Lecture 19 Nov. 1 -2006

Data Visualization

STAT 442 / 890, CM 462

Lecture: Ali Ghodsi Scribes: Stefan Pintilie

1 Clustering

Clustering (also known as segmentation) is the task of grouping a collection of objects into groups or clusters such that objects within each cluster are more similar than objects assigned to different clusters.

The notion of similarity (or dissimilarity) is central in all clustering algorithms. We usually use distance as a measure of dissimilarity.

What is a distance measure? We can say that a distance measure must satisfy the following three properties:

1. $d_{ij} \geq 0$. All distances are greater or equal to zero.

2. $d_{ij} = d_{ji}$. The distance from i to j is the same as the distance from j to i.

3. $d_{ij} = 0$ if $(x_i = x_j)$.

If the Triangular Inequality $d_{ij} \leq d_{ik} + d_{kj}$ is also satisfied then we can call this distance measure a metric.

A common distance is $||x_i - x_j||^2$. This is called $l2$. This is a popular distance measure because we can easily take the derivative of this distance. An alternative to this distance measure is called $l1$ and is:

$$l|x_i - x_j| = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|$$

Of course, there are many other distance measures that satisfy the three conditions listed above.

Clustering is an ill defined problem. Suppose there exists a clustering function f which takes distance D and returns partitions Γ. Distance D is defined on set S and Γ is a partitioning on this set.

$$\Gamma = f(D)$$

In order for this to be a good clustering algorithm we would like to satisfy a number of conditions. The following three conditions are considred desirable properties of a clustering algorithm.

1. Scale Invariant. We want for any scale α that $f(D) = f(\alpha D)$.

2. Richness. Any possible partitioning Γ on S should be a possible outcome of f.

3. Consistency. Γ is the outcome of f when applied to D, $(\Gamma = f(D))$. Now consider that we produce D' such that the distances between all points in each cluster are reduced and distances between points assigned to different clusters are enlarged then $\Gamma = f(D')$.

Impossibility Theorem:

It is impossible to satisfy all three of these conditions at the same time. It is possible to satisfy two of the above conditions at a time but there does not exist an algorithm that can satisfy all three.

There are three different types of clustering algorithms.

1. Combinatorial Algorithms.

2. Mixture Models.

3. Model Based.

Suppose we have n data points that are indexed $1, 2, \ldots, n$. Suppose we need K clusters. Take $i \epsilon \{1, 2, \ldots n\}$ and $k \epsilon \{1, 2, \ldots K\}$. We need to assign each point to one cluster. $k = C(i)$. Where C is called an encoder.

The goal is to find a grouping of data such that the distances between points within a cluster tend to be small and distances between points in different clusters tend to be large.