ΤΕΙ ΚΕΝΤΡΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΑΝΟΙΚΤΑ ΑΚΑΔΗΜΑΙΚΑ ΜΑΘΗΜΑΤΑ

Ειδικά Θέματα Βάσεων Δεδομένων

Ενότητα 15: Εξόρυξη Δεδομένων (Data Mining)

Δρ. Τσιμπίρης Αλκιβιάδης

Τμήμα Μηχανικών Πληροφορικής ΤΕ

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ & ΘΡΗΣΚΕΥΜΑΤΩΝ, ΠΟΛΙΤΙΣΜΟΥ & ΑΘΛΗΤΙΣΜΟΥ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ

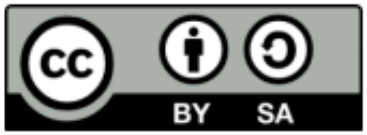
ΕΣΠΑ 2007-2013
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.

- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.



Χρηματοδότηση

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

Ενότητα 15

Εξόρυξη Γνώσης (Data Mining)

Δρ. Τσιμπίρης Αλκιβιάδης

Περιεχόμενα ενότητας

- Εξόρυξη Δεδομένων (DATA MINING)
- KDD ΚΑΙ DATA MINING
- Μέθοδοι DATA MINING
- Συσταδοποίηση (CLUSTERING)
- Ιεραρχικό , Διαμεριστικό ΚΑΙ Ασαφές CLUSTERING
- Διαμεριστικοί Αλγόριθμοι
- K-means clustering

Σκοποί ενότητας

Ο όρος εξόρυξη δεδομένων αναφέρεται στην διαδικασία ανάλυσης μεγάλων βάσεων δεδομένων για εύρεση χρήσιμων μοτίβων. Σχετίζεται με την ανακάλυψη της γνώσης σε βάσεις δεδομένων. Στην ενότητα αυτή γίνεται μια αναφορά σε βασικές τεχνικές εξόρυξης δεδομένων όπως συσταδοποίηση, κατηγοριοποίηση, δέντρα αποφάσεων και κανόνες συσχέτισης. Παρουσιάζονται στον σπουδαστή νέα επιστημονικά πεδία ερευνητικής και επαγγελματικής δράσης και γίνεται επεξήγηση αντιπροσωπευτικών αλγορίθμων.

Εισαγωγή στην KDD

- Στις μέρες μας και σε ένα πολύ μεγάλο εύρος πεδίων η συλλογή, καταγραφή και επεξεργασία δεδομένων γίνεται με δραματικούς ρυθμούς. Δημιουργούνται συνεχώς θεωρίες και εργαλεα τα οποία βοηθούν στην συλλογή χρήσιμων πληροφοριών από τα ολοένα αυξανόμενα δεδομένα. Αυτές οι θεωρίες και τα εργαλεία είναι το αντικείμενο της Ανακάλυψης Γνώσης σε Βάσεις Δεδομένων γνωστή και ως KDD (Knowledge Discovery in Databases). Η KDD διαδικασία ασχολείται με την δημιουργία και μελέτη μεθόδων και τεχνικών που εφαρμόζονται στα δεδομένα με σκοπό την εξαγωγή χρήσιμων πληροφοριών.

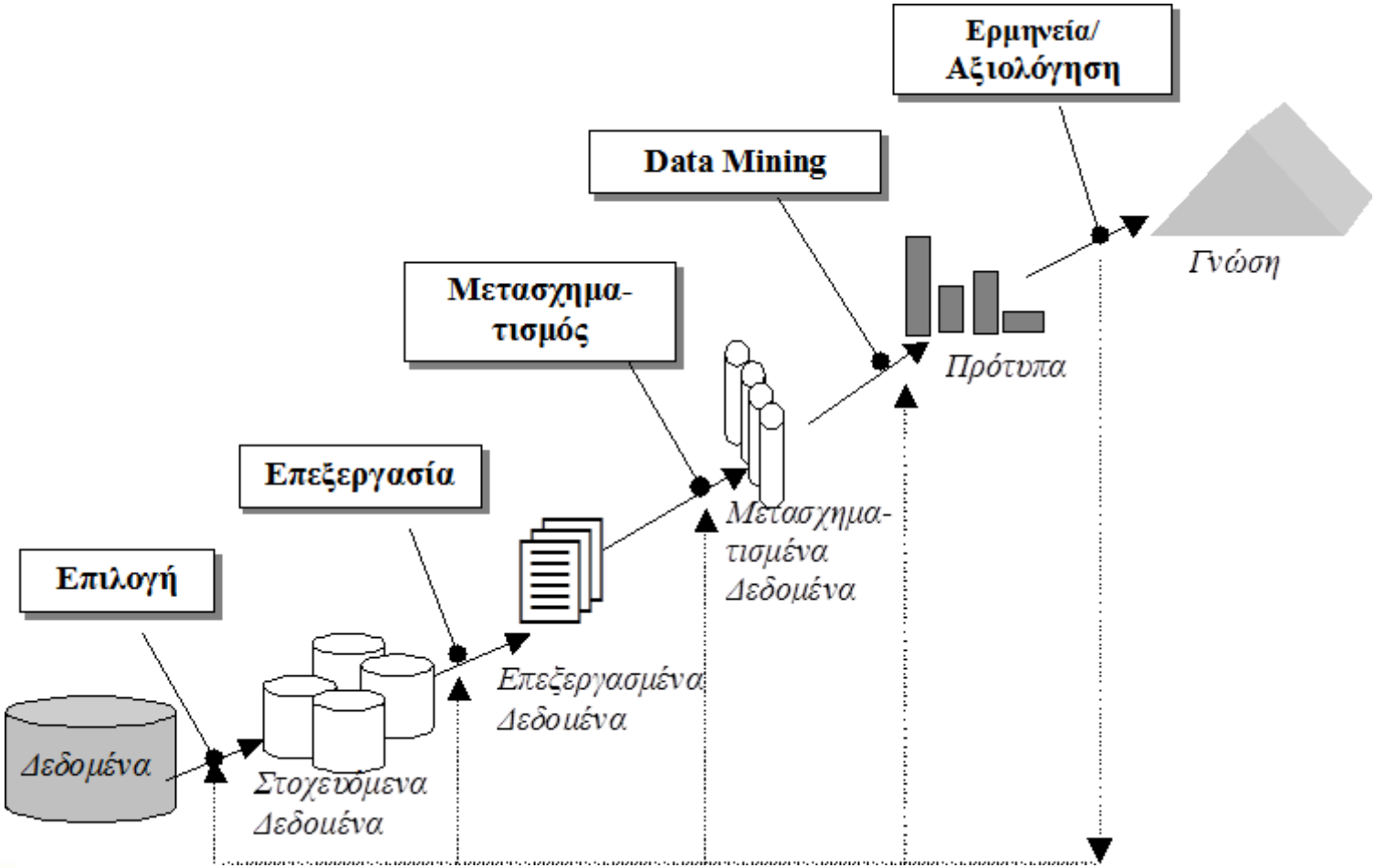
Χρήση των τεχνικών KDD

- Marketing
- Επενδύσεις (Investment)
- Fraud Detection
- Επικοινωνίες και Data Cleaning
- Ιατρική

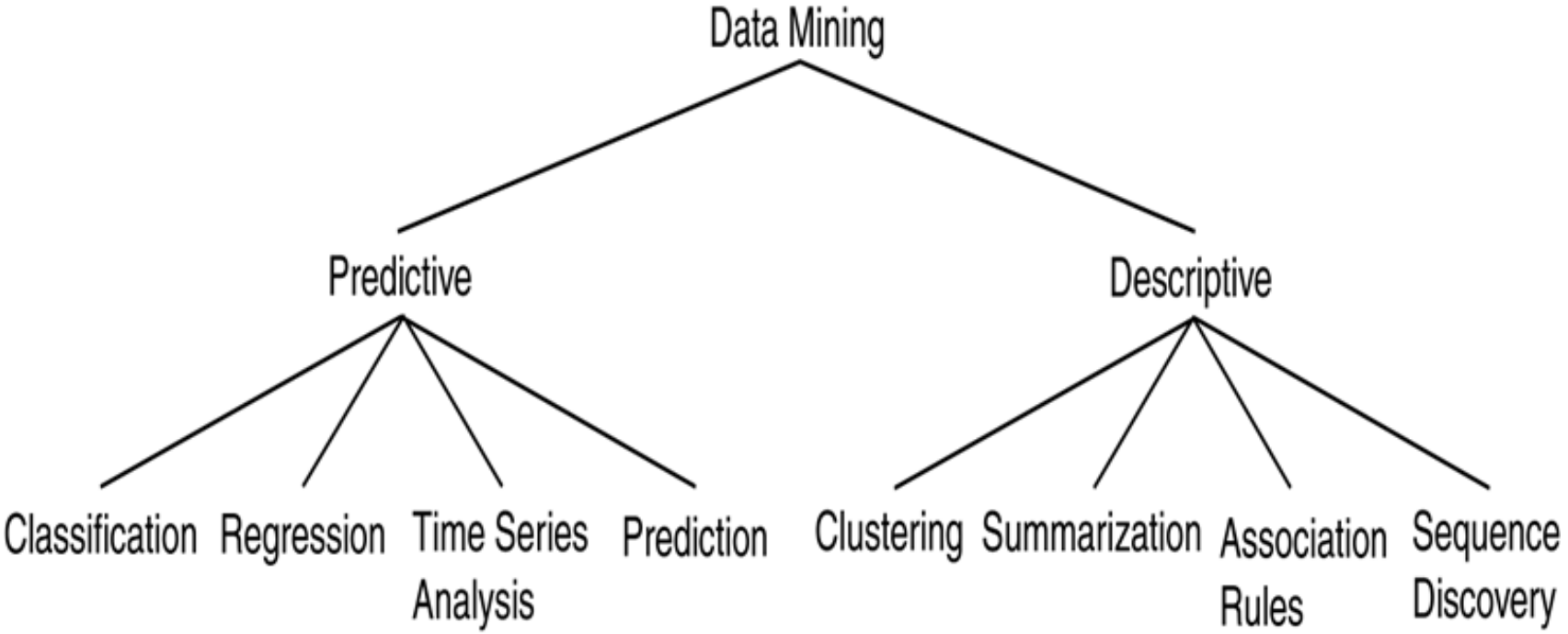
KDD - Data Mining

- Γενικά ο όρος KDD αναφέρεται στη συνολική διαδικασία την εύρεσης γνώσης από σύνολα δεδομένων και ο όρος Data Mining αναφέρεται σε ένα μέρος αυτής της διαδικασίας. Data Mining είναι η εφαρμογή συγκεκριμένου αλγορίθμου για την εξαγωγή μοντέλου από τα δεδομένα.
- Επιπλέον βήματα στην KDD διαδικασία είναι: προετοιμασία των δεδομένων, επιλογή δεδομένων, καθαρισμός δεδομένων (data cleaning), ενσωμάτωση κατάλληλης γνώσης και σωστή μετάφραση και μελέτη των αποτελεσμάτων

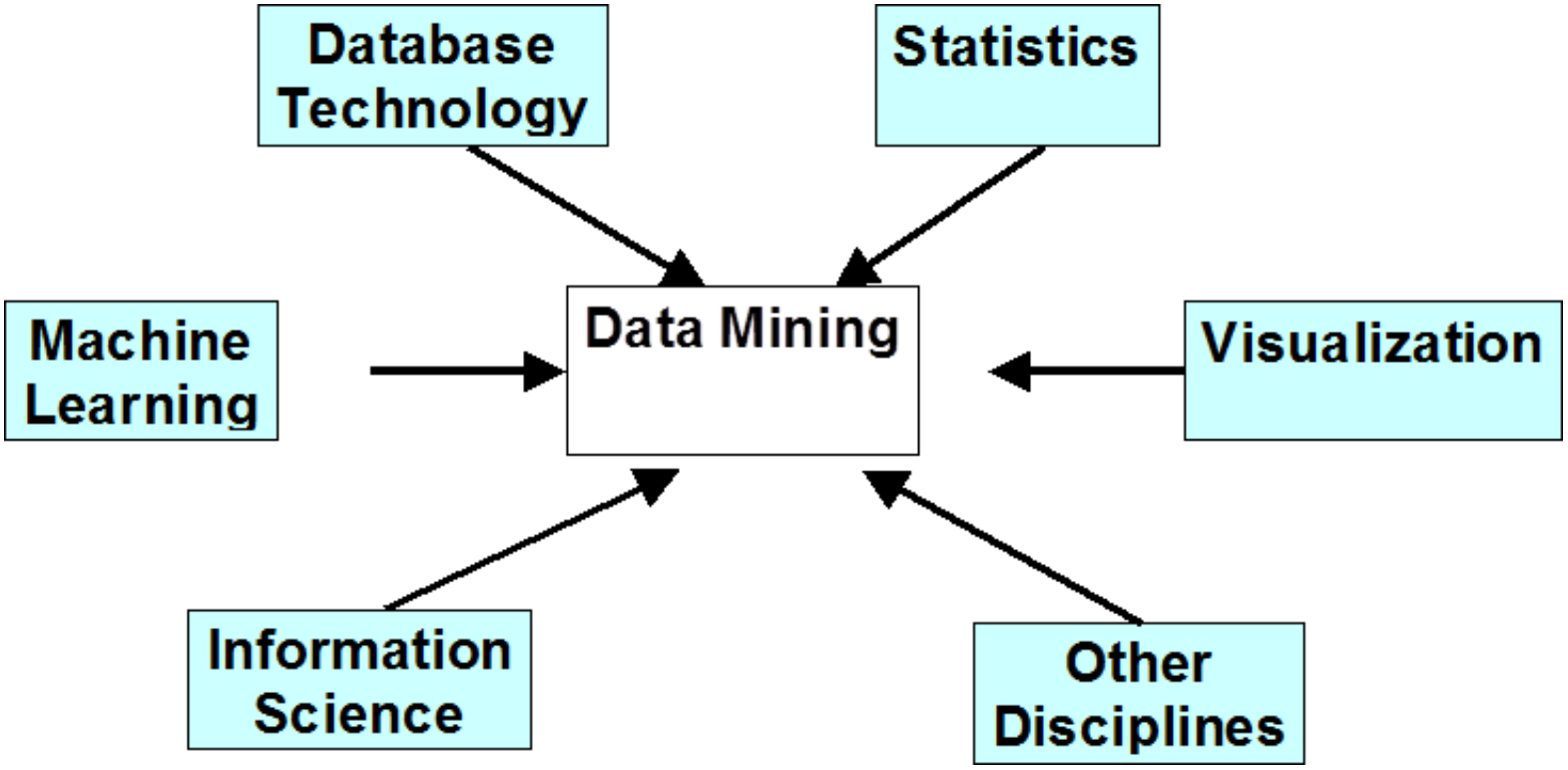
KDD - Data Mining



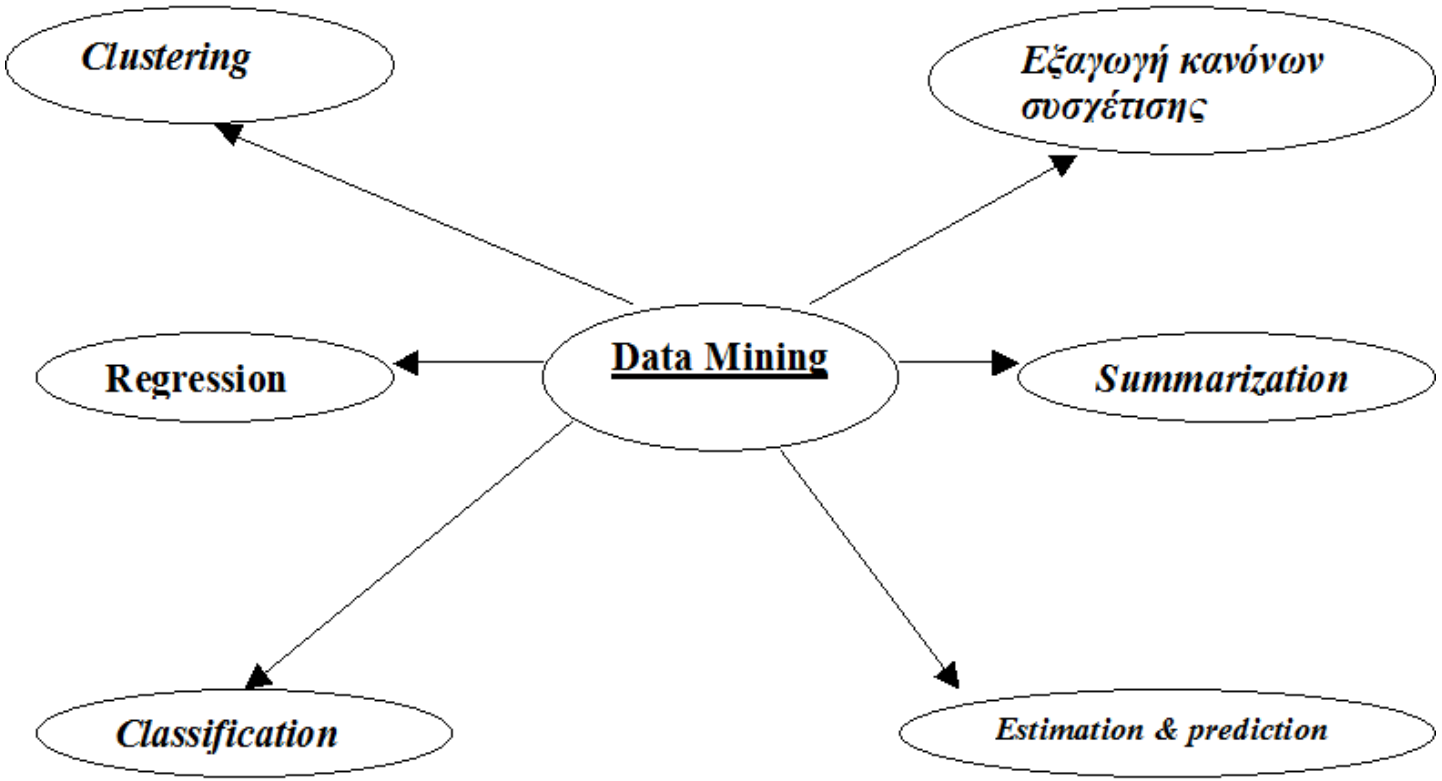
Data Mining


Data Mining
Predictive
Descriptive
Classification
Regression
Time Series Analysis
Prediction
Clustering
Summarization
Association Rules
Sequence Discovery

Data Mining


Database Technology
Statistics
Machine Learning
Data Mining
Visualization
Information Science
Other Disciplines

Data Mining


Clustering
Εξαγωγή κανόνων συσχέτισης
Regression
Data Mining
Summarization
Classification
Estimation & prediction

Συσταδοποίηση- κατηγοριοποίηση

Clustering (συσταδοποίηση)

Το clustering είναι η εργασία του μερισμού ενός συνόλου δεδομένων σε ομάδες ομοίων στοιχείων, clusters. Τα δεδομένα *ομαδοποιούνται σε σύνολα με βάση κάποιο κριτήριο ομοιότητας.* Το clustering *δεν βασίζεται σε προκαθορισμένες κλάσεις*.

Classification (κατηγοριοποίηση)

Η διαδικασία κατηγοριοποίησης των δεδομένων σε κάποια από τις προκαθορισμένες κλάσεις. Συχνά η διαδικασία του classification περιγράφεται σαν μία συνάρτηση μάθησης (learning function), η οποία ταξινομεί (classifies) κάθε αντικείμενο του συνόλου δεδομένων σε μία από τις προκαθορισμένες κατηγορίες. Η διαδικασία του classification χαρακτηρίζεται από: Ένα σύνολο καλά ορισμένων κατηγοριών, ένα training set . ***Στόχος:*** Ο ορισμός ενός μοντέλου το οποίο μπορεί να κατηγοριοποιεί νέα δεδομένα από ένα test set

Κανόνες συσχέτισης - εκτίμηση και πρόβλεψη

Εξαγωγή κανόνων συσχέτισης (association rules extraction)

Προσδιορισμός και εξαγωγή των συσχετίσεων ή προτύπων τα οποία υπάρχουν σε μία συλλογή αντικειμένων. Τα πρότυπα μπορούν να εκφραστούν με κανόνες, των οποίων η γενική μορφή είναι ***"If X then Y".*** Κριτήρια εγκυρότητας και σημαντικότητας κανόνων: *support factor, confidence factor*

Estimation & prediction (εκτίμηση και πρόβλεψη).

Περιλαμβάνει τεχνικές εκτίμησης και πρόβλεψης μελλοντικών *τάσεων ή τιμών.* Ο στόχος εδώ είναι να κατασκευάσουμε ένα μοντέλο που θα επιτρέπει την τιμή μιας μεταβλητής να προβλεφθεί από τις γνωστές τιμές άλλων μεταβλητών

Παλινδρόμηση - Συνάθροιση

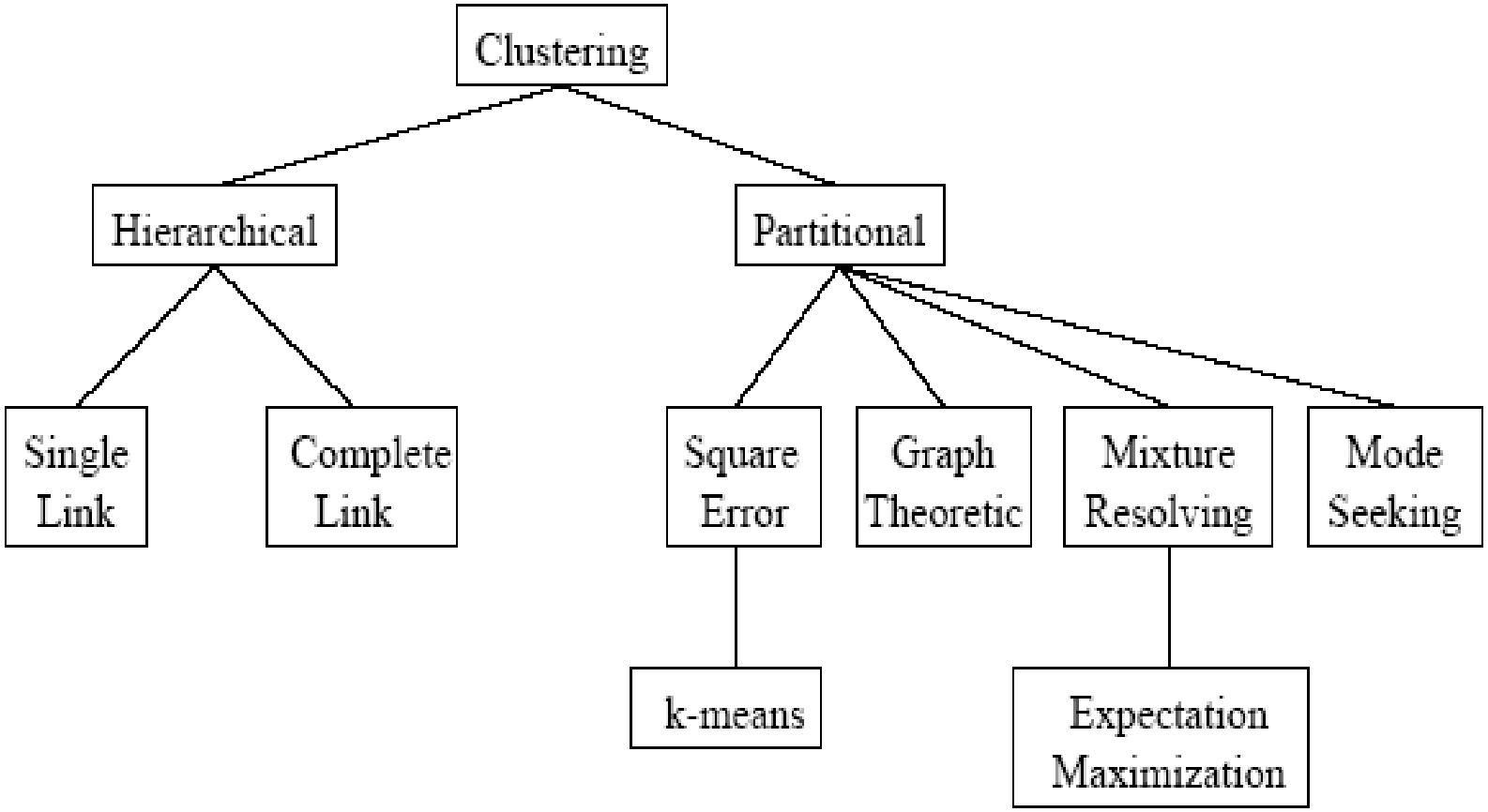
Regression (παλινδρόμηση).

Αντιστοιχεί τα αντικείμενα από ένα σύνολο δεδομένων στην τιμή μίας μεταβλητής πρόβλεψης

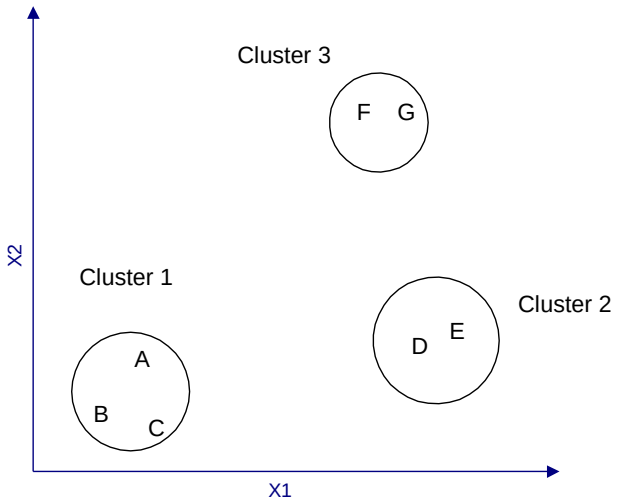
Summarization (Συνάθροιση)

Περιλαμβάνει μεθόδους για την περιγραφή ενός υποσυνόλου δεδομένων. Π.χ. η εκτίμηση της μέσης και της τυπικής απόκλισης για όλα τα πεδία, reports, τεχνικές παρουσίασης, την παραγωγή συνοπτικών κανόνων.

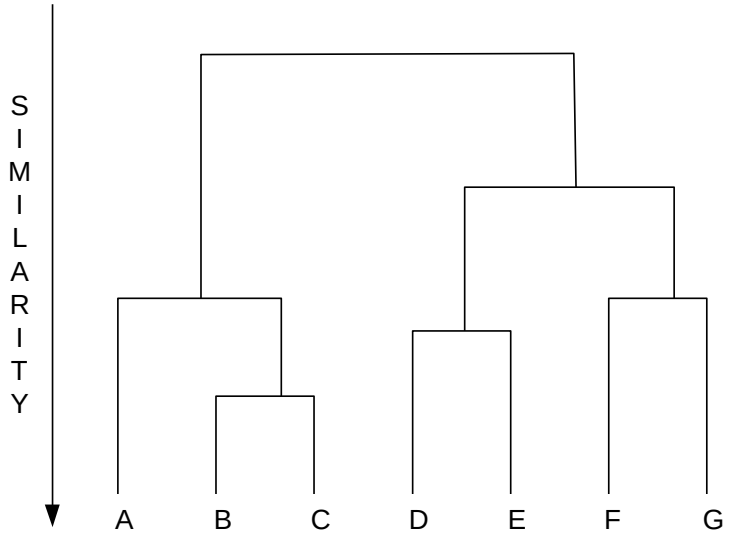
Clustering


Clustering
Hierarchical
Partitional
Single Link
Complete Link
Square Error
Graph Theoretic
Mixture Resolving
Mode Seeking
k-means
Expectation Maximization

Clustering



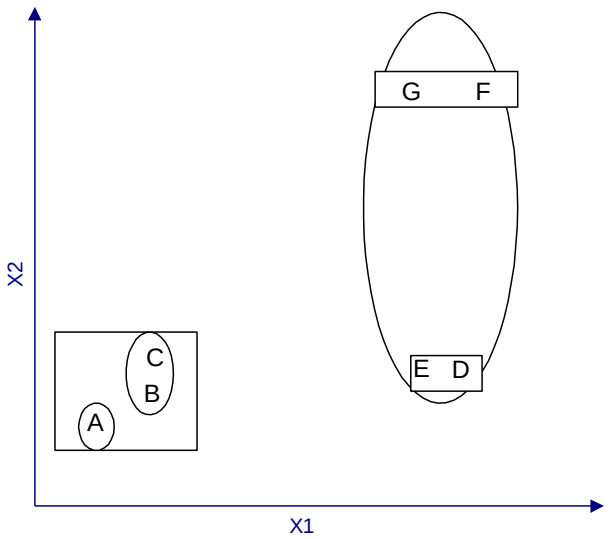
Σημεία σε τρία clusters



Δενδοδιάγραμμα ιεραρχίας

K-means clustering

1. Επιλογή k κεντροειδών cluster τα οποία αποτελούν και τα μόνα στοιχεία των k επελεγμένων clusters.
2. Τοποθέτησε κάθε στοιχείο στο πιο κοντινό cluster μετά από υπολογισμό της απόστασης του σημείου από το κεντροειδές του cluster.
3. Υπολόγισε το νέο κεντροειδές.
4. Αν το κριτήριο τερματισμού δεν ικανοποιείται πήγαινε στο βήμα 2.



Ευαισθησία του αλγορίθμου k-means στην αρχική επιλογή clusters.