CSE 575 - Statistical Machine Learning

Project Part 2: Unsupervised Learning (K-means)

Skanda Suresh

{1217132644}

27th October, 2019

1 Problem Statement

Implement the K-means algorithm and apply the implementation on the given dataset of 2D points. The algorithm is to be implemented with 2 different initialization strategies for the K-means algorithm, which are described in the sections below. An objective function is computed, and it's behavior is observed against the number of clusters.

2 Dataset

The dataset provided is a set of 2D points. There are 300 rows thus giving it dimensions 300 x 2.

3 K-means algorithm

The K-means algorithm is an unsupervised learning algorithm. It can work on unlabeled data and cluster groups of data samples similar to one and other based on a distance metric. The metric used in this project is Euclidean distance. The 'K' refers to number of clusters. The algorithm works towards finding 'K' centroids such that each sample is centred around these centroids. For a given sample the Euclidean distance is computed with each of the centroids, and it is assigned to the one with minimum distance. The new centroids are now computed by taking the mean of the samples corresponding to each centroid. The above process is repeated until there is no change in the values of the centroids. This algorithm is guaranteed to converge. The Euclidean distance formulation is as shown below.

$$||x_i - \mu||^2 \tag{1}$$

Where x_i refers to a data sample, and μ is a centroid.

3.1 Initialization Strategies

One the most important steps of the K-means algorithm is the intialization of centroids. A bad initialization could lead to poor clusters. The project covers two initialization strategies as shown below.

3.1.1 Strategy 1 - Random Initialization

In this strategy, the initial 'K' centroids are picked at random. Choosing centroids that are very close to each other initially may lead to poor clustering resulting in empty clusters.

3.1.2 Strategy 2 - Maximizing Centre Distances

In this method, the first centre is picked at random. The next centre is picked such that it is farthest apart from the first one, and the third such that it's average distance from the previous two are maximized and so on. This intuitively results in a set of initial centroids such that they are maximally spaced out. Formallt, while picking ith, ensure that it's average distance to previous $i-1$ centres is maximal.

4 Objective Function

The working of K-means aims to optimize an objective function given k clusters, D_i sets of data, and mean vectors $\mu_1, \mu_2, ...\mu_k$ which are nothing but the centroids. The objective function is as defined below

$$\sum_{i=1}^{k} \sum_{i \in D_i} ||x - \mu_i||^2 \tag{2}$$

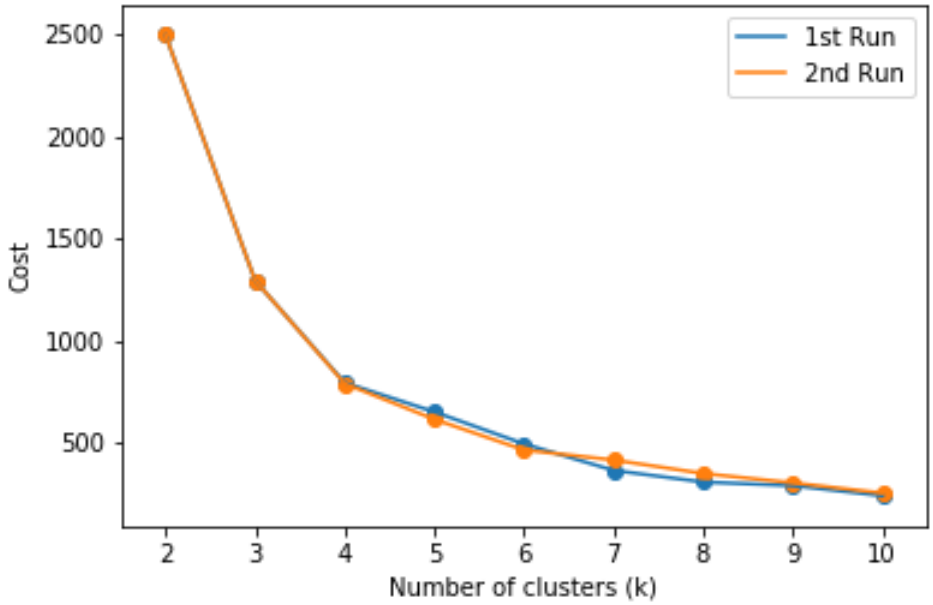
This objective function should decrease or stay the same as the number of clusters, k increases. We demonstrate this by plotting the objective function vs the number of clusters k ranging from 2 to 10. **Note:** when $k = 1$, the objective function transforms into the variance of the dataset multiplied by the number of samples, $(\sigma^2 \times N)$.

4.1 Objective Function vs Number of Clusters

We observe the behavior of the objective function with number of clusters varying from $(2, 10)$, under each initialization strategy.

4.1.1 Random Initialization

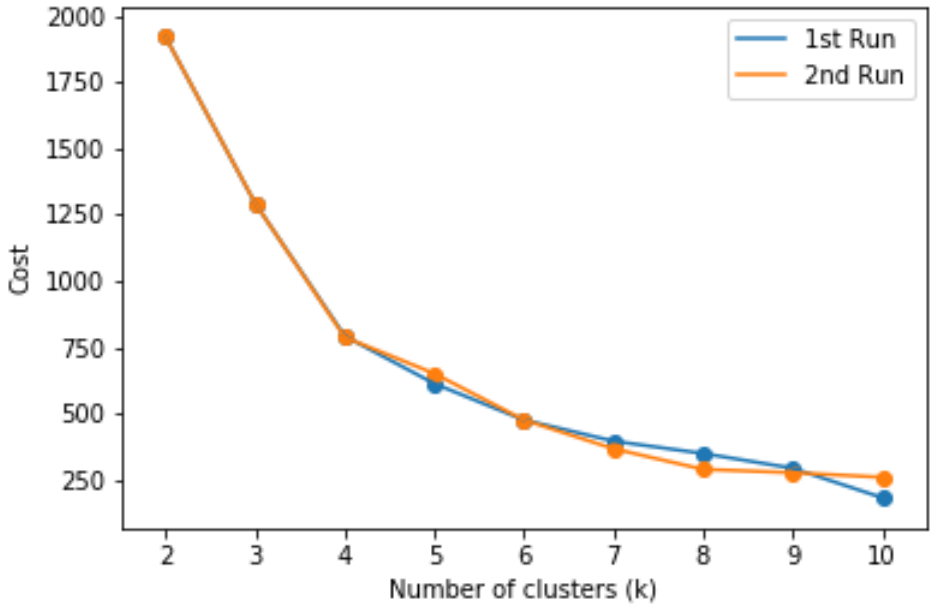
We plot the graph of the objective function for two runs of the K-means algorithm with random initialization.



We observe for $K = 2$, the objective function takes values 2498.11 and 2497.99 and for $K = 10$ it reduces to 238.55 and 251.74 for each run respectively.

4.1.2 Maximizing Initial Centre Distances

For strategy 2 we observe the initial cost for $K = 2$ is lesser with value approximately 1921.033 and it attains a value of 264.903 and 261.2525 with $K = 10$ for each run respectively.



5 Conclusions

Thus from the plots of Objective function vs No.of Clusters we observe that Strategy 2 allows for lesser initial value of objective function, i.e when $K = 2$. This may be due to the maximal spacing of the initial centroids. However since the first centroid is picked at random, the objective function may have some variance in results for different runs with suitable initialization. However we notice that in both cases, with Number of clusters, $K - 10$, the function converges at a value between $250 - 270$, more or less similar. Using these plots we may also find the optimal number of clusters (Elbow method) which is out scope of this project.