CSE 347/447 Data Mining: Homework 2

Due on 11:59 PM, September 36, 2017

Submission Instructions

The submission should be one .zip file containing: one .pdf file with your answers to Q1 ~ Q3, and one .py file with your answer to Q4. Please name your file as "HW2.zip".

Q1 (5pt)

Hierarchical clustering is sometimes used to generate K clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: $\{6, 12, 18, 24, 30, 42, 48\}$.

(a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i. $\{18, 45\}$ ii. $\{15, 40\}$

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

(c) What are the two clusters produced by single link?

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

(e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)

(f) What well-known characteristic of the K-means algorithm explains the previous behavior?

Q2 (3pt) Prove Equation 8.15 in TSK (our textbook).

CSE 347/447 Data Mining: Homework 2

Q3 (6pt)

On the space of nonnegative integers, which of the following functions are distance measures? If so, prove it; if not, prove that it fails to satisfy one or more of the axioms.

(a) $\max(x, y) =$ the larger of x and y.

(b) $\text{diff}(x, y) = |x - y|$ (the absolute magnitude of the difference between x and y).

(c) $\text{sum}(x, y) = x + y$.

Q4 (6pt) (Programming)

In this exercise, we will try to build a simple meta-clustering algorithm that combines the results of multiple individual clustering algorithms. Specifically, two data points are clustered together *only if* all individual clustering algorithms say they should belong to the same cluster.

The individual clustering algorithms we consider include: *K-Means*, *Ward hierarchical clustering*, and *DBSCAN*. We will visualize the clustering results to examine the difference.

data.txt: data points to be clustered. Each row represents a 2-D data point. The dataset consists of 2K points. The number of real clusters is 5.

meta_cluster.py: a partially finished code snippet. Necessary libraries have been included. The setting of *K-Means, Ward*, and *DBSCAN* is provided.

You are required to finish *meta_cluster.py* such that it performs ensemble clustering and plots out the results. You can run the code by: python *meta_cluster.py < data.txt*

Hints: (1) each algorithm returns the cluster labels of data points. However, they may be inconsistent even if they agree with each other; for example, one may label two points as cluster 3, while the other may label them as cluster 4. (2) We may consider each data point as a node in an undirected graph, two nodes are adjacent if and only if they belong to the same cluster; thus, each connected component in the graph corresponds to one distinct cluster.