BIOST 2079 Homework 2

Distributed: 11/15/2023
Deadline: 11/24/2023

Theory

1. This question is modified from Question 1 in Chapter 10.7 in ISLR. This problem involves the K-means clustering algorithm. Let $(C_1, \dots, C_K)$ be a clustering of n data points, the objective function of K-means is

$$\min_{(C_1,\dots,C_K)} \sum_{k=1}^{K} WCSS(C_k) = \min_{(C_1,\dots,C_K)} \sum_{k=1}^{K} \left[\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2\right]$$

$$\text{(or equivalently in vector form)} = \min_{(C_1,\dots,C_K)} \sum_{k=1}^{K} \left[\frac{1}{|C_k|}\sum_{i,i'\in C_k}\|\boldsymbol{x}_i-\boldsymbol{x}_{i'}\|^2\right]$$

where $|C_k|$ means the number of data points in C_k. (5 points)

(a) Given that the center of C_k is $\mu_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}\, , j = 1, \dots, p$, (or equivalently, $\boldsymbol{\mu}_k = \frac{1}{|C_k|}\sum_{i\in C_k}\boldsymbol{x}_i$), prove that the above objective is equivalent to

$$\min_{(C_1,\dots,C_K)} \sum_{k=1}^{K} \left[2\cdot\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\mu_{kj}\right)^2\right],$$

$$\text{(or equivalently)} \min_{(C_1,\dots,C_K)} \sum_{k=1}^{K} \left[2\cdot\sum_{i\in C_k}\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|^2\right].$$

(b) On the basis of the identity proved in (a), argue that the K-means (Lloyd's) clustering algorithm decreases the above objective function at each iteration.

2. Question 3 in Chapter 10.7 in ISLR. In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are as follows. (6 points)

Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Plot the observations.
(b) Randomly assign a cluster label to each observation. You can use sample() command in R to do this. Report the cluster labels for each observation.
(c) Compute the centroid for each cluster.
(d) Assign each observation to the centroids to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.
(e) Repeat (c) and (d) until the answers obtained stop changing.
(f) In your plot from (a), color the observations according to the cluster labels obtained.

Computing

* You are encouraged to use R Markdown (template provided) to generate pdf reports with embedded R codes and outputs.

3. This question is modified from Question 9 in Chapter 10.7 in ISLR. Consider the USArrests data set from the R package ISLR. Write a data analysis report to address the following problems. (9 points)

(a) Apply PCA to the dataset. Plot the data in the first two PCs labeled by state names. Do the states appear to have any clustering patterns in the plot?

(b) Run K-means with K from 1 to 6 and plot the associated within cluster sum of squares (WCSS). Note that when K = 1, WCSS is actually the total sum of squares (TSS) since there is no between cluster sum of square (i.e. BCSS = 0).

(c) Visualize your clustering for when K = 3 in the plot generated in (a).

(d) Now, using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.

(e) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

(f) Compare the clustering from K-means (K = 3) and the hierarchical clustering in a confusion table. Do the clustering results from two methods agree? Compute the Rand index between the two clustering results.

(g) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

(h) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Bonus questions

4. If each feature vector is standardized to mean zero and variance one, show that Pearson correlation (r) and Euclidean distance (d) are related: $\frac{d^2}{n-1} = 2 - 2 \cdot r$. In other words, large Pearson correlation results in small Euclidean distance. One may use $1 - r$ as the dissimilarity measure for clustering. (2 points)