ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΤΜΗΜΑ ΜΗΧ/ΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ

Εισαγωγή στην Μηχανική Μάθηση

Β. Μεγαλοοικονόμου

(εν μέρη βασισμένο σε σημειώσεις των J. Han και M. Kamber, του C. Faloutsos, και των S.J. Russell και P. Norvig)

Θεματολογία

- Εισαγωγή
- Προεπεξεργασία δεδομένων
- Κατηγοριοποίηση
- Πρόβλεψη
- Συσταδοποίηση

Εξέλιξη της τεχνολογίας διαχείρισης και ανάλυσης δεδομένων

- 1960:
  - Συλλογή δεδομένων, δημιουργία βάσης δεδομένων, συστήματα διαχείρισης πληροφορίας (IMS) και δικτυωτές βάσεις δεδομένων (network DBMS)
- 1970:
  - Σχεσιακό μοντέλο δεδομένων, υλοποίηση σχεσιακού ΣΔΒΔ
- 1980:
  - RDBMS, προηγμένα μοντέλα δεδομένων (extended-relational, OO, deductive,κτλ.) και συστήματα προσανατολισμένα στην εφαρμογή (spatial, scientific, engineering, κτλ.)
- 1990—2000:
  - Εξόρυξη δεδομένων και αποθηκών δεδομένων (Data mining, data warehousing), βάσεις δεδομένων πολυμέσων, βάσεις δεδομένων στο Παγκόσμιο Ιστό (Web databases)
- 2000 - σήμερα:
  - NoSQL databases, NewSQL, Hadoop, DSMS, big data analytics & Machine Learning

Μηχανική Μάθηση (Machine Learning)

- ✓ Ασχολείται με την ανάπτυξη και τη μελέτη στατιστικών αλγορίθμων που μπορούν να μάθουν από δεδομένα και να γενικεύσουν σε δεδομένα που δεν έχουν δει και έτσι να εκτελέσουν εργασίες χωρίς ρητές οδηγίες
- ✓ Η πρόοδος στον τομέα της βαθιάς μάθησης επέτρεψε στα νευρωνικά δίκτυα να ξεπεράσουν πολλές προηγούμενες προσεγγίσεις στην απόδοση
- ✓ Εφαρμογές σε πολλούς τομείς: επεξεργασία φυσικής γλώσσας, υπολογιστική όραση, αναγνώριση ομιλίας, φιλτράρισμα email, η γεωργία, η ιατρική, κ.α.
- ✓ Η εφαρμογή της ML σε επιχειρηματικά προβλήματα είναι γνωστή ως predictive analytics.
- ✓ Η εξόρυξη δεδομένων είναι ένα σχετικό πεδίο μελέτης, που εστιάζει στην διερευνητική ανάλυση δεδομένων μέσω μάθησης χωρίς επίβλεψη

Προσεγγίσεις Μηχανικής Μάθησης

- Τρεις ευρείες κατηγορίες: αντιστοιχούν σε παραδείγματα μάθησης, ανάλογα με τη φύση της «ανάδρασης» που διατίθεται στο σύστημα εκμάθησης:
- **Εποπτευόμενη μάθηση**: Παρουσιάζονται παραδείγματα εισόδων και επιθυμητών εξόδων, που δίνονται από έναν «δάσκαλο» και ο στόχος είναι να μάθουμε έναν γενικό κανόνα που αντιστοιχίζει τις εισόδους στις εξόδους.
- **Μη εποπτευόμενη μάθηση**: Δεν δίνονται ετικέτες στον αλγόριθμο εκμάθησης, αφήνοντάς τον μόνο του να βρει δομή στα δεδομένα εισόδου. Η μάθηση μπορεί να είναι ένας στόχος από μόνος του (ανακάλυψη κρυφών μοτίβων στα δεδομένα) ή ένα μέσο για την επίτευξη ενός σκοπού (εκμάθηση χαρακτηριστικών).
- **Ενισχυτική μάθηση**: Ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτύχει έναν συγκεκριμένο στόχο (όπως η οδήγηση ενός οχήματος ή το να παίξει ένα παιχνίδι εναντίον ενός αντιπάλου). Καθώς περιηγείται στον χώρο του προβλήματος, το πρόγραμμα λαμβάνει ανατροφοδότηση που είναι ανάλογη με τις ανταμοιβές, τις οποίες προσπαθεί να μεγιστοποιήσει.

Διαφορά Μηχανικής Μάθησης από Ανακάλυψη Γνώσης-Εξόρυξης δεδομένων ή γνώσης

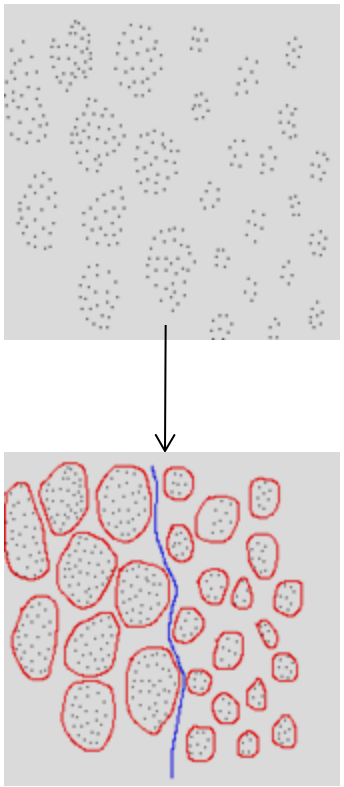
- ✓ Ανακάλυψη γνώσης από βάσεις δεδομένων (knowledge discovery from databases - KDD):
  - – Αποκάλυψη ή παραγωγή λειτουργικής γνώσης, μέσω της ανάλυσης δεδομένων από μεγάλες αποθήκες δεδομένων.
  - – Εύρεση δομών γνώσης που αναδεικνύουν γνώση (π.χ. συσχετίσεις ή κανόνες) που είναι κρυμμένη μέσα στα δεδομένα και δεν μπορούν να εξαχθούν από τον άνθρωπο με ευκολία.
  - – Αναφέρεται στην όλη διαδικασία.

Ανακάλυψη γνώσης από δεδομένα

Ορισμός

- ✓ «KDD είναι η ντετερμινιστική διαδικασία αναγνώρισης έγκυρων, καινοτόμων, ενδεχομένως χρήσιμων και εν τέλει κατανοητών προτύπων στα δεδομένα.» (*Frawley, Piatesky-Shaphiro and Matheus, 1991*).

Δεδομένα: Οντότητες ή συσχετίσεις του πραγματικού κόσμου (π.χ. εγγραφές συναλλαγών τραπέζης, supermarket κλπ).
Πρότυπο: Μια έκφραση σε μια γλώσσα που χαρακτηρίζει ένα υποσύνολο των δεδομένων (π.χ. ένας κανόνας).
Εγκυρότητα: Το πρότυπο είναι συνεπές σε νέα δεδομένα.
Πιθανή χρησιμότητα: Να μπορεί να χρησιμοποιηθεί για κάποιο σκοπό (π.χ. λήψη αποφάσεων).
Τελικά κατανοητό: Κοινώς κατανοητό, ώστε να είναι κοινώς χρήσιμο.



Βασικά Βήματα Διαδικασίας Μηχανικής Μάθησης

- *Επιλογή* (*Selection*): Απόκτηση δεδομένων από διάφορες ετερογενείς πηγές.
- *Προεπεξεργασία* (*Preprocessing*): Εσφαλμένα, προβληματικά ή ελλείποντα δεδομένα τακτοποιούνται (μπορεί να απαιτήσει το 60% της προσπάθειας!)
- *Μετασχηματισμός* (*Transformation*): Μετατροπή ετερογενών δεδομένων σε κοινή τυποποίηση για επεξεργασία.
- *Εφαρμογή αλγορίθμων για παραγωγή μοντέλου*
- *Ερμηνεία/Αξιολόγηση* (*Interpretation/Evaluation*): Παρουσίαση των αποτελεσμάτων στους χρήστες για αξιολόγηση (χρήση μεθόδων οπτικοποίησης και GUI).

Εξαγωγή Χαρακτηριστικών (Features)

Τα χαρακτηριστικά διακρίνονται σε:

- Ποσοτικά χαρακτηριστικά (Quantitative features):
  - continuous values (π.χ. βάρος),
  - discrete values (π.χ. ο αριθμός των υπολογιστών),
  - interval values (π.χ. η διάρκεια ενός γεγονότος).
- Ποιοτικά χαρακτηριστικά (Qualitative features):
  - Ονομαστικά (nominal) or unordered (π.χ. χρώμα),
  - Ordinal (π.χ. στρατιωτική διάκριση ή ποιοτική αξιολόγηση της θερμοκρασίας (“ζεστό” or “κρύο”) ή της έντασης του ήχου (“ήσυχα” or “δυνατά”)).
- Δομικά χαρακτηριστικά (Structured features):
  - trees
  - symbolic objects

Προσεγγίσεις Αλγορίθμων Μηχανικής Μάθησης

- ✓ Σύμφωνα με τα παρακάτω:
  - ✓ Συνθετικά μέρη
  - ✓ Περιγραφή μοντέλου
    - Τύπος ή στόχος μοντέλου (π.χ. ταξινόμηση, συσταδοποίηση)
    - Μορφή μοντέλου (π.χ. δέντρα, κανόνες, νευρωνικά δίκτυα)
  - ✓ Κριτήρια αξιολόγησης μοντέλου (εγκυρότητα, ακρίβεια κλπ)
  - ✓ Τεχνική αναζήτησης:
    - Παραμέτρων ή
    - Μοντέλου

Τύποι Μοντέλων

- ✓Προβλεπτικό (predictive)
  - Κάνει πρόβλεψη συμπεριφοράς κάποιων μεταβλητών που παρουσιάζουν ενδιαφέρον και οι οποίες βασίζονται στη συμπεριφορά άλλων μεταβλητών. Η πρόβλεψη μπορεί να στηρίζεται σε ιστορικά δεδομένα.
- ✓Περιγραφικό (descriptive)
  - Βρίσκει πρότυπα (patterns) ή σχέσεις (relationships) που ενυπάρχουν στα δεδομένα. Ερευνά τις ιδιότητες των υπό εξέταση δεδομένων, εξηγεί τη συμπεριφορά τους.

Διεργασίες Μηχανικής Μάθησης

Προβλεπτικού τύπου ή μοντέλου

- ✓ Κατηγοριοποίηση (Classification)
- ✓ Παλινδρόμηση ή Παρεμβολή (Regression)
- ✓ Ανάλυση χρονοσειρών (Time series analysis)
- ✓ Πρόβλεψη (Prediction)

Περιγραφικού τύπου ή μοντέλου

- ✓ Συσταδοποίηση (Clustering)
- ✓ Σύνοψη (Summarization) ή Γενίκευση (Generalization)
- ✓ Εύρεση Κανόνων Συσχέτισης (Association Rules)
- ✓ Ανακάλυψη Συσχετίσεων σε Ακολουθίες (Pattern Discovery in Sequences)

Απαιτήσεις Αλγορίθμων

- Χειρισμός διαφορετικών τύπων δεδομένων
- Απόδοση και εξελιξιμότητα των αλγορίθμων
- Χρησιμότητα, εγκυρότητα και εκφραστικότητα των αποτελεσμάτων
- Διαφορετικού τύπου εκφράσεις ερωτήσεων και αποτελεσμάτων.
- Σύνθεση διαφορετικών πηγών δεδομένων

Κατηγοριοποίηση Δεδομένων

- ✓ *Κατηγοριοποίηση δεδομένων* (*data classification*):
  - ✓ Βασίζεται στην εξέταση των χαρακτηριστικών ενός νέου αντικειμένου (*μη κατηγοριοποιημένο*) το οποίο με βάση τα χαρακτηριστικά αυτά αντιστοιχίζεται σε ένα προκαθορισμένο σύνολο κλάσεων.
  - ✓ Τα προς κατηγοριοποίηση αντικείμενα αναπαριστάνονται γενικά από τις εγγραφές της βάσης δεδομένων
  - ✓ Η διαδικασία της κατηγοριοποίησης αποτελείται από την ανάθεση κάθε εγγραφής σε κάποιες από τις προκαθορισμένες κατηγορίες.
- ✓ <u>Βασικές μέθοδοι</u>: Μέθοδος Bayes, Δέντρα Αποφάσεων, Νευρωνικά Δίκτυα

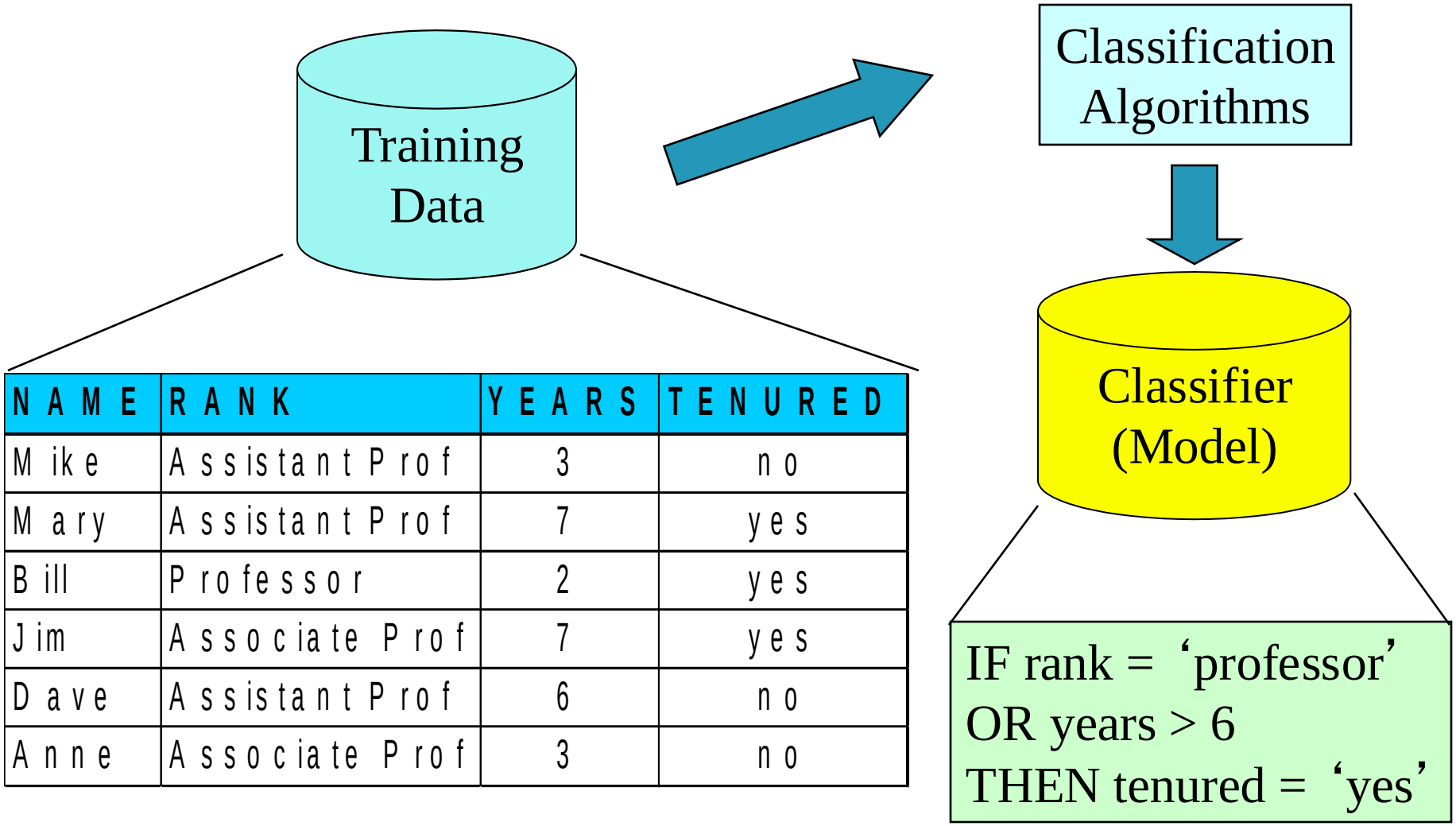
Κατηγοριοποίηση – Πρόβλεψη

- Κατηγοριοποίηση:
  - Προβλέπει την κατηγορία (predicts categorical class labels)
  - Κατηγοριοποιεί δεδομένα (κατασκευάζοντας ένα μοντέλο) βασισμένο σε training set και τις τιμές (ετικέτες κατηγορίας class labels) του χαρακτηριστικού κατηγορίας και το χρησιμοποιεί για κατηγοριοποίηση νέων δεδομένων
- Πρόβλεψη:
  - Μοντελοποίηση συνεχών συναρτήσεων, π.χ., πρόβλεψη άγνωστων ή απώντων τιμών
- Τυπικές Εφαρμογές
  - credit approval
  - target marketing
  - medical diagnosis
  - treatment effectiveness analysis

Κατηγοριοποίηση — Μια διαδικασία δύο βημάτων

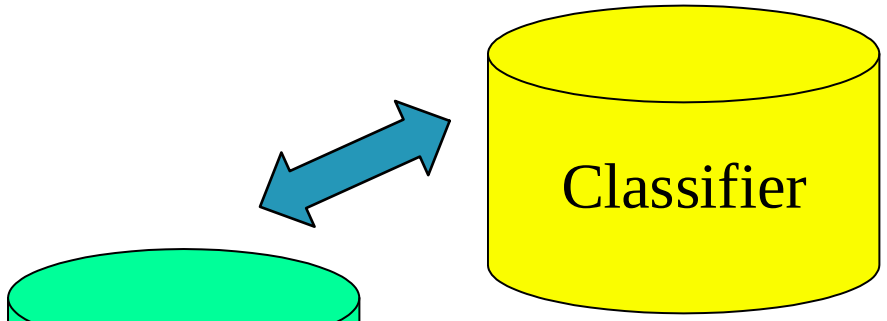
- Κατασκευή μοντέλου: περιγράφοντας ένα σύνολο προκαθορισμένων κατηγοριών
  - Καθε μια από τις πλειάδες (tuples) υποτίθεται ότι ανήκει σε μια προκαθορισμένη κατηγορία, όπως καθορίζεται από το γνώρισμα κατηγορίας (κλάσης) (**supervised learning**)
  - Το σύνολο των πλειάδων που χρησιμοποιείται για κατασκευή μοντέλου: training set
  - Το μοντέλο αναπαρίσταται σαν classification rules, decision trees, ή μαθηματικές εξισώσεις
- Χρήση μοντέλου: για κατηγοριοποίηση άγνωστων αντικειμένων
  - Υπολογισμός ακρίβειας του μοντέλου κάνοντας χρήση ενός test set
    - Η γνωστή ετικέτα του test sample συγκίνεται με το αποτέλεσμα κατηγοριοποίησης του μοντέλου
    - Η τιμή της ακρίβειας είναι το ποσοστό των test set samples που κατηγοριοποιήθηκαν σωστά απο το μοντέλο
    - Το test set είναι ανεξάρτητο από το training set, αλλιώς μπορεί να συμβεί over-fitting

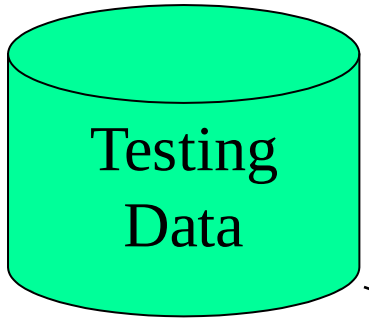
Διαδικασία Κατηγοριοποίησης: Κατασκευή Μοντέλου

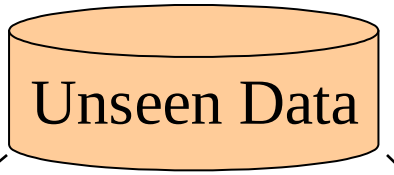


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Διαδικασία Κατηγοριοποίησης: Χρήση Μοντέλου για Πρόβλεψη







(Jeff, Professor, 4)

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Tenured?

Yes

Συσταδοποίηση Δεδομένων (1)

- ✓ Η συσταδοποίηση δεδομένων (data clustering) είναι η εργασία του καταμερισμού ενός ετερογενούς πληθυσμού σε ένα σύνολο συστάδων (clusters).
- ✓ Στην συσταδοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες. Οι εγγραφές ομαδοποιούνται σε σύνολα με βάση την ομοιότητα που παρουσιάζουν μεταξύ τους. Επαφίεται σε εμάς να καθορίσουμε την σημασία που θα έχει κάθε μία από τις ομάδες που προκύπτουν. Για παράδειγμα ομάδες που περιλαμβάνουν τα χαρακτηριστικά που σχετίζονται με τα φύλλα και τον καρπό φυτών μπορεί να υποδεικνύουν διαφορετικές ποικιλίες ενός φυτού.
- ✓ Αντιπροσωπευτικοί αλγόριθμοι: K-Means και παραλλαγές, PAM, COBWEB

Συσταδοποίηση Δεδομένων (2)

- Διαιρετική συσταδοποίηση (Partitioning Clustering)
- Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)
- Ασαφής συσταδοποίηση (Fuzzy Clustering)
- Μη ασαφής συσταδοποίηση (Crisp Clustering)
- Συσταδοποίηση βασισμένη στα δίκτυα Kohonen (Kohonen Net Clustering)
- Συσταδοποίηση βασισμένη στην πυκνότητα (Density-based Clustering)
- Συσταδοποίηση βασισμένη σε πλέγμα (Grid-based Clustering)
- Συσταδοποίηση υποχώρων (Subspace Clustering).

Συσταδοποίηση Δεδομένων (3)

- ✓ Οι περισσότεροι αλγόριθμοι υποθέτουν μία μεγάλη δομή δεδομένων που είναι memory resident.
- ✓ Η ομαδοποίηση μπορεί να πραγματοποιηθεί πρώτα σε ένα δείγμα της Β.Δ. και στη συνέχεια να εφαρμοστεί σε όλη τη Β.Δ.
- ✓ Αλγόριθμοι
  - BIRCH
  - DBSCAN
  - CURE

Συσταδοποίηση Δεδομένων (4)

- ✓ Ένα αντικείμενο **x** είναι ένα ξεχωριστό-μοναδικό data item που χρησιμοποιείται από τον clustering αλγόριθμο.

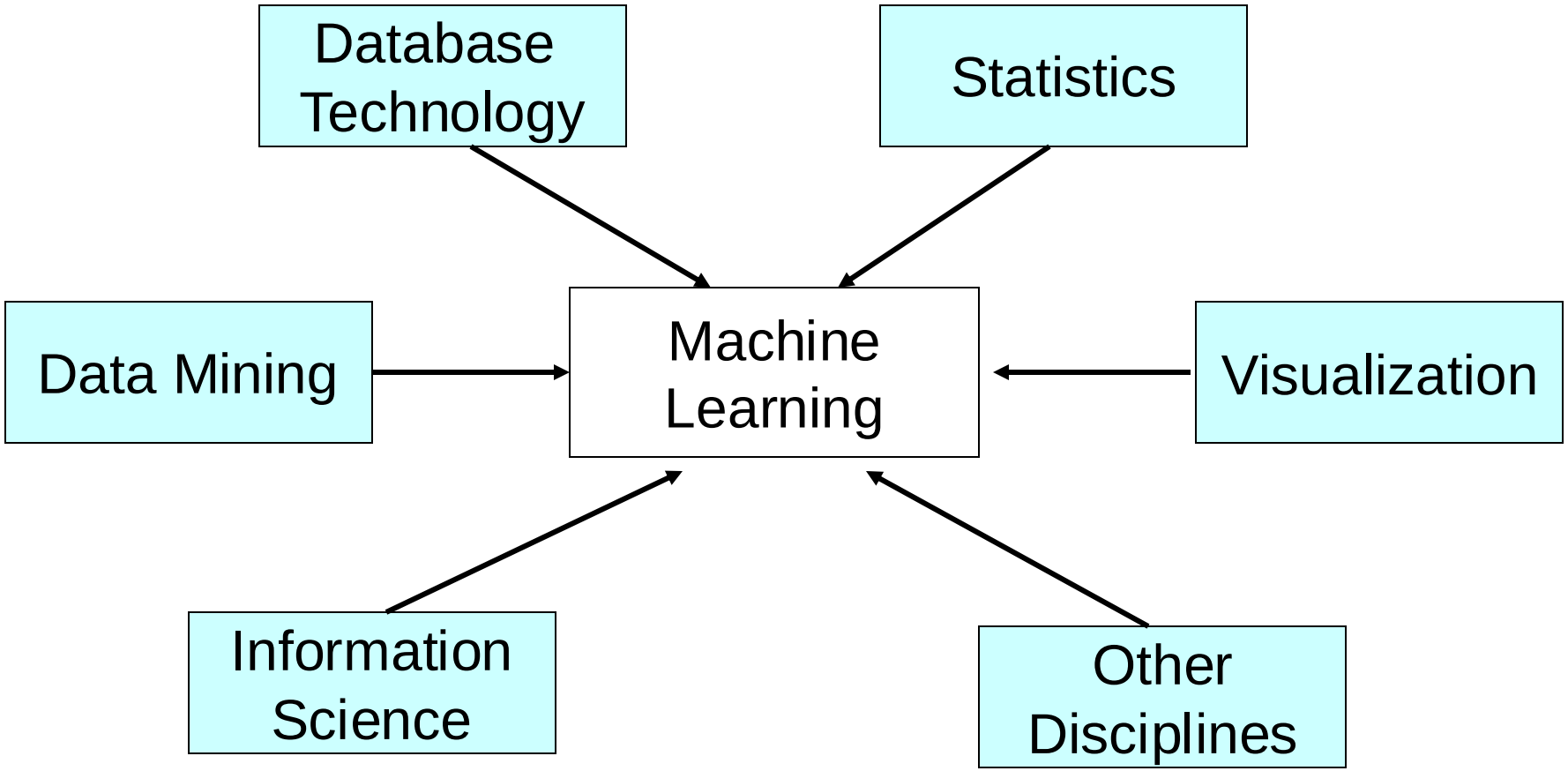
- ✓ Αποτελείται από ένα διάνυσμα **d** μετρικών: $\mathbf{x = (x_1, \ldots x_d)}$, όπου x_i είναι το i-οστό χαρακτηριστικό (feature) του αντικειμένου και *d* η διάσταση του αντικειμένου ή του χώρου που δημιουργείται από τα αντικείμενα.

Συσταδοποίηση Δεδομένων (5)

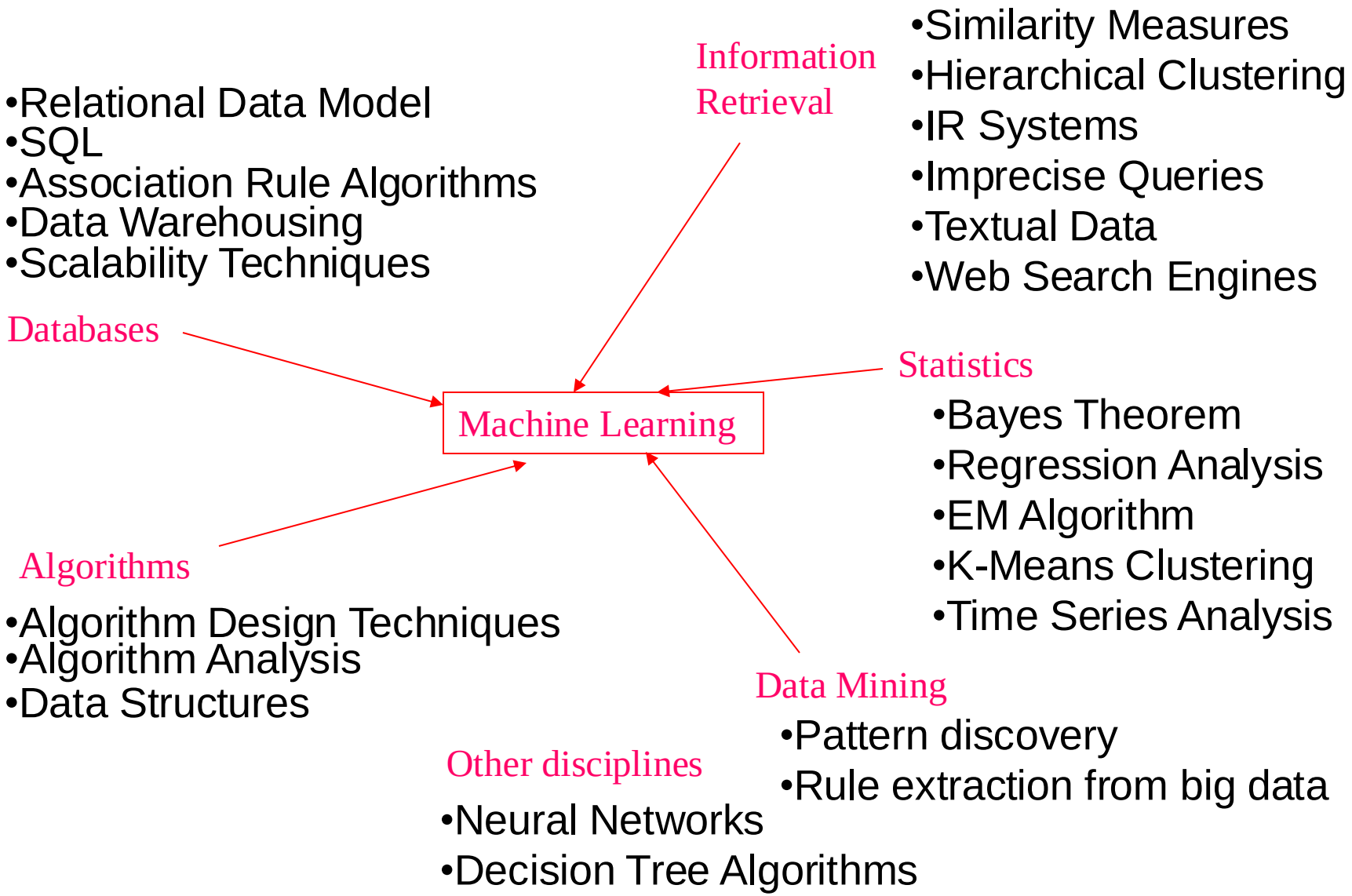
- ✓ *Hard* clustering τεχνικές: αναθέτουν μία ετικέτα κλάσης(class) *li* σε κάθε αντικείμενο **x***i*, που είναι προσδιοριστική για την κλάση του.

- ✓ *Fuzzy* clustering διαδικασίες: αναθέτουν σε κάθε pattern **xi** που δίνεται ως είσοδος ένα «ποσοστό συμμετοχής» *fij* σε κάθε παραγόμενο cluster *j*.

Γειτονικές Περιοχές



Machine Learning Development



Είδη Μηχανικής Μάθησης

- Μάθηση υπό επίβλεψη ή από παραδείγματα:
  - Μάθηση εννοιών (Concept Learning)
  - Δέντρα Ταξινόμησης ή Απόφασης (Classification or Decision Trees)
  - Μάθηση Κανόνων (Rule Learning)
  - Μάθηση κατά Bayes
  - Γραμμική Παλινδρόμηση/Παρεμβολή (Linear Regression)
  - Νευρωνικά Δίκτυα (Neural Networks)
  - Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)
- Μάθηση χωρίς επίβλεψη ή από παρατήρηση:
  - Εξαγωγή Ομάδων/Συστάδων (clustering)
  - Εξαγωγή συσχετίσεων (association rules)

Μηχανική Μάθηση και Εξόρυξη Γνώσης

- Ο στόχος της Μηχανικής Μάθησης είναι η εύρεση τεχνικών μάθησης ως μίμηση της ανθρώπινης συμπεριφοράς (δηλ. η παραγωγή προγραμμάτων που μαθαίνουν), ενώ της Εξόρυξης Γνώσης η ανακάλυψη πληροφορίας που θα είναι χρήσιμη στον άνθρωπο. (Dunham, 2003)
- Η Εξόρυξη Γνώσης χρησιμοποιεί μεγαλύτερες, ετερογενείς και «ακατέργαστες» βάσεις δεδομένων, ενώ η Μηχανική Μάθηση μικρότερες, ομογενείς και κατάλληλης μορφής.
- Αρκετοί αλγόριθμοι Εξόρυξης Γνώσης χρησιμοποιούνται στην Μηχανική Μάθηση

Προεπεξεργασία Δεδομένων

Γιατί προ-επεξεργασία δεδομένων;

- Τα πραγματικά δεδομένα έχουν διάφορα προβλήματα
  - Ελλειπή:
    - έλλειψη τιμών των χαρακτηριστικών,
    - έλλειψη ορισμένων χαρακτηριστικών ενδιαφέροντος, ή
    - περιέχουν μόνο συναθροιστικά δεδομένα
  - Θορυβώδη: περιέχουν σφάλματα ή ακραίες τιμές (outliers)
  - Ασυνεπή: περιέχουν ασυνέπειες στους κωδικούς ή στα ονόματα
- Μη ποιοτικά δεδομένα, μη ποιοτικά αποτελέσματα εξόρυξης!
  - Οι ποιοτικές αποφάσεις βασίζονται σε ποιοτικά δεδομένα
  - Οι αποθήκες δεδομένων χρειάζονται συνεπή ενοποίηση ποιοτικών δεδομένων

Πολύ-επίπεδη μετρική ποιότητας δεδομένων

- Μια αποδεκτή πολυδιάστατη άποψη:
  - Ακρίβεια
  - Πληρότητα
  - Συνέπεια
  - Προσβασιμότητα
  - Timeliness
  - Believability
  - value added
  - interpretability
- Διευρυμένες κατηγορίες:
  - intrinsic, contextual, representational, accessibility.

Βασικές Εργασίες στην Προεπεξεργασία Δεδομένων

- Καθαρισμός Δεδομένων
  - Συμπλήρωση ελλιπών τιμών
  - εξομάλυνση δεδομένων που έχουν θόρυβο
  - αναγνώριση ή απομάκρυνση των ακραιών τιμών, και
  - επίλυση ασυνεπειών
- Ενοποίηση Δεδομένων
  - Ενοποίηση πολλαπλών βάσεων δεδομένων, κύβων δεδομένων, αρχείων, ή σημειώσεων

Βασικές Εργασίες στην Προεπεξεργασία Δεδομένων (2)

- Μετασχηματισμός Δεδομένων
  - Κανονικοποίηση (κλιμάκωση σε ένα συγκεκριμένο εύρος)
  - Συνάθροιση (aggregation)
- Μείωση Δεδομένων
  - Απόκτηση μειωμένων αναπαραστάσεων σε όγκο αλλά παραγωγή των ίδιων ή παρόμοιων αναλυτικών αποτελεσμάτων
  - Διακριτοποίηση δεδομένων: με συγκεκριμένη σημασία, ειδικά για αριθμητικά δεδομένα
  - Συνάθροιση δεδομένων, μείωση διαστατικότητας, συμπίεση δεδομένων, γενίκευση

Καθαρισμός Δεδομένων

- Δραστηριότητες καθαρισμού δεδομένων:
  - Συμπλήρωση ελλιπών τιμών
  - Αναγνώριση ακραίων τιμών (outliers)
  - Εξομάλυνση δεδομένων που έχουν θόρυβο
  - Διόρθωση ασυνεπών δεδομένων

Ελλιπή Δεδομένα

- Τα δεδομένα δεν είναι πάντα διαθέσιμα
  - Π.χ., αρκετές πλειάδες δεν έχουν καταγεγραμμένες τιμές για ορισμένα χαρακτηριστικά, όπως ο μισθός ενός πελάτη σε δεδομένα πωλήσεων
- Τα ελλιπή δεδομένα μπορεί να οφείλονται σε
  - δυσλειτουργία του εξοπλισμού
  - ασυνέπειες με άλλα καταγεγραμμένα δεδομένα και συνεπώς διαγραφή τους
  - μη καταχώρηση δεδομένων
  - μη καταγεγραμμένο ιστορικό ή αλλαγές δεδομένων
  - ορισμένα δεδομένα μπορεί να μην θεωρήθηκαν σημαντικά την στιγμή της εκχώρησης
- Τα ελλιπή δεδομένα ίσως χρειάζεται να παραχθούν με βάση τα υπάρχοντα (με συμπερασμό)

Πως χειριζόμαστε τα ελλιπή δεδομένα?

- Αναγνώριση της πλειάδας:

  συνήθως γίνεται όταν λείπει η ετικέτα κλάσης (υποθέτοντας ότι ο στόχος είναι η κατηγοριοποίηση— μη αποτελεσματικό σε ορισμένες περιπτώσεις)

- Συμπλήρωση των ελλιπών τιμών χειρωνακτικά:

  κουραστικό/ανέφικτο

- Χρήση μιας καθολικής σταθεράς:

  π.χ., “unknown”, μια νέα κλάση?!

- Χρήση της μέσης τιμής του χαρακτηριστικού

- Χρήση της μέσης τιμής των δειγμάτων της ίδιας κλάσης

- Χρήση της πιο πιθανής τιμής:

  βασισμένη σε regression, Bayesian εξίσωση, δένδρα απόφασης, κ.α.

Πως χειριζόμαστε τα ελλιπή δεδομένα?

- Χρήση της μέσης τιμής των δειγμάτων της ίδιας κλάσης (… εξυπνότερο)
- Χρήση της πιο πιθανής τιμής:
  - βασισμένη στο συμπερασμό με τεχνικές όπως regression, Bayesian εξίσωση, δένδρα απόφασης, κ.α.

Δεδομένα με θόρυβο

- Τι είναι θόρυβος?
  - Τυχαίο σφάλμα σε μια μεταβλητή μέτρησης
- Λανθασμένες τιμές χαρακτηριστικών λόγω:
  - Προβληματικών εργαλείων συλλογής δεδομένων
  - Προβλημάτων εγγραφής δεδομένων
  - Προβλημάτων μετάδοσης δεδομένων
  - Περιορισμών λόγω συγκεκριμένης τεχνολογίας
  - Ασυνεπειών στην μετατροπή ονομάτων
- Άλλα προβλήματα δεδομένων που απαιτούν καθαρισμό
  - διπλοεγγραφές
  - μη ολοκληρωμένα δεδομένα
  - μη συνεπή δεδομένα

Πως χειριζόμαστε τα δεδομένα με θόρυβο;

- Μέθοδος Binning:
  - Αρχικά ταξινόμησε τα δεδομένα και διαχώρισέ τα σε (ίσου-βάθους) bins (δοχεία)
  - Τότε μπορεί να εφαρμοστεί η εξής ομαλοποίηση: smooth by bin means, smooth by bin median, smooth by bin boundaries, κτλ.
  - Χρησιμοποιείται και για διακριτοποίηση (συζητείται αργότερα)
- Συσταδοποίηση
  - Ανίχνευση και απομάκρυνση των ακραίων τιμών (outliers)
- Ημι-αυτοματοποιημένη μέθοδος: συνδυάζει τον υπολογιστή με την ανθρώπινη διαίσθηση
  - Ανίχνευση υπόπτων τιμών και χειρωνακτικός έλεγχος
- Παλινδρόμηση (Regression)
  - Εξομάλυνση με το ταίριασμα των δεδομένων σε συναρτήσεις παλινδρόμησης

Απλές Μέθοδοι Διακριτοποίησης: Binning

- Διαμερισμός ίσου-πλάτους (απόσταση):
  - Διαιρεί το εύρος σε *N* διαστήματα ίσου μεγέθους: ομοιόμορφο grid
  - Αν *A* και *B* είναι η χαμηλότερη και η υψηλότερη τιμή ενός χαρακτηριστικού αντίστοιχα, το πλάτος των διαστημάτων θα είναι: $W = (B-A)/N$.
  - Η πιο απλή μέθοδος
  - Οι outliers μπορεί να συνεχίσουν να υπάρχουν
  - Τα ασύμμετρα δεδομένα (skewed data) δεν διαχειρίζονται σωστά

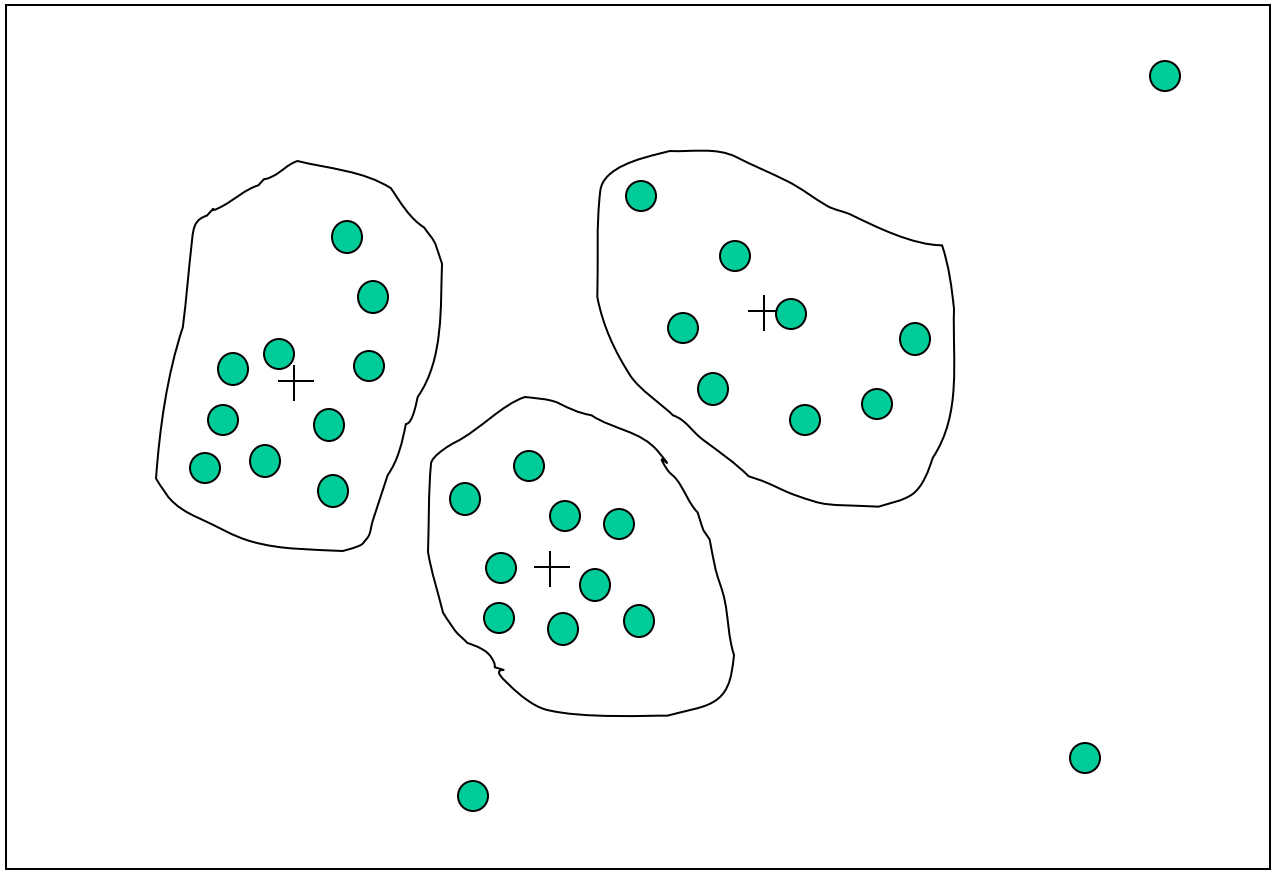
Απλές Μέθοδοι Διακριτοποίησης: Binning

- Διαμερισμός ίσου-βάθους (συχνότητα):
  - Διαιρεί το εύρος σε *N* διαστήματα, που περιέχουν περίπου τον ίδιο αριθμό δειγμάτων
  - Καλή κλιμάκωση των δεδομένων
  - Η διαχείριση των κατηγορικών χαρακτηριστικών μπορεί να αποβεί περίπλοκη

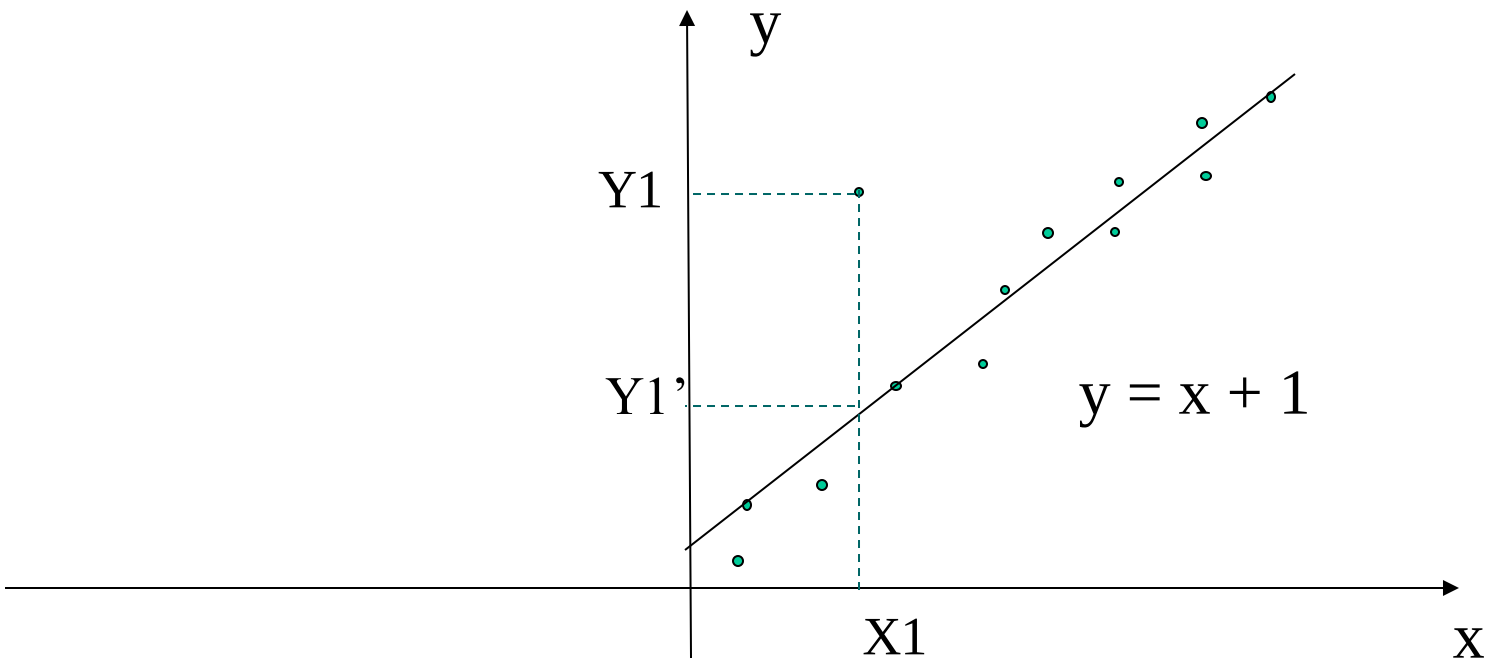
Μέθοδοι Binning για Εξομάλυνση Δεδομένων

* Δεδομένα ταξινομημένα με βάση την τιμή (σε δολλάρια): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Διαμερισμός σε bins equi-depth:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Ομαλοποίηση (smoothing) με βάση τον μέσο όρο των bins:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Ομαλοποίηση (smoothing) με βάση τα όρια τιμών των bins:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Ανάλυση Συσταδοποίησης



Regression (Παλινδρόμηση)



- Γραμμική παλινδρόμηση (η βέλτιστη γραμμή που περιγράφει δύο μεταβλητές)
- Πολλαπλή γραμμική παλινδρόμηση (για πιο πολλές από δύο μεταβλητές, ταίριασμα σε μια πολυδιάστατη επιφάνεια)

Πως χειριζόμαστε τα Ασυνεπή Δεδομένα?

- Χειρωνακτική διόρθωση χρησιμοποιώντας εξωτερικές αναφορές
- Ημι-αυτόματη διόρθωση χρησιμοποιώντας διάφορα εργαλεία
  - Για την ανίχνευση της παραβίασης γνωστών *συναρτησιακών εξαρτήσεων* και *περιορισμών* των δεδομένων
  - Για την διόρθωση *πλεοναζόντων δεδομένων*

Ενοποίηση Δεδομένων

- Ενοποίηση δεδομένων:
  - Συνδυάζει δεδομένα από πολλαπλές πηγές σε μια συνεκτική αποθήκευση
- Ενοποίηση Σχήματος (Schema integration)
  - Ενοποίηση μεταδεδομένων από διαφορετικές βάσεις δεδομένων (πηγές)
  - Πρόβλημα ταυτοποίησης οντότητας (entity identification problem):
    - αναγνώριση πραγματικών οντοτήτων από πολλαπλές πηγές δεδομένων, π.χ., A.cust-id ≡ B.cust-#
- Ανίχνευση και ανάλυση της σύγκρουσης των τιμών δεδομένων
  - για την ίδια πραγματική οντότητα, οι τιμές των χαρακτηριστικών από διαφορετικές πηγές διαφέρουν
  - πιθανές αιτίες: διαφορετικές αναπαραστάσεις, διαφορετικές κλίμακες, π.χ., Ελληνικές vs. Βρετανικές μονάδες μέτρησης

Διαχείριση Πλεοναζόντων Δεδομένων στην Ενοποίηση Δεδομένων

- Τα πλεονάζοντα δεδομένα εμφανίζονται συχνά όταν συνενώνονται πολλαπλές ΒΔ
  - Το ίδιο γνώρισμα μπορεί να έχει διαφορετικά ονόματα σε διαφορετικές βάσεις δεδομένων
  - Ένα γνώρισμα μπορεί να είναι "παραγόμενο" γνώρισμα σε έναν άλλο πίνακα, π.χ., annual revenue
- Τα πλεονάζοντα δεδομένα μπορεί να ανιχνευτούν με ανάλυση συσχετίσεων $r_{A,B} = \frac{\Sigma(A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B}$
- Η προσεκτική ενοποίηση μπορεί να μειώσει/αποφύγει περιττές πληροφορίες και ασυνέπειες, να βελτιώσει την ταχύτητα εξόρυξης και να αυξήσει την ποιότητα

Μετασχηματισμός Δεδομένων

- Εξομάλυνση: απομάκρυνση θορύβου από δεδομένα (binning, συσταδοποίηση, regression)
- Συνάθροιση: σύνοψη, data cube construction
- Γενίκευση: ανάβαση ιεραρχίας εννοιών
- Κανονικοποίηση: κλιμάκωση σε ένα μικρό, περιορισμένο εύρος
  - min-max κανονικοποίηση
  - z-score κανονικοποίηση
  - κανονικοποίηση με δεκαδική κλίμακα
- Δημιουργία γνωρίσματος/χαρακτηριστικού
  - Νέα χαρακτηριστικά κατασκευάζονται από τα δοσμένα

Μετασχηματισμός Δεδομένων: Κανονικοποίηση

Ιδιαίτερα χρήσιμη για κατηγοριοποίηση (NNs, μετρήσεις απόστασης, NN κατηγοριοποίηση, κτλ)

- min-max κανονικοποίηση

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score κανονικοποίηση

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- κανονικοποίηση με δεκαδική κλίμακα

$$v' = \frac{v}{10^j}$$

Όπου j είναι ο μικρότερος ακέραιος τέτοιος ώστε $\text{Max}(|v'|)<1$

Μείωση Δεδομένων

- Πρόβλημα:
  - Οι Αποθήκες Δεδομένων αποθηκεύουν terabytes/petabytes δεδομένων: Η ανάλυση/εξόρυξη σύνθετων δεδομένων μπορεί να χρειαστεί πολύ χρόνο για να τρέξει σε ολόκληρο το σύνολο δεδομένων

- Πιθανή Λύση;
  - Μείωση Δεδομένων…

Μείωση Δεδομένων

- Παράγει μια μειωμένη αναπαράσταση του συνόλου δεδομένων η οποία είναι αρκετά μικρότερη σε όγκο αλλά ωστόσο παράγει τα ίδια (ή περίπου το ίδια) αναλυτικά αποτελέσματα

- Στρατηγικές Μείωσης Δεδομένων
  - Συνάθροιση Κύβου Δεδομένων (Data cube aggregation)
  - Μείωση Διαστατικότητας (Dimensionality reduction)
  - Συμπίεση Δεδομένων (Data compression)
  - Μείωση πολυαριθμίας (Numerosity reduction)
  - Διακριτοποίηση και τεχνικές ιεραρχικής γενίκευσης

Συνάθροιση Κύβου Δεδομένων

- Πολλαπλά επίπεδα συνάθροισης σε κύβους δεδομένων
  - Επιπλέον μείωση του μεγέθους των δεδομένων που χειριζόμαστε
- Αναφορά σε κατάλληλα επίπεδα
  - Χρήση των μικρότερων αναπαραστάσεων για την επίλυση του προβλήματος
- Τα ερωτήματα σχετικά με τη συναθροισμένη πληροφορία μπορούν να απαντηθούν χρησιμοποιώντας κύβους δεδομένων, όταν αυτό είναι δυνατό

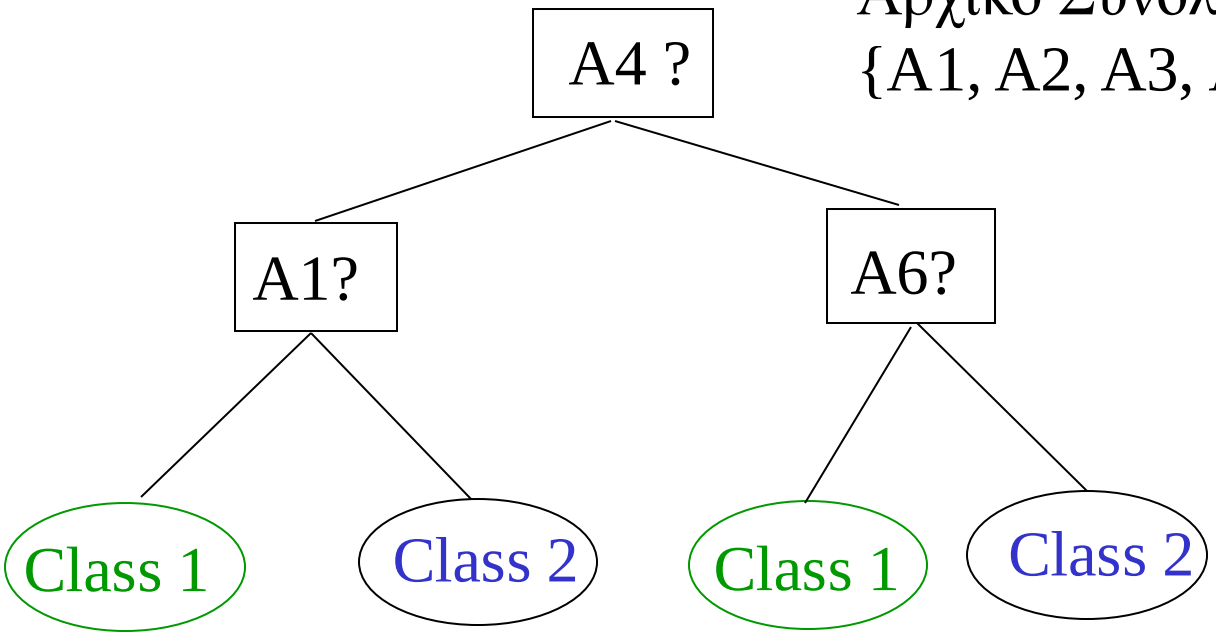
Μείωση Διαστατικότητας

- Πρόβλημα: Επιλογή Χαρακτηριστικών (π.χ., επιλογή υποσυνόλου χαρακτηριστικών):
  - Επέλεξε ένα ελάχιστο σύνολο χαρακτηριστικών τέτοιο ώστε η κατανομή πιθανότητας των διαφορετικών κλάσεων δεδομένων των τιμών αυτών των χαρακτηριστικών να είναι όσο το δυνατόν πιο κοντά στην αρχική κατανομή δεδομένων των τιμών όλων των χαρακτηριστικών
  - Ωφέλιμο side-effect: μειώνει το # των χαρακτηριστικών στα ανακαλυφθέντα πρότυπα (διευκολύνοντας την κατανόηση)
- Λύση: Ευρετικές μέθοδοι (λόγω του εκθετικού # των επιλογών) συνήθως άπληστες (greedy):
  - step-wise προς-τα-εμπρός επιλογή
  - step-wise προς-τα-πίσω περιορισμός
  - Συνδυασμός προς-τα-εμπρός επιλογής και προς-τα-πίσω περιορισμού
  - Συμπερασμός με δένδρο απόφασης

Παράδειγμα Συμπερασμού με Δένδρο Απόφασης

Εσωτερικοί κόμβοι: tests
Κλαδιά: outcomes of tests
Κόμβοι φύλλα: class prediction

Αρχικό Σύνολο Γνωρισμάτων:
{A1, A2, A3, A4, A5, A6}



------> Μειωμένο Σύνολο Γνωρισμάτων : {A1, A4, A6}

Ευρετικές Μέθοδοι Επιλογής Χαρακτηριστικών

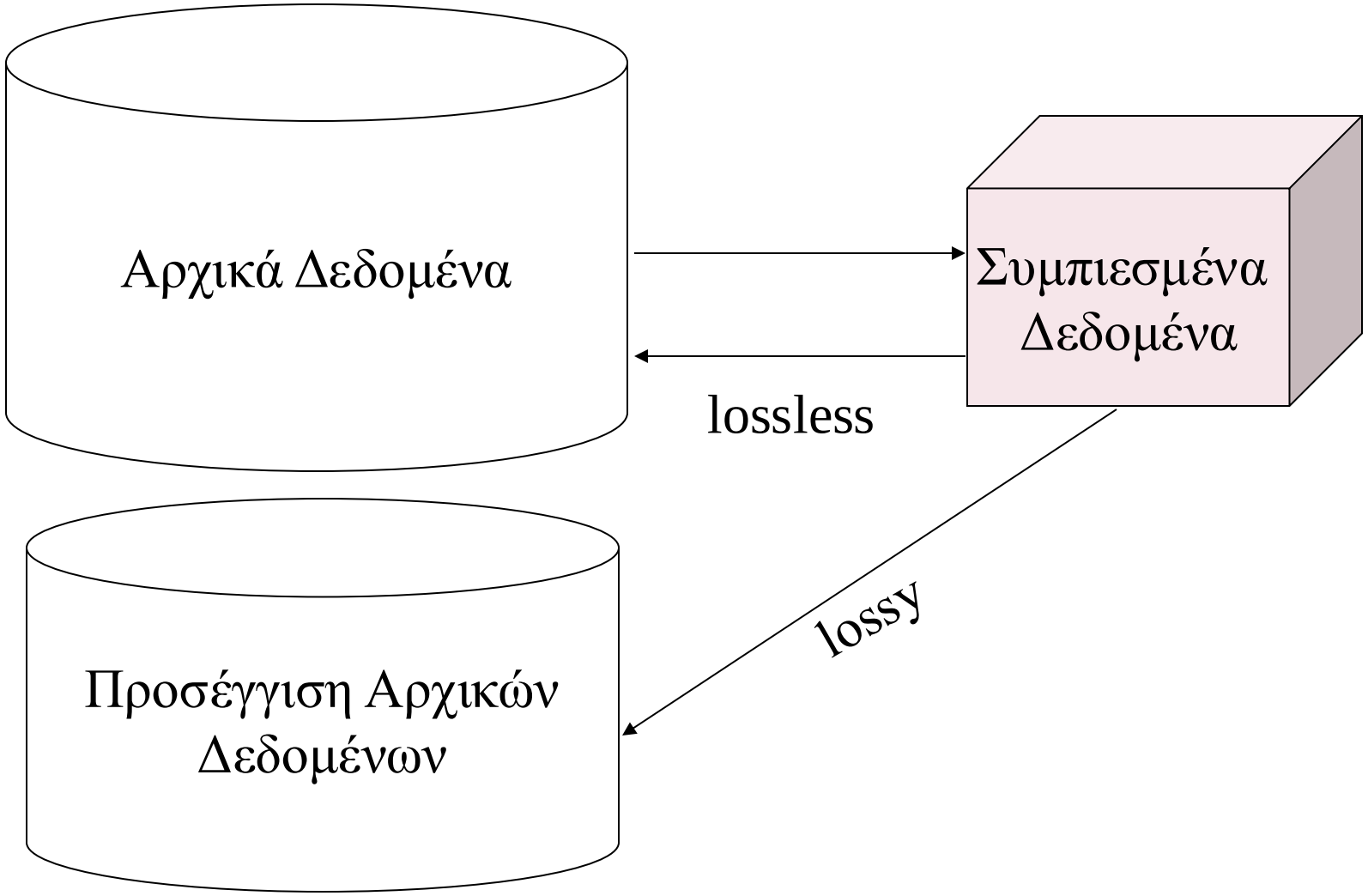
- Πόσα είναι τα πιθανά υπό-χαρακτηριστικά των d χαρακτηριστικών; 2^d
- Διάφορες μέθοδοι επιλογής χαρακτηριστικών:
  - Βέλτιστα χαρακτηριστικά υπό την υπόθεση της ανεξαρτησίας των χαρακτηριστικών: επιλογή με τεστ σημαντικότητας
  - Step-wise επιλογή χαρακτηριστικών:
    - Το βέλτιστο χαρακτηριστικό επιλέγεται αρχικά
    - Η επόμενη βέλτιστη συνθήκη χαρακτηριστικού με βάση το πρώτο, ...
  - Step-wise περιορισμός χαρακτηριστικών :
    - Επαναληπτικός περιορισμός των χειρότερων χαρακτηριστικών
  - Συνδυασμός επιλογής χαρακτηριστικών και περιορισμού χαρακτηριστικών :

    Βέλτιστο branch και bound:
    - Χρήση περιορισμού χαρακτηριστικών και προς-τα-πίσω-ανίχνευσης (backtracking)

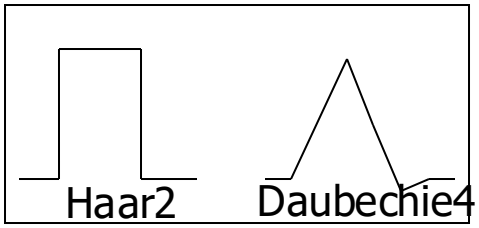
Συμπίεση Δεδομένων

- Συμπίεση αλφαριθμητικού
  - Υπάρχουν εκτενείς θεωρίες και well-tuned αλγόριθμοι
  - Συνήθως lossless (δηλ. δεν χάνεται πληροφορία)
  - Ωστόσο, περιορισμένη διαχείριση είναι δυνατή χωρίς επέκταση/αποσυμπίεση
- Audio/video, συμπίεση εικόνας
  - Συνήθως lossy συμπίεση, με προοδευτική βελτίωση
  - Μερικές φορές τμήματα του σήματος μπορούν να κατασκευαστούν χωρίς πλήρη ανακατασκευή
- Η χρονική ακολουθία δεν είναι audio
  - Συνήθως σύντομη και μεταβάλλεται αργά με το χρόνο

Συμπίεση Δεδομένων



Μετασχηματισμοί Κυματιδίου (Wavelet Transforms)



- Διακριτός μετασχηματισμός κυματιδίου DWT): γραμμική επεξεργασία σήματος
- Συμπιεσμένη προσέγγιση:
  - αποθήκευση μόνο ενός μικρού τμήματος των ισχυρότερων wavelet συντελεστών
- Παρόμοιος με το Διακριτό Μετασχηματισμό Fourier (DFT), αλλά
  - καλύτερη lossy συμπίεση, εντοπισμένη στο χώρο (συγκρατεί τοπική πληροφορία)
- Μέθοδος (ιεραρχικός αλγόριθμος πυραμίδας):
  - Το μήκος, L, πρέπει να είναι ακέραιος, δύναμη του 2 (συμπλήρωση με 0, αν χρειάζεται)
  - Κάθε μετασχηματισμός έχει 2 λειτουργίες:
    - Εξομάλυνση (π.χ., άθροισμα, βεβαρυμένος μέσος),
    - βεβαρυμένη διαφορά
  - Εφαρμόζεται σε ζεύγη δεδομένων,καταλήγοντας σε δύο σύνολα δεδομένων μήκους L/2
  - Εφαρμόζει δύο λειτουργίες αναδρομικά, μέχρι να φτάσει στο επιθυμητό μήκος

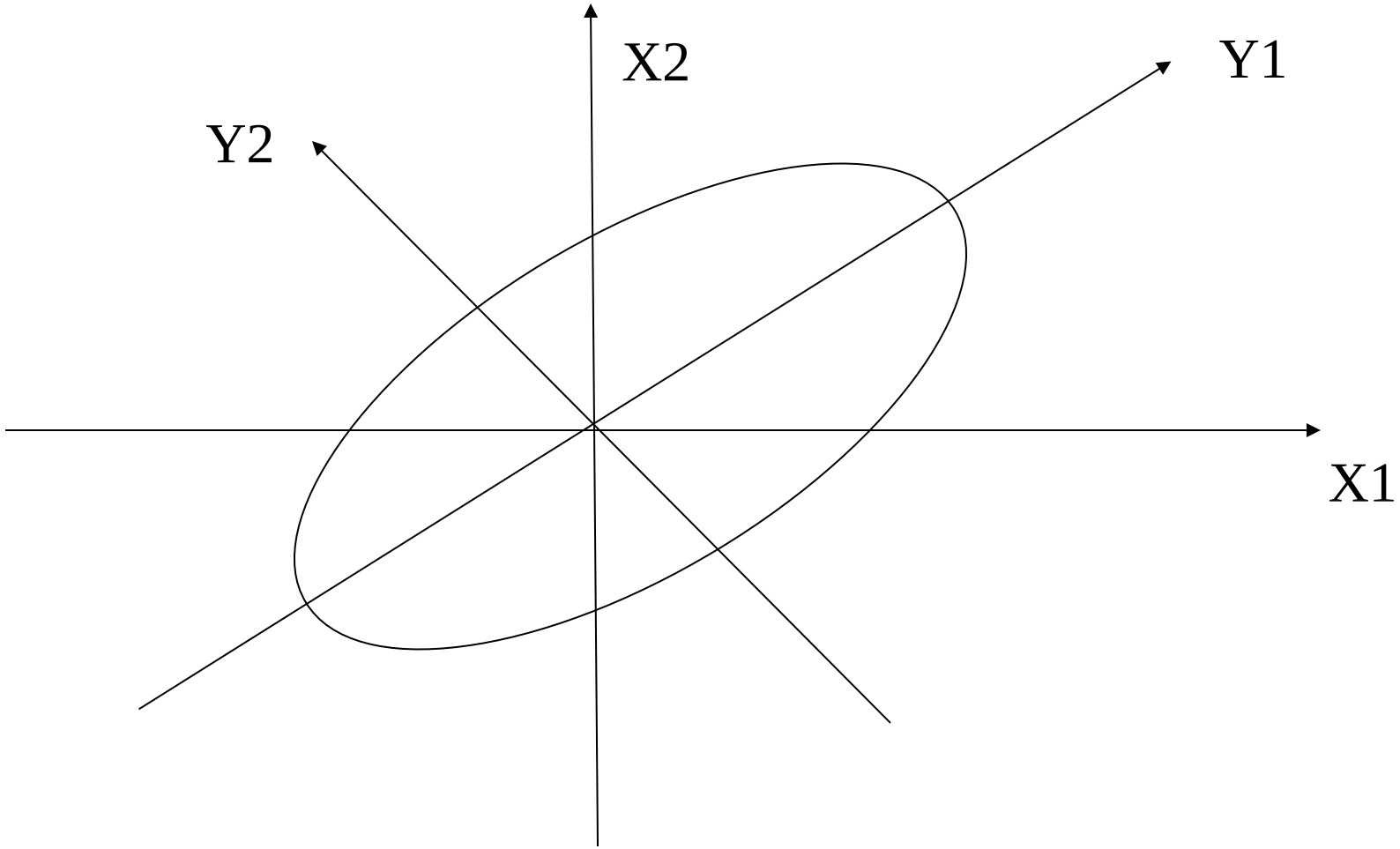
Ανάλυση Βασικών Συνιστωσών (Principal Component Analysis, PCA), Μέθοδος Karhunen-Loeve (K-L)

- Δεδομένων *N* διανυσμάτων δεδομένων *k*-διαστάσεων, βρες $c <= k$ ορθογώνια διανύσματα τα οποία μπορούν να χρησιμοποιηθούν για τη βέλτιστη αναπαράσταση των δεδομένων
  - Το αρχικό σύνολο δεδομένων μειώνεται (προβάλλεται) σε ένα νέο το οποίο αποτελείται από N διανυσμάτα δεδομένων πάνω σε c βασικές συνιστώσες (μειωμένες διαστάσεις)
- Κάθε διάνυσμα δεδομένων είναι γραμμικός συνδυασμός των *c* διανυσμάτων βασικών συνιστωσών
- Δουλεύει για διατεταγμένα και μη διατεταγμένα χαρακτηριστικά
- Χρησιμοποιείται όταν ο αριθμός των διαστάσεων είναι μεγάλος

Ανάλυση Βασικών Συνιστωσών

• Η ανάλυση σε συνιστώσες (νέο σύνολο αξόνων) προσφέρει σημαντική πληροφορία σχετικά με τη διασπορά

• Χρησιμοποιώντας τις ισχυρότερες συνιστώσες, μπορούμε να ανακατασκευάσουμε με μια καλή προσέγγιση το αρχικό σήμα



Μείωση Πολυαριθμίας

- Παραμετρικές μέθοδοι
  - Υποθέτουν ότι τα δεδομένα ταιριάζουν σε ένα μοντέλο, υπολογίζουν τις παραμέτρους του μοντέλου, αποθηκεύουν μόνο τις παραμέτρους, και ΟΧΙ ΤΑ ΙΔΙΑ τα δεδομένα (εκτός πιθανών outliers)
  - Π.χ.: Log-linear μοντέλα: αποκτούν τιμή σε ένα σημείο του m-Δ χώρου ως γινόμενο των αντίστοιχων ορισμένων (marginal) υποχώρων
- Μη-Παραμετρικές μέθοδοι
  - Δεν χρησιμοποιούν μοντέλα
  - Βασικές οικογένειες: ιστογράμματα, συσταδοποίηση, δειγματοληψία

Παλινδρόμηση και Log-Linear Μοντέλα

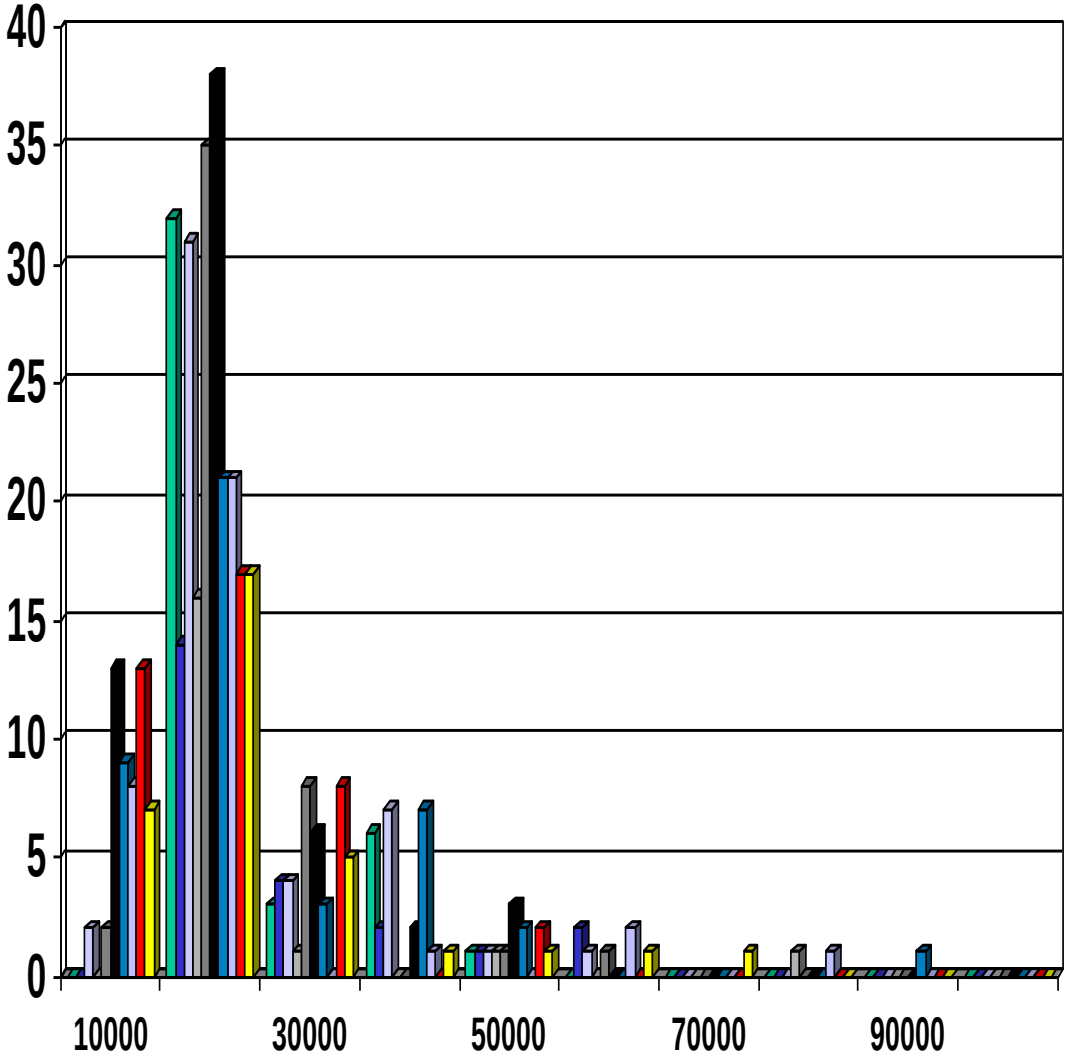
- Γραμμική παλινδρόμηση: Τα δεδομένα μοντελοποιούνται ώστε να ταιριάζουν σε μια ευθεία γραμμή:
  - Συχνά χρησιμοποιεί τη μέθοδο ελαχίστων τετραγώνων για το ταίριασμα στη γραμμή
- Πολλαπλή παλινδρόμηση: επιτρέπει μια μεταβλητή απόκρισης y να μοντελοποιηθεί ως γραμμική συνάρτηση πολυδιάστατων διανυσμάτων επιλογής (predictor variables)
- Log-linear μοντέλα: προσεγγίζουν διακριτές πολυδιάστατες κατανομές από-κοινού πιθανοτήτων

Ανάλυση Παλινδρόμησης και Log-Linear Μοντέλα

- Γραμμική παλινδρόμηση: $Y = \alpha + \beta X$
  - Δύο παράμετροι, α και β καθορίζουν τη γραμμή και υπολογίζονται με βάση τα διαθέσιμα δεδομένα
  - Χρησιμοποιώντας το κριτήριο ελαχίστων τετραγώνων στις γνωστές τιμές των $Y1, Y2, \ldots, X1, X2, \ldots$.
- Πολλαπλή παλινδρόμηση: $Y = b0 + b1\ X1 + b2\ X2$
  - Αρκετές μη γραμμικές συναρτήσεις μπορούν να μετασχηματιστούν στην παραπάνω
- Log-linear μοντέλα:
  - Ο multi-way πίνακας των από-κοινού πιθανοτήτων προσεγγίζεται με το γινόμενο των πινάκων χαμηλότερης τάξης
  - Πιθανότητα: $p(a, b, c, d) = \alpha ab\ \beta ac \chi ad\ \delta bcd$

Ιστογράμματα

- Προσεγγίζουν κατανομές δεδομένων
- Χωρίζουν τα δεδομένα σε buckets και αποθηκεύουν το μέσο όρο (ή άθροισμα) για κάθε bucket
- Ένα bucket αναπαριστά ένα ζευγάρι τιμής χαρακτηριστικού /συχνότητας
- Μπορεί να κατασκευαστεί βέλτιστα σε μια διάσταση με χρήση δυναμικού προγραμματισμού
- Σχετίζεται με προβλήματα ποσοτικοποίησης



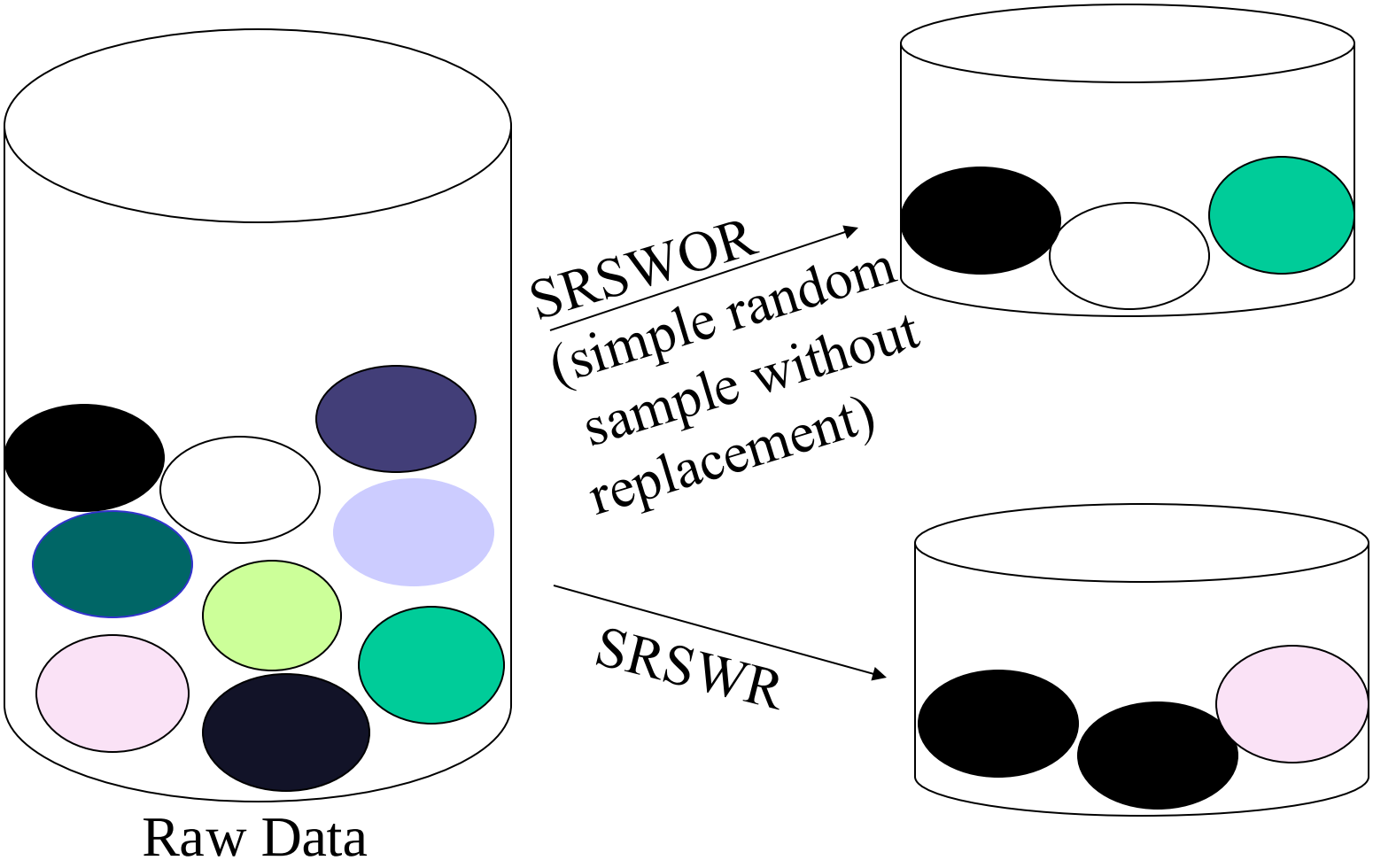
Συσταδοποίηση

- Διαχωρισμός των δεδομένων σε συστάδες, και αποθήκευση μόνο των αναπαραστάσεων των συστάδων
- Η ποιότητα των συστάδων ποσοτικοποιείται με τη διάμετρό τους (μέγιστη απόσταση μεταξύ δύο οποιωνδήποτε αντικειμένων της συστάδας) ή με την κεντροειδή απόσταση (μέση απόσταση κάθε αντικειμένου της συστάδας από το κεντροειδές της)
- Αρκετά αποδοτική αν τα δεδομένα είναι συσταδοποιημένα αλλά όχι αν τα δεδομένα είναι “smeared”
- Μπορεί να έχει ιεραρχική συσταδοποίηση (πιθανώς αποθηκευμένη σε πολυδιάστατες δενδρικές δομές δεικτοδότησης (B+-δένδρο, R- δένδρο, quad-δένδρο, κτλ))
- Υπάρχουν αρκετές επιλογές για τον ορισμό της συσταδοποίησης και για αλγορίθμους συσταδοποίησης (λεπτομέρειες στη συνέχεια)

Δειγματοληψία

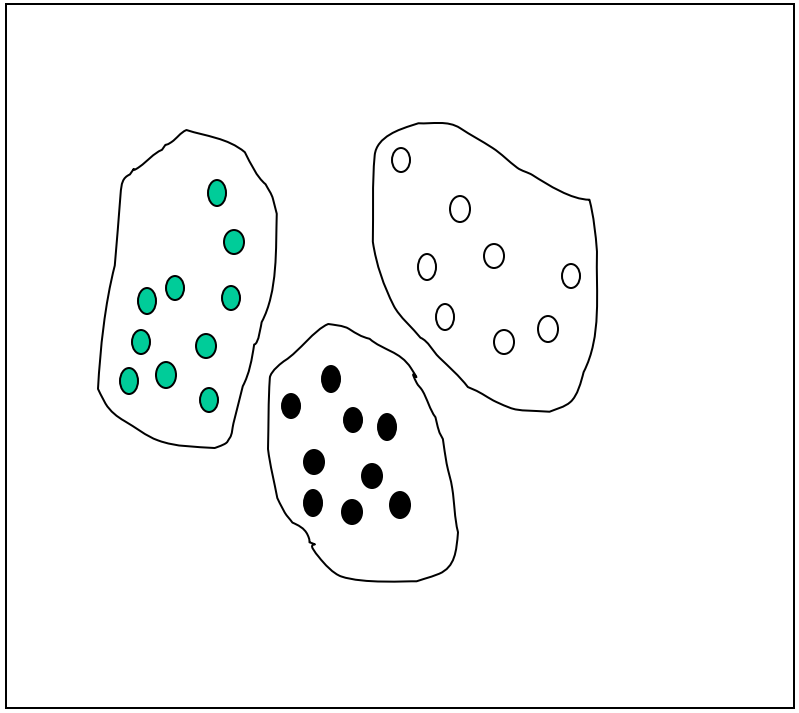
- Επιτρέπει σε ένα αλγόριθμο εξόρυξης να τρέξει σε πολυπλοκότητα η οποία είναι πιθανώς υπό-γραμμική (sublinear) ως προς το μέγεθος των δεδομένων
- Κόστος δειγματοληψίας: ανάλογο με το μέγεθος του δείγματος, αυξάνεται γραμμικά με το πλήθος των διαστάσεων
- Επιλογή ενός αντιπροσωπευτικού υποσυνόλου των δεδομένων
  - Η απλή, τυχαία δειγματοληψία μπορεί να έχει χαμηλή απόδοση όταν υπάρχουν skews
- Ανάπτυξη προσαρμοζόμενων μεθόδων δειγματοληψίας
  - Stratified δειγματοληψία:
    - Προσέγγισε το ποσοστό κάθε κλάσης (ή υποπληθυσμών ενδιαφέροντος) στα συνολικά δεδομένα
    - Χρησιμοποιείται σε συνδυασμό με skewed δεδομένα
- Η δειγματοληψία μπορεί να μειώσει τα I/Os στη βάση δεδομένων
- Δειγματοληψία: φυσική επιλογή για προοδευτική βελτίωση ενός ελαττωμένου συνόλου δεδομένων

Δειγματοληψία

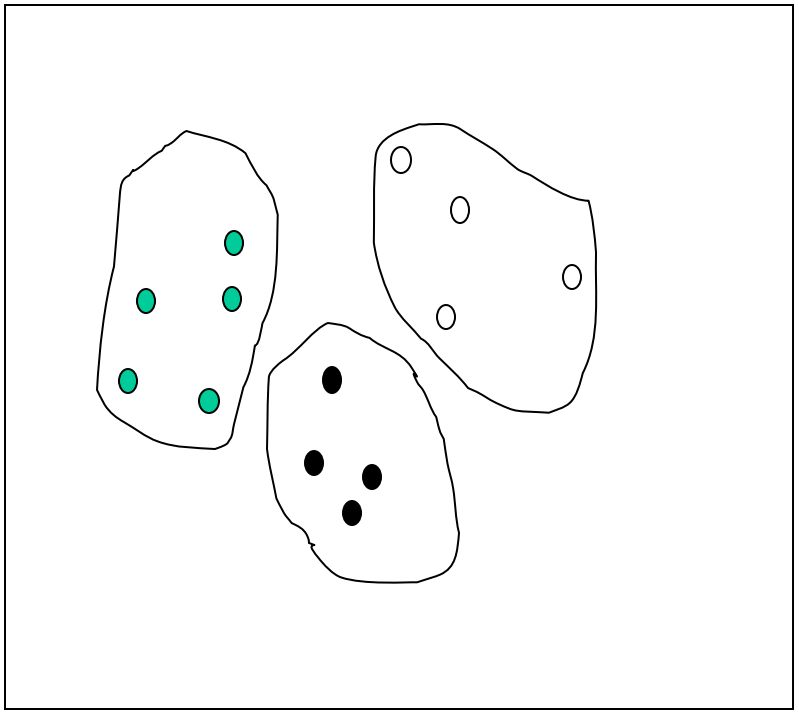


Δειγματοληψία

Ανεπεξέργαστα Δεδομένα



Συστάδες/Stratified Δείγμα



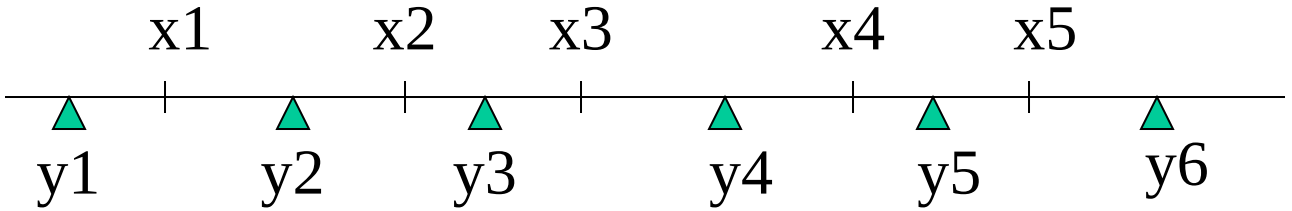
Ιεραρχική Μείωση

- Χρήση δομής πολυ-ανάλυσης με διαφορετικούς βαθμούς μείωσης
- Η ιεραρχική συσταδοποίηση συχνά εφαρμόζεται αλλά τείνει να ορίζει διαχωρισμούς των συνόλων δεδομένων παρά “συστάδες”
- Παραμετρικές μέθοδοι δεν υπόκεινται συνήθως στην ιεραρχική συσταδοποίηση
- Ιεραρχική συνάθροιση
  - Ένας δενδρικός δείκτης χωρίζει ιεραρχικά ένα σύνολο δεδομένων σε διαμερισμούς με βάση το εύρος τιμών των χαρακτηριστικών
  - Κάθε διαχωρισμός μπορεί να θεωρηθεί ως ένα bucket
  - Συνεπώς, ένας δενδρικός δείκτης με συναθροίσεις που αποθηκεύεται σε κάθε κόμβο είναι ένα ιεραρχικό ιστόγραμμα

Διακριτοποίηση/Ποσοτικοποίηση

- Τρεις τύποι γνωρισμάτων:
  - Ονομαστικός— τιμές από ένα μη διατεταγμένο σύνολο
  - Διατακτικός— τιμές από ένα διατεταγμένο σύνολο
  - Συνεχής— πραγματικοί αριθμοί
- Διακριτοποίηση/Ποσοτικοποίηση:

  Διαίρεσε το εύρος ενός συνεχούς χαρακτηριστικού σε διαστήματα



  - Κάποιοι αλγόριθμοι κατηγοριοποίησης δέχονται μόνο κατηγορικά χαρακτηριστικά
  - Μείωση μεγέθους δεδομένων με διακριτοποίηση
  - Προετοιμασία για επιπλέον ανάλυση

Διακριτοποίηση και Ιεραρχική Λογική

- Διακριτοποίηση
  - Μείωση του αριθμού των τιμών για ένα δεδομένο συνεχές χαρακτηριστικό με διαχωρισμό του εύρους του χαρακτηριστικού σε διαστήματα.
  - Οι ετικέτες των διαστημάτων μπορούν να χρησιμοποιηθούν για να αντικαταστήσουν τις πραγματικές τιμές των δεδομένων
- Ιεραρχίες Εννοιών
  - Μείωση δεδομένων με συλλογή και αντικατάσταση εννοιών χαμηλού επιπέδου (όπως αριθμητικές τιμές για το χαρακτηριστικό ηλικία) με έννοιες υψηλότερου επιπέδου (όπως νέος, μεσήλικας, ή ηλικιωμένος)

Διακριτοποίηση και Παραγωγή Ιεραρχίας Εννοιών για Αριθμητικά Δεδομένα

- Ιεραρχική και αναδρομική ανάλυση με χρήση:
  - Binning (εξομάλυνση δεδομένων)
  - Ανάλυση Ιστογραμμάτων (μείωση πολυαριθμίας)
  - Ανάλυση Συσταδοποίησης (μείωση πολυαριθμίας)
- Διακριτοποίηση βασισμένη στην Εντροπία
- Κατάτμηση με φυσικό διαχωρισμό

Διακριτοποίηση βασισμένη στην Εντροπία

- Δεδομένου ενός συνόλου δειγμάτων S, αν το S διαχωρίζεται σε δύο διαστήματα S1 και S2 χρησιμοποιώντας το κατώφλι T στις τιμές του χαρακτηριστικού A, το κέρδος πληροφορίας που προκύπτει από το διαχωρισμό είναι:

$$I(S,T) = \frac{|S_1|}{|S|} E(S_1) + \frac{|S_2|}{|S|} E(S_2)$$

όπου η συνάρτηση εντροπίας E για ένα δεδομένο σύνολο υπολογίζεται με βάση την κατανομή κλάσης των δειγμάτων στο σύνολο. Δεδομένων m κλάσεων η εντροπία του S1 είναι: $E(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$

όπου pi είναι η πιθανότητα της κλάσης i στο S1.

- Το κατώφλι T που μεγιστοποιεί το κέρδος πληροφορίας σε σχέση με όλα τα πιθανά κατώφλια, επιλέγεται ως δυαδική διακριτοποίηση
- Η διαδικασία εφαρμόζεται αναδρομικά σε διαχωρισμούς μέχρι να ικανοποιηθεί ένα κριτήριο τερματισμού, π.χ., $E(S) - I(S,T) < \delta$
- Συμπέρασμα από πειραματικές μελέτες:
  - η διακριτοποίηση μπορεί να μειώσει το μέγεθος των δεδομένων και να βελτιώσει την ακρίβεια της κατηγοριοποίησης

Σύνοψη

- Η προετοιμασία των δεδομένων είναι ένα σημαντικό ζήτημα τόσο για την αποθήκευση όσο και για την εξόρυξη (60% της προσπάθειας)
- Η προετοιμασία των δεδομένων περιλαμβάνει:
  - Καθαρισμό και ενοποίηση των δεδομένων
  - Μείωση των δεδομένων και επιλογή χαρακτηριστικών
  - Διακριτοποίηση
- Ένα πλήθος μεθόδων έχει αναπτυχθεί αλλά η προεπεξεργασία δεδομένων παραμένει μια ανοιχτή ερευνητική περιοχή

Μηχανική Μάθηση

Κατηγοριοποίηση: Δέντρα Απόφασης

(εν μέρη βασισμένο σε σημειώσεις του C. Faloutsos)

Περισσότερες Λεπτομέρειες… Δένδρα Απόφασης

- Δέντρα Απόφασης
  - Πρόβλημα
  - Προσέγγιση

- Κατηγοριοποίηση μέσω δένδρων
- Φάση Κατασκευής – Πολιτικές Διαχωρισμού
- Φάση Κλάδευσης (για αποφυγή over-fitting)

Δένδρα Απόφασης

- Πρόβλημα: Κατηγοριοποίηση– δηλ., δεδομένου ενός συνόλου εκπαίδευσης (Ν πλειάδες, με Μ χαρακτηριστικά, και ενός επιπλέον χαρακτηριστικού-κλάσης) βρες κανόνες για να προβλέψεις την κλάση νέων δειγμάτων

Σχηματικά:

Δένδρα Απόφασης

Age	Chol-level	Gender	…	CLASS-ID
30	150	M		+
				…
				-
				??

Δένδρα Απόφασης

- Θέματα:
  - Ελλειπείς τιμές
  - θόρυβος
  - ‘σπάνια’ γεγονότα

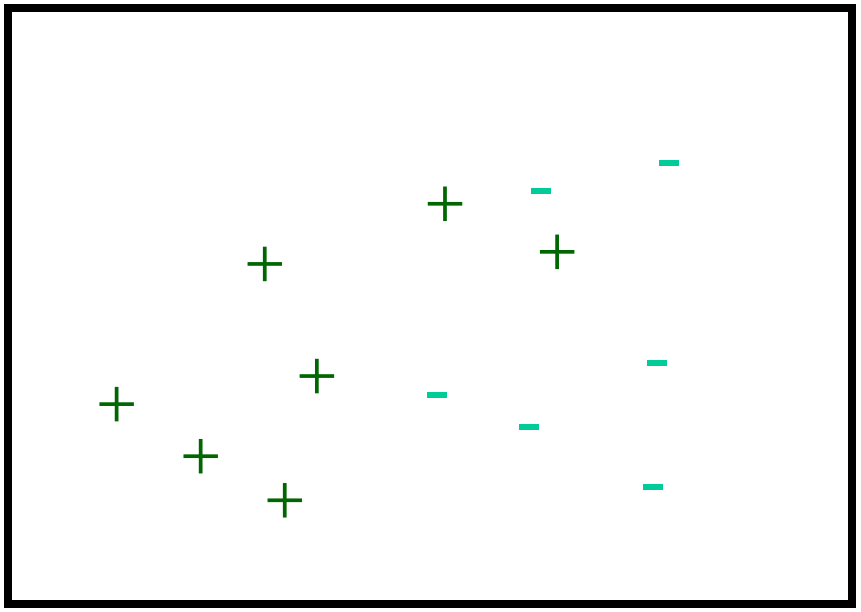
Δένδρα Απόφασης

- Τύποι χαρακτηριστικών:
  - Αριθμητικός (= συνεχής) – π.χ.: ‘μισθός’
  - Διατακτικός (= ακέραιος) – π.χ.: ‘# τέκνων’
  - Ονομαστικός (= κατηγορικός) – π.χ.: ‘μάρκα αυτοκινήτου’

Δένδρα Απόφασης

- Σχηματικά, έχουμε:

num. attr#2
(e.g., chol-level)

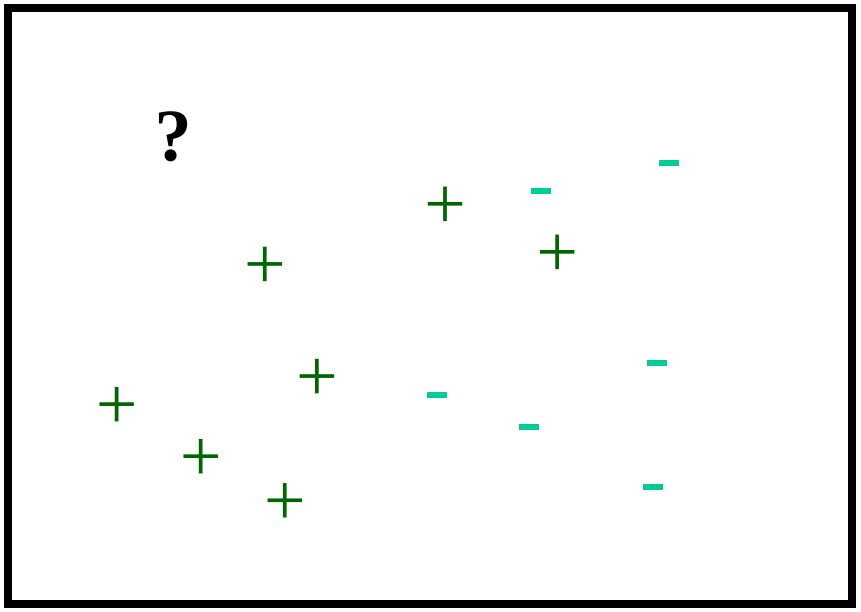


num. attr#1 (e.g., ‘age’)

Δένδρα Απόφασης

- Και θέλουμε την κλάση του '**?**'

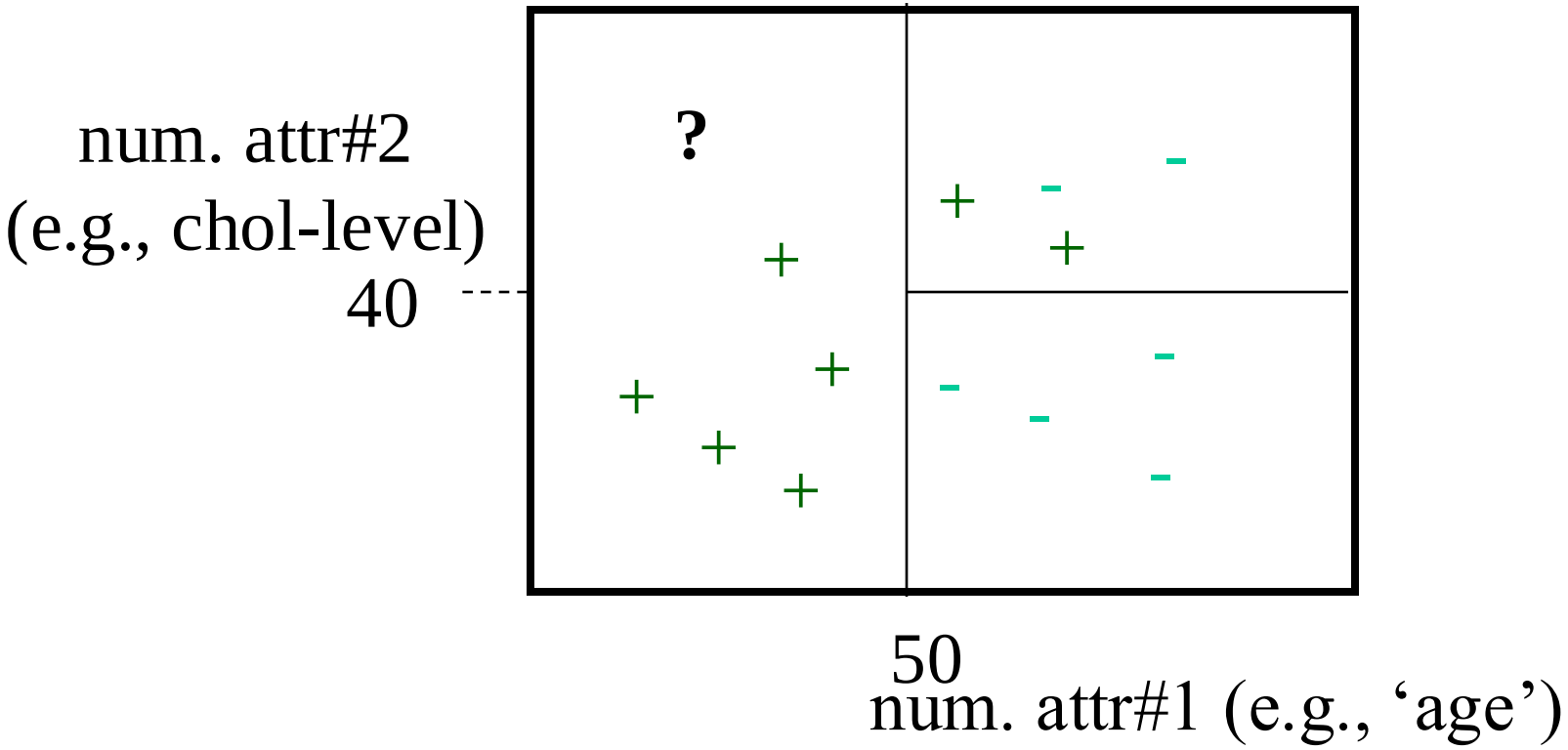
num. attr#2
(e.g., chol-level)



num. attr#1 (e.g., 'age')

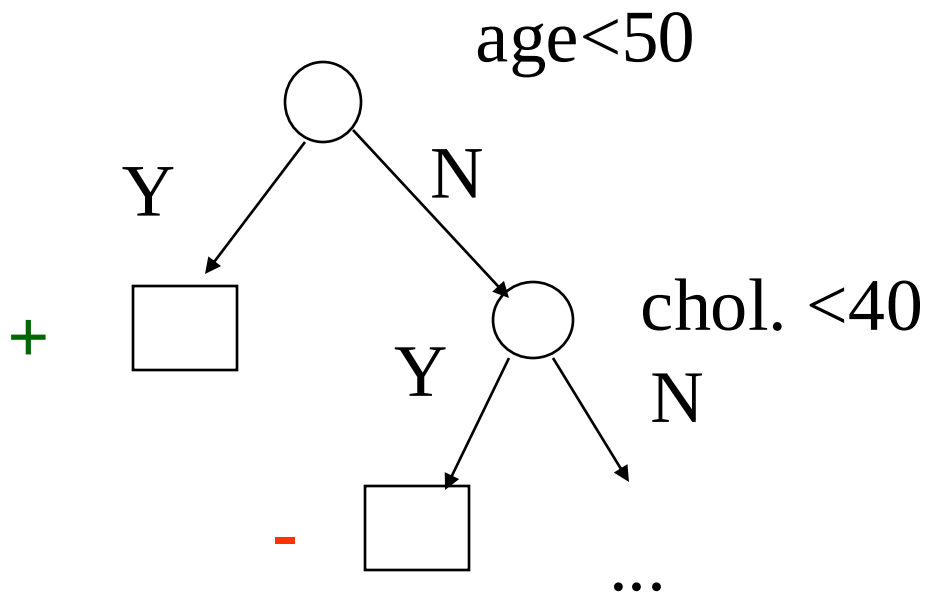
Δένδρα Απόφασης

- Άρα κατασκευάζουμε ένα δένδρο απόφασης:



Δένδρα Απόφασης

- Η κατασκευή του δένδρου απόφασης:

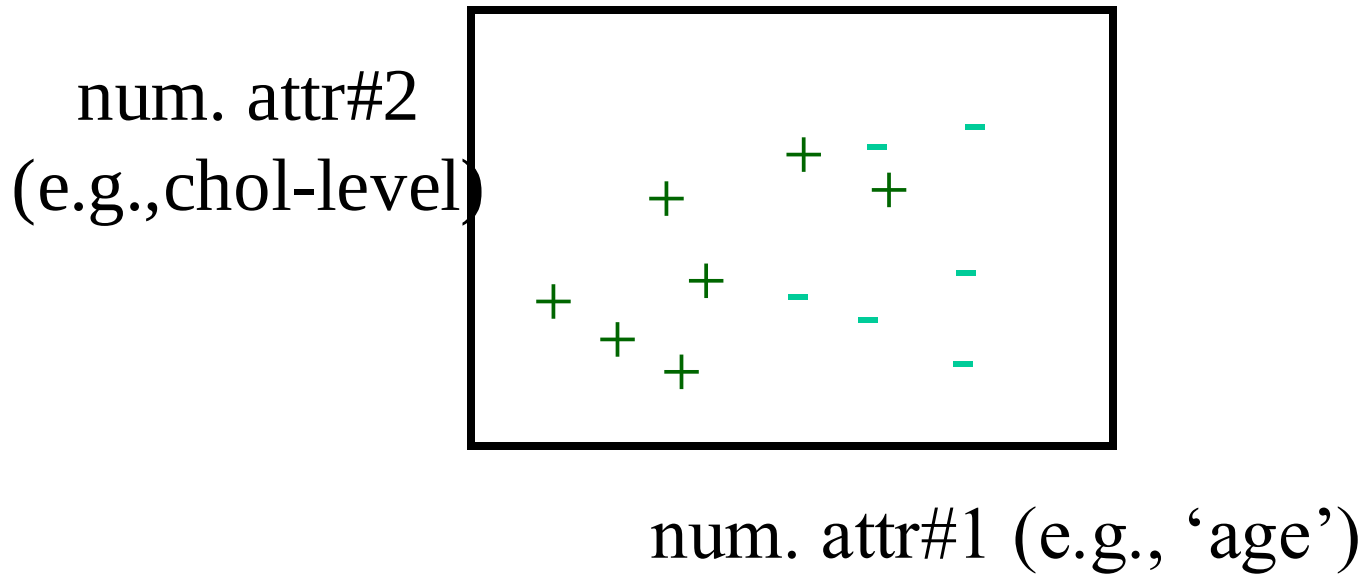


Δένδρα Απόφασης

- Συνήθως, δύο βήματα:
  - Κατασκευή δένδρου
  - Κλάδευση δένδρου (για αποφυγή over-training/over-fitting)

Κατασκευή Δένδρου

- Πως;



Κατασκευή Δένδρου

- Πως;
- Απ.: Διαχωρισμός, αναδρομικά
- Ψευδοκώδικας:

  Partition (dataset S)

  if all points in S have same label

  then return

  evaluate splits along each attribute A

  pick best split, to divide S into S1 and S2

  Partition(S1); Partition(S2)
- Ερ. 1: πως διαχωρίζουμε το εύρος τιμών ενός χαρακτηριστικού A_i ?
- Ερ. 2: πως αξιολογούμε τον διαχωρισμό?

Κατασκευή Δένδρου

- Ερ. 1: πως διαχωρίζουμε το εύρος τιμών ενός χαρακτηριστικού A_i ?
- Απ.:
  - Για αριθμητικά χαρακτηριστικά:
    - Δυαδικός διαχωρισμός, ή
    - Πολλαπλός διαχωρισμός
  - Για κατηγορικά χαρακτηριστικά :
    - Υπολογισμός όλων των υποσυνόλων (ακριβός!), ή
    - Χρήση μιας άπληστης (greedy) προσέγγισης

Κατασκευή Δένδρου

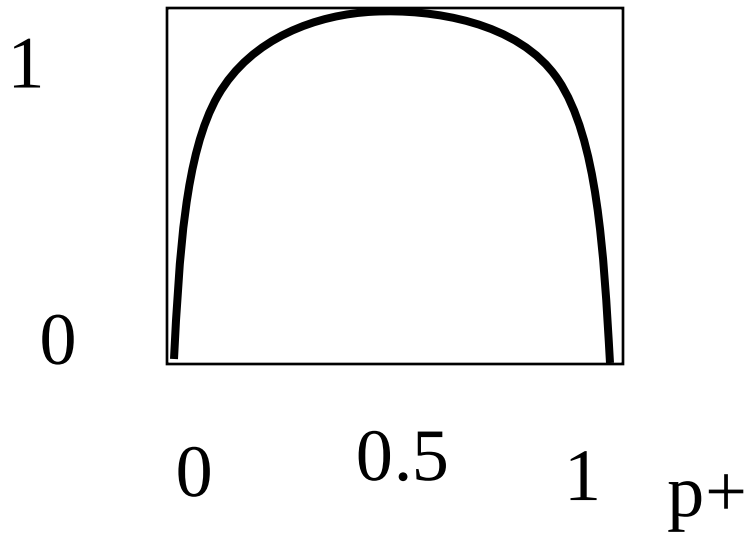
- Ερ. 2: Πως αξιολογούμε τον διαχωρισμό?
- Απ. : Από το πόσο κοντά στο κανονικό (uniform) είναι κάθε υποσύνολο - πχ., χρειαζόμαστε μια μετρική της ανομοιομορφίας:

Κατασκευή Δένδρου

Εντροπία: H(p+, p-)

Υπάρχουν άλλες μετρικές;

p+: σχετική συχνότητα της κλάσης + στο S

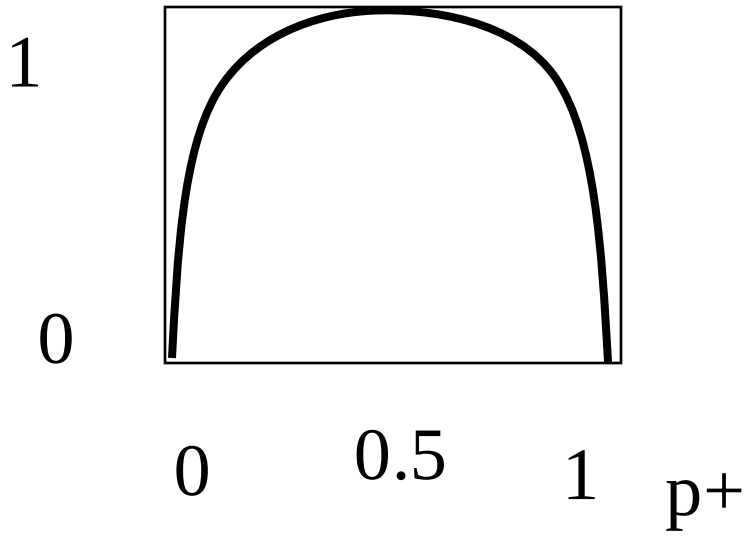


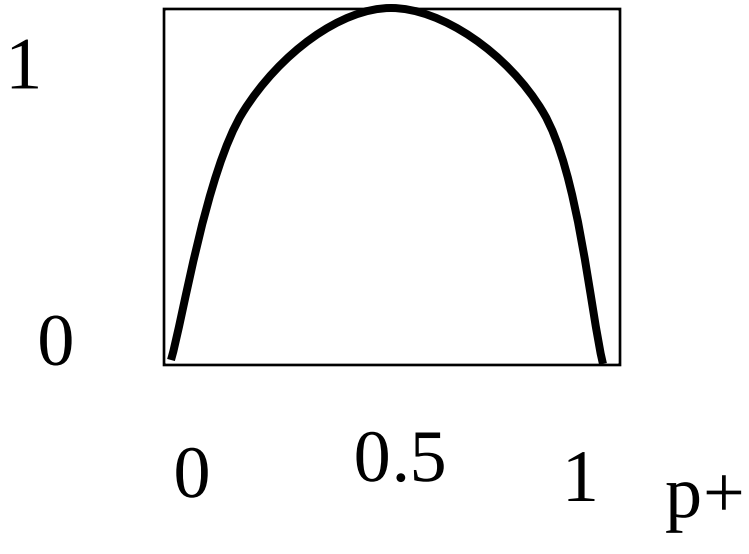
Κατασκευή Δένδρου

Εντροπία : $H(p_+, p_-)$

Δείκτης ‘gini’ : $1-p_+^2 - p_-^2$

p+: σχετική συχνότητα της κλάσης + στο S





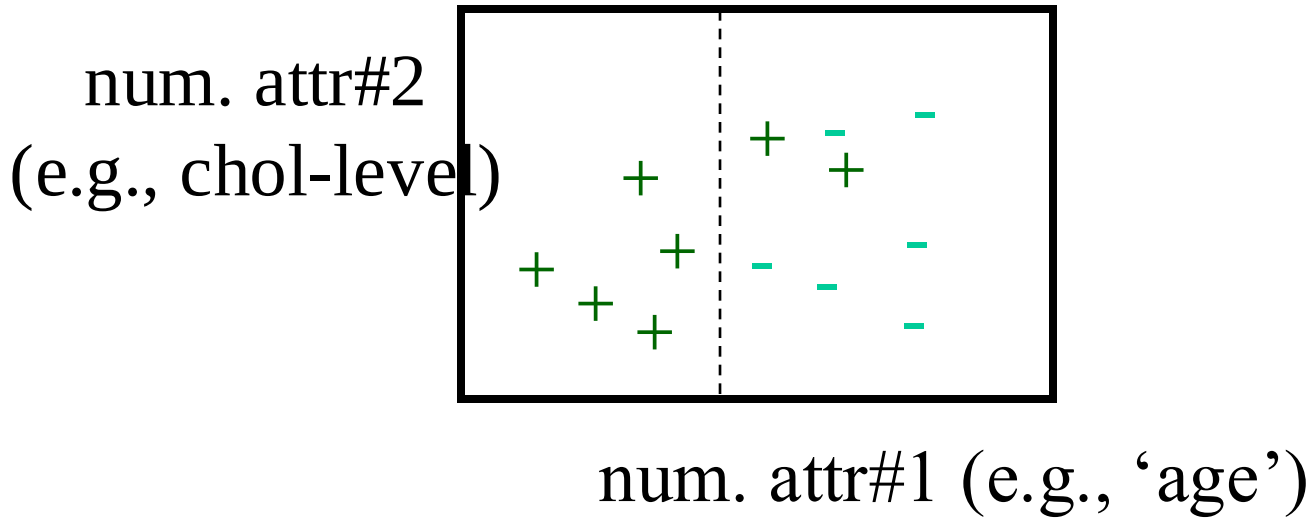
Κατασκευή Δένδρου

Διαισθητικά...:

- Εντροπία: #bits για την κωδικοποίηση της ετικέτας κλάσης
- gini: σφάλμα κατηγοριοποίησης, αν τυχαία μαντέψουμε ‘+’ με πιθανότητα p_+

Κατασκευή Δένδρου

Συνεπώς, επιλέγουμε το διαχωρισμό που μειώνει κατά το μέγιστο την εντροπία/σφάλμα-κατηγοριοποίησης: Π.χ.:



Κατασκευή Δένδρου

- Πριν τον διαχωρισμό χρειαζόμαστε:

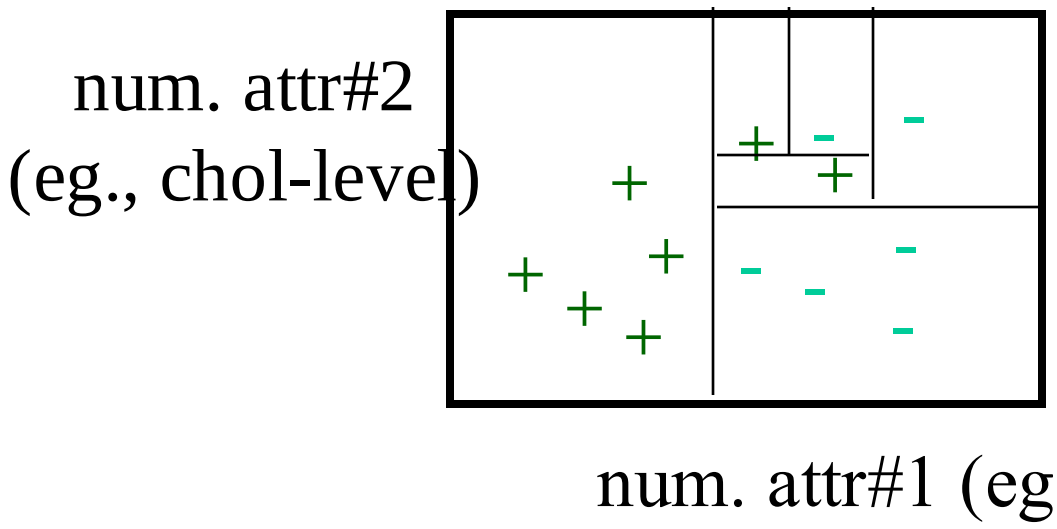
  $(n_+ + n_-) * H(p_+, p_-) = (7+6) * H(7/13, 6/13)$

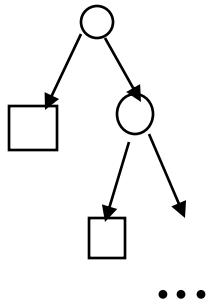
  bits συνολικά, για την κωδικοποίηση όλων των ετικετών κλάσης
- Μετά τον διαχωρισμό χρειαζόμαστε:

  0 bits για το πρώτο μισό και (2+6) * H(2/8, 6/8) bits για το δεύτερο μισό

Κλάδευση Δένδρου

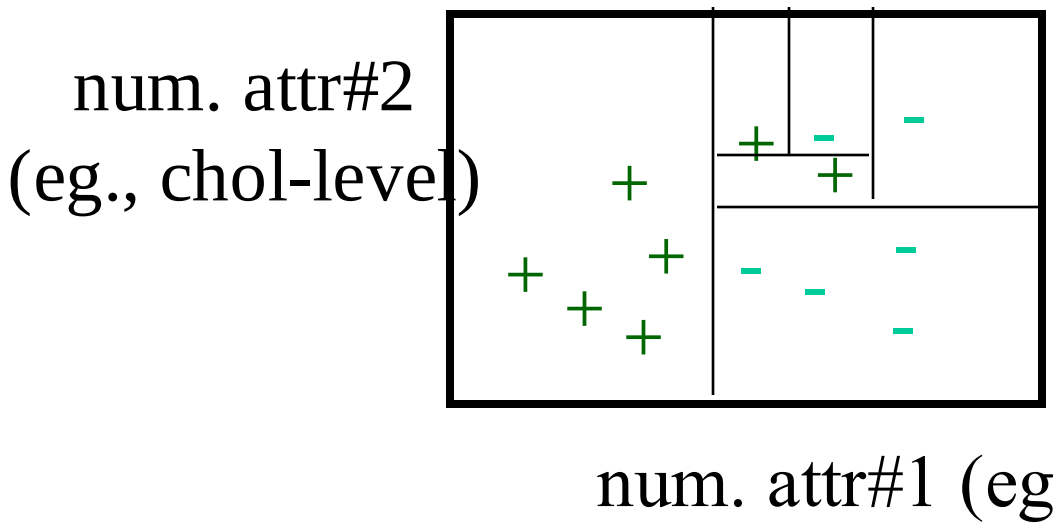
- Για ποιο λόγο;

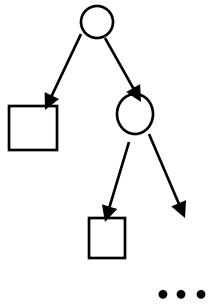




Κλάδευση Δένδρου

- Ερ.: Πως πραγματοποιείται;





Κλάδευση Δένδρου

- Ερ.: Πως πραγματοποιείται?
- Απ.1: χρήση ενός συνόλου ‘εκπαίδευσης’ και ενός συνόλου ‘ελέγχου’ – κλάδευση των κόμβων που βελτιώνουν την κατηγοριοποίηση στο σύνολο ‘εκπαίδευσης’. (Μειονεκτήματα;)

Κλάδευση Δένδρου

- Ερ.: Πως πραγματοποιείται?
- Απ.1: χρήση ενός συνόλου ‘εκπαίδευσης’ και ενός συνόλου ‘ελέγχου’ – κλάδευση των κόμβων που βελτιώνουν την κατηγοριοποίηση στο σύνολο ‘εκπαίδευσης’. (Μειονεκτήματα;)
- Απ.2: ή, βασιζόμαστε στον MDL (= Minimum Description Language) – πιο λεπτομερώς:

Κλάδευση Δένδρου

- Αντιμετωπίζουμε το πρόβλημα ως πρόβλημα συμπίεσης
- ...και προσπαθούμε να ελαχιστοποιήσουμε το # των bits για τη συμπίεση

  (a) των ετικετών των κλάσεων ΚΑΙ

  (b) της αναπαράστασης του δένδρου απόφασης

Μηχανική Μάθηση

Κατηγοριοποίηση και Πρόβλεψη

(εν μέρη βασισμένο σε σημειώσεις των J. Han και M. Kamber)

Κατηγοριοποίηση και Πρόβλεψη

- Κατηγοριοποίηση:
  - Προβλέπει κατηγορικές ετικέτες κλάσης
  - Κατηγοριοποιεί δεδομένα (κατασκευάζει μοντέλο) χρησιμοποιώντας το σύνολο εκπαίδευσης και τις τιμές (ετικέτες κλάσης) του προς κατηγοριοποίηση χαρακτηριστικού και με βάση αυτά κατηγοριοποιεί τα νέα δεδομένα
- Πρόβλεψη:
  - Μοντελοποιεί συνεχείς συναρτήσεις, π.χ. προβλέπει άγνωστες ή χαμένες τιμές
- Βασικές Εφαρμογές:
  - Πιστοληπτική απόφαση
  - Εύρεση στοχευόμενου κοινού (target marketing)
  - Ιατρική διάγνωση
  - Ανάλυση απόδοσης διαχείρισης
- Μεγάλα σύνολα δεδομένων: αποθήκευση στο δίσκο αντί αποθήκευσης στην κύρια μνήμη

Κατηγοριοποίηση —Μία διαδικασία δύο βημάτων

- Κατασκευή Μοντέλου: η περιγραφή ενός συνόλου προκαθορισμένων κλάσεων
  - Κάθε πλειάδα θεωρείται ότι ανήκει σε μια προκαθορισμένη κλάση, που καθορίζεται από το χαρακτηριστικό της ετικέτας κλάσης (επιβλεπόμενη μάθηση – supervised learning)
  - Σύνολο εκπαίδευσης (training set): το σύνολο των πλειάδων που χρησιμοποιείται για την κατασκευή του μοντέλου
  - Το μοντέλο αναπαρίστανται ως κανόνες κατηγοριοποίησης, δέντρα απόφασης ή μαθηματικοί τύποι
- Εφαρμογή Μοντέλου: για την κατηγοριοποίηση νέων αντικειμένων
  - Αξιολόγηση της απόδοσης του μοντέλου χρησιμοποιώντας ένα σύνολο ελέγχου (test set)
    - Η γνωστή κλάση του δείγματος ελέγχου συγκρίνεται με το αποτέλεσμα της κατηγοριοποίησης
    - Ακρίβεια (accuracy) είναι το ποσοστό των δειγμάτων ελέγχου που κατηγοριοποιήθηκαν ορθά από το μοντέλο
    - Το σύνολο ελέγχου πρέπει να είναι ανεξάρτητο από το σύνολο εκπαίδευσης για αποφυγή over-fitting

Επιβλεπόμενη και Μη Επιβλεπόμενη Μάθηση

- Επιβλεπόμενη μάθηση (κατηγοριοποίηση)
  - Επίβλεψη: Τα δεδομένα εκπαίδευσης (παρατηρήσεις, μετρήσεις, κ.α.) συνοδεύονται από ετικέτες που δείχνουν την κλάση τους
  - Τα νέα δεδομένα κατηγοριοποιούνται βάση του συνόλου εκπαίδευσης
- Μη επιβλεπόμενη μάθηση (συσταδοποίηση)
  - Οι ετικέτες κλάσης του συνόδου εκπαίδευσης είναι άγνωστες
  - Δοσμένου ενός συνόλου μετρήσεων, παρατηρήσεων, κτλ, ο στόχος είναι η εύρεση της ύπαρξης κλάσεων ή συστάδων μεταξύ των δεδομένων

Θεματολογία

- Τι είναι Κατηγοριοποίηση; Τι είναι Πρόβλεψη;
- Ζητήματα κατηγοριοποίησης και πρόβλεψης
- Κατηγοριοποίηση με επαγωγή δένδρου απόφασης
- Μπεϋζιανή Κατηγοριοποίηση
- Κατηγοριοποίηση με πίσω διάδοση (backpropagation)
- Κατηγοριοποίηση βασισμένη σε έννοιες από την εξόρυξη κανόνων συσχετίσεων
- Λοιπές μέθοδοι Κατηγοριοποίησης
- Πρόβλεψη
- Ακρίβεια κατηγοριοποίησης (accuracy)
- Σύνοψη

Ζητήματα κατηγοριοποίησης και πρόβλεψης: **Αξιολόγηση Μεθόδων Κατηγοριοποίησης**

- Ακρίβεια πρόβλεψης
- Ταχύτητα και κλιμάκωση
  - Χρόνος κατασκευής του μοντέλου
  - Χρόνος εφαρμογής του μοντέλου
  - Αποδοτικότητα σε βάσεις δεδομένων αποθηκευμένες στο δίσκο (disk-resident)
- Ανθεκτικότητα (robustness)
  - Διαχείριση θορύβου και χαμένων τιμών
- Ερμηνευσιμότητα (interpretability):
  - Κατανόηση και διορατικότητα που προσφέρει το μοντέλο
- Ποιότητα των κανόνων
  - Μέγεθος του δένδρου απόφασης
  - Περιεκτικότητα των κανόνων κατηγοριοποίησης

Κατηγοριοποίηση με Επαγωγή Δένδρου Απόφασης

- Βασικά χαρακτηριστικά δένδρου απόφασης (καλύφθηκαν προηγουμένως)
- Μέτρα επιλογής χαρακτηριστικών:
  - Κέρδος πληροφορίας (ID3/C4.5)
    - Όλα τα χαρακτηριστικά θεωρούνται κατηγορικά
    - Δυνατότητα τροποποίησης για συνεχή χαρακτηριστικά
  - Δείκτης Gini (IBM IntelligentMiner)
    - Όλα τα χαρακτηριστικά θεωρούνται συνεχείς μεταβλητές
    - Θεωρεί ότι υπάρχουν διάφορες πιθανές τιμές διαχωρισμού για κάθε χαρακτηριστικό
    - Η εύρεση των τιμών διαχωρισμού μπορεί να απαιτεί χρήση εργαλείων όπως π.χ. συσταδοποίηση
    - Δυνατότητα τροποποίησης για κατηγορικά χαρακτηριστικά
- Αποφυγή overfitting
- Εξαγωγή κανόνων κατηγοριοποίησης από δένδρα

Δείκτης Gini (IBM IntelligentMiner)

- Για ένα σύνολο δεδομένων T, που περιέχει παραδείγματα από n κλάσεις, ο δείκτης gini index, $gini(T)$, ορίζεται ως

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

όπου p_j είναι η σχετική συχνότητα της κλάσης j στο T.

- Για ένα σύνολο δεδομένων T, που αποτελείται από δύο υποσύνολα T_1 και T_2 με μέγεθος N_1 και N_2 αντίστοιχα, ο δείκτης *gini* των διαχωρισμένων δεδομένων περιέχει παραδείγματα από n κλάσεις και ορίζεται ως

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

- Το γνώρισμα που παρέχει το ελάχιστο $gini_{split}(T)$ επιλέγεται για τον διαχωρισμό του κόμβου (*απαιτείται η απαρίθμηση όλων των πιθανών σημείων διαχωρισμού για κάθε γνώρισμα*)

Προσεγγίσεις για τον Καθορισμό του Τελικού Μεγέθους του Δένδρου

- Διαχωρισμός συνόλου εκπαίδευσης (2/3) και συνόλου ελέγχου (1/3)
- Χρήση σταυρωτής επικύρωσης (cross validation), π.χ., 10-fold cross validation
- Χρήση όλων των δεδομένων για εκπαίδευση
  - αλλά εφαρμογή ενός στατιστικού test (π.χ., chi-square) για την εκτίμηση του αν η διεύρυνση ή η περικοπή ενός κόμβου θα βελτιώσει τη συνολική κατανομή
- Χρήση της Αρχής Ελάχιστου Μήκους Περιγραφής (Minimum Description Length (MDL) principle)
  - αναστολή της ανάπτυξης του δένδρου όταν η κωδικοποίηση ελαχιστοποιείται

Μπεϋζιανή Κατηγοριοποίηση: Γιατί;

- Πιθανοτική Μάθηση:
  - Υπολόγισε άμεσα τις πιθανότητες για την υπόθεση
  - Μεταξύ των πλέον πρακτικών μεθόδων για συγκεκριμένους τύπους προβλημάτων μάθησης
- Επαυξησιμότητα:
  - Κάθε παράδειγμα εκπαίδευσης μπορεί επαυξητικά να αυξήσει/μειώσει την πιθανότητα της ορθότητας της υπόθεσης
  - Προηγούμενη γνώση μπορεί να συνδυαστεί με τα παρατηρημένα δεδομένα
- Πιθανοτική πρόβλεψη:
  - Προβλέπει πολλαπλές υποθέσεις, σταθμισμένες με βάση τις πιθανότητές τους
- Τυπικά:
  - Ακόμα και αν οι Μπεϋζιανές μέθοδοι είναι υπολογιστικά intractable, μπορούν να παρέχουν ένα standard επίπεδο βέλτιστης λήψης αποφάσεων σε σύγκριση με άλλες μεθόδους που μπορούν να μετρηθούν

Μπεϋζιανό Θεώρημα

- Δοσμένου ενός σύνολο εκπαίδευσης *D*, *η εκ των υστέρων πιθανότητα (posteriori probability) της υπόθεσης h*, P(h|D) ακολουθεί το θεώρημα *του* Bayes:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- MAP (maximum posteriori) υπόθεση:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Πρακτικές δυσκολίες:
  - Απαιτούν αρχική γνώση αρκετών πιθανοτήτων
  - Σημαντικό υπολογιστικό κόστος

Αφελής (Naive) Μπεϋζιανός Κατηγοριοποιητής

- Απλοποιημένη υπόθεση: τα χαρακτηριστικά είναι υπό-συνθήκη ανεξάρτητα:

$$P(C_j|V) \propto P(C_j)\prod_{i=1}^{n} P(v_i|C_j)$$

όπου V είναι απλά δείγματα, v_i είναι η τιμή του χαρακτηριστικού i στο δείγμα και C_j είναι η j-οστή κλάση

- Μειώνει σημαντικά το υπολογιστικό κόστος, υπολογίζοντας μόνο την κατανομή κλάσης

Αφελής (Naive) Μπεϋζιανός Κατηγοριοποιητής

- Δοσμένου ενός συνόλου εκπαίδευσης, μπορούμε να υπολογίσουμε τις πιθανότητες

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

Μπεϋζιανή Κατηγοριοποίηση

- Το πρόβλημα της κατηγοριοποίησης μπορεί να διατυπωθεί χρησιμοποιώντας εκ των υστέρων πιθανότητες:
- P(C|X) = πιθανότητα ότι το δείγμα X=<x_1,…,x_k> ανήκει στην κλάση C
- Π.χ. P(class=N | outlook=sunny,windy=true,…)
- Ιδέα: ανάθεσε στο δείγμα X την κλάση C ώστε η πιθανότητα P(C|X) είναι μέγιστη

Υπολογισμός των εκ των υστέρων πιθανοτήτων

- Μπεϋζιανό θεώρημα:

$$P(C|X) = P(X|C)\cdot P(C) / P(X)$$

- P(X) είναι σταθερή για κάθε κλάση
- P(C) = σχετική συχνότητα των δειγμάτων της κλάσης C
- Επιλογή του C έτσι ώστε P(C|X) να μεγιστοποιείται = C έτσι ώστε P(X|C)·P(C) να μεγιστοποιείται
- Πρόβλημα: ο υπολογισμός της P(X|C) είναι ανέφικτος!

Αφελής Μπεϋζιανή Κατηγοριοποίηση

- Αφελής υπόθεση: ανεξαρτησία των χαρακτηριστικών

$$P(x_1,\ldots,x_k|C) = P(x_1|C)\cdot\ldots\cdot P(x_k|C)$$

- Αν το i-οστό χαρακτηριστικό είναι κατηγορικό: $P(x_i|C)$ υπολογίζεται ως η σχετική συχνότητα των δειγμάτων που έχουν την τιμή x_i ως το i-οστό χαρακτηριστικό στην κλάση C
- Αν το i-οστό χαρακτηριστικό είναι συνεχές: $P(x_i|C)$ υπολογίζεται μέσω μιας Γκαουσιανής συνάρτησης πυκνότητας πιθανότητας
- Υπολογιστικά εύκολο, και στις δύο περιπτώσεις !!!

Play-tennis παράδειγμα: Υπολογισμός $P(x_i|C)$

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

outlook	
P(sunny|p) = 2/9	**P(sunny|n) = 3/5**
P(overcast|p) = 4/9	**P(overcast|n) = 0**
P(rain|p) = 3/9	**P(rain|n) = 2/5**
temperature	
P(hot|p) = 2/9	**P(hot|n) = 2/5**
P(mild|p) = 4/9	**P(mild|n) = 2/5**
P(cool|p) = 3/9	**P(cool|n) = 1/5**
humidity	
P(high|p) = 3/9	**P(high|n) = 4/5**
P(normal|p) = 6/9	**P(normal|n) = 1/5**
windy	
P(true|p) = 3/9	**P(true|n) = 3/5**
P(false|p) = 6/9	**P(false|n) = 2/5**

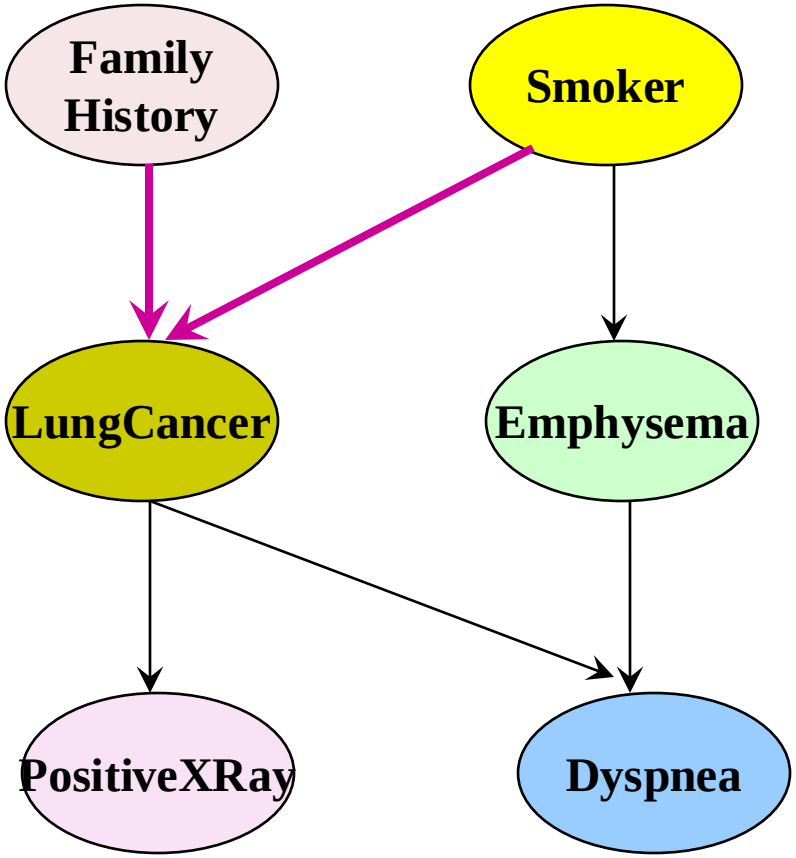
Play-tennis παράδειγμα: Κατηγοριοποίηση του X

- Ένα άγνωστο δείγμα X = <rain, hot, high, false>
- $P(X|p)\cdot P(p) = P(rain|p)\cdot P(hot|p)\cdot P(high|p)\cdot P(false|p)\cdot P(p) = 3/9\cdot 2/9\cdot 3/9\cdot 6/9\cdot 9/14 = 0.010582$
- $P(X|n)\cdot P(n) = P(rain|n)\cdot P(hot|n)\cdot P(high|n)\cdot P(false|n)\cdot P(n) = 2/5\cdot 2/5\cdot 4/5\cdot 2/5\cdot 5/14 = 0.018286$
- Το δείγμα X κατηγοριοποιείται στην κλάση n (don't play)

Η υπόθεση ανεξαρτησίας…

- … καθιστά τον υπολογισμό δυνατό
- … αποφέρει βέλτιστους κατηγοριοποιητές όταν ικανοποιείται
- … ωστόσο, σπάνια ικανοποιείται στην πράξη, καθώς τα χαρακτηριστικά (μεταβλητές) συχνά συσχετίζονται
- Προσπάθειες να υπερπηδήσουν αυτόν τον περιορισμό:
  - Μπεϋζιανά δίκτυα, τα οποία συνδυάζουν την Μπεϋζιανή λογική με αιτιατές σχέσεις μεταξύ των χαρακτηριστικών
  - Δέντρα απόφασης, τα οποία χειρίζονται ένα γνώρισμα κάθε στιγμή, ξεκινώντας από τα πλέον σημαντικά γνωρίσματα

Δίκτυα Μπεϋζιανής Λογικής



Δίκτυο Μπεϋζιανής Λογικής

	(FH, S)	(FH, ~S)	(~FH, S)	(~FH, ~S)
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

Ο πίνακας με τις υπό-συνθήκη πιθανότητες για τη μεταβλητή LungCancer

Δίκτυα Μπεϋζιανής Λογικής

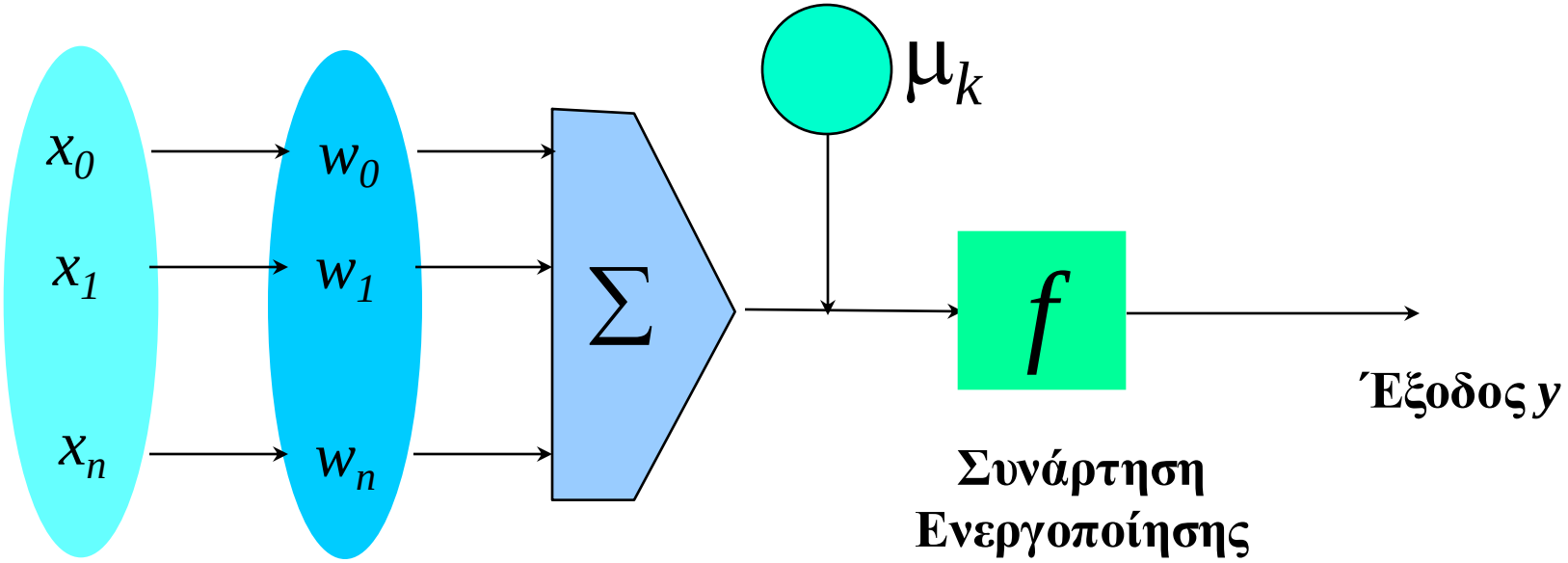
- Τα δίκτυα Μπεϋζιανής λογικής θεωρούν ότι ένα υποσύνολο των μεταβλητών είναι υπό-συνθήκη ανεξάρτητο
- Ένα γραφικό μοντέλο των αιτιατών σχέσεων
- Διάφορες περιπτώσεις μάθησης των Μπεϋζιανών δικτύων
  - Δεδομένου μιας δομής δικτύου και όλων των μεταβλητών: εύκολο
  - Δεδομένου μιας δομής δικτύου αλλά μερικών μεταβλητών
  - Όταν η δομή του δικτύου δεν είναι γνωστή εξ αρχής
- Η διαδικασία κατηγοριοποίησης επιστρέφει μια κατανομή πιθανότητας για όλες τις ετικέτες του χαρακτηριστικού κλάσης (όχι μόνο για μία ετικέτα κλάσης)

Κατηγοριοποίηση με πίσω διάδοση: Νευρωνικά Δίκτυα

Ένα σύνολο συνεκτικών μονάδων εισόδου/εξόδου όπου κάθε σύνδεση έχει ένα συντελεστή βαρύτητας

- Πλεονεκτήματα
  - η ακρίβεια πρόβλεψης είναι γενικά υψηλή
  - robust, δουλεύει όταν τα παραδείγματα του συνόλου εκπαίδευσης περιέχουν λάθη
  - η έξοδος μπορεί να είναι διακριτή, συνεχών τιμών, ή ένα διάνυσμα διαφόρων διακριτών ή συνεχών τιμών χαρακτηριστικών
  - γρήγορη αξιολόγηση της μαθημένης συνάρτησης στόχου
- Μειονεκτήματα
  - μεγάλος χρόνος εκπαίδευσης
  - απαιτούν παραμέτρους (που συνήθως καθορίζονται εμπειρικά) (π.χ. η τοπολογία του δικτύου)
  - δύσκολη κατανόηση της μαθημένης συνάρτησης (βάρη)
  - δύσκολη ενσωμάτωση της γνώσης του εκάστοτε γνωστικού πεδίου

Ένας νευρώνας

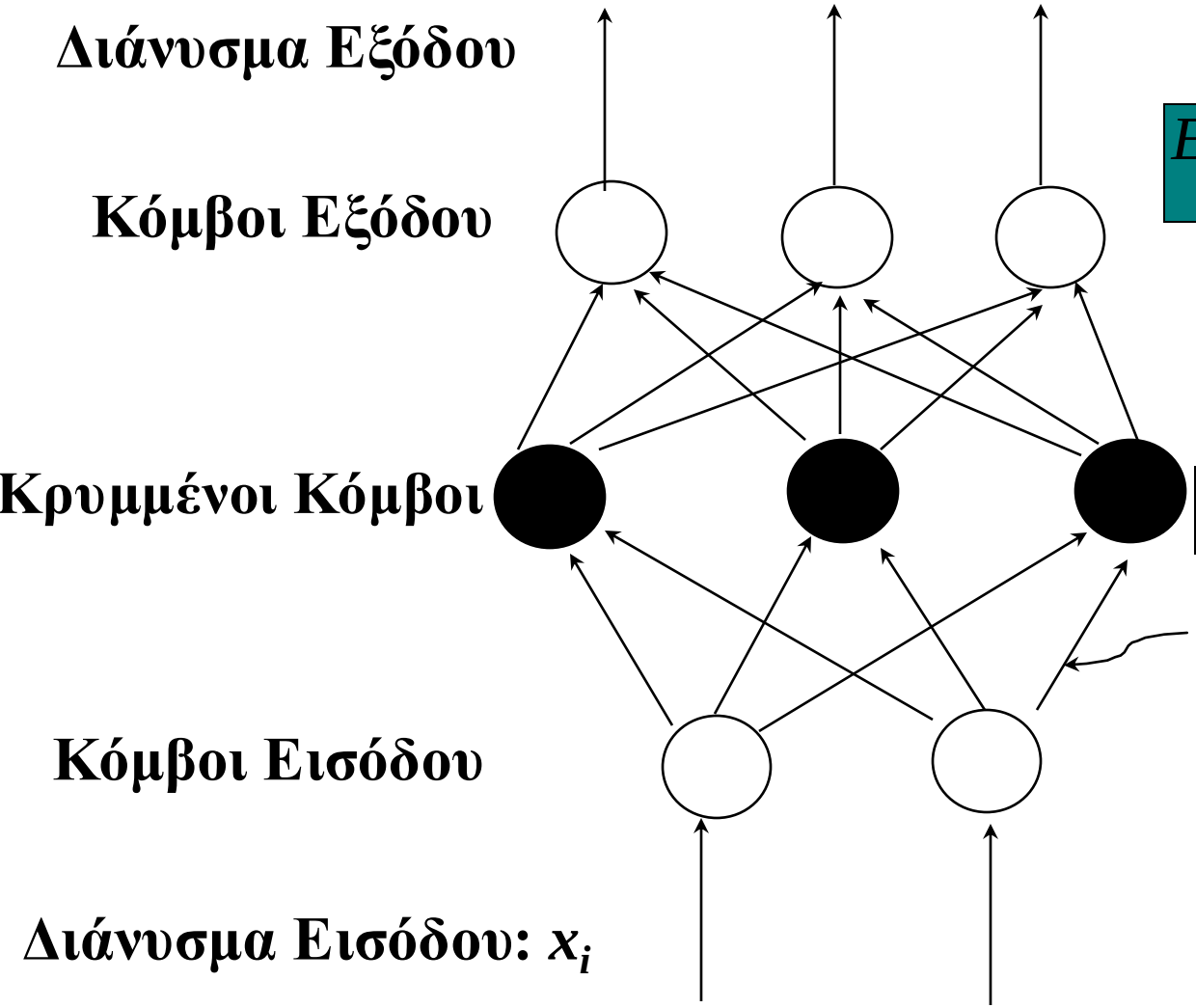


- Το n-διαστάσεων διάνυσμα εισόδου x αντιστοιχίζεται στη μεταβλητή y μέσω ενός βαθμωτού γινομένου και μιας μη-γραμμικής συνάρτησης αντιστοίχισης

Εκπαίδευση Δικτύου

- Ο απώτερος στόχος της εκπαίδευσης
  - Εύρεση ενός συνόλου βαρών τα οποία εφαρμοζόμενα σε (σχεδόν) όλες τις πλειάδες του συνόλου εκπαίδευσης, τις κατηγοριοποιούν σωστά
- Βήματα
  - Αρχικοποίηση βαρών με τυχαίες τιμές
  - Τροφοδοσία των πλειάδων εισόδου στο δίκτυο μία προς μία
  - Για κάθε μονάδα
    - Υπολόγισε την συνολική είσοδο της μονάδας ως το γραμμικό συνδυασμό όλων των εισόδων της μονάδας
    - Υπολόγισε την τιμή εξόδου χρησιμοποιώντας τη συνάρτηση ενεργοποίησης
    - Υπολόγισε το σφάλμα
    - Ανανέωσε τα βάρη και τα bias

Πολυ-επίπεδο δίκτυο Perceptron



$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\theta_j = \theta_j + (l)Err_j$$

$$w_{ij} = w_{ij} + (l)Err_j O_i$$

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

$$O_j = \frac{1}{1+e^{-I_j}}$$

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

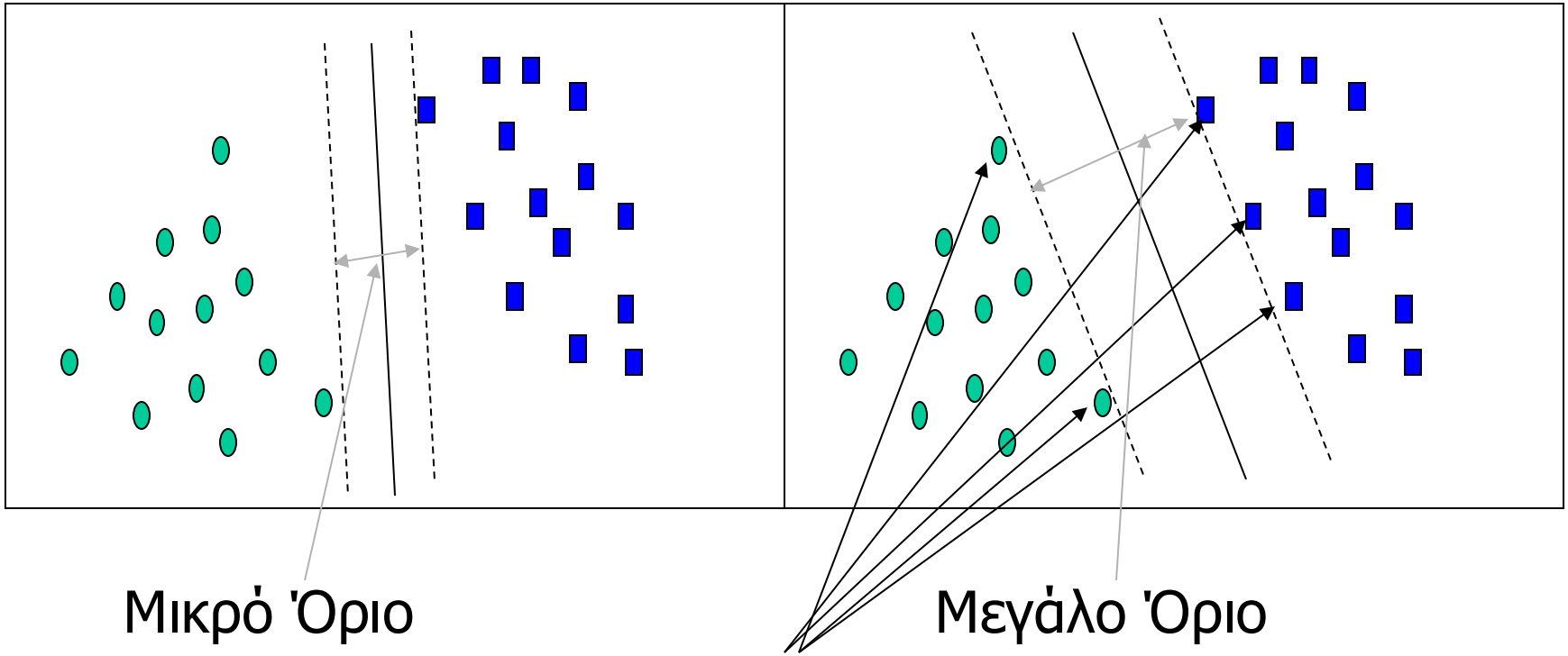
SVM- Support Vector Machines
Μηχανές Διανυσμάτων Υποστήριξης

- Μια νέα μέθοδος κατηγοριοποίησης για γραμμικά και μη γραμμικά δεδομένα
- Χρησιμοποιεί μη γραμμική αντιστοίχιση για να μετασχηματίσει τα αρχικά δεδομένα εκπαίδευσης σε δεδομένα υψηλότερης διάστασης
- Με βάση τη νέα διάσταση, ψάχνει για γραμμικώς διαχωριζόμενα υπερεπίπεδα (δηλ., "όρια απόφασης")
- Με μια κατάλληλη μη γραμμική αντιστοίχιση σε μια επαρκώς υψηλότερη διάσταση, τα δεδομένα από δύο κλάσεις μπορούν να διαχωριστούν από ένα υπερεπίπεδο
- Η μέθοδος SVM βρίσκει αυτό το υπερεπίπεδο χρησιμοποιώντας
  - Διανύσματα υποστήριξης (support vectors) ("κρίσιμες" πλειάδες εκπαίδευσης) και
  - Όρια - margins (που καθορίζονται από τα διανύσματα υποστήριξης)

SVM—Ιστορία και Εφαρμογές

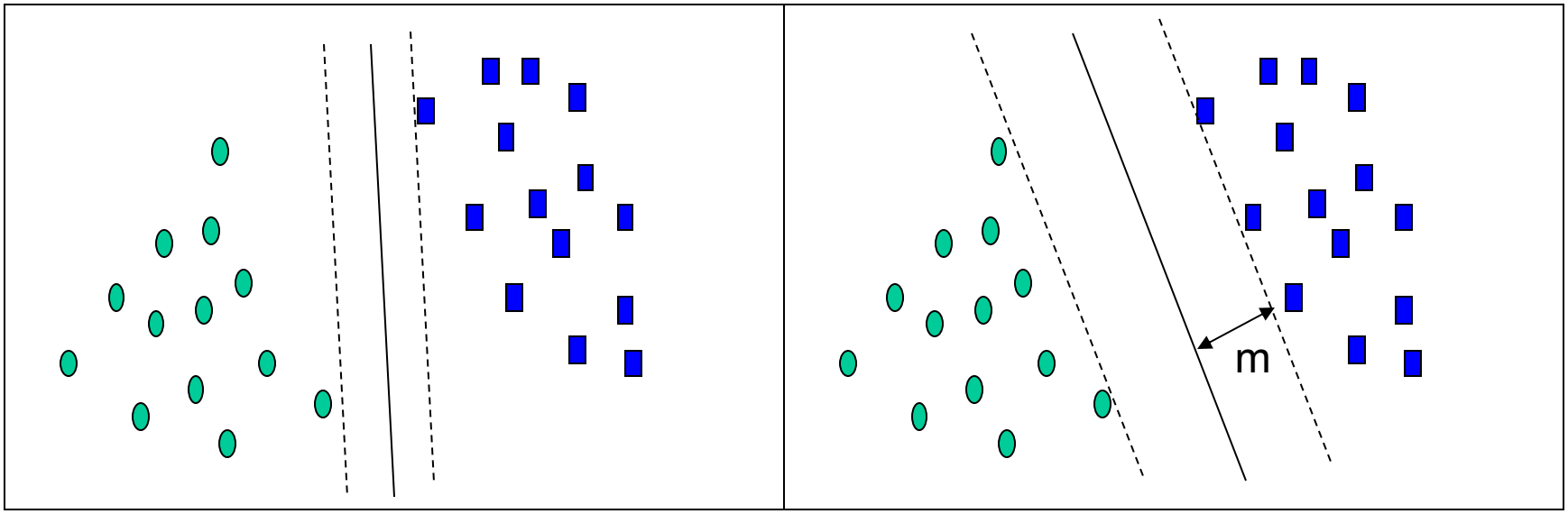
- Vapnik et al. (1992)—βασίστηκαν στην θεωρία στατιστικής εκμάθησης τωνVapnik & Chervonenkis (δεκαετία 1960)
- Χαρακτηριστικά:
  - η εκπαίδευση μπορεί να είναι αργή
  - η ακρίβεια είναι υψηλή χάρη στην ικανότητα μοντελοποίησης σύνθετων, μη γραμμικών ορίων απόφασης (μεγιστοποίηση ορίων)
- Χρησιμοποίηση για κατηγοριοποίηση και πρόβλεψη
- Εφαρμογές:
  - αναγνώριση χειρόγραφων, αναγνώριση αντικειμένων, ταυτοποίηση ομιλίας, έλεγχοι πρόβλεψης χρονοσειρών benchmarking

SVM—Γενική Φιλοσοφία



Διανύσματα Υποστήριξης

SVM—Όταν τα δεδομένα είναι γραμμικώς διαχωριζόμενα



Έστω το σύνολο δεδομένων D που αποτελείται από $(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_{|D|}, y_{|D|})$, όπου $\mathbf{X}_i$ είναι το σύνολο των πλειάδων εκπαίδευσης που σχετίζονται με τις ετικέτες κλάσης y_i

Υπάρχουν άπειρες γραμμές (υπερεπίπεδα) που διαχωρίζουν τις δύο κλάσεις αλλά δεν είναι βέλτιστες. Στόχος είναι η εύρεση του βέλτιστου υπερεπιπέδου (αυτό που ελαχιστοποιεί το σφάλμα κατηγοριοποίησης στα άγνωστα δεδομένα)

Η μέθοδος SVM αναζητά το υπερεπίπεδο με το μέγιστο όριο, π.χ., **maximum marginal hyperplane** (MMH)

Το πρόβλημα γίνεται ένα τετραγωνικό πρόβλημα βελτιστοποίησης με περιορισμούς (convex) : τετραγωνική συνάρτηση στόχου και γραμμική περιορισμοί -> Quadratic

SVM—Συναρτήσεις Πυρήνα (Kernel functions)

- Αντί του υπολογισμού του εσωτερικού γινομένου στις μετασχηματισμένες πλειάδες δεδομένων, είναι μαθηματικά ισοδύναμη η εφαρμογή συναρτήσεων πυρήνα $K(\mathbf{X_i}, \mathbf{X_j})$ στα αρχικά δεδομένα, π.χ., $K(\mathbf{X_i}, \mathbf{X_j}) = \Phi(\mathbf{X_i})\ \Phi(\mathbf{X_j})$
- Τυπικές Συναρτήσεις Πυρήνα

$$\text{Polynomial kernel of degree } h: \quad K(X_i, X_j) = (X_i \cdot X_j + 1)^h$$

$$\text{Gaussian radial basis function kernel}: \quad K(X_i, X_j) = e^{-\|X_i - X_j\|^2/2\sigma^2}$$

$$\text{Sigmoid kernel}: \quad K(X_i, X_j) = \tanh(\kappa X_i \cdot X_j - \delta)$$

- Η μέθοδος SVM μπορεί να χρησιμοποιηθεί για κατηγοριοποίηση περισσότερων (> 2) κλάσεων και για ανάλυση παλινδρόμησης (regression analysis) με χρήση επιπλέον παραμέτρων

SVM - Σχετικά Links

- SVM Website
  - http://www.kernel-machines.org/
- Αντιπροσωπευτικές υλοποιήσεις
  - **LIBSVM**: μια αποδοτική υλοποίηση της SVM, πολλαπλής κλάσης κατηγοριοποιήσεις, nu-SVM, one-class SVM, περιλαμβάνει επίσης διάφορες διεπαφές με java, python, κτλ.
  - **SVM-light**: απλούστερη αλλά με απόδοση όχι καλύτερη της LIBSVM, υποστηρίζει μόνο δυαδική κατηγοριοποίηση και μόνο τη γλώσσα C
  - **SVM-torch**: μια άλλη πρόσφατη υλοποίηση γραμμένη σε C

SVM—Εισαγωγική Βιβλιογραφία

- "Statistical Learning Theory" by Vapnik: δυσνόητο, περιέχει αρκετά λάθη
- C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2(2), 1998.
  - Καλύτερο από το βιβλίο του Vapnik, ωστόσο δυσνόητο για εισαγωγή και τα παραδείγματα δεν είναι τόσο διαισθητικά
- "An Introduction to Support Vector Machines" by N. Cristianini and J. Shawe-Taylor
  - Επίσης δυσνόητο για εισαγωγή, αλλά η εξήγηση του θεωρήματος mercer's theorem είναι η βέλτιστη μεταξύ των ανωτέρω βιβλιογραφικών πηγών
- The neural network book by Haykins
  - Περιέχει ένα καλογραμμένο κεφάλαιο για εισαγωγή στις SVMs

Άλλες Μέθοδοι Κατηγοριοποίησης

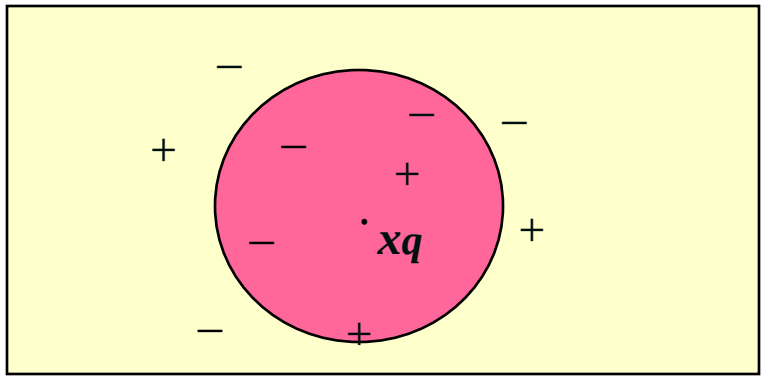
- Κατηγοριοποιητής k-πλησιέστερων γειτόνων
- Συλλογιστική περιπτώσεων (case-based reasoning)
- Γενετικοί αλγόριθμοι
- Rough set προσέγγιση
- Fuzzy set προσεγγίσεις

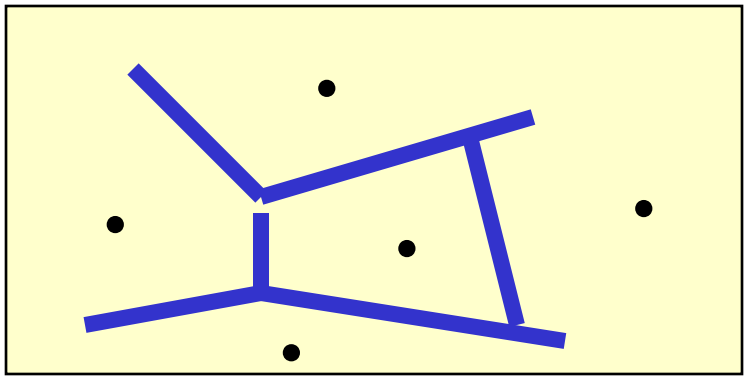
Μέθοδοι βασισμένοι σε περιπτώσεις/στοιχεία (Instance based)

- Μάθηση βασισμένη σε περιπτώσεις (ή μάθηση με ΑΝΑΛΟΓΙΑ):
  - Αποθήκευσε τα παραδείγματα εκπαίδευσης και καθυστέρησε την επεξεργασία (“lazy evaluation”) μέχρι ένα νέο δείγμα/στοιχείο να χρειάζεται κατηγοριοποίηση
- Βασικές προσεγγίσεις
  - *Προσέγγιση k-πλησιέστερων γειτόνων*
    - Τα στοιχεία αναπαρίστανται ως σημεία στον Ευκλείδειο χώρο
  - *Τοπικά βεβαρυμένη παλινδρόμηση (Locally weighted regression)*
    - Κατασκευάζει τοπική προσέγγιση
  - *Συλλογιστική Περιπτώσεων (Case-based reasoning)*
    - Χρησιμοποιεί συμβολικές αναπαραστάσεις και knowledge-based συμπερασμό

Ο αλγόριθμος των *k*-Πλησιέστερων Γειτόνων

- Όλα τα στοιχεία αντιστοιχίζονται σε σημεία στον n-Δ χώρο
- Οι πλησιέστεροι γείτονες υπολογίζονται με την βοήθεια της Ευκλείδειας απόστασης
- Η συνάρτηση στόχος μπορεί να έχει διακριτές ή συνεχείς τιμές
- Για διακριτές τιμές, η μέθοδος *k*-NN επιστρέφει την πιο κοινή τιμή μεταξύ των *k* παραδειγμάτων εκπαίδευσης (training) που είναι πλησιέστερα στο x_q
- Vonoroi διάγραμμα: η επιφάνεια απόφασης παράγεται από τον 1-NN για ένα τυπικό σύνολο δεδομένων εκπαίδευσης





Συζήτηση πάνω στον αλγόριθμο *k*-NN

- Ο αλγόριθμος k-NN για συνεχείς συναρτήσεις στόχου
  - Υπολόγισε τις μέσες τιμές των *k* πλησιέστερων γειτόνων
- Distance-weighted nearest neighbor αλγόριθμος
  - Αναθέτει βάρος στην συνεισφορά καθενός από τους k γείτονες ανάλογα με την απόσταση τους από το σημείο-ερώτημα x_q $w \equiv \frac{1}{d(x_q, x_i)^2}$
    - Αναθέτει μεγαλύτερο βάρος στους πλέον κοντινούς γείτονες
  - Ομοίως, για συνεχείς συναρτήσεις στόχου
- Robust σε θορυβώδη δεδομένα βρίσκοντας τη μέση τιμή των k-πλησιέστερων γειτόνων
- Curse of dimensionality: η απόσταση μεταξύ γειτόνων μπορεί να κυριαρχείται από μη σχετικά χαρακτηριστικά
  - Για την αποφυγή του, οι άξονες «επιμηκύνουν» ή «περιορίζουν» τα λιγότερο σχετικά χαρακτηριστικά

Πρόβλεψη

- Η πρόβλεψη παρουσιάζει ομοιότητα με την κατηγοριοποίηση:
  - Αρχικά, κατασκευάζεται ένα μοντέλο
  - Έπειτα, το μοντέλο χρησιμοποιείται για την πρόβλεψη των άγνωστων τιμών
    - Βασική μέθοδος για πρόβλεψη είναι η παλινδρόμηση (regression)
      - Γραμμική και πολλαπλή παλινδρόμηση
      - Μη γραμμική παλινδρόμηση
- Η πρόβλεψη διαφέρει από την κατηγοριοποίηση:
  - Η κατηγοριοποίηση αναφέρεται στην πρόβλεψη των κατηγορικών ετικετών κλάσης
  - Η πρόβλεψη μοντελοποιεί συναρτήσεις συνεχών τιμών

Προβλεπτική Μοντελοποίηση

- Προβλεπτική μοντελοποίηση:
  - Προβλέπει τιμές δεδομένων ή κατασκευάζει γενικευμένα γραμμικά μοντέλα βασισμένα στη βάση δεδομένων
  - Προβλέπει πεδία τιμών ή κατανομές κατηγοριών
- Περιγραφή μεθόδου:
  - Ελάχιστη γενίκευση
  - Ανάλυση συσχετίσεων μεταξύ των γνωρισμάτων
  - Κατασκευή γενικευμένου γραμμικού μοντέλου
  - Πρόβλεψη
- Καθορισμός των σημαντικών παραγόντων οι οποίοι επηρεάζουν την πρόβλεψη
  - Ανάλυση συσχετίσεων των δεδομένων: μέτρηση αβεβαιότητας, ανάλυση εντροπίας, κρίση ειδικών, κτλ.
- Πολύ-επίπεδη πρόβλεψη: drill-down και roll-up ανάλυση (data cubes)

Ανάλυση Παλινδρόμησης και Log-Linear Μοντέλα για Πρόβλεψη

- Γραμμική Παλινδρόμηση: $Y = \alpha + \beta X$
  - Δύο παράμετροι, α και β καθορίζουν τη Y-intercept και κλίση της γραμμής και υπολογίζονται με βάση τα δεδομένα.
  - Χρησιμοποιώντας το κριτήριο των ελαχίστων τετραγώνων στις γνωστές τιμές των (X_1, Y_1), (X_2, Y_2), …, (X_s, Y_s)

$$\beta = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}, \quad \alpha = \bar{y} - \beta\,\bar{x}$$

Regress Ανάλυση Παλινδρόμησης και Log-Linear Μοντέλα για Πρόβλεψη

- Πολλαπλή παλινδρόμηση: $Y = a + b1\ X1 + b2\ X2$.
  - –Περισσότερες από μία μεταβλητές πρόβλεψης
  - –Αρκετές μη γραμμικές συναρτήσεις μπορούν να μετασχηματιστούν όπως παραπάνω
- Μη γραμμική παλινδρόμηση: $Y = a + b1\ X + b2\ X^2 + b3\ X^3$.
- Log-linear μοντέλα:
  - –Προσεγγίζουν διακριτές πολυδιάστατες κατανομές πιθανότητας (multi-way πίνακας με από-κοινού πιθανότητες) μέσω ενός γινομένου χαμηλής τάξης πινάκων
  - –Πιθανότητα: $p(a, b, c, d) = \alpha_{ab}\ \beta_{ac}\ \chi_{ad}\ \delta_{bcd}$

Τοπικά Βεβαρυμένη Παλινδρόμηση

- Κατασκευάζει μια explicit προσέγγιση της f πάνω σε μια τοπική περιοχή γύρω από το ερώτημα x_q.
- Τοπικά βεβαρυμένη γραμμική παλινδρόμηση:
  - Η συνάρτηση στόχος f προσεγγίζεται κοντά στο x_q χρησιμοποιώντας τη γραμμική συνάρτηση $\hat{f}(x) = w_0 + w_1 a_1(x) + \cdots + w_n a_n(x)$
  - Ελαχιστοποιεί το τετραγωνικό σφάλμα: distance-decreasing weight K

$$E(x_q) \equiv \frac{1}{2} \sum_{x \in k\_nearest\_neighbors\_of\_x_q} (f(x) - \hat{f}(x))^2 K(d(x_q, x))$$

  - Ο κανόνας εκπαίδευσης φθίνουσας κλίσης:

$$\Delta w_j \equiv \eta \sum_{x \in k\_nearest\_neighbors\_of\_x_q} K(d(x_q, x))((f(x) - \hat{f}(x))a_j(x)$$

- Στις περισσότερες περιπτώσεις, η συνάρτηση στόχος προσεγγίζεται από σταθερά, γραμμική ή τετραγωνική συνάρτηση

Μετρικές Ακρίβειας Κατηγοριοποιητή

	C_1	C_2
C_1	True positive	False negative
C_2	False positive	True negative

Classes model ->	buy_computer = yes	buy_computer = no	total	recognition(%)
buy_computer = yes	6954	46	7000	99.34
buy_computer = no	412	2588	3000	86.27
total	7366	2634	10000	95.52

- Η Ακρίβεια (Accuracy) ενός κατηγοριοποιητή M, acc(M): το ποσοστό των πλειάδων ελέγχου που είναι σωστά κατηγοριοποιημένα από το μοντέλο M
  - Ποσοστό Λάθους (Error rate) του M = 1 – acc(M)
  - Δοσμένων *m* κλάσεων, $CM_{i,j}$, μια εγγραφή στον **confusion matrix**, δείχνει το πλήθος των πλειάδων στην κλάση *i* τα οποία είναι κατηγοριοποιημένα από το μοντέλο στην κλάση *j*
- Εναλλακτικές Μετρικές Ακρίβειας (π.χ., για διάγνωση καρκίνου)

  sensitivity = t-pos/pos /* true positive recognition rate */

  specificity = t-neg/neg /* true negative recognition rate */

  precision = t-pos/(t-pos + f-pos)

  accuracy = sensitivity * pos/(pos + neg) + specificity * neg/(pos + neg)
  - Αυτό το μοντέλο μπορεί επίσης να χρησιμοποιηθεί ανάλυση cost-benefit

Μέτρηση Σφάλματος Προβλέπτη (Predictor)

- Μέτρηση ακρίβειας προβλέπτη: μετρά πόσο απέχει η τιμή πρόβλεψης από την πραγματική (γνωστή) τιμή
- **Συνάρτηση Απώλειας (Loss function)**: μετρά το σφάλμα μεταξύ y_i και της τιμής πρόβλεψης y_i'
  - Απόλυτο λάθος: $|y_i - y_i'|$
  - Τετραγωνικό λάθος: $(y_i - y_i')^2$
- Σφάλμα ελέγχου (γενίκευση σφάλματος): η μέση απώλεια στο σύνολο ελέγχου
  - Mean absolute error: $\frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$ Mean squared error: $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$
  - Relative absolute error: $\frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}$ Relative squared error: $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}$

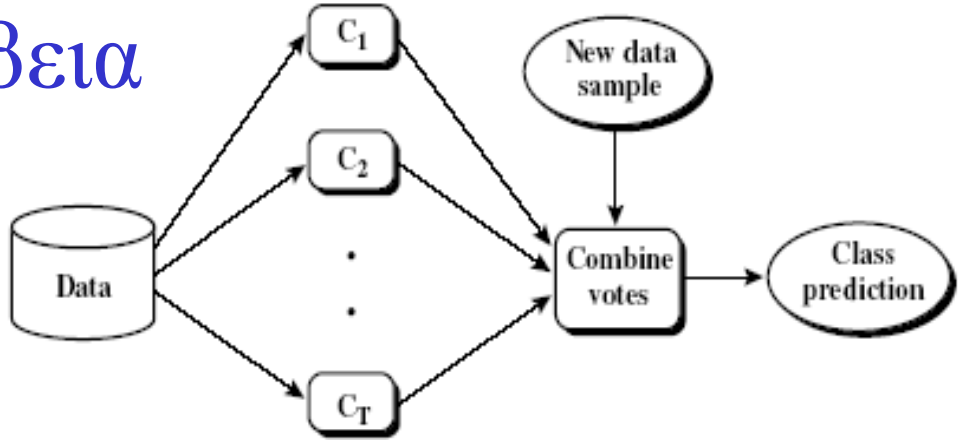
  Το μέσο τετραγωνικό λάθος υπερβάλει την παρουσία των outliers

  Πιο δημοφιλείς μετρικές: mean-square error και root relative squared error

Αξιολόγηση της Ακρίβειας ενός Κατηγοριοποιητή ή Προβλέπτη (I)

- Holdout μέθοδος
  - Τα δεδομένα τυχαία διαχωρίζονται σε δύο ανεξάρτητα σύνολα
    - **Σύνολο εκπαίδευσης** (π.χ., 2/3) για κατασκευή μοντέλου
    - **Σύνολο ελέγχου** (π.χ., 1/3) για αξιολόγηση ακρίβειας
  - Τυχαία δειγματοληψία: μια παραλλαγή της holdout
    - Επανέλαβε την τεχνική holdout k φορές, συνολική ακρίβεια (accuracy) = μ.ο. της ακρίβειας των επιμέρους εκτελέσεων
- Σταυρωτή επικύρωση (Cross-validation) (*k-αναδιπλώσεις (k*-fold), όπου συνήθως k = 10)
  - Διαχώρισε τυχαία τα δεδομένα σε k αμοιβαία διαχωριζόμενα υποσύνολα, περίπου ισομεγέθη
  - Στην *i*-οστή επανάληψη, χρησιμοποίησε το D_i ως σύνολο ελέγχου και τα υπόλοιπα ως σύνολο εκπαίδευσης
  - Leave-one-out: k αναδιπλώσεις όπου k = # των πλειάδων, για μικρού όγκου δεδομένα
  - Stratified cross-validation: οι αναδιπλώσεις καθορίζονται έτσι ώστε η κατανομή κλάσεων σε κάθε αναδίπλωση να είναι περίπου ίδια με την αυτή στα αρχικά δεδομένα

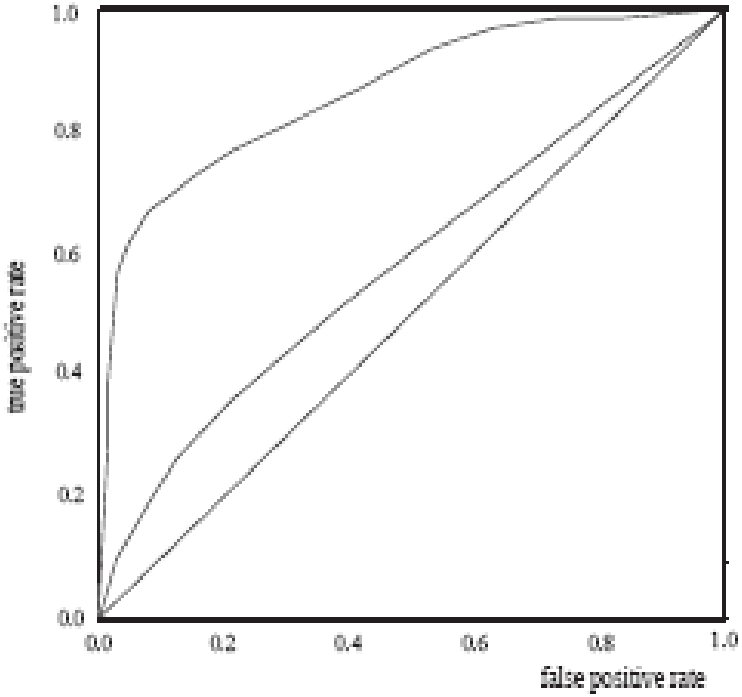
Μέθοδοι Συνένωσης: Αυξάνοντας την ακρίβεια



- Μέθοδοι συνένωσης
  - Χρήση ενός συνδυασμού από μοντέλα με σκοπό την αύξηση της ακρίβειας
  - Συνδυασμός μια σειράς από k εκπαιδευμένα μοντέλα, M_1, M_2, …, M_k, για την δημιουργία ενός βελτιωμένου μοντέλου M*
- Συνήθεις μέθοδοι συνένωσης
  - Bagging: υπολογισμός μέσου όρου της πρόβλεψης πάνω σε μια συλλογή κατηγοριοποιητών
  - Boosting: βεβαρυμένη συμμετοχή σε μια συλλογή κατηγοριοποιητών
  - Ensemble: συνδυασμός ενός συνόλου ετερογενών κατηγοριοποιητών

Επιλογή Μοντέλου: ROC Καμπύλες

- ROC (Receiver Operating Characteristic) καμπύλες: για οπτική σύγκριση των μοντέλων κατηγοριοποίησης
- Προέρχεται από την θεωρία ανίχνευσης σημάτων
- Δείχνει το trade-off μεταξύ true positive rate και false positive rate
- Η επιφάνεια κάτω από την ROC καμπύλη είναι μετρική της ακρίβειας του μοντέλου
- Ταξινόμησε τις πλειάδες ελέγχου σε φθίνουσα σειρά: αυτές που είναι πιο πιθανό να ανήκουν στη θετική κλάση εμφανίζονται στην κορυφή της λίστας
- Όσο πιο κοντά στη διαγώνιο (π.χ., εμβαδόν περιοχής περίπου 0.5), τόσο λιγότερο ακριβές είναι το μοντέλο



- Κάθετος άξονας: true positive rate
- Οριζόντιος άξονας: false positive rate
- Διαγώνια γραμμή;
- Μοντέλο με τέλεια ακρίβεια: εμβαδόν 1.0

Σύνοψη (I)

- Κατηγοριοποίηση και πρόβλεψη είναι δύο μορφές ανάλυσης δεδομένων που μπορούν να χρησιμοποιηθούν για την εξαγωγή μοντέλων που περιγράφουν σημαντικά δεδομένα κλάσεων ή για την πρόβλεψη μελλοντικών τάσεων
- Αποδοτικές και κλιμακώσιμες μέθοδοι έχουν αναπτυχθεί όπως δέντρα απόφασης, Αφελής Μπεϋζιανός κατηγοριοποιητής, Δίκτυα Μπεϋζιανής Λογικής, κατηγοριοποιητές βασισμένοι σε κανόνες, πίσω μετάδοση σφάλματος, Μηχανές Διανυσμάτων Υποστήριξης (SVM), συσχετιστική κατηγοριοποίηση, κατηγοριοποιητές πλησιέστερων γειτόνων, case-based συλλογιστική, και άλλες μέθοδοι κατηγοριοποίησης όπως γενετικοί αλγόριθμοι, rough set και fuzzy set προσεγγίσεις.
- Γραμμικά, μη γραμμικά και γενικευμένα γραμμικά μοντέλα παλινδρόμησης μπορούν να χρησιμοποιηθούν για πρόβλεψη. Αρκετά μη γραμμικά προβλήματα μπορούν να μετατραπούν σε γραμμικά προβλήματα εκτελώντας μετασχηματισμούς στις μεταβλητές πρόβλεψης. Δένδρα παλινδρόμησης και μοντέλα δένδρων έχουν χρησιμοποιηθεί επίσης για πρόβλεψη.

Σύνοψη(II)

- Η stratified k-αναδιπλώσεων σταυρωτή επικύρωση (k-fold cross-validation) είναι μια μέθοδος που ενδείκνυται για τον υπολογισμό της ακρίβειας. Οι τεχνικές Bagging και boosting μπορούν να χρησιμοποιηθούν για την αύξηση της συνολικής ακρίβειας εκπαιδεύοντας και συνδυάζοντας μια σειρά επιμέρους μοντέλων.
- Significance tests και ROC curves είναι χρήσιμα εργαλεία για επιλογή μοντέλων
- Έχουν προταθεί πολυάριθμες συγκρίσεις διαφορετικών μεθόδων κατηγοριοποίησης και πρόβλεψης, ωστόσο το ερευνητικό ζήτημα παραμένει ανοιχτό
- Καμία μέθοδος δεν έχει βρεθεί να υπερτερεί των υπολοίπων για όλα τα σύνολα δεδομένων
- Ζητήματα όπως η ακρίβεια, ο χρόνος εκπαίδευσης, οι ιδιότητες robustness και interpretability, καθώς και η δυνατότητα κλιμάκωσης μπορούν να θεωρηθούν ως trade-offs, περιπλέκοντας ακόμα την εύρεση ενός καθολικά αποδοτικότερου μοντέλου

Μηχανική Μάθηση

Συσταδοποίηση (Clustering)

(based on notes by Jiawei Han and Micheline Kamber)

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms (characterization and classification – operate on clusters)

Typical Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
  - e.g., land use, city planning, earth-quake studies
- Image Processing
- Economic Science (especially market research) – e.g., marketing, insurance
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Requirements of Clustering

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape (not just spherical clusters)
- Minimal requirements for domain knowledge to determine input parameters (such as # of clusters)
- Able to deal with noise and outliers
- Insensitive to order of input records
- Incremental clustering
- High dimensionality (especially very sparse and highly skewed data)
- Incorporation of user-specified constraints
- Interpretability and usability (close to semantics)

Types of Data and Data Structures

- Data matrix (object-by-variable structure)
  - (two modes)

  n objects, p variables

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix (object-by-object structure)
  - (one mode) between all pairs of n objects

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric: $d(i, j)$
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for *interval-scaled, boolean, categorical, ordinal* and *ratio variables*.
- Weights should be associated with different variables based on applications and data semantics.
- Hard to define "similar enough" or "good enough"
  - highly subjective.

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

- If $q = 1$, d is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Similarity and Dissimilarity Between Objects

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

- Properties
  - $d(i,j) \geq 0$
  - $d(i,i) = 0$
  - $d(i,j) = d(j,i)$, *symmetry*
  - $d(i,j) \leq d(i,k) + d(k,j)$, *triangular inequality*
- Also one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures.

Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion (k-means, k-medoids)
- Hierarchical algorithms: Create a hierarchical decomposition (agglomerative (bottom-up) or divisive(top-down)) of the set of data (or objects) using some criterion (CURE, Chameleon, BIRCH)
- Density-based: based on connectivity and density functions – grow a cluster as long as density in the neighborhood exceeds a threshold (DBSCAN, CLIQUE)
- Grid-based: based on a multiple-level grid structure (i.e., quantized space) (STING, CLIQUE)
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of the data to the given model (EM)

Partitioning Algorithms: Basic Concept

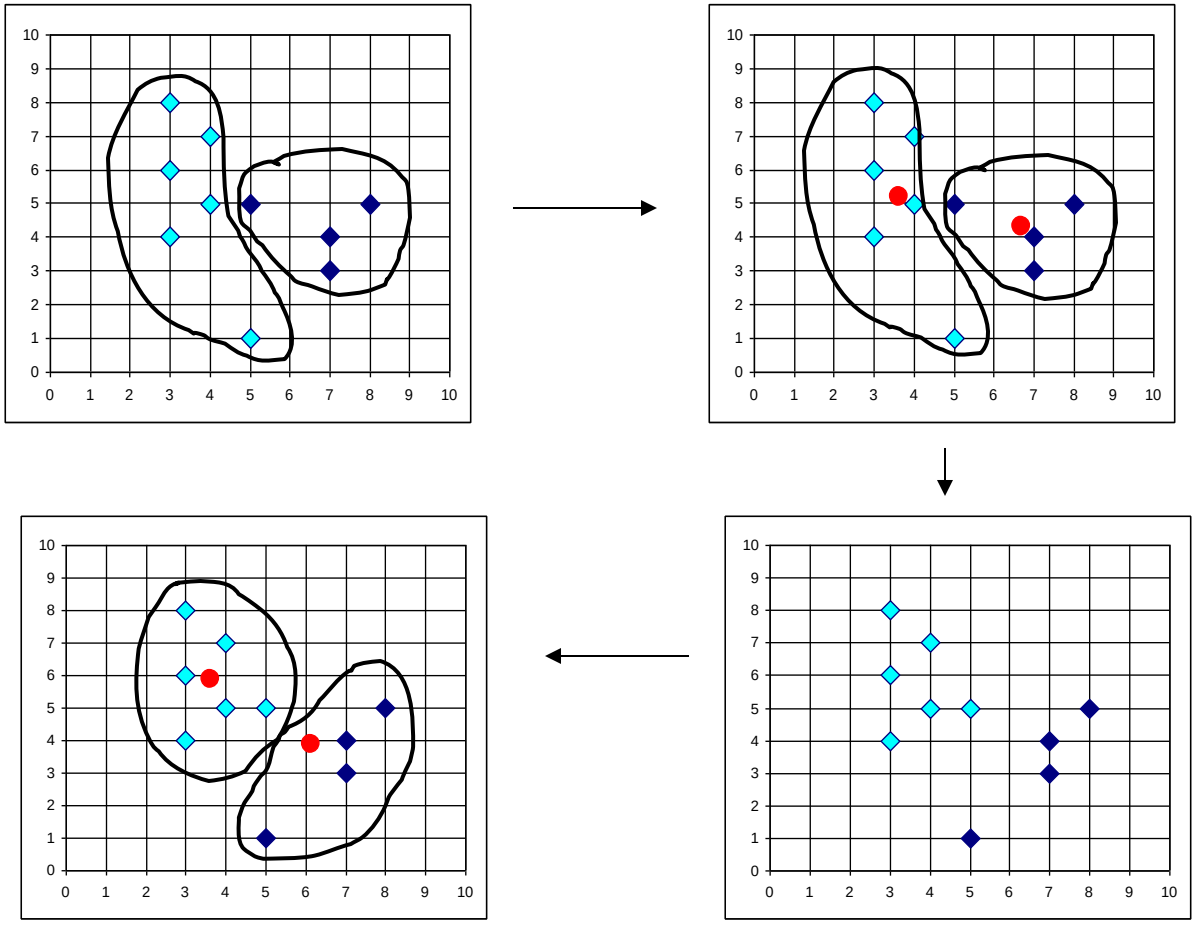
- Partitioning method: Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67): Each cluster is represented by the center of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

The *K-Means* Clustering Method

- Given *k*, *k-means* is implemented in 4 steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  3. Assign each object to the cluster with the nearest seed point.
  4. Go back to Step 2, stop when no more new assignment (or fractional drop of SSE or MSE is less than a threshold).

The *K-Means* Clustering Method

- Example



Comments on the *K-Means* Method

- Strengths
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t << n$.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weaknesses
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k*, the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes* (Huang'98)
  - Replacing means of clusters with modes
  - Using new dissimilarity measures to deal with categorical objects
  - Using a frequency-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

The *K-Medoids* Clustering Method

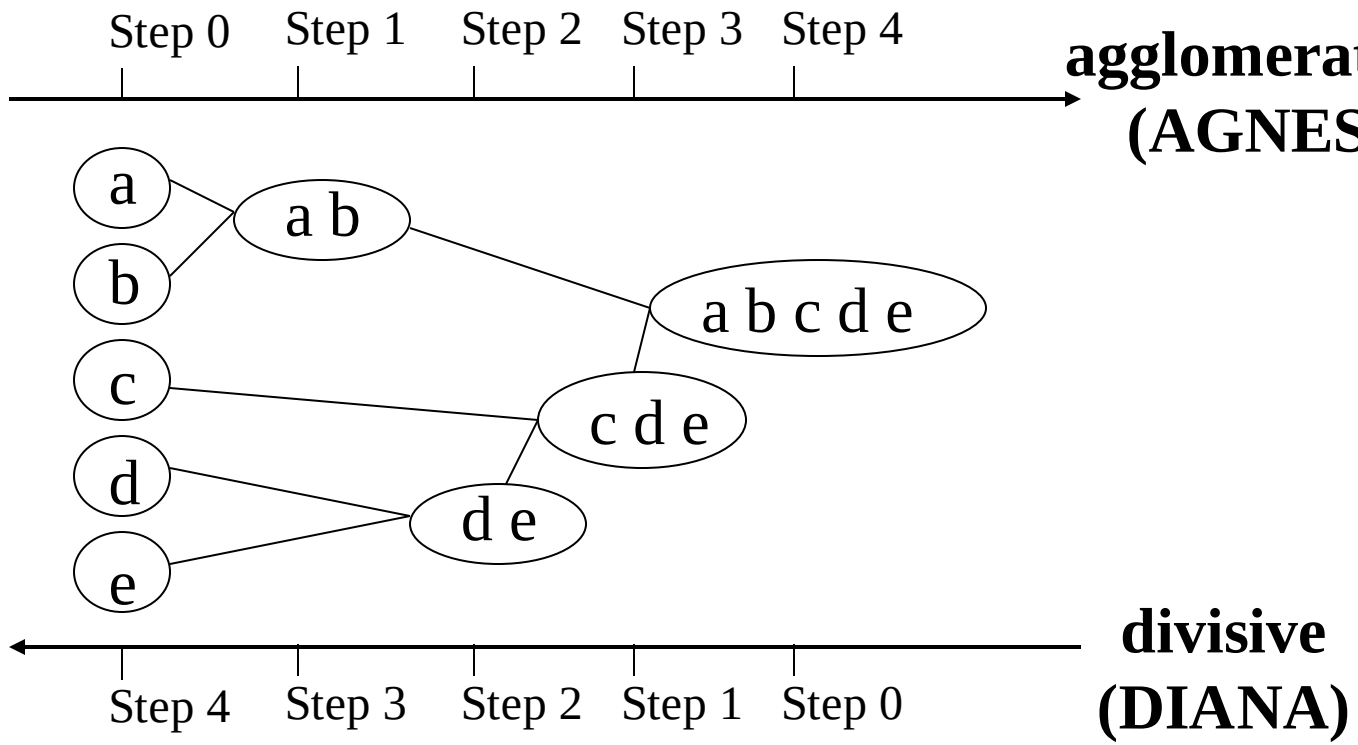
- Find *representative* (i.e., the most centrally located) objects, called medoids, in clusters
  - *PAM* (Partitioning Around Medoids, 1987)
    - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
    - works effectively for small data sets; does not scale well for large data sets
    - More robust than k-means
    - Complexity: $O(k(n-k)^2)$ for n objects and k clusters
  - *CLARA* (Kaufmann & Rousseeuw, 1990): Uses multiple samples
  - *CLARANS* (Ng & Han, 1994): Randomized sampling

PAM (Partitioning Around Medoids) (1987)

- PAM (Kaufman and Rousseeuw, 1987), built in Splus
- Use real object to represent the cluster
  1. Select ***k*** representative objects arbitrarily
  2. For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost $\boldsymbol{Tc_{ih}}$
  3. For each pair of ***i*** and ***h***,
     - If $TC_{ih} < 0$, ***i*** is replaced by ***h***
     - Then assign each non-selected object to the most similar representative object
  4. repeat steps 2-3 until there is no change

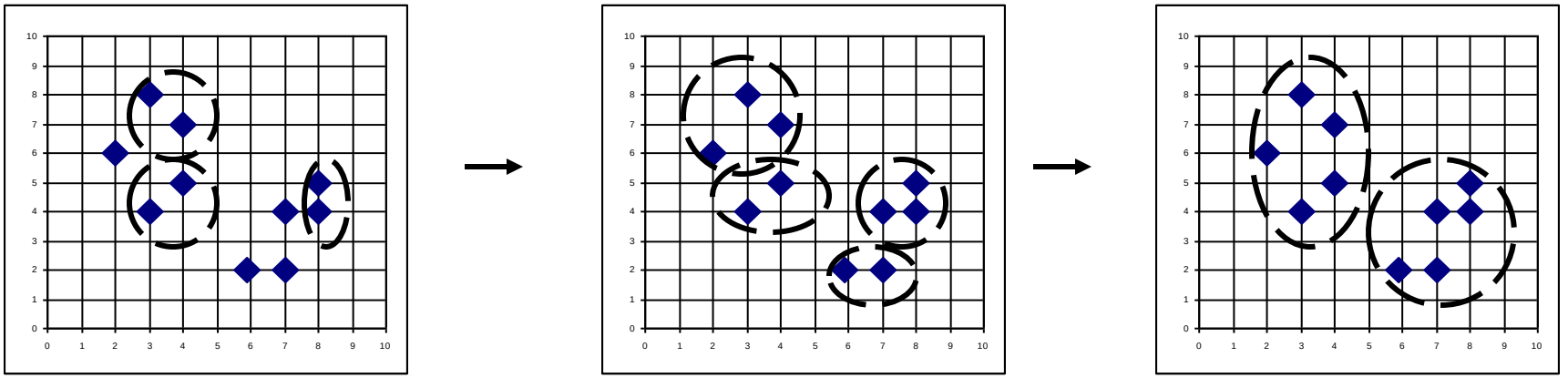
Hierarchical Clustering

- Use distance matrix as clustering criteria.
- It does not require the number of clusters ***k*** as input, but needs a termination condition



AGNES (Agglomerative Nesting)

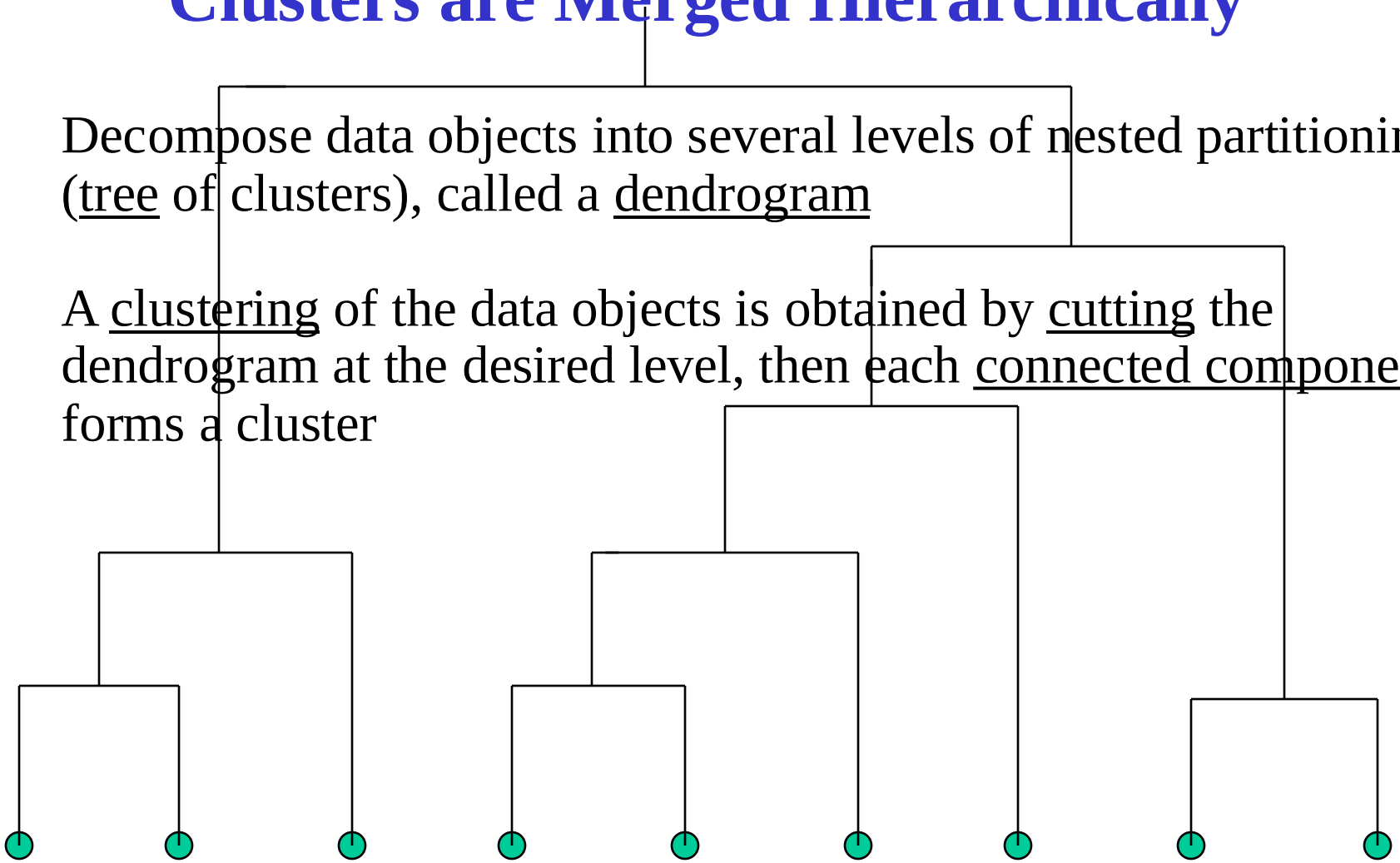
- Introduced in Kaufmann and Rousseeuw (1990), implemented in statistical analysis packages, e.g., Splus
  - Use the Single-Link method and the dissimilarity matrix:
    - each cluster is represented by all of its objects
    - the similarity between clusters is measured by the similarity of the closest pair of data points belonging to different clusters
  - Merge nodes that have the least dissimilarity
  - Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



A *Dendrogram* Shows How the Clusters are Merged Hierarchically

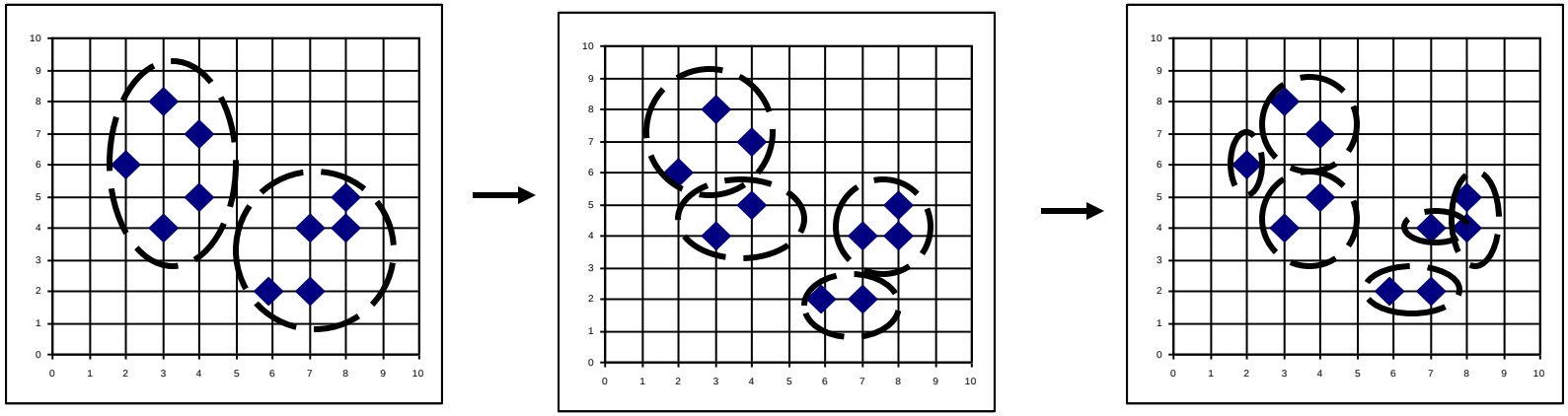
Decompose data objects into several levels of nested partitioning (tree of clusters), called a dendrogram

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990), implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



More on Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo (backtrack) what was done previously
- Integration of hierarchical with distance-based clustering
  - BIRCH (1996): uses CF-tree (clustering feature tree) and incrementally adjusts the quality of sub-clusters
  - CURE (1998): selects well-scattered (representative) points from the cluster and then shrinks them towards the center of the cluster by a specified fraction
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling

BIRCH (1996)

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies, by Zhang, Ramakrishnan, Livny (SIGMOD'96)
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
  - Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
- *Scales linearly*: finds a good clustering with a single scan and improves the quality with a few additional scans
- *Weakness:* handles only numeric data, and is sensitive to the order of the data record.

Clustering Feature Vector

Clustering Feature: $CF = (N, \overrightarrow{LS}, SS)$

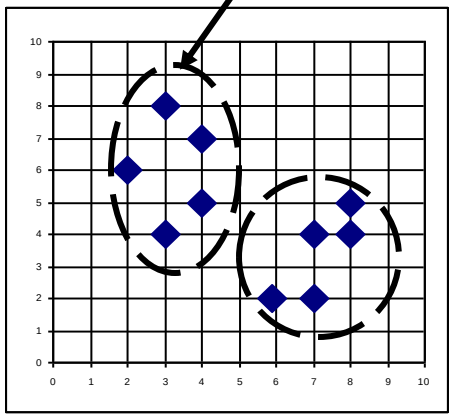
N: **Number of data points**

$LS: \sum_{i=1}^{N} \overrightarrow{X_i}$ *(linear sum)*

$SS: \sum_{i=1}^{N} \overrightarrow{X_i^2}$ *(squared sum)*

Clustering features are additive

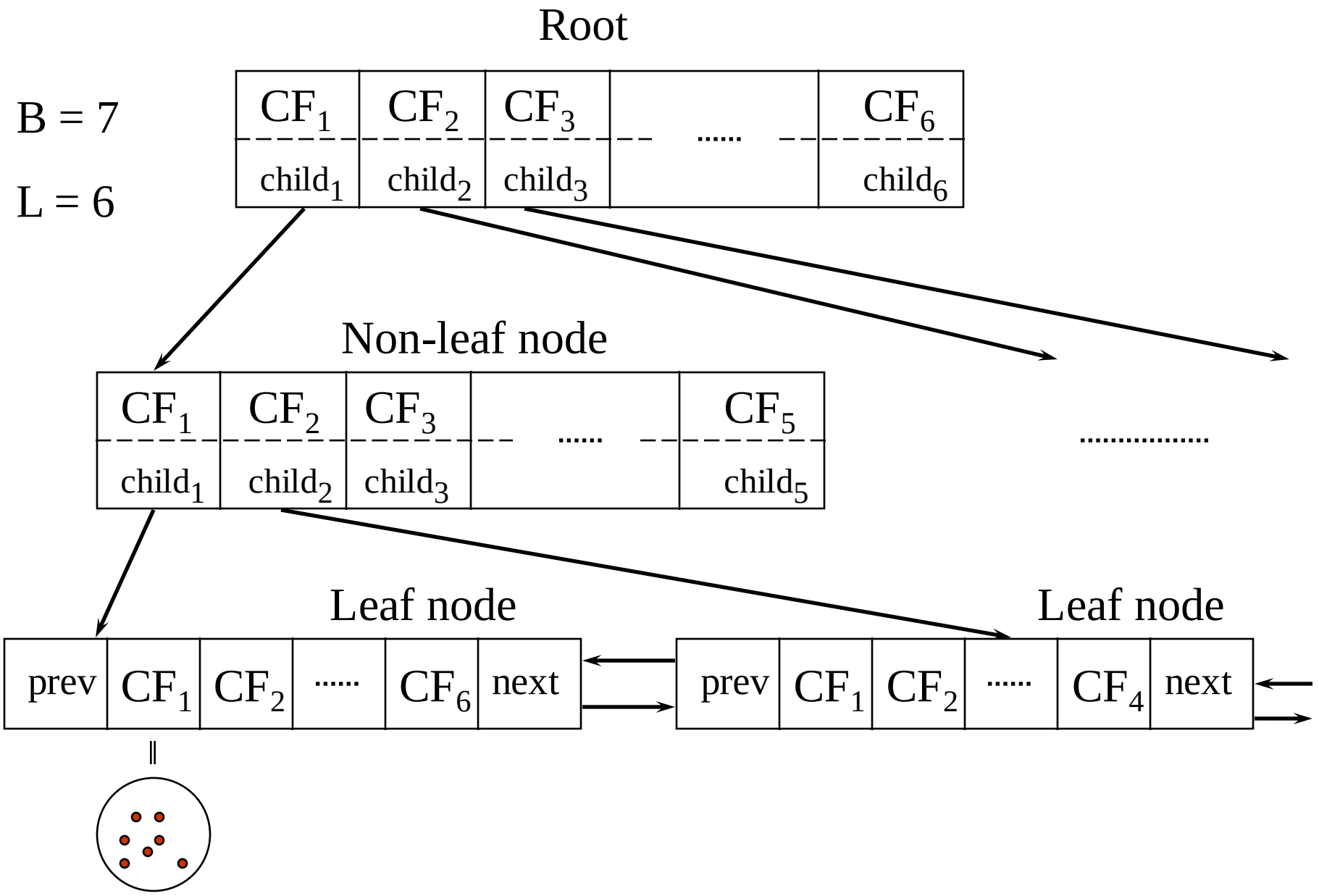
CF = (5, (16,30),(54,190))



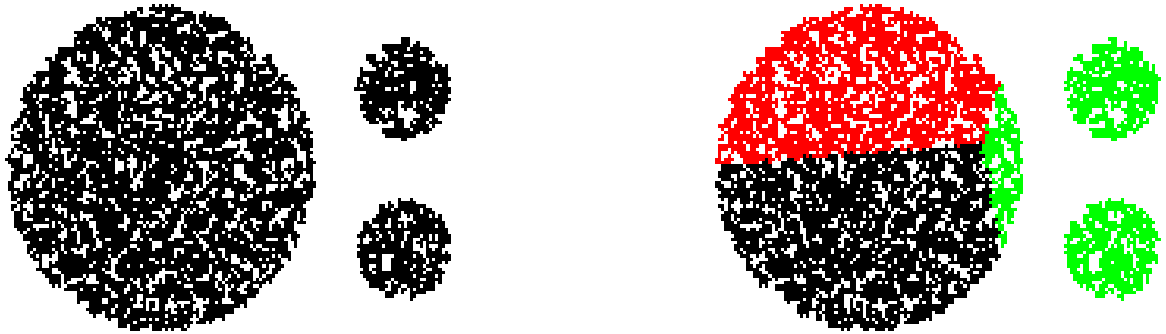
(3,4)
(2,6)
(4,5)
(4,7)
(3,8)

CF Tree

B: **Branching factor**: max # children per nonleaf node
L: **Threshold**: max diameter of subclusters at leaf nodes

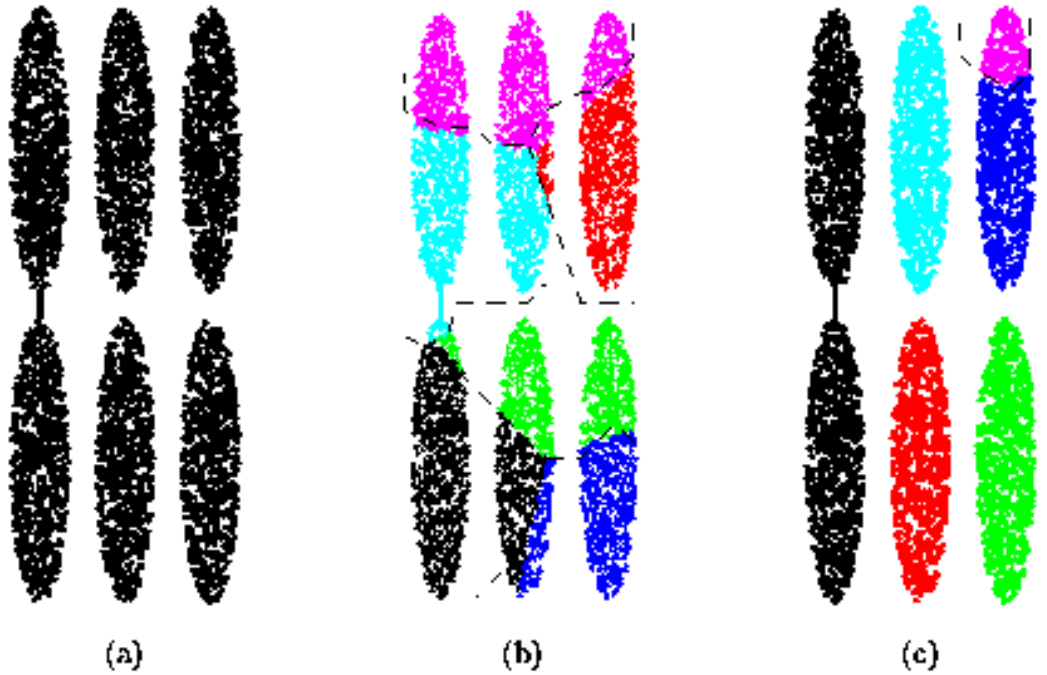


CURE (Clustering Using REpresentatives)



- CURE: proposed by Guha, Rastogi & Shim, 1998 (O(n))
  - Stops the creation of a cluster hierarchy if a level consists of *k* clusters
  - Uses multiple representative points to evaluate the distance between clusters
  - Adjusts well to arbitrary shaped clusters
  - Avoids single-link effect and is more robust to outliers

Drawbacks of Distance-Based Method



- Drawbacks of squared-error based clustering method
  - Consider only one point as representative of a cluster
  - Good only for convex shaped, similar size and density, and if *k* can be reasonably estimated

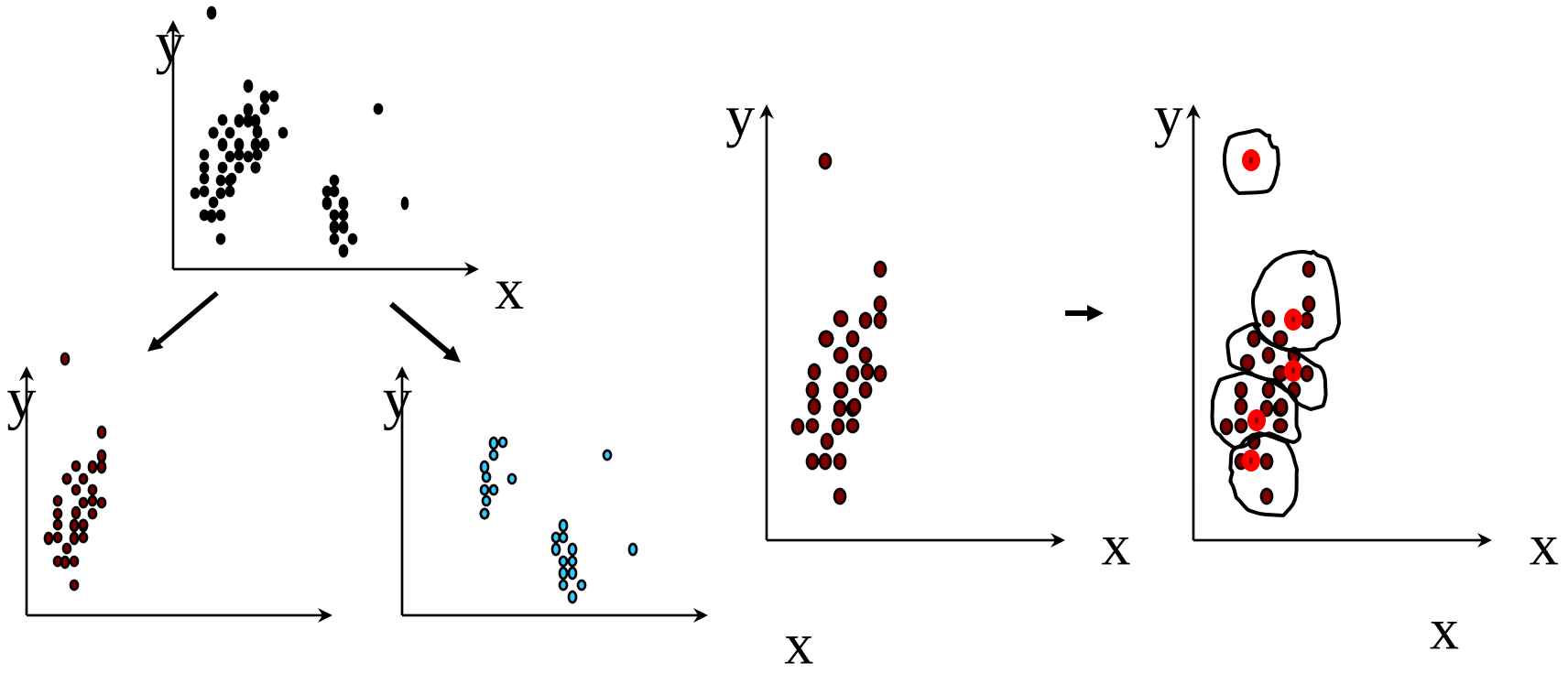
CURE: The Algorithm

- Draw random sample *s*
- Partition sample to *p* partitions with size *s/p*
- Partially cluster partitions into *s/pq* clusters
- Eliminate outliers
  - By random sampling
  - If a cluster grows too slow, eliminate it
- Cluster partial clusters
- Label data in disk

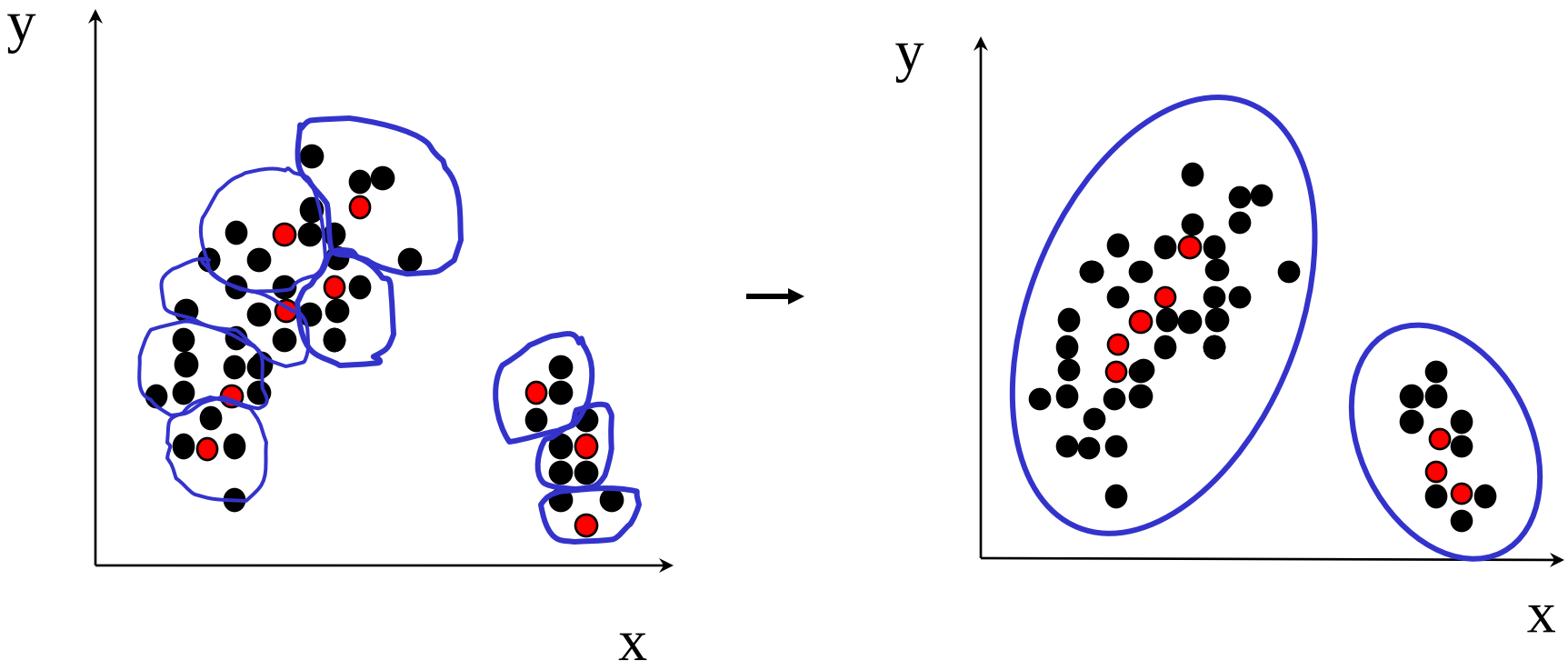
Data Partitioning and Clustering

- s = 50
- p = 2
- s/p = 25

- s/pq = 5



Cure: Shrinking Representative Points



- Shrink the multiple representative points towards the gravity center by a fraction of α.
- Multiple representatives capture the shape of the cluster

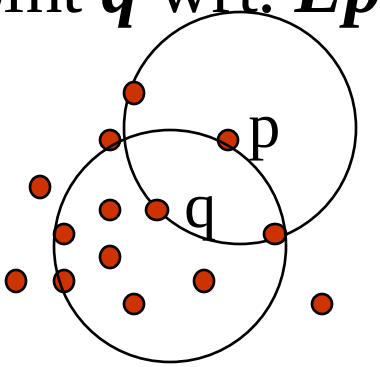
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

Density-Based Clustering: Background

- Two parameters**:**
  - ***Eps***: Maximum radius of the neighborhood
  - ***MinPts***: Minimum number of points in an Eps-neighborhood Neps of that point
- $N_{Eps}(p)$: ***{q belongs to D | dist(p,q) <= Eps}***
- Directly density-reachable**:** A point ***p*** is directly density-reachable from a point ***q*** wrt. ***Eps***, ***MinPts*** if
  - 1) ***p*** belongs to $N_{Eps}(q)$
  - 2) core point condition:

$$|N_{Eps}(q)| >= MinPts$$

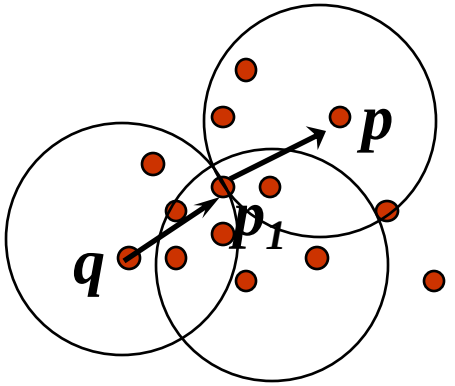


MinPts = 5

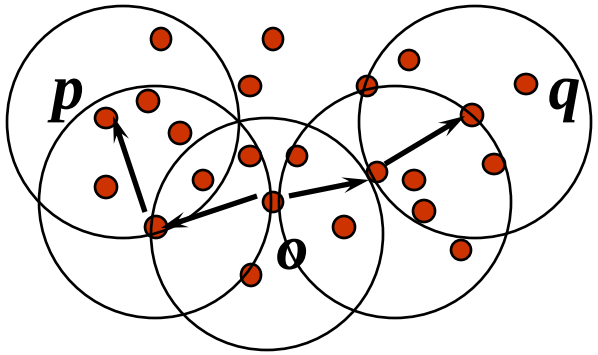
Eps = 1 cm

Density-Based Clustering: Background

- Density-reachable:
  - A point *p* is density-reachable from a point *q* wrt *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i (*assymetric relationship*)

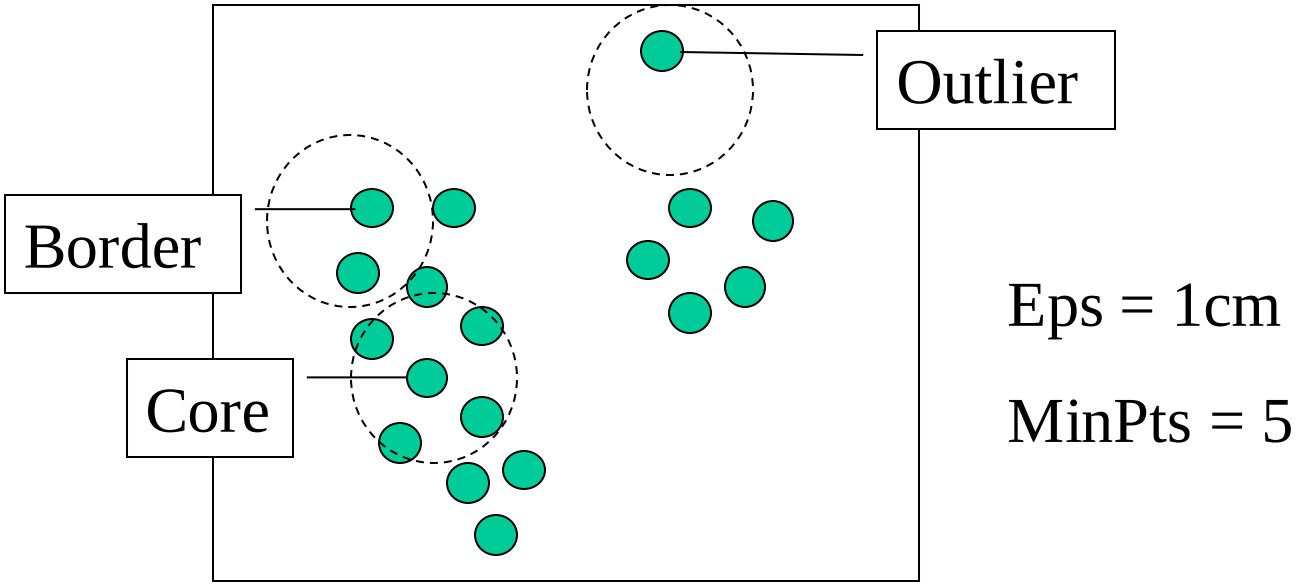


- Density-connected:
  - A point *p* is density-connected to a point *q* wrt. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts* (*symmetric relationship*).



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- ***Density-based cluster***: A maximal set of density-connected points; points not contained in the cluster are considered to be noise
- Discovers clusters of arbitrary shape in spatial databases with noise
- The user selects certain parameters



DBSCAN: The Algorithm

- Steps of the Algorithm:
  - Arbitrary select a point ***p***
  - Retrieve all points density-reachable from ***p*** wrt ***Eps*** and ***MinPts***.
  - If ***p*** is a core point, a cluster is formed.
  - If ***p*** is a border point, no points are density-reachable from ***p*** and DBSCAN visits the next point of the database.
  - Continue the process until all of the points have been processed and no new point can be added to any cluster.
- **$O(n^2)$ ->O(nlogn) with spatial indexing**

Agenda

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis
- Summary

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches
- There are still lots of research issues on cluster analysis, such as constraint-based clustering