EEL4930/EEL5840 Fall 2016 – Homework 5
Clustering and Cluster Validity

November 3, 2016

Due: November 11, 2016, 11:59 PM

Instructions

For this homework, please show any plots and tables. Label your plots' axes and include plot titles. As well, state all assumptions that you made. Do not include code. You should mention whether you programmed the solutions yourself or if you downloaded a package online and from which website.

Remember that commenting your results is very important. It is expected of you to systematically discuss your results. If no explanation is given, your grade will be penalized.

Your homework submission must cite any references used, including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you should program the algorithms yourself. If you do work with others, you must list the people you worked with. Submit your solutions as a single PDF file to the course website at http://elearning.ufl.edu/.

If you have any questions, then address them to:

- Catia Silva (TA) – catiaspsilva@ufl.edu
- Isaac Sledge (TA) – isledge@ufl.edu

Problems

In this homework, you will be implementing different unsupervised pattern recognition approaches for distinguishing between and evaluating species of flowers, which include Iris setosa, Iris virginica, and Iris versicolor. The feature data are provided on the eLearning website.

The dataset is composed of 150 samples of different flower specimens. There are four features that were captured for each specimen. These features include anatomical properties: sepal length, sepal width, petal length, and petal width. Your goal is to apply unsupervised pattern recognition approaches to determine if there is enough cluster structure to distinguish between the three flower species using these provided features. The species of flowers are given as trinary labels in the first column of the provided dataset. The remaining four columns are the real-valued features.

Complete the following tasks:

1) (5 points) For a given set of feature data $X \in \mathbb{R}^{n \times p}$, the objective function for the k-means algorithm is given by

$$\min_{U,V} \left\{ J(U,V) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{j,i} \|x_i - v_j\|^2 \right\}.$$

Here, $U \in \mathbb{R}^{k \times n}$ is a binary-valued membership matrix that denotes to which of the k clusters each sample $x_i \in \mathbb{R}^p$ belongs. A sample can only belong to a single cluster, not multiple clusters; this implies that each row of the membership function will sum to one. The variable $V \in \mathbb{R}^{k \times p}$ holds the k means for the different clusters.

The minimization of the k-means objective function can be performed using an alternating-optimization-based approach. For k-means, this involves updating the cluster means $v_j \in \mathcal{R}^p$

$$v_j = \sum_{i=1}^{n} u_{j,i} x_i \Big/ \sum_{q=1}^{n} u_{j,q}$$

while keeping the membership values constant. After all k means have been updated, they are kept fixed while updating the binary membership values $u_{j,i} \in \mathbb{R}$

$$u_{j,i} = \begin{cases} 1, & \|x_i - v_j\| \leq \|x_i - v_q\|, \ q \neq j \\ 0, & \text{otherwise} \end{cases}.$$

Here, ties are broken arbitrarily. Once the memberships have been updated, they are again fixed while the means are updated. This process is repeated until either the change in the cluster means or the change in the membership values is negligible.

Implement the k-means algorithm. Use the Euclidean distance. Assume that you initialize all of the mean values to the zero vector. Terminate the k-means update process when the membership values for all samples do not change across a pair of consecutive iterations. Provide a plot of the objective function magnitude versus the number of iterations. Comment on whether this plot is monotonically decreasing or not. Repeat these steps three more times, but under the assumption that you randomly initialize the mean values in a neighborhood of the data. Comment on if you are able to distinguish between the three classes of flowers based upon the cluster structure.

2) (5 points) For a given membership matrix, Dunn gave a geometrically motivated cluster validity index. He proposed, for any non-empty subsets S, T of $\mathbb{R}^p$, to define two quantities: the set diameter $\Delta(S) = \max_{x,y \in S}\{\|x - y\|\}$ and the set distance $\delta(S,T) = \min_{x \in S, y \in T}\{\|x - y\|\}$. He used these quantities to define a real-valued separation index V_d:

$$V_d(U;X) = \min_{1 \leq i \leq k} \left\{ \min_{1 \leq j \leq k, j \neq i} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \leq p \leq k} \Delta(X_k)} \right\} \right\}$$

where X_i denotes those subset of samples that belong to cluster i. It can be seen that the quantity $\delta(X_i, X_j)$ measures the distance between clusters i and j directly on the plots in the clusters. The quantity $\Delta(X_k)$ is a measure of the scatter volume for cluster k. Since the measures of separation and compactness in V_d occur inversely to their appearance, large values of V_d correspond to good clusters. Hence, the number of clusters that maximizes V_d is taken as the best solution.

Another geometrically motivated cluster validity index was proposed by Davies and Bouldin. The Davis-Bouldin criterion is based on a ratio of within- and between-cluster distances:

$$V_{db}(U;X) = \frac{1}{k}\sum_{j=1}^{k} \max_{i \neq j} \left\{ \frac{d_i + d_j}{d_{j,i}} \right\}.$$

Here d_i is the average distance between each point in the ith cluster and the mean of the ith cluster. Likewise, d_j is the average distance between each point in the jth cluster and the mean of the jth cluster. The variable $d_{j,i}$ is the distance between the means of the ith and jth clusters. The maximum value of V_{db} represents the worst-case within-to-between cluster ratio for each cluster. The optimal number of clusters thus corresponds to the minimum value of V_{db}.

Implement Dunn's index and the Davies-Bouldin index. Use the Euclidean distance. Use your implementation of k-means from the first problem to generate partitions of the feature data for $k = 2$ to $k = 10$. Assume that the mean vectors for k-means are randomly initialized in a neighborhood of the data. Provide either a table or a plot of the Dunn's index and Davies-Bouldin index values for each of these partitions. Comment on if the best number of clusters, as selected by the indicies, corresponds to the number of classes in the data.

As well, construct a dendrogram of the data using the unweighted, average Euclidean distance; use only up to 30 leaf nodes for the dendrogram. The height of the dendrogram leaves corresponds to the average of the unweighted Euclidean distances between all pairs of samples in a given set. Relate the best number of clusters that you found from Dunn's index and the Davis-Bouldin index to where you would partition this dendrogram to produce the same number of clusters. Comment on if this dendrogram cut-off produces clusters that have a high average, unweighted Euclidean distance between each other.