Assignment #4

Elements of Machine Learning

Saarland University – Winter Semester 2024/25

Rabin Adhikari
7072310
raad00002@stud.uni-saarland.de

Dhimitrios Duka
7059153
dhdu00001@stud.uni-saarland.de

1 Problem 1 (K-means)

1.1 Given a dataset D of 4 points...

First, we start by calculating the distance of the given points to the initial centroids $\bar{x}_1$ and $\bar{x}_2$.

$$
\begin{aligned}
d_{x_1\bar{x}_1} &= 0 \\
d_{x_2\bar{x}_1} &= \sqrt{0^2 + 3^2} = 3 \\
d_{x_3\bar{x}_1} &= \sqrt{0^2 + 3^2} = 3 \\
d_{x_4\bar{x}_1} &= \sqrt{3^2 + 3^2} = 3\sqrt{2} \\
d_{x_1\bar{x}_3} &= 3 \\
d_{x_2\bar{x}_3} &= \sqrt{3^2 + 3^2} = 3\sqrt{2} \\
d_{x_3\bar{x}_3} &= \sqrt{0^2 + 3^2} = 0 \\
d_{x_4\bar{x}_3} &= \sqrt{0^2 + 3^2} = 3
\end{aligned}
\tag{1}
$$

Therefore, we can conclude that the first cluster would contain the data points x_1 and x_2 and the second cluster would include the remaining data points x_3 and x_4.

Now for the second iteration, first we have to calculate the new centroids.

$$
\begin{aligned}
\bar{x}_1 &= (\frac{1+1}{2}, \frac{1+4}{2}) = (1, \frac{5}{2}) \\
\bar{x}_2 &= (\frac{4+4}{2}, \frac{1+4}{2}) = (4, \frac{5}{2})
\end{aligned}
\tag{2}
$$

Now, we calculate the distances of every point to the new centroids.

$$
\begin{aligned}
d_{x_1\bar{x}_1} &= \sqrt{0^2 + (\frac{5}{2} - 1)^2} = \frac{3}{2} \\
d_{x_2\bar{x}_1} &= \sqrt{0^2 + (4 - \frac{5}{2})^2} = \frac{3}{2} \\
d_{x_3\bar{x}_1} &= \sqrt{(4-1)^2 + (1 - \frac{5}{2})^2} = \frac{3}{2}\sqrt{5} \\
d_{x_4\bar{x}_1} &= \sqrt{(4-1)^2 + (4 - \frac{5}{2})^2} = \frac{3}{2}\sqrt{5} \\
d_{x_1\bar{x}_2} &= \sqrt{(4-1)^2 + (\frac{5}{2} - 1)^2} = \frac{3}{2}\sqrt{5} \\
d_{x_2\bar{x}_2} &= \sqrt{(4-1)^2 + (\frac{5}{2} - 4)^2} = \frac{3}{2}\sqrt{5} \\
d_{x_3\bar{x}_2} &= \sqrt{0^2 + (\frac{5}{2} - 1)^2} = \frac{3}{2} \\
d_{x_4\bar{x}_2} &= \sqrt{0^2 + (\frac{5}{2} - 4)^2} = \frac{3}{2}
\end{aligned}
\tag{3}
$$

From the above calculations, we can see that there is no reassignment of any of the data points to a new cluster. As there are no reassignments, the cluster centroids would remain the same. This indicates that the model has stabilized, and further iterations would yield no changes. Therefore, the algorithm has reached convergence.

1.2 Below, you are given a plot representing the within-cluster variation (also known as inertia or within-cluster sum-of-squares, WCSS) for different numbers of clusters (k) in k-means.

1.2.1 According to the elbow heuristic, what is the optimal number of clusters for this dataset? Explain why did you choose this value.

Based on the provided graph and the intuition behind the elbow heuristic, we would choose a value of $k = 3$[1]. This choice is based on the fact that for values smaller than 3, the WCSS decreases significantly. However, for values larger than 3, the decrease is much slower, suggesting diminishing returns.

1.2.2 Intuitively explain how the within-cluster variation changes as the number of clusters increases.

As the number of clusters k increases, each cluster becomes smaller and more specific, thus containing fewer data samples. As a result, the data samples within a cluster are closer together, reducing the within-cluster variation. However, based on the above exercise, we can see that the within-cluster variation follows an elbow curve. This means that after a certain point, the improvement of the within-cluster variation becomes smaller. This is because new clusters start splitting data points that are already well-grouped together.

1.2.3 Intuitively explain under what conditions the within-cluster variation equals to zero.

If we suppose that $k = N$, where N is the number of data points, and we have N distinct data points, the within-cluster variation would be zero since each cluster would contain only one sample and the distance of that sample from itself, which is the center of the cluster, would be zero.

[1]The value $k = 4$ would also be acceptable depending on the problem that we are dealing with.

1.2.4 The figure below shows the resulting clusters for a random dataset using both k-means and k-medoids. Identify which of the two plots corresponds to k-medoids and explain your reasoning.

From the given plots, we can conclude that Plot 2 is the plot that corresponds to the k-medoids clustering algorithm. This is because, in k-medoids, the center of the cluster is one of the data samples itself, while in the k-means clustering algorithm, the center of the cluster is not necessarily a sample point of the cluster.