Week 9 Exercise Sheet Solutions

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time or effort to solve compared to an exercise with (***), which is a more complex exercise.

Unsupervised Learning

1. (**) Consider the following dataset:

$$\begin{aligned} \mathbf{x}_1 &= (1,1) & \mathbf{x}_2 &= (2,2) \\ \mathbf{x}_3 &= (3,1) & \mathbf{x}_4 &= (4,2) \\ \mathbf{x}_5 &= (5,1) & \mathbf{x}_6 &= (6,2) \end{aligned}$$

Perform the K-means algorithm on this data to find 2 clusters. Initialise your centroids to $\mathbf{m}_1 = (0,0)$ and $\mathbf{m}_2 = (7,2)$, which datapoints are assigned to each cluster in the first iteration? What are the values of the centroids after the first iteration and then after the second iteration?

Solution:
When assigning the data points to a cluster it is based on which is the shortest distance. For example, datapoint 4:

$$\mathbf{x}_4 - \mathbf{m}_1 = (4,2) \rightarrow |\mathbf{x}_4 - \mathbf{m}_1| = \sqrt{4^2 + 2^2} = 4.4721$$
$$\mathbf{x}_4 - \mathbf{m}_2 = (-3,0) \rightarrow |\mathbf{x}_4 - \mathbf{m}_2| = 3$$

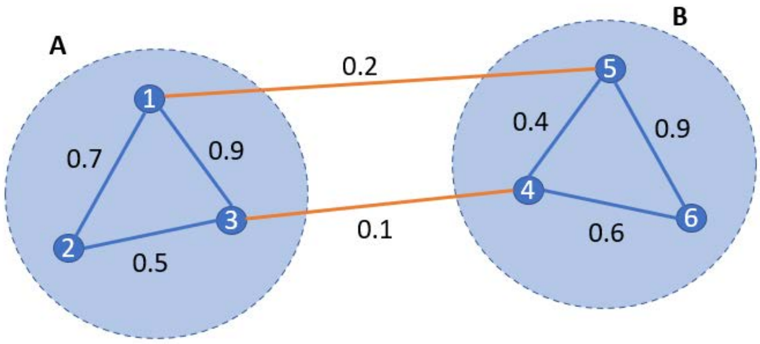
so it is closer to cluster 2. So datapoints 1, 2 and 3 are assigned to cluster 1 and 4, 5 and 6 are assigned to cluster 2. Using this clustering the new centroids are

$$\mathbf{m}_1 = \frac{\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3}{3} = (2, 1.3333)$$
$$\mathbf{m}_2 = \frac{\mathbf{x}_4 + \mathbf{x}_5 + \mathbf{x}_6}{3} = (5, 1.6667)$$

In the second iteration, the datapoints are assigned to the same clusters as the first iteration and so the centroids will not change any further. This means the algorithm has converged.

---

2. (**) For the graph below, compute the normalised cut, $\mathrm{Ncut}(A, B)$.



Solution:
First we can calculate the node degrees using $d_i = \sum_j W_{ij}$:

$$
\begin{aligned}
d_1 &= 0.7 + 0.9 + 0.2 = 1.8 \\
d_2 &= 0.7 + 0.5 = 1.3 \\
d_3 &= 0.5 + 0.9 + 0.1 = 1.5 \\
d_4 &= 0.1 + 0.4 + 0.6 = 1.1 \\
d_5 &= 0.2 + 0.4 + 0.9 = 1.5 \\
d_6 &= 0.9 + 0.6 = 1.5
\end{aligned}
$$

Using these values we can calculate the the volumes using $\mathrm{vol}(A) = \sum_{i \in A} d_i$:

$$
\begin{aligned}
\mathrm{vol}(A) &= d_1 + d_2 + d_3 = 1.8 + 1.3 + 1.5 = 4.6 \\
\mathrm{vol}(B) &= d_4 + d_5 + d_6 = 1.1 + 1.5 + 1.5 = 4.1
\end{aligned}
$$

The cut is the sum of the connections between the sets $\mathrm{cut}(A, B) \sum_{i \in A} \sum_{j \in B} W_{ij}$:

$$\mathrm{cut}(A, B) = W_{15} + W_{34} = 0.2 + 0.1 = 0.3$$

Finally the normalised cut is given by

$$
\begin{aligned}
\mathrm{Ncut}(A, B) = \mathrm{cut}(A, B)\left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(B)}\right) &= \mathrm{cut}(A, B)\left(\frac{\mathrm{vol}(A) + \mathrm{vol}(B)}{\mathrm{vol}(A)\mathrm{vol}(B)}\right) \\
= 0.3 \times \left(\frac{1}{4.6} + \frac{1}{4.1}\right) &= 0.3 \times \left(\frac{4.6 + 4.1}{4.6 \times 4.1}\right) \\
= 0.138388123
\end{aligned}
$$

3. (***) In spectral clustering, the graph partitioning is solved through a generalised eigenvalue equation of the graph Laplacian

$$(\mathbf{D} - \mathbf{W})\mathbf{y} = \lambda \mathbf{D}\mathbf{y} \tag{1}$$

where $\mathbf{W}$ is the graph connection matrix, $\mathbf{D}$ is the degree matrix with diagonal entries $D_{ii} = d_i = \sum_j W_{ij}$. Show that $\mathbf{y} = \mathbf{1}$ (a vector of all ones) is an eigenvector of this equation and that its eigenvalue is $\lambda = 0$. What is the significance of this solution?

Solution:
To show that $\mathbf{y} = \mathbf{1}$ is an eigenvector we can first recognise that

$$\mathbf{W}\mathbf{y} = \mathbf{W}\mathbf{1} = \mathbf{d},$$

as it will perform a sum along each row of $\mathbf{W}$ which is the definition of $\mathbf{d}$. $\mathbf{D}$ is a diagonal matrix with d_i elements along the diagonal so

$$\mathbf{D}\mathbf{y} = \mathbf{D}\mathbf{1} = \mathbf{d}.$$

This means that

$$\begin{aligned}
(\mathbf{D} - \mathbf{W})\mathbf{y} &= \lambda \mathbf{D}\mathbf{y} \\
\mathbf{D}\mathbf{1} - \mathbf{W}\mathbf{1} &= \lambda \mathbf{D}\mathbf{1} \\
\mathbf{d} - \mathbf{d} &= \lambda \mathbf{d},
\end{aligned}$$

This equation is satisfied is $\lambda = 0$ showing that $\mathbf{y} = \mathbf{1}$ is an eigenvector with zero as the eigenvalue. The significance of this is that all data points belong to the same cluster so there is no cut being performed.
