WORKSHOP II SOLUTION - POINT 1

SPECTRAL CLUSTERING

a) In which cases might it be more useful to apply?

Spectral clustering is most useful for application in few clusters, uniform cluster size, non-planar geometry. It has its application in many areas including image segmentation, educational data mining, entity resolution, speech separation, spectral clustering of protein sequences, text image segmentation.

As spectral partitioning allows processing large-scale data, this method is often used for marketing purposes. Companies use this algorithm to segment their targets according to their expectations, needs, profile, maturity level, etc.

b) What are the mathematical fundamentals of it?

In spectral clustering, the data points are treated as nodes of a graph. Thus, clustering is treated as a graph partitioning problem. The nodes are then mapped to a low-dimensional space that can be easily segregated to form clusters. An important point to note is that no assumption is made about the shape/form of the clusters.

Given an enumerated set of data points, the similarity matrix can be defined as a symmetric matrix A, where $A_{ij} \geq 0$ represents a measure of similarity between data points with indices i and j. The general approach to spectral clustering is to use a standard clustering method on the relevant eigenvectors of a Laplacian matrix of A. There are many ways to define a Laplacian that have different mathematical interpretations, so the clustering will also have different interpretations. The eigenvectors that are relevant are those that correspond to several smaller eigenvalues of the Laplacian except for the smallest eigenvalue which will have a value of 0. For computational efficiency, these eigenvectors are often computed as the eigenvectors corresponding to the several largest eigenvalues of a function of the Laplacian.

MATHEMATICAL FUNDAMENTALS:

Graph distance.

Linear algebra:

Adjacency and Affinity Matrix (A): square matrix used to represent a finite graph.

Degree Matrix (D): diagonal matrix, where the degree of a node (i.e. values) of the diagonal is given by the number of edges connected to it.

Laplacian Matrix (L) = D – A: Spectral clustering is well known to relate to partitioning of a mass-spring system, where each mass is associated with a data point and each spring stiffness corresponds to a weight of an edge describing a similarity of the two related data points, as in the spring system.

Eigenvectors:

The cost of computing the n-by-k (with $k \ll n$) matrix of selected eigenvectors of the graph Laplacian is normally proportional to the cost of multiplication of the n-by-n graph Laplacian matrix by a vector, which varies greatly whether the graph Laplacian matrix is dense or sparse.

c) What is the algorithm to compute it?

Though spectral clustering is a technique based on graph theory, the approach is used to identify communities of vertices in a graph based on the edges connecting them. This method is flexible and allows us to cluster non-graph data as well either with or without the original data.

We assume that our data consists of n "points" x1,...,xn which can be arbitrary objects. We measure their pairwise similarities sij = s(xi, xj) by some similarity function which is symmetric and non-negative, and we denote the corresponding similarity matrix by S = (sij)i,j=1...n.

Unnormalized spectral clustering:

Input: Similarity matrix $S \in n \times n$, number k of clusters to construct.

1. Transform the distance matrix into a graph using the affinity matrix A (or similarity matrix, adjacent matrix).
2. Compute the degree matrix D and the Laplacian matrix L = D – A.
3. Find the eigenvalues and eigenvectors of L.
4. With the eigenvectors of k largest eigenvalues computed from the previous step form a matrix.
5. Normalize the vectors.
6. Cluster the data points in k-dimensional space.

d) Does it hold any relation to some of the concepts previously mentioned in class? Which, and how?

Spectral clustering is closely related to nonlinear dimensionality reduction. It is similar to the TSNE algorithm in that both construct a similarity matrix, project the data into a lower dimensional space and cluster them. The difference is that in spectral clustering points that are connected or immediately next to each other are placed in the same cluster. Even if the distance between 2 points is smaller, if they are not connected, they are not clustered.