[ML20] Assignment 10

Danny Radosevich

Due: Apr 20 (b 8:45am)

Implement K-means algorithm from scratch and apply it to cluster the entire data set.

[1] Visualize your clustering result in a 2D feature space obtained by PCA. (You need to first cluster instances in the original feature space, and then plot their clustering result in the 2D space.) Mark instances in one cluster by one color, and instances in different clusters by different colors, e.g., all instances in Cluster 1 are red and all in Cluster 2 are blue. Choose your own value of K to get a clustering result that can be easily explained, e.g., if we observe two distant groups of instances in the 2D space, but you set K = 1, then it is hard to explain why these two groups are assigned to the same cluster.

Figure 1 is an example plot with K = 3 on a different data set. You don't have to plot cluster centroids, and your data distribution and clustering result will probably look different from it.

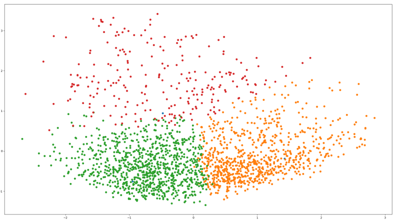

Fig. 1. Clustering Result of K-means (K = 3).

[2] Choose a proper K so that you can run the same algorithm multiple times to obtain different clustering results. Show two different clustering results in Figure 2 and Figure 3, respectively. (Same form as Figure 1.)

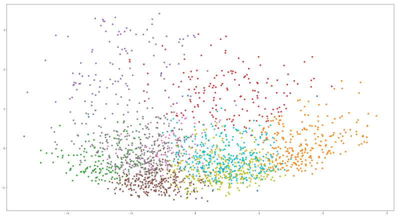

Fig. 2. Clustering Result of Trial 1. (K = 9)

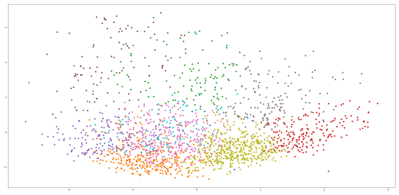

Fig. 3. Clustering Result of Trial 2. (K = 9)

[3] Vary K to get different clustering result. For each result, evaluate its quality using two metrics

$$J_1 = \frac{1}{K}\sum_{k=1}^{K}\sum_{x_i \in C_k} ||x_i - \mu_k||^2 \qquad \text{and} \qquad J_2 = \sum_{k=1}^{K}\sum_{x_i \in C_k} ||x_i - \mu_k||^2. \tag{1}$$

Plot two figures. Figure 4 shows a curve of J_1 versus K, and Figure 5 shows a curve of J_2 versus K. (Thus y-axis is J and x-axis is K.) Choose 7 values of K by yourself for both figures, and try to cover the behaviors of J_1 and J_2 as comprehensive as possible. (Think about how J will behave as K decreases and increases.)

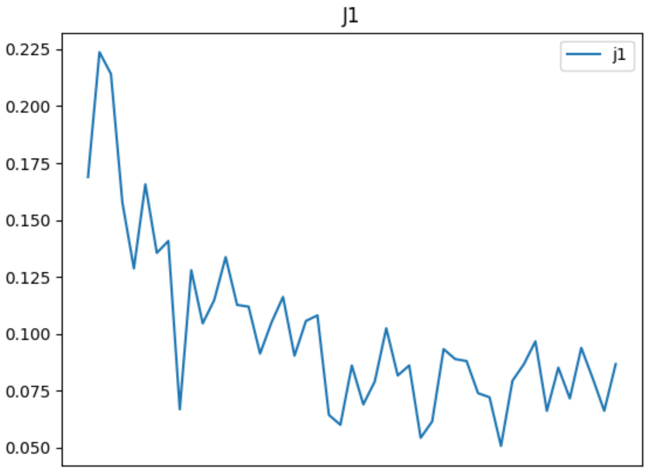

Fig. 4. Clustering Quality J_1 versus K.

[4] Apply GMM to cluster the same data set and visualize your result in Figure 6. (Same form as Figure 1.) No need to implement GMM from scratch – just use any GMM library. Use the same K as in task [1]. (Compare this result with your K-means clustering result by yourself.)

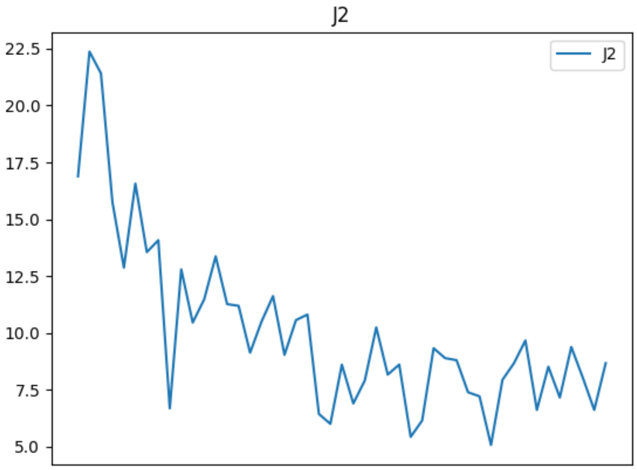

Fig. 5. Clustering Quality J_2 versus K.

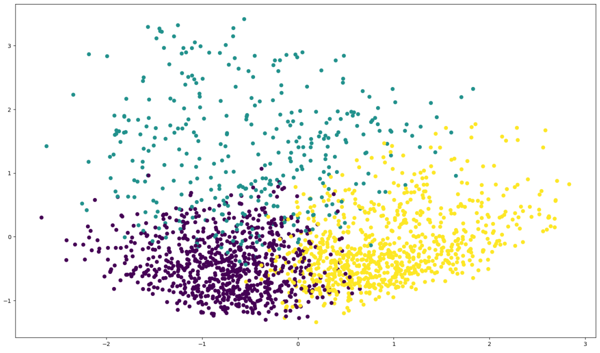

Fig. 6. Clustering Result of GMM. (K = 3)