4033/5033: Assignment 6

Your Name

Due: Nov 28 (by 11:59pm)

In this assignment, we play with K-means.
Let $x_1, \ldots, x_n$ be a set of instances in $\mathbb{R}^p$ and $c_1, \ldots, c_K$ be a set of cluster centers in the same space.
Let $\mathbb{I}_{ij}$ be a variable such that $\mathbb{I}_{ij} = 1$ if x_i is assigned to cluster centered at c_j and $\mathbb{I}_{ij} = 0$ otherwise.

[1] In theory, we know K-means is equivalent to minimizing the following objective

$$\sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{I}_{ij} ||x_i - c_j||^2. \tag{1}$$

Specifically, its step that fixes clustering center and assign instances to their nearest centers is equivalent to fixing c_j while optimizing $\mathbb{I}_{ij}$ in (1), and its step that fixes cluster assignment and updates cluster centers is equivalent to fixing $\mathbb{I}_{ij}$ while optimizing c_j in (1). Based on this, answer the following question.

Suppose we want to modify the K-means algorithm so it is equivalent to minimizing the following objective

$$\sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{I}_{ij} w_i ||x_i - c_j||^2, \tag{2}$$

where w_i is a weight of instance x_i. Explain your modified algorithm. In the answer, clearly state how each step in K-means is modified (if necessary) e.g., Step X is not changed, and Step Y is changed to ....

[2] Implement standard K-means algorithm from scratch, evaluate it on the diabetes data set and visualize the clustering result in 2-dimensional space using the PCA technique.

Specifically, you need to draw three figures. Figure 1 plots the data distribution with K = 2. Figure 2 plots the data distribution with K = 3. Figure 3 plots the data distribution with K = 5. (Pick proper color for each cluster yourself.)

Also, report the random index and Davies–Bouldin index of each clustering result. You can use existing functions to evaluate these indices. `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`.

Tip: we only run clustering on the features not on labels. The labels are only used for external evaluation. Practically, exclude the last column in 'data' (which is label) when running clustering.

Submission Instruction

Please submit two files to Canvas. (Do not zip them. Upload them separately.)

(i) All your mathematical and experimental results should be presented in a single pdf file named as 'hw6.pdf'.

(ii) A Python source code for the implementation of K-means named 'hw6_Kmeans.py'