University of Pittsburgh
CS1675 Introduction to Machine Learning — Handout 10
Professor Milos Hauskrecht — March 28, 2019

Problem assignment 9

Due: Thursday, April 4, 2019

Problem 1. K-means clustering

Assume we have a 2-dimensional space and four points $(0, 0), (0, 5), (6, 7)$ and $(7, 0)$ and we want to cluster the examples into two groups using the k-means algorithm and the Euclidean distance.

Part a. Let us assume the algorithm is initialized with means $(0, 0)$ and $(7, 0)$. What are the values of the two means the algorithm converges to. How are datapoints divided into groups?

Part b. Now let us assume the algorithm is initialized from the means $(3, 3)$ and $(7, 0)$. What are the values of the two means the algorithm converges to. How are datapoints divided into groups?

Problem 2. K-means clustering experiments

Please load the dataset *clustering_data.txt*.

Part a. Run the k-means algorithm (implemented in Matlab in the function kmeans) for finding 3 clusters. Use Euclidean distance to define the differences in between the points. Report the sizes of the three groups found by the kmeans. Use scatter function to plot the data in the dataset and the means of the clusters. Please use colors to distinguish data that were assigned different groups. Use a separate color to show the cluster centers (means). Include the plot in your report.

Part b. Repeat the setup in Part a. but now assume the number of means is 4. Again report the sizes of the groups and plot the data (with different group colors) and the means found by kmeans.

Part c. The kmeans procedure (if initial means seeds are not set) uses a random set of seeds in each run. Rerun kmeans algorithm for $k = 4$ (the same as Part b). The means found are likely to change. If they did not, try to rerun the procedure again till you see the change in the means. Show the scatter plot of the results when the centers changed.

Part d. Let us assume the two runs of the k-means lead to two different clusterings. Write a math expression that would let you compare these different clusterings and pick the best one. Hint: what criterion does the k-means optimize?

Part e. Run the kmeans procedure (in the default mode) with $k = 4$ 30 times. Report the cluster sizes found for these different runs? Use formula from Part d to decide which clustering is the best. Show the scatter plot of the best clustering.

Problem 3. Hierarchical clustering experiments

Please load the dataset *clustering_data.txt* and keep it in variable Y.

Part a. Run matlab's linkage function to create a hierarchy of clusters (cluster tree):
Z = linkage(Y,'complete','euclidean');
This will create a hierarchy of clusters using the euclidean distance for pairs of points and 'max' distance for linkages. Plot the dendrogram of the full cluster tree using function dendrogram(Z,0);
Include the graph in the report.

Part b. The cluster tree can be used to define clusterings with the different number of groups. Use function cluster:
C = cluster(Z,'maxclust',4);
to assign data instances to four clusters. Using scatter function plot the results obtained by the hierarchical clustering. Similarly to Problem 2 use colors to distinguish the groups found. Include the graph in your report. Compare the results to Problem 2. part e. Are the clusters the same or different?