ECE 532, Spring 2019

Homework #4

Due: Tue, Mar 26 @ 2:30 pm

The file `seeds.csv` contains data about three different varieties of wheat, based on physical characteristics of the wheat kernels (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove). The first seven columns provide the values of these measurements, and the last column is a categorical variable recording the variety of wheat. We will use this data set to visualize the performance of clustering algorithms in Problems 1–3; hence, we will only focus on the measurements in the 5th and 7th columns, corresponding to the width of kernel and length of kernel groove. (You should ignore the measurements in all other columns, and perform clustering in only two dimensions.)

Problem 1.

(a) Use a 2D scatter plot to visualize the data according to the measurements in the 5th and 7th columns. Use a different marker for each variety.

(b) Using the first three data points as the initial cluster centers for K-means, implement the K-means algorithm with $K = 3$ from scratch. How many iterations does the algorithm take to converge?

(c) Provide 2D scatter plots showing the locations of the cluster means for each of the first three iterations (where the initialization counts as the 0th iteration, so you should actually have four scatter plots). Also use three different markers to indicate the cluster assignments of the points on each iteration. Finally, provide a 2D scatter plot showing the final cluster assignment and cluster centers at convergence.

(d) Repeat parts (b) and (c), but this time initializing the cluster centers as the 1st, 80th, and 150th data points (which are all in different clusters, according to the ground truth). Does K-means converge faster?

Problem 2. We will now experiment with different choices of K.

(a) Using the built-in Matlab function `kmeans`, run K-means for $K \in \{1, 2, 3, \ldots, 10\}$. Specify the initialization to be randomly sampled data points by using `'Start', 'sample'`. Also use 100 different random initializations for each value of K, by using `'Replicates'`. Provide an elbow plot for the value of the objective function vs. K. Which point(s) would you consider to be an elbow?

(b) Visualize the final output of K-means for $K \in \{2, 3, 4, 5, 6\}$, including both cluster centers and cluster assignments.

(c) Now use the Matlab function `silhouette` to create silhouette plots for the K-means clustering outputs visualized in (b). Report the average silhouette score $\frac{1}{n}\sum_{i=1}^{n} s(i)$. Comment on the result.

Problem 3. We now artificially introduce a few outliers to the data set. Augment the data matrix by adding two observations with measurements $(-1, -1)$ and $(10, 0)$.

(a) Run K-means on the contaminated data set, with $K = 3$ and 100 randomly subsampled initializations. Visualize the final output, including both cluster assignments and cluster centers.

(b) Now run K-medians on the contaminated data set, using the `cityblock` option in the Matlab function `kmeans`. Also use $K = 3$ and 100 randomly subsampled initializations. Visualize the final output, including both cluster assignments and cluster centers.

(c) Finally, use the Matlab function `kmedoids` to run K-medoids with $K = 3$ and 100 randomly subsampled initializations (use the `cityblock` option to make sure the objective function uses the ℓ_1-norm). Visualize the final output, including both cluster assignments and cluster centers.

(d) Comment on the differences between the outputs in (a), (b), and (c), and the relative robustness of the algorithms.

Problem 4. The file `nutrients.xlsx` contains nutrition information for 7637 common foods, providing the nutritional composition per 100g of edible portion among 14 nutritional categories. (For more information, you can view the file `sr28_doc.pdf`.)

(a) Perform K-means clustering in 14 dimensions (ignore the food ID in the final column), with $K \in \{2, 3, \ldots, 10\}$. In each case, initialize using 100 randomly subsampled iterations. Provide an elbow plot. Based on the elbow plot, what value(s) of K might you choose?

(b) Provide silhouette plots for the values of K near the elbow.

(c) Consider the set of cluster centers computed for each value of K in (a). Find the closest food item (in terms of Euclidean distance) to each of the cluster centers, and report a list of the K representative food items for each value of K.

(d) (Somewhat subjective): Combining your domain knowledge (of food) with the results from (a), (b), and (c), which value of K would you choose?

Problem 5. The file `spectral.csv` contains 2D data from two clusters, with the third column being a categorical variable recording the true cluster assignments.

(a) Provide a scatter plot of the data, with different markers for the two categories.

(b) Run K-means clustering with $K = 2$ and 100 randomly subsampled initializations, and visualize the output, including both cluster assignments and cluster centers. What fraction of points in each category are classified incorrectly?

(c) Construct the ϵ-neighborhood graph with $\epsilon \in \{0.5, 1, 2\}$, and draw the edges on the scatter plot. Run spectral clustering with $K = 2$ on each graph, and provide a scatter plot with the cluster assignments. What fraction of points in each category are classified incorrectly?

(d) Repeat part (c), this time performing spectral clustering using a similarity graph with a Gaussian kernel, with $\sigma \in \{0.1, 0.5, 1\}$.