MLAI Week 8 Exercise: Unsupervised Learning

Note: An indicative mark is in front of each question. The total mark is 13. You may mark your own work when we release the solutions.

$\boxed{1}$ 1. Consider 30-bit deep colour images of size 1200 × 1200. How many possible images of this size and bit depth are there?

Solution:

For a 30-bit image, each pixel in the image can be represented by a 30-bit integer. The 30-bit integer can have 2^{30} possible values. Therefore, the total number of distinct images with a size of 1200×1200 (1200×1200 pixels) is calculated as follows.

$$\#\text{Distinct Images} = (2^{30})^{(1200\times1200)} = 2^{30\times1200\times1200} \tag{1}$$

$\boxed{2}$ 2. We are using PCA to reduce data dimensionality from 3 to 2. The top two eigenvectors are $\begin{pmatrix} 0.4729 & -0.8817 \\ -0.8817 & -0.4719 \\ 0 & 0 \end{pmatrix}$ where each column is an eigenvector. Use this PCA transformation to reduce the dimensionality of two data points $\mathbf{x}_1 = (2,3,3)^\top$ and $\mathbf{x}_2 = (4,1,0))^\top$ to 2 as $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. Show the procedures to compute $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. Assume that all data points are centred already.

Solution:
Let $\mathbf{R} = \begin{pmatrix} 0.4729 & -0.8817 \\ -0.8817 & -0.4719 \\ 0 & 0 \end{pmatrix}$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$. The projection of data points $\mathbf{x}_1$ and $\mathbf{x}_2$ from the 3-dimensional space to the 2-dimensional subspace spanned by the eigenvectors is calculated as follows.

$$\hat{\mathbf{X}} = (\hat{\mathbf{x_1}}, \hat{\mathbf{x_2}}) = \mathbf{R}^\top\mathbf{X} = \begin{pmatrix} 0.4729 & -0.8817 & 0 \\ -0.8817 & -0.4719 & 0 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 3 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} -1.6993 & 1.0099 \\ -3.1791 & -3.9987 \end{pmatrix} \tag{2}$$

"Assume that all data points are centred already." was added while preparing the solution. Centring, i.e. subtracting the mean (from the training data), before projection is a good and common practice.

$\boxed{2}$ 3. Given a dataset $\{0, 2, 4, 6, 24, 26\}$, initialise the k-means clustering algorithm with 2 cluster centres $c_1 = 3$ and $c_2 = 4$. What are the values of c_1 and c_2 after one iteration of k-means? What are the values of c_1 and c_2 after the second iteration of k-means?

Solution:
We define the centre values of cluster 1 and cluster 2 as c_1 and c_2. Let $\mathbf{X} = \{0, 2, 4, 6, 24, 26\}$. There are two steps in each iteration of K-means algorithm, aiming to find:

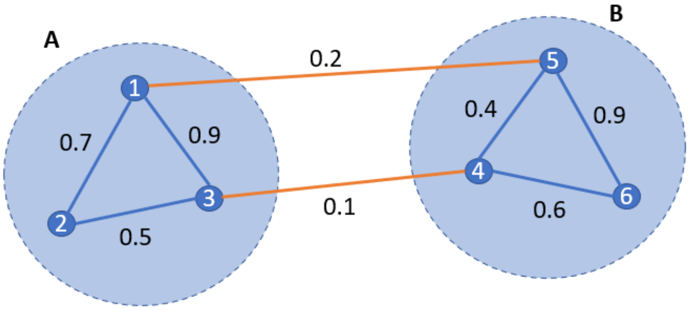
$$\min \sum_{j=1}^{2} \sum_{x^{(i)} \in \mathbf{X} \text{ allocated to } j} \left(x^{(i)} - c_j \right)^2 \tag{3}$$

In the first step we group the data points to a cluster whose centre they are closest to in terms of distance. With an initialisation of $c_1 = 3$ and $c_2 = 4$ in the first iteration, we allocate 0 and 2 to cluster 1 and 4, 6, 24, 26 to cluster 2. In the second step, we set new centres to the clusters found in the first step. In this step, we simply set the centre to the mean value of data points in the same cluster:

$$c_i = \mathbf{E}[X_i] = \frac{1}{|X_i|} \sum_{j}^{|X_i|} x_i^{(j)} \quad , x_i^{(j)} \in X_i, \ i = 1, 2 \tag{4}$$

where $X_1 = \{0, 2\}$, $X_2 = \{4, 6, 24, 26\}$. Therefore the centres of cluster 1 and cluster 2 are updated to $c_1 = 1$ and $c_2 = 15$ in Eqn. (4). In the next iteration, we repeat step 1 and step 2. With the $c_1 = 1$ and $c_2 = 15$ from previous iteration, cluster 1 and cluster 2 will contain data points $X_1 = \{0, 2, 4, 6\}$, $X_2 = \{24, 26\}$ in the second iteration. The new centres for cluster 1 and cluster 2 are updated to $c_1 = 3$ and $c_2 = 25$.

$\boxed{2}$ 4. For the graph below, compute the normalised cut $Ncut(A, B)$.



Solution:
$Ncut(A, B) = cut(A, B)\frac{Vol(A)+Vol(B)}{Vol(A)Vol(B)}$

We first show the similarity matrix for the data $S = \begin{pmatrix} 1 & 0.7 & 0.9 & 0 & 0.2 & 0 \\ 0.7 & 1 & 0.5 & 0 & 0 & 0 \\ 0.9 & 0.5 & 1 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 0.4 & 0.6 \\ 0.2 & 0 & 0 & 0.4 & 1 & 0.9 \\ 0 & 0 & 0 & 0.6 & 0.9 & 1 \end{pmatrix}$.
Note that the similarity of data between itself is 1 because the Gaussian Kernel in Eqn. (5) is equal to 1 when $x_i = x_j$. The similarity of data points without direct link in the graph is set to 0 because we assume $|x_i - x_j|^2 \to \infty$.

$$\mathbf{W}(i, j) = \exp \frac{-|x_i - x_j|^2}{\sigma^2} \tag{5}$$

By the definition of volume in page 38 of week 8 slide, $Vol(A) = \sum s_{ij},\ i = 1,2,3; j = 1,2,3,4,5,6$ with s_{ij} is an element in S and i and j are the entries of S. So $Vol(A) = 7.5$. Similarly, $Vol(B) = \sum s_{ij},\ i = 4,5,6; j = 1,2,3,4,5,6$ with $Vol(B) = 7.1$. Finally, the $cut(A,B) = \sum s_{ij},\ i = 1,2,3; j = 4,5,6$ with $cut(A,B) = 0.3$. Note that S is symmetrical, so $cut(A,B)$ can be also calculated by $cut(A,B) = \sum s_{ij},\ i = 4,5,6; j = 1,2,3$ with the same value.

With $Vol(A), Vol(B), cut(A,B)$ calculated above, the normalised cut $Ncut(A,B)$ is obtained by:

$$Ncut(A,B) = cut(A,B)\frac{Vol(A)+Vol(B)}{Vol(A)Vol(B)} = 0.3 \times \frac{7.5+7.1}{7.5 \times 7.1} = 0.08225 \tag{6}$$

3 5. An alternative to derive PCA is to minimize the reconstruction error (Slide 26) for all N data samples $\mathbf{x}^{(i)}, i = 1, \cdots, N$, assuming that the mean $\boldsymbol{\mu} = \sum_i \mathbf{x}^{(i)}$ is zero. Take this approach to derive the first principal component (as the first eigenvector of the data matrix).

Solution: The most elegant proof is from https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/dimensionality/paper-1x2.pdf.

Let us denote an **orthonormal** projection vector as $\mathbf{u}$. It will project an input vector $\mathbf{x}$ to a scalar $y = \mathbf{u}^\top\mathbf{x}$. Using this scalar to reconstruct $\mathbf{x}$ as $\hat{\mathbf{x}} = \mathbf{u}y = \mathbf{u}\mathbf{u}^\top\mathbf{x}$.

$$\begin{aligned}
& \text{Reconstruction error} \\
= & \sum_{i=1}^{N} \left\|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)})\right\|^2 && (7) \\
= & \sum_{i=1}^{N} \left\|\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right\|^2 && (8) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)})^\top(\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right) && (9) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top} - \mathbf{x}^{(i)^\top}\mathbf{u}\mathbf{u}^\top)(\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right) && (10) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top}\mathbf{x}^{(i)} - \mathbf{x}^{(i)^\top}\mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)} - \mathbf{x}^{(i)\top}\mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)} + \mathbf{x}^{(i)^\top}\mathbf{u}\mathbf{u}^\top\mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right) && \text{note } \mathbf{u}^\top\mathbf{u}\text{=1} \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top}\mathbf{x}^{(i)} - \mathbf{x}^{(i)^\top}\mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right) && (11) \\
= & \text{ constant} - \sum_{i=1}^{N} \left(\mathbf{x}^{(i)^\top}\mathbf{u}\mathbf{u}^\top\mathbf{x}^{(i)}\right) && (12) \\
= & \text{ constant} - \sum_{i=1}^{N} \left(\mathbf{u}^\top\mathbf{x}^{(i)}\right)^2 && (13)
\end{aligned}$$

Note $\mathbf{u}^\top\mathbf{x}^{(i)}$ is the projection $y^{(i)} = \mathbf{u}^\top\mathbf{x}^{(i)}$ so the summation in Eqn. (13) is the variance. Maximising the variance minimises the reconstruction error so we have the same solution as that by variance maximisation.

$\boxed{3}$ 6. In spectral clustering, show that the smallest eigenvalue for the formulated generalized eigenvalue problem on Slide 41 is 0 with the corresponding generalized eigenvector $\mathbf{y} = \mathbf{1}$, hence the same "representation/embedding" for all nodes.

Solution:
$(D - W)y = \lambda Dy$
$(D - W)y = \lambda D^{\frac{1}{2}} D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)y = \lambda D^{\frac{-1}{2}} D^{\frac{1}{2}} D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} D^{\frac{1}{2}} y = \lambda I D^{\frac{1}{2}} y$
Make the substitution of $z = D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} z = \lambda z$
If we set y to $\mathbf{1}$ we get
$z = D^{\frac{1}{2}}\mathbf{1}$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} D^{\frac{1}{2}}\mathbf{1} = \lambda D^{\frac{1}{2}}\mathbf{1}$
$D^{\frac{-1}{2}}(D - W)I\mathbf{1} = \lambda D^{\frac{1}{2}}\mathbf{1}$
If we observe $(D - W)\mathbf{1}$, we can see that it's a summation of the rows of $D - W$
In a row of the Laplacian, $D - W$, we have the degree of the ith node, d_i on the diagonal, and all of the negative weights of the edges connected to node i filling the rest of the row. Therefore, adding across a row gives us:
$d_i + \sum_{j=1}^{n}(-w_{i,j}) = \sum_{j=1}^{n} w_{i,j} + \sum_{j=1}^{n}(-w_{i,j}) = 0$
Which means $(D - W)\mathbf{1} = \mathbf{0}$ and therefore the eigenvector corresponds to the eigenvalue $\lambda = 0$.