4033/5033 Assignment: K-Means Clustering

Sasank Sribhashyam

In this assignment, we will analyze weighted K-means clustering algorithm, and implement the standard K-means clustering algorithm and evaluate it on the Diabetes data set. Task 1. Consider a weighted K-means clustering algorithm which aims to find K cluster centers $c_1, \ldots, c_K \in \mathbb{R}^p$ and cluster memberships for a set of instances $x_1, \ldots, x_n \in \mathbb{R}^p$ that minimize the following objective.

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

where $w_i \in \mathbb{R}$ is a weight for x_i (assumed given), $c_j \in \mathbb{R}^p$ is center of cluster j (to optimize), and δ_{ij} is an indicator function outputting 1 if x_i is assigned to cluster j and outputting 0 otherwise (to optimize). Below is an incomplete description of this weighted K-means clustering algorithm. Its Step 2 (cluster center update) and Step 3 (cluster membership update) are missing. Please complete them. Note: You may explain the two steps outside the algorithm environment. Your explanation should be concise, mathematical and offers proper justification.

Algorithm 1 Weighted K-Means Clustering

Input: a set of instances $x_1, \ldots, x_n$, number of clusters K

1: randomly initialize cluster centers $c_1, \ldots, c_K$.

while cluster membership is updated in the previous round **do**

2: please explain how to update cluster membership based on the current $c_1, \ldots, c_K$

3: please explain how to update cluster centers based on the current cluster membership

end while

Step 2: Update Cluster Membership

Given current cluster centers $c_1, \ldots, c_K$, update the cluster memberships δ_{ij} as follows:

$$\delta_{ij} = \begin{cases} 1 & \text{if } j = \arg\min_k w_i ||x_i - c_k||^2 \\ 0 & \text{otherwise} \end{cases}$$

Justification for Step 2:

Minimizing weighted squared Euclidean distance:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \delta_{ij} w_i ||x_i - c_j||^2$$

Step 3: Update Cluster Centers

Given updated cluster memberships, update cluster centers $c_1, \ldots, c_K$ as the weighted mean:

$$c_j = \frac{\sum_{i=1}^{n} \delta_{ij} w_i x_i}{\sum_{i=1}^{n} \delta_{ij} w_i}$$

Justification for Step 3:

Minimizing J with respect to c_j:

$$\frac{\partial J}{\partial c_j} = -2 \sum_{i=1}^{n} \delta_{ij} w_i (x_i - c_j) = 0 \implies c_j = \frac{\sum_{i=1}^{n} \delta_{ij} w_i x_i}{\sum_{i=1}^{n} \delta_{ij} w_i}$$

These updates align with minimizing the objective function J with respect to cluster memberships and centers.

Task 2. Implement the standard K-means clustering algorithm from scratch.

Task 3. Apply K-means to cluster the Diabetes data set into k groups, and visualize the clustering results with $k = 2$ and 3 in Figure 1 and 2, respectively. To get each figure, you should first apply K-means to cluster data, then apply PCA to reduce feature dimension to 2, and finally plot the projected data and mark the two clusters using different colors.

Task 4. Evaluate your clustering results based on the Randn index and report result versus K in Figure 3. Pick five values of K yourself.

Task 5. Evaluate your clustering results based on the DaviesâBouldin index and report result versus K in Figure 4. Pick five values of K yourself.

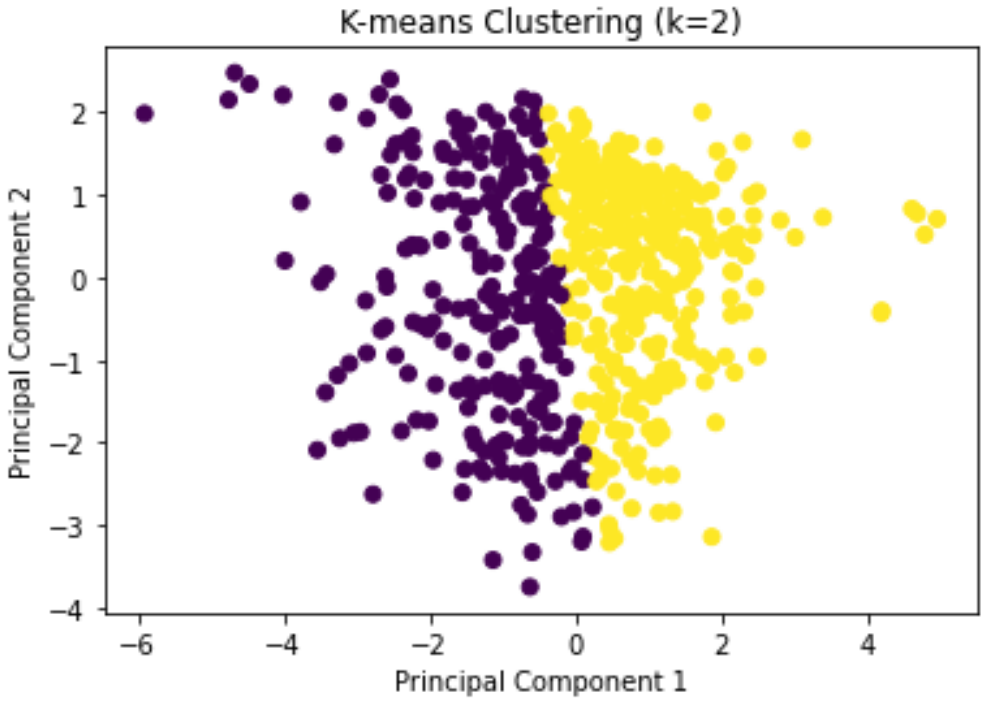

Fig. 1. K-Means Clustering with $k = 2$.

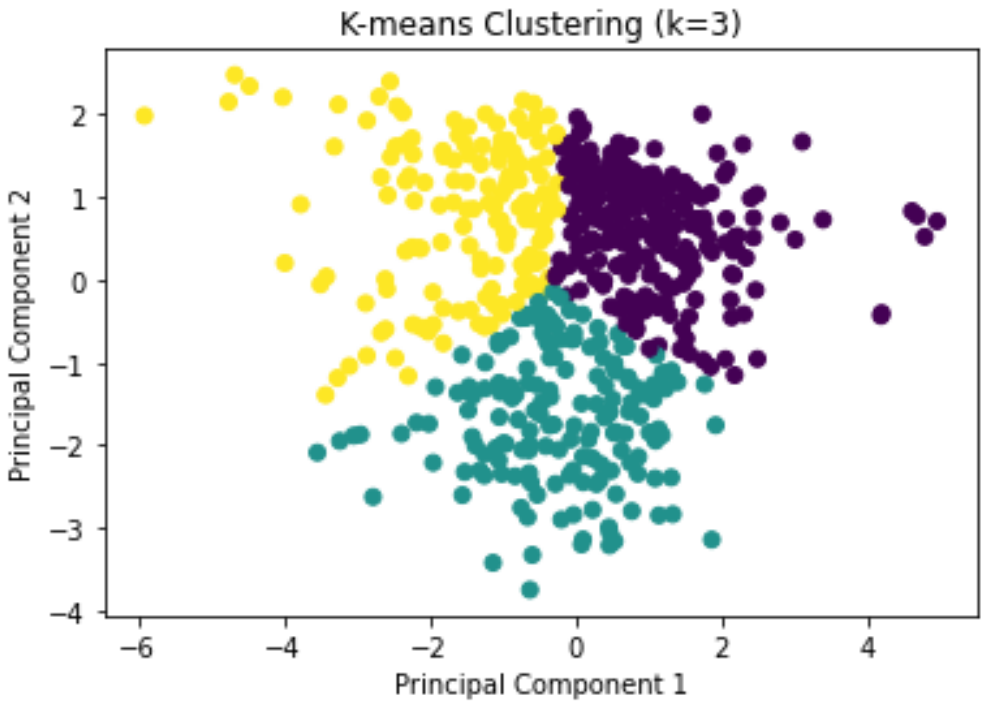

Fig. 2. K-Means Clustering with $k = 3$.

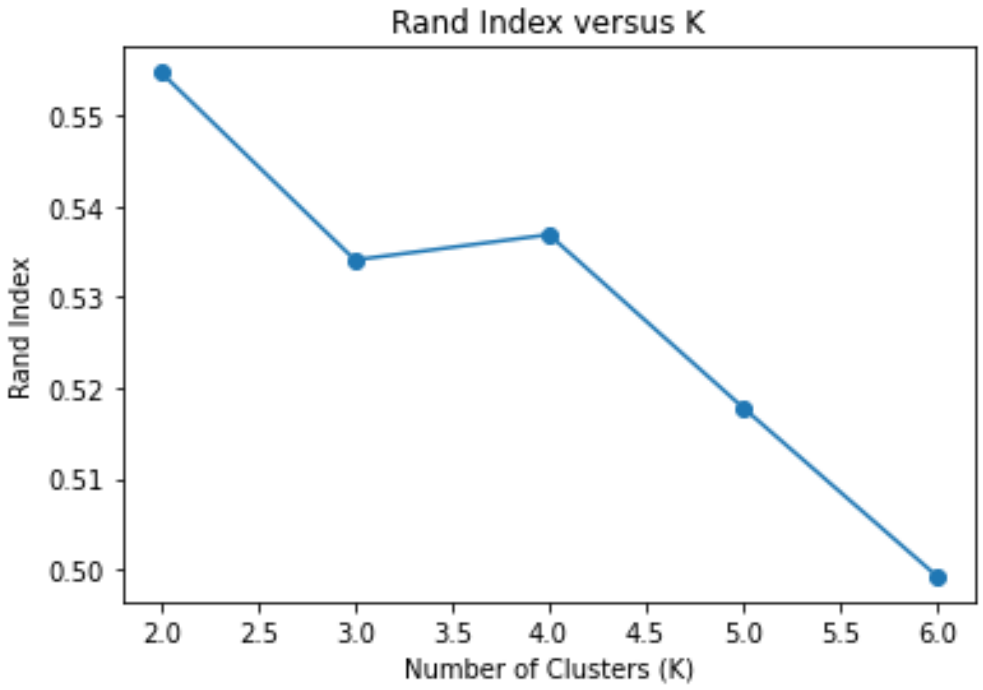

Fig. 3. Randn Index versus K.

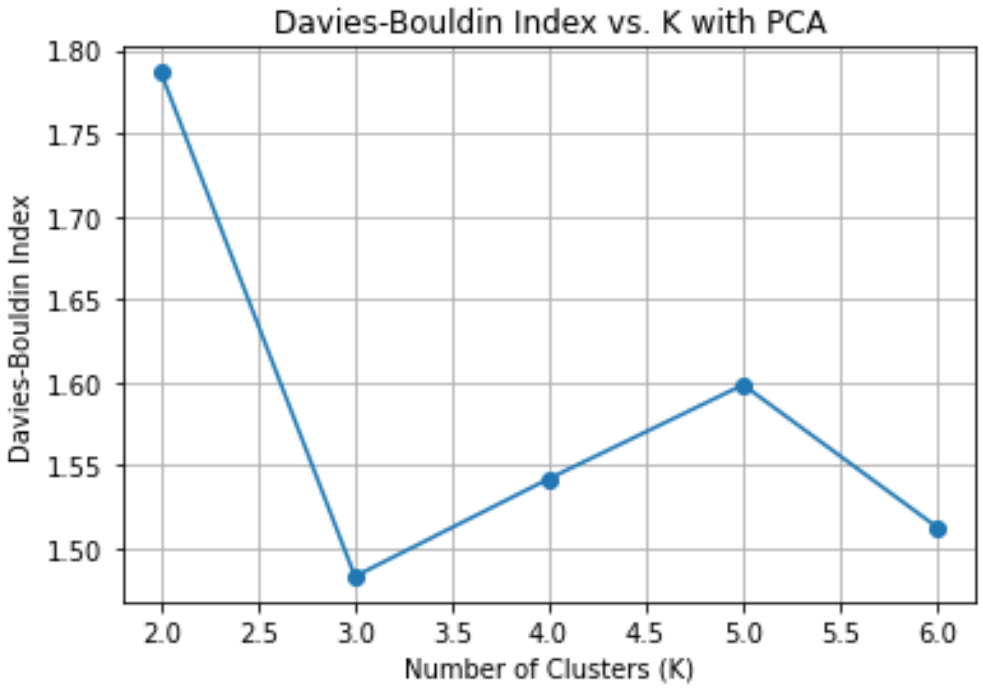

Fig. 4. DaviesâBouldin Index versus K.