Statistics 202: Data Mining

Clustering
Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

December 2, 2012
Clustering

- Goal: Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

- An unsupervised problem that tries to produce “labelled” data from “unlabelled” data.

- Many different techniques, but most revolve around forming groups with small “within cluster” distances relative to “between cluster” distances.
Cluster analysis

What is Cluster Analysis?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

Inter-cluster distances are maximized.

Intra-cluster distances are minimized.

Inter-cluster distances are maximized.
Clustering

Applications

- Understanding of some structure in a dataset.
  - Group related documents for browsing;
  - Group genes and proteins that have similar functionality;
  - Group stocks with similar price fluctuations

- Summarization: reduce the size of large data sets
  (sometimes known as vector quantization . . . )
Cluster analysis

FIGURE 14.14. DNA microarray data: average linkage hierarchical clustering has been applied independently to the rows (genes) and columns (samples), determining the ordering of the rows and columns (see text). The colors range from bright green (negative, under-expressed) to bright red (positive, over-expressed).
Cluster analysis

Original image . . .
Cluster analysis

Some compression . . .
Cluster analysis

Too much compression?
Cluster analysis

Clustering can be ambiguous. How many clusters?

- Four Clusters
- Two Clusters
- Six Clusters

Clusters can be ambiguous.
Clustering

Types of clustering

**Partitional**  A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

**Hierarchical**  A set of nested clusters organized as a hierarchical tree. Each data object is in exactly one subset for any horizontal cut of the tree . . .
Cluster analysis

Figure 14.4. Simulated data in the plane, clustered into three classes (represented by orange, blue and green) by the $K$-means clustering algorithm that at each level of the hierarchy, clusters within the same group are more similar to each other than those in different groups.

Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consists of a set distinct subgroups, each group representing objects with substantially different properties. This latter goal requires an assessment of the degree of difference between the objects assigned to the respective clusters.

Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it. This can only come from subject matter considerations. The situation is somewhat similar to the specification of a loss or cost function in prediction problems (supervised learning). There the cost associated with an inaccurate prediction depends on considerations outside the data.

Figure 14.4 shows some simulated data clustered into three groups via the popular $K$-means algorithm. In this case two of the clusters are not well separated, so that "segmentation" more accurately describes the part of this process than "clustering.

$K$-means clustering starts with guesses for the three cluster centers. Then it alternates the following steps until convergence:

1. For each data point, the closest cluster center (in Euclidean distance) is identified;
2. Recompute the cluster center as the mean of all data points in the cluster;
3. Repeat steps 1 and 2 until convergence.

A partitional example
Cluster analysis

A hierarchical example
Clustering

Other distinctions

**Exclusivity**  Are points in only one cluster?

**Soft vs. hard**  Can we give a “score” for each case and each cluster?

**Partial vs. complete**  Do we cluster all points, or only some?

**Heterogeneity**  Are the clusters similar in size, shape, etc.
Types of clusters: well-separated

This type of cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
This type of cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster.
Types of clusters: center-based

The center of these clusters is usually, the average of all the points in the cluster, or a medoid, the most “representative” point of a cluster.
Types of clusters: contiguity-based

This type of cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.
Types of clusters: contiguity-based

These types of clusters are made up of dense regions of points, often described by one of the other cluster types, but separated by low-density regions, often in the form of noise.
Clustering

Mathematical characterizations

- Most clustering algorithms are based on a dissimilarity measure $d$.
- Data may be of mixed type so some of the similarities we saw earlier may be used.
- Most clustering algorithms do not insist that the dissimilarity is truly a distance.
- *(Local)* Given two elements of the partition $C_r, C_s$, we might consider

$$d_{SL}(C_r, C_s) = \min_{x \in C_r, y \in C_s} d(x, y)$$

These types of measures are often used in *hierarchical clustering* algorithms.
Clustering

Mathematical characterizations

- (Global) A clustering is a partition $C = \{C_1, \ldots, C_k\}$ of the cases. It writes

$$T = \frac{1}{2} \sum_{i,j=1}^{n} d^2(x_i, x_j)$$

$$= W(C) + B(C)$$

and tries to minimize $W(C)$.

- This problem is NP Hard . . .

- These types of measures are often used in *partitional clustering* algorithms.
Clustering

Most common models

Partitional $K$-means / medoid ; mixture models
Hierarchical agglomerative (bottom-up) hierarchical clustering.

We’ll spend a some time on each of these . . .
Cluster analysis

Cluster analysis is also used to form descriptive statistics to ascertain whether or not the data consists of a set distinct subgroups, each group representing objects with substantially different properties. This latter goal requires an assessment of the degree of difference between the objects assigned to the respective clusters.

Central to all of the goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects being clustered. A clustering method attempts to group the objects based on the definition of similarity supplied to it. This can only come from subject matter considerations. The situation is somewhat similar to the specification of a loss or cost function in prediction problems (supervised learning). There the cost associated with an inaccurate prediction depends on considerations outside the data.

Figure 14.4 shows some simulated data clustered into three groups via the popular \(K\)-means algorithm. In this case two of the clusters are not well separated, so that "segmentation" more accurately describes the part of this process than "clustering."

\(K\)-means clustering starts with guesses for the three cluster centers. Then it alternates the following steps until convergence:

1. for each data point, the closest cluster center (in Euclidean distance) is identified;

... steps until convergence:
Cluster analysis

FIGURE 14.12. Dendrogram from agglomerative hierarchical clustering with average linkage to the human tumor microarray data.

Technical structure produced by the algorithm. Hierarchical methods impose hierarchical structure whether or not such structure actually exists in the data. The extent to which the hierarchical structure produced by a dendrogram actually represents the data itself can be judged by the cophenetic correlation coefficient. This is the correlation between the $N(N-1)/2$ pairwise observation dissimilarities $d_{ii}'$ input to the algorithm and their corresponding cophenetic dissimilarities $C_{ii}'$ derived from the dendrogram. The cophenetic dissimilarity $C_{ii}'$ between two observations $(i, i')$ is the intergroup dissimilarity at which observations $i$ and $i'$ are first joined together in the same cluster.

The cophenetic dissimilarity is a very restrictive dissimilarity measure. First, the $C_{ii}'$ over the observations must contain many ties, since only $N-1$ of the total $N(N-1)/2$ values can be distinct. Also these dissimilarities obey the ultrametric inequality

$$C_{ii}' \leq \max\{C_{ik}, C_{i'k}\} \quad (14.40)$$