MTH793P **Advanced Machine Learning, Semester B, 2023/24**
Coursework 3 solution

---

1 k-means clustering

1. Perform two steps of k-means clustering by hand for the ten data points $x_1 = -3$, $x_2 = 2$, $x_3 = -1$, $x_4 = 7$, $x_5 = 11$, $x_6 = 6$, $x_7 = -30$, $x_8 = 0$, $x_9 = -50$ and $x_{10} = 15$. Assume $k = 3$ clusters and initialise your centroids as $\mu_1^0 = -4$, $\mu_2^0 = 0$ and $\mu_3^0 = 1$, respectively

$$\mu^0 := \begin{pmatrix} -4 & 0 & 1 \end{pmatrix} .$$

   For each iteration update the assignment variable z^l first, and then μ^l. Here $l \in \{1, 2\}$ denotes the iteration index.

2. Did the method converge after two iterations?

3. Perform k-means clustering by hand for the five data points

$$x_1 = \begin{pmatrix} -3 \\ 6 \end{pmatrix}, x_2 = \begin{pmatrix} 2 \\ -30 \end{pmatrix}, x_3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, x_4 = \begin{pmatrix} 7 \\ -50 \end{pmatrix} \quad \text{and} \quad x_5 = \begin{pmatrix} 11 \\ 15 \end{pmatrix} .$$

   Assume $k = 3$ clusters and initialise your centroids as

$$\mu_1^0 := \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \mu_2^0 := \begin{pmatrix} 3 \\ -5 \end{pmatrix} \quad \text{and} \quad \mu_3^0 := \begin{pmatrix} 10 \\ 15 \end{pmatrix} .$$

   Perform as many iterations as are required to guarantee convergence.

Solution:

1. First iteration: the (Euclidean) distances between the data points and the initial cen-

troids are

$$
\begin{pmatrix}
|-3+4| & |2+4| & |-1+4| & |7+4| & |11+4| \\
|-3| & |2| & |-1| & |7| & |11| \\
|-3-1| & |2-1| & |-1-1| & |7-1| & |11-1|
\end{matrix}
$$

$$
\begin{matrix}
|6+4| & |-30+4| & |0+4| & |-50+4| & |15+4| \\
|6| & |-30| & |0| & |-50| & |15| \\
|6-1| & |-30-1| & |0-1| & |-50-1| & |15-1|
\end{pmatrix}
$$

$$
= \begin{pmatrix}
1 & 6 & 3 & 11 & 15 & 10 & 26 & 4 & 46 & 19 \\
3 & 2 & 1 & 7 & 11 & 6 & 30 & 0 & 50 & 15 \\
4 & 1 & 2 & 6 & 10 & 5 & 31 & 1 & 51 & 14
\end{pmatrix},
$$

leading to the following assignment:

$$
z_1 = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}.
$$

Hence, we update the centroids to

$$
(\mu_1)_1 = \frac{x_1 + x_7 + x_9}{3} = \frac{-3-30-50}{3} \approx -27.67
$$

$$
(\mu_1)_2 = \frac{x_3 + x_8}{2} = \frac{-1+0}{2} = -\frac{1}{2}
$$

$$
(\mu_1)_3 = \frac{x_2 + x_4 + x_5 + x_6 + x_{10}}{5} = \frac{2+7+11+6+15}{5} = 8.2\,,
$$

which we can write in vectorial form as

$$
\mu_1 = \begin{pmatrix} -83/3 & -1/2 & 41/5 \end{pmatrix}.
$$

Second iteration: the (Euclidean) distances between the data points and the centroids from the first iteration are

$$
\begin{pmatrix}
\frac{74}{3} & \frac{89}{3} & \frac{80}{3} & \frac{104}{3} & \frac{116}{3} & \frac{101}{3} & \frac{7}{3} & \frac{83}{3} & \frac{67}{3} & \frac{128}{3} \\
\frac{5}{2} & \frac{5}{2} & \frac{1}{2} & \frac{15}{2} & \frac{23}{2} & \frac{13}{2} & \frac{59}{2} & \frac{1}{2} & \frac{99}{2} & \frac{31}{2} \\
\frac{56}{5} & \frac{31}{5} & \frac{46}{5} & \frac{6}{5} & \frac{14}{5} & \frac{11}{5} & \frac{191}{5} & \frac{41}{5} & \frac{291}{5} & \frac{34}{5}
\end{pmatrix}
$$

leading to the following assignment:

$$
z_2 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1
\end{pmatrix}.
$$

Hence, we update the centroids to

$$(\mu_2)_1 = \frac{x_7 + x_9}{3} = \frac{-30 - 50}{2} = -40$$

$$(\mu_2)_2 = \frac{x_1 + x_2 + x_3 + x_8}{4} = \frac{-3 + 2 - 1 + 0}{4} = -0.5$$

$$(\mu_2)_3 = \frac{x_4 + x_5 + x_6 + x_{10}}{4} = \frac{7 + 11 + 6 + 15}{4} = 9.75\,,$$

which we can write in vectorial form as

$$\mu_2 = \begin{pmatrix} -40 & -0.5 & 9.75 \end{pmatrix}\,.$$

2. The method converges after two iterations, which can be seen by the fact that z_3 is the same as z_2, which means that μ_3 will be equal to μ_2. This is left as an exercise for the reader, but please note that you would have to verify this statement in an exam in order to obtain full marks.

3. Similar to the first exercise, we compute the Euclidean distances of the data points with respect to the initial centroids:

$$\begin{pmatrix} \sqrt{29} & \sqrt{970} & 1 & \sqrt{2665} & \sqrt{340} \\ \sqrt{157} & \sqrt{626} & \sqrt{41} & \sqrt{2041} & \sqrt{464} \\ \sqrt{250} & \sqrt{2089} & \sqrt{346} & \sqrt{4234} & 1 \end{pmatrix}\,,$$

leading to the following assignment:

$$z_1 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}\,.$$

Hence, we update the centroids to

$$\mu_1^1 = \frac{x_1 + x_3}{2} = \begin{pmatrix} -2 \\ 3 \end{pmatrix}$$

$$\mu_2^1 = \frac{x_2 + x_4}{2} = \begin{pmatrix} \frac{9}{2} \\ -40 \end{pmatrix}$$

$$\mu_3^1 = x_5 = \begin{pmatrix} 11 \\ 15 \end{pmatrix}\,.$$

Second iteration: the (Euclidean) distances between the data points and the centroids from the first iteration are

$$\begin{pmatrix} \sqrt{10} & \sqrt{1105} & \sqrt{10} & \sqrt{2890} & \sqrt{313} \\ \frac{\sqrt{8689}}{2} & \frac{\sqrt{425}}{2} & \frac{\sqrt{6521}}{2} & \frac{\sqrt{425}}{2} & \frac{\sqrt{12269}}{2} \\ \sqrt{277} & \sqrt{2106} & \sqrt{369} & \sqrt{4241} & 0 \end{pmatrix}$$

leading to the following assignment:

$$z_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

We see that $z^2 = z^1$, hence $\mu^2 = \mu^1$ and we have converged.