Types of clustering methods:

1. Partitioning - Distance based, effective for small to medium data, simple. Brute force approach, heuristic methods, given n objects
2. Hierarchical - Dendrogram, agglomerative, decisive
3. Grid-based - Fast processing time, can be integrated with other clustering methods
4. Density - Can find arbitrary shapes, noise tolerant, single scan, adjustable density parameters, local clusters with high density
5. Probabilistic - Hidden categories/cluster models, mixture models

Classification methods

1. Decision Tree Induction - Top-down, recursive, attribute selection and split
2. Bayesian Classification - Probability, naive assumption, belief network
3. Support Vector Machines - SVM - Objects with class labels, classification for both linear and nonlinear. Separating hyperplane: maximum margin, max margin hyperplane, support vector
4. Neural Networks - Connected input/output units, each connection has a weight associated with it.

Model Evaluation

- Holdout - The given data are randomly partitioned into two independent sets, a training and test set.
- Random sampling A variation of the holdout method in which the holdout method is repeated k times.
- K-folds cross validation - The initial data are randomly partitioned into k mutually exclusive subsets
- Bootstrapping - Samples the given training tuples uniformly with replacement

Anomaly Detection methods

- Statistical methods - Makes assumptions of data normality. Data not following the model are outliers
- Proximity-based methods - Assume that an object is an outlier if the nearest neighbors of the object are far away in feature space
- Clustering-based methods - Assumes that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or do not belong to any clusters.
  OR
- Supervised methods - Model data normality and abnormality. Domain experts examine and label a sample of the underlying data. Outlier detection can then be modeled as a classification problem
- Semi-supervised methods - Where Only a small amount of data is labeled.
- Unsupervised methods - It expects that normal objects follow a pattern far more frequently than outliers.

Types of outliers/anomalies

- Global - It deviates significantly from the rest of the data set. Simplest type of outlier.
- Contextual - Deviatessignificantly with respect to a specific context of the object.
- Collective - A subset of data objects forms a collective outlier if the objects as a whole deviate significantly from the entire data set. The individual data objects may not be outliers.

Apriori Algorithm Challenges - Multiple scans of the whole dataset. Huge number of candidates

Apriori Algorithm Improvements

- Partitioning - Partitioning the data to find candidate itemsets
- Sampling - Mining on a subset of the given data
- Transaction reduction - Reducing the number of transactions scanned in future iterations

Monotonic - if an itemset satisfies the rule constraint so do all its supersets.
Antimonotonic - If an itemset does not satisfy the rule constraint none of its supersets can satisfy the constraint.
Supervised learning classification - Predefined classes, training data with ground truth label.
Unsupervised learning classification - Clustering, no predefined classes, Items to identify potential clusters/patterns.
FP Growth - method to mine frequent itemsets within a database without generating candidate sets explicitly.uses a divide-and-conquer strategy.

Clustering with Expectation Maximization (EM) - good performance in many applications, easy to implement, converges quickly, may not be optimal, not good if objects are small. Computation-intensive for large number of clusters.

Information Gain - The difference between the original info requirement and new requirement. Gain(A) = Info(D) - InfoA(D)
Association rule - patterns that reflect items that are frequently associated together. Association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

Correlation analysis - Given two attributes, correlation analysis can measure how strongly one attribute implies the other, based on the available data. X^2 (chi-square) - used with nominal data. Correlation coefficient and covariance - used with numerical attributes.