CSE 881: Data Mining (Fall 2013) Homework 5

Due date: Nov 21, 2013

1. k-means does not always converge to a globally optimal solution as it is sensitive to the choice of initial centroids. The initial centroids are typically chosen **by randomly sample k of the data points** to be clustered.

   (a) If the data set contains 100 data points, how many distinct ways are there to initialize the cluster centroids? Assume that the number of clusters is equal to 5.

   (b) If the data set contains 100 data points, how many distinct clustering solutions are there (assuming there are 5 clusters)?

   (c) If the data set contains 100 data points, how many times do you need to repeat the k-means algorithm (each time with a different set of initial centroids) to ensure that there is a 30% chance the algorithm will converge to a globally optimal solution? Assume that the number of clusters is equal to 5. Note: you may assume each initial centroid configuration yields a different clustering solution.

   (d) If the number of data points increases from 100 to 200 (but number of clusters is still 5), will it improve or diminish your chance of finding the optimal k-means solution by randomly choosing a subset of the data points to be the initial centroids?

   (e) If the number of data points is fixed but the number of clusters (and number of initial centroids) increases, will it generally improve or diminish your chances of finding an optimal k-means solution by randomly choosing a subset of the data points to be the initial centroids?

2. You have been hired as a data mining consultant for a large automobile company. Your first task is to apply clustering to segment the customers who bought a vehicle from the company. The customer data set contains only 10 categorical attributes (gender, marital status, occupation, highest education level, etc).

   (a) You had recalled how entropy can be used to split data points with categorical attributes into homogeneous classes (see lecture 7 on decision stumps). So you decide to develop an entropy-based divisive hierarchical clustering algorithm to handle the categorical attributes. Your algorithm starts with 1 cluster containing all the points. It then selects the best attribute to partition the cluster into two smaller clusters in a way that minimizes the total entropy. It then repeats this partitioning step on each cluster until every cluster contains only a single data point. You present the idea to your supervisor, who immediately shoots it down. Explain why such an approach conceptually will not work.

(b) Since clustering algorithms such as k-means works with continuous variables, you decide to binarize the data (i.e., transform each categorical attribute with k possible values into k binary attributes). You then apply the single-link (MIN) hierarchical clustering with Euclidean distance to segment the customers, but your supervisor argued that you should be using cosine similarity instead since the attributes are asymmetric binary. Your colleague, however, told you it does not matter whether you use Euclidean distance or cosine similarity as the proximity measure because the results you get should be the same when applied to the binarized data you had created. Who should you trust, your colleague or your supervisor? Why?

(c) After applying the hierarchical clustering algorithm on the binarized customer data, you show the results to your supervisor who seems disappointed because the clusters do not make any sense. Your supervisor asks you to re-check your results. You noticed that the data actually contains a lot of missing values coded as "N/A" (not applicable). Explain how the missing values affect the results of your hierarchical clustering algorithm. Explain how would you could handle the missing value problem without estimating the missing value nor discarding the data points with missing values.

(d) After addressing the missing value problem, you re-run the single-link hierarchical clustering algorithm on the binarized data and show the dendrogram to your supervisor. Your supervisor is still not satisfied because she prefers to be given a final set of clusters (instead of a nested hierarchy of clusters). So you examine the y-axis of the dendrogram and determine where to "cut" the dendrogram to produce an optimal set of clusters. Let $d(k-1)$ be the distance shown on the dendrogram when two of the k clusters are merged to produce $k-1$ clusters. For example, consider the dendrogram for 6 data points as shown in Figure 1. For the dendrogram shown, d(1) = 0.245, d(2) = 0.161, d(3) = 0.149, d(4) = 0.115, and d(5) = 0.080. You plan to use the gap between $d(k-1)$ and $d(k)$ to determine the number of clusters. Since the gap between d(1) and d(2) is largest, you decided to cut the dendrogram at this point and create 2 clusters. For the binarized data, what are are the minimum and maximum possible values for the gap, $d(k-1) - d(k)$? Explain the limitation of using the widest gap between $d(k-1)$ and $d(k)$ to determine the right number of clusters for this data set.

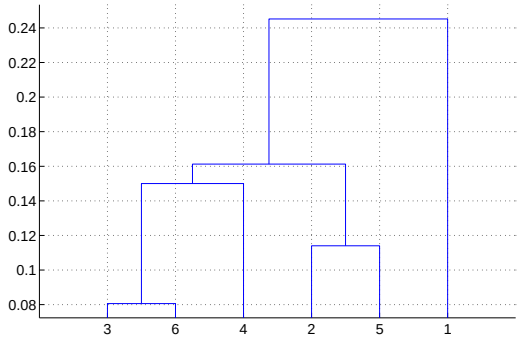

Figure 1: Dendrogram of hierarchical clustering.

3. Consider the following set of one-dimensional data points:

$$0.6,\ \ 1.2,\ \ 1.8,\ \ 2.4,\ \ 3.0,\ \ 4.2,\ 4.8$$

   (a) Suppose we apply kmeans clustering to obtain two clusters. If the initial centroids are located at 1.8 and 4.5, show the cluster assignments and locations of the centroids after the algorithm converges. Compute the total sum-of-squared errors of the clusters.

   (b) Repeat the previous question using 1.5 and 4.0 as the initial centroids. Show the cluster assignment and locations of centroids after the algorithm converges. Compute the total sum-of-squared errors of the clusters.

   (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.)

4. Consider the data set shown in Figure 2. Suppose we apply DBScan algorithm with Eps = 0.15 (in Euclidean distance) and MinPts = 3.

   (a) List all the core points in the diagram (you can use the labels of the data points in the diagram). Note: a point is considered a core point if there are **more than MinPts** number of points (including the point itself) within a neighborhood of radius Eps.

   (b) List all the border points in the diagram.

   (c) List all the noise points in the diagram.

   (d) Using the DBScan algorithm, how many clusters will be obtained from the data set?

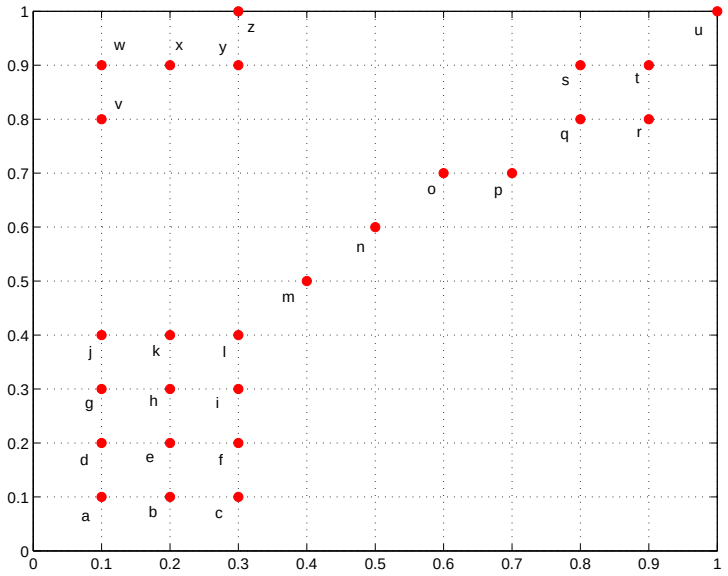

Figure 2: DBScan clustering.

5. **[7 points]**

Consider the confusion matrices for two clustering solutions as shown below, where the rows correspond to the clusters and the columns correspond to the ground truth classes. Note that solution 2 simply partitions the first cluster of solution 1 into two smaller sub-clusters.

	Solution 1	
	Ground truth class	
	Class 1	Class 2
Cluster 1	40	20
Cluster 2	10	30

	Solution 2	
	Ground truth class	
	Class 1	Class 2
Cluster 1	35	15
Cluster 2	5	5
Cluster 3	10	30

Each entry n_{ij} in the matrix corresponds to the number of data points assigned to cluster i that belong to class j. Furthermore, let $n_{i+} = \sum_j n_{ij}$ (i.e., the sum of all entries in row i) be the number of points in cluster i, $n_{+j} = \sum_i n_{ij}$ (i.e., the sum of all entries in column j) be the number of data points that belong to class j, and $N = \sum_{ij} n_{ij}$ (i.e., the sum of all entries in the table) be the total number of data points. In this exercise, you will compare the performance of the two clustering solutions using the following measures:

- Entropy, $e = \sum_i \frac{n_{i+}}{N} e_i$, where $e_i = -\sum_j \frac{n_{ij}}{n_{i+}} \log \frac{n_{ij}}{n_{i+}}$ is the entropy of cluster i

- Normalized mutual information

$$NMI = \frac{2\sum_{i,j} \frac{n_{ij}}{N} \log \frac{n_{ij}N}{n_{i+}n_{+j}}}{H_1 + H_2},$$

where $H_1 = -\sum_i \frac{n_{i+}}{N} \log \frac{n_{i+}}{N}$, and $H_2 = -\sum_j \frac{n_{+j}}{N} \log \frac{n_{+j}}{N}$.

Answer the following questions:

(a) Compute the values of entropy and NMI when the clusters are pure (i.e., contains only data points from a single class). Assume the number of clusters is the same as number of classes (i.e., $k = 2$).

(b) Compute the entropy for both solutions. Which solution is better?

(c) Compute the NMI for both solutions. Which solution is better?

(d) Based on your answers above, state which supervised measure do you think is better and explain why.

6. **[7 points]**

Download the S&P-500 stock market time series data from the class web page. There are 3 files provided:

- *prices.txt*, which contains the normalized closing prices of the stocks from January 1, 2007 until December 31, 2012.
- *sp500.class*, which contains the category ID of each stock. There are 10 distinct categories.
- *classes.txt*, which contains the mapping from category ID to category name.

In this exercise, you will investigate the feasibility of applying k-means clustering algorithm to the data.

(a) Load the prices.txt data into Matlab.

(b) Which proximity measure do you think is more appropriate to cluster the data—Euclidean distance or correlation? Explain why.

(c) Run k-means clustering with $k = 10$ using the proximity measure you had chosen. Type `help kmeans` to determine how to set the appropriate measure.

(d) Compute the 10×10 confusion matrix (using the stock categories as ground truth). Type `help confusionmat` to determine how to create the confusion matrix. Do the clusters correctly represent the stock categories?

(e) Create 3 10×10 matrices, called mincorr, maxcorr, and avgcorr. The mincorr matrix computes the minimum correlation between stocks from one category to another. For example mincorr(4,6) gives the minimum correlation between stocks that belong to categories 4 and

6 while mincorr(5,5) gives the minimum correlation between stocks that belong to category 5 (excluding self-correlation, i.e., correlation between the time series of a stock with itself). Similarly, maxcorr and avgcorr are matrices that encode the corresponding maximum and average correlation between categories. Use the matrices to help explain what you observed in part (d).

Deliverables: You should report the values of the confusion matrix obtained using k-means. Also report the values of the mincorr, maxcorr, and avgcorr matrices.