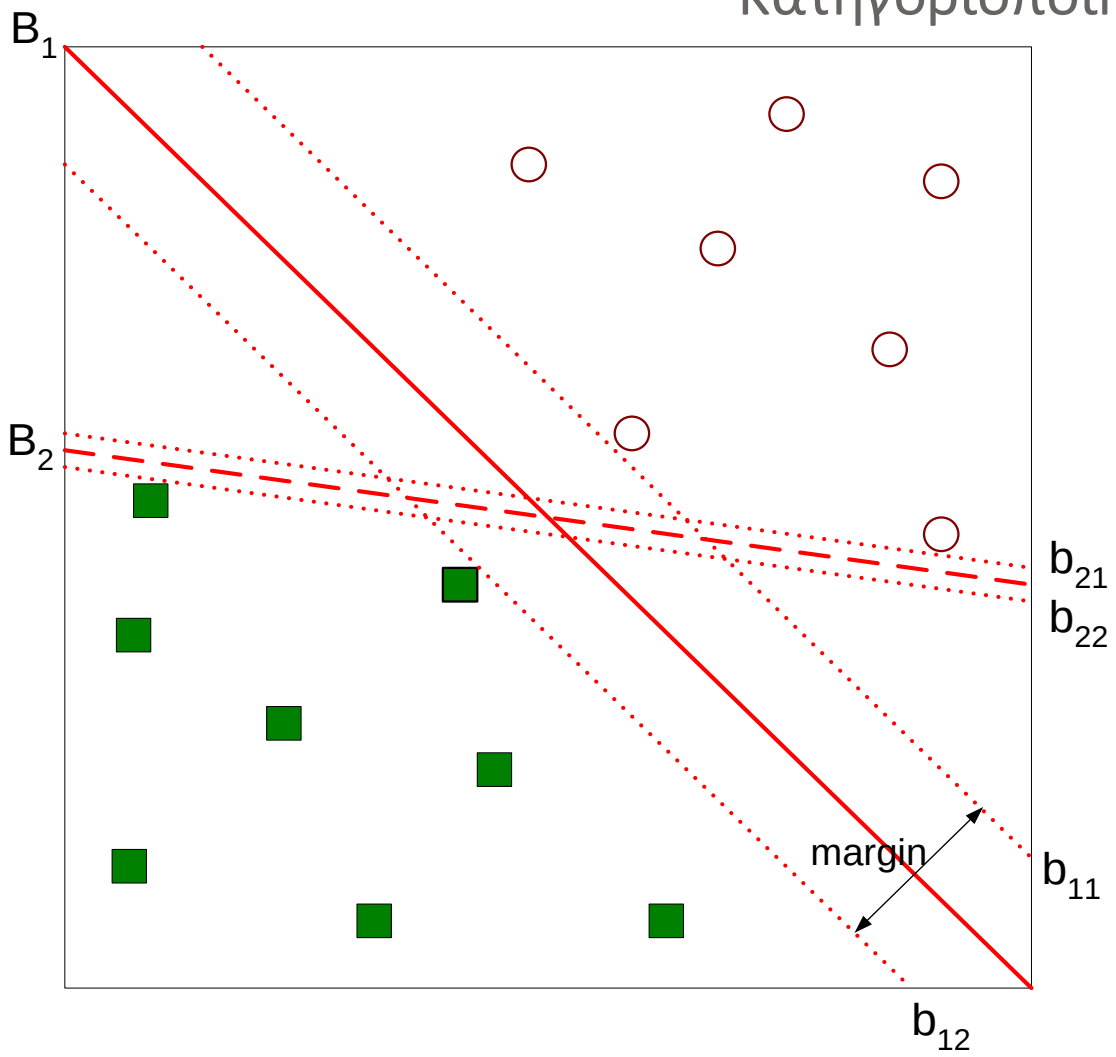
- Άλλες πιθανές λύσεις

Κατηγοριοποιητές SVM



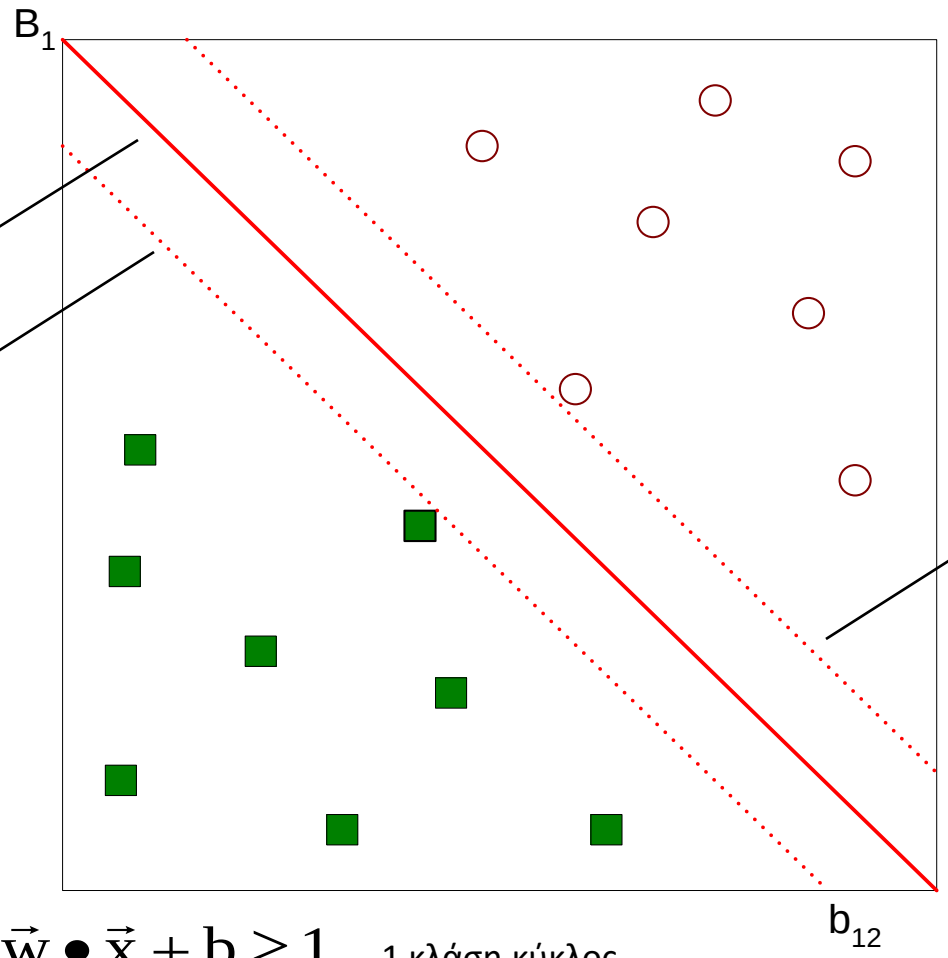
- Ποια είναι καλύτερη η B1 ή η B2?
- Πώς ορίζεται το καλύτερη; Με ποιο κριτήριο;

Κατηγοριοποιητές SVM



- Το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο (margin) => το B1 είναι καλύτερο από το B2 (χωρητικότητα)

Κατηγοριοποιητές SVM



$\vec{w} \bullet \vec{x} + b = 0$

$\vec{w} \bullet \vec{x} + b = -1$

$\vec{w} \bullet \vec{x} + b = +1$

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

1 κλάση κύκλος

-1 κλάση τετράγωνο

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

Κατηγοριοποιητές SVM

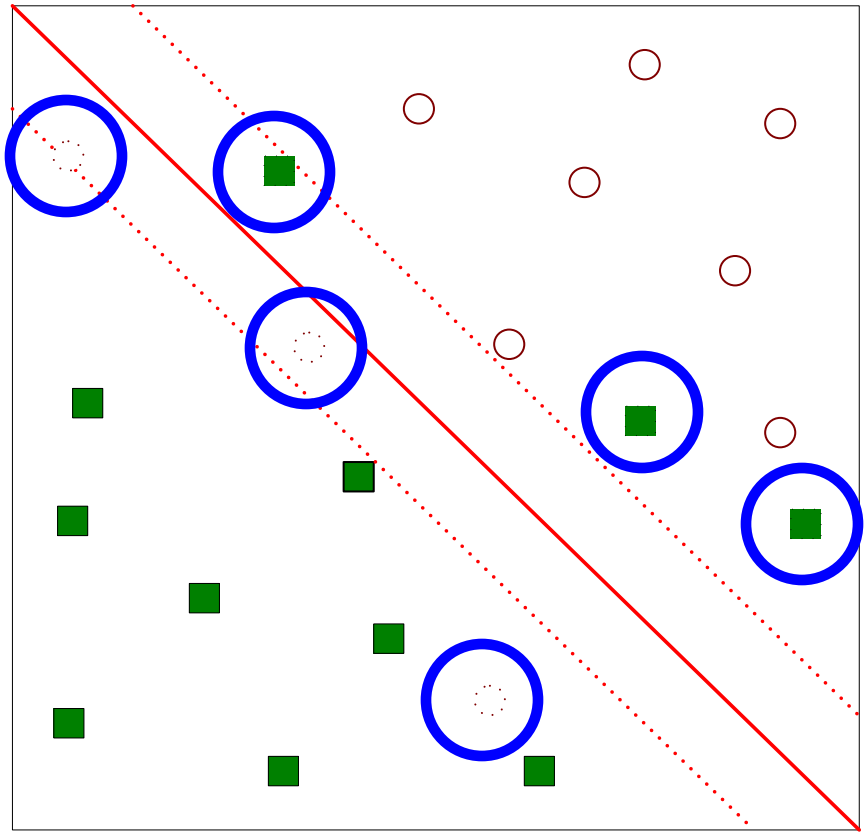
- Θέλουμε να μεγιστοποιήσουμε: $$\text{Margin} = \frac{2}{\| \vec{w} \|^2}$$
  - Το οποίο είναι ισοδύναμο με το να ελαχιστοποιήσουμε: $$L(w) = \frac{\| \vec{w} \|^2}{2}$$
  - Με βάση τους παρακάτω περιορισμούς (constraints):

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

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

Κατηγοριοποιητές SVM

- Τι συμβαίνει αν το πρόβλημα δεν είναι γραμμικώς διαχωρίσιμο



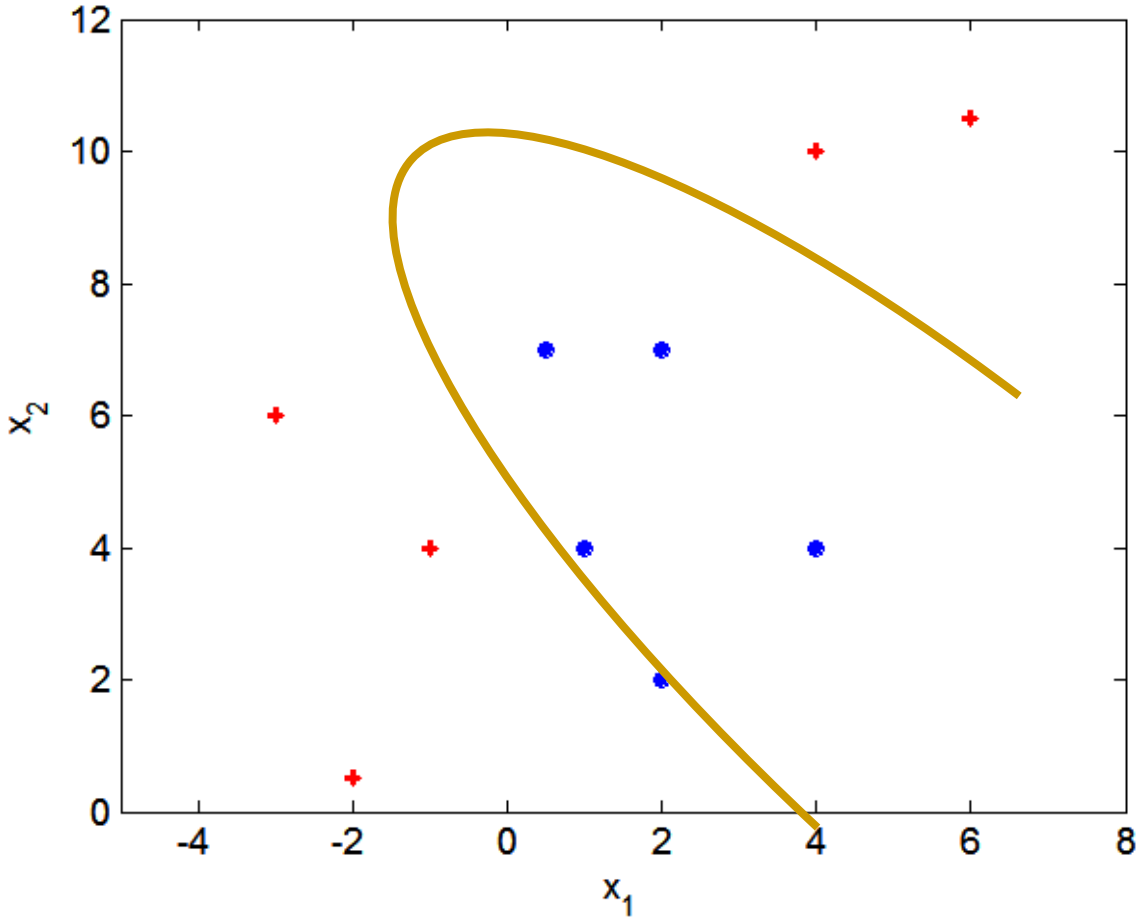
- Εισαγωγή χαλαρών μεταβλητών (slack variables)
  - Ελαχιστοποίηση:

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

  - Με τους περιορισμούς:

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

Κατηγοριοποιητές SVM

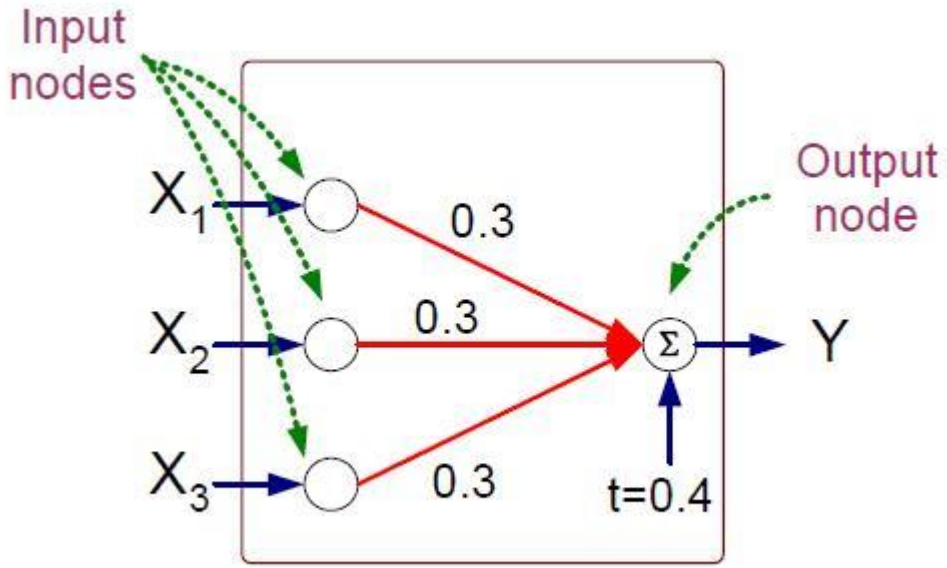


Νευρωνικά δίκτυα

- Ξεκίνησε από την προσπάθεια για προσομοίωση βιολογικών νευρωνικών συστημάτων.
- Ο ανθρώπινος εγκέφαλος περιέχει νευρικά κύτταρα που ονομάζονται νευρώνες (neurons) και συνδέονται με νευρίτες (νηματοειδείς ίνες).
- Ένας νευρώνας είναι μια στοιχειώδης υπολογιστική μονάδα, η οποία δέχεται τιμές εισόδου και υπολογίζει μια τιμή εξόδου.
- Οι νευρώνες συνδέονται μεταξύ τους με κατευθυνόμενα βέλη ή συνδέσεις.

Νευρωνικά δίκτυα

- Η επεξεργασία που διενεργεί ένας νευρώνας i ολοκληρώνεται σε δύο στάδια.
  - Στο πρώτο στάδιο αθροίζονται οι τιμές εισόδου. Οι τιμές εισόδου ισούνται με τις τιμές εξόδου των συνδεδεμένων νευρώνων, πολλαπλασιασμένες με τα βάρη των αντίστοιχων συνδέσεων.
  - Στο δεύτερο στάδιο, μετασχηματίζεται το άθροισμα των τιμών εισόδου, με χρήση μιας συνάρτησης γνωστής ως συνάρτηση ενεργοποίησης (activation function) ή συνάρτηση μετασχηματισμού.



$$\widehat{y} = \begin{cases} 1, \text{ ан } 0.3x_1 + 0.3x_2 + 0.3x_3 - 0.4 > 0; \\ -1, \text{ ан } 0.3x_1 + 0.3x_2 + 0.3x_3 - 0.4 < 0 \end{cases}$$

Επιπλέον θέματα…

- Χρονικά δεδομένα (temporal data) είναι συχνά σε εφαρμογές εξόρυξης δεδομένων.
  - Π.χ. διακυμάνσεις μετοχών σε μια χρονική περίοδο

- Χωρικά δεδομένα
  - Είναι η διαδικασία ανακάλυψης δυνητικά χρήσιμων μοτίβων από μεγάλα σύνολα χωρικών δεδομένων

- Γενετικοί αλγόριθμοι
  - Βρίσκουν τη βέλτιστη λύση σε ένα πρόβλημα εξετάζοντας πολύ μεγάλο αριθμό εναλλακτικών λύσεων σ' αυτό
  - Βασίζεται σε τεχνικές εμπνευσμένες από την εξελικτική βιολογία: κληρονομικότητα, μετάλλαξη, φυσική επιλογή κ.λπ.
  - Χρησιμοποιούνται για τη λύση περίπλοκων προβλημάτων που είναι πολύ δυναμικά και σύνθετα με εκατοντάδες ή χιλιάδες μεταβλητές ή μαθηματικούς τύπους

Online βιβλία στα Ελληνικά

- Β. Βερύκιος κ.α. Η ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕΣΑ ΑΠΟ ΤΗ ΓΛΩΣΣΑ R, (Κεφ. 1, 5, 6) http://repository.kallipos.gr/handle/11419/2965
- Ε. Κύρκος, Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων (Κεφ. 6, 7, 9, 11) http://repository.kallipos.gr/handle/11419/1226

Οδηγοί χρήσης εργαλείων (στα Ελληνικά)

- Γλώσσα R: http://repository.kallipos.gr/handle/11419/2965
- Waikato Environment of Knowledge Analysis – WEKA: http://repository.kallipos.gr/handle/11419/1226 (κεφ. 13)

Πηγές που χρησιμοποιήθηκαν

- Βιβλίο «Εισαγωγή στην Εξόρυξη δεδομένων», Tan κ.ά.
- Β. Βερύκιος κ.α. Η ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕΣΑ ΑΠΟ ΤΗ ΓΛΩΣΣΑ
- Ε. Κύρκος, Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων
- Διαφάνειες Β. Βερύκιου από ΕΑΠ & ΑΠΚΥ
- Διαφάνειες Ε. Πιτουρά από Παν. Ιωαννίνων