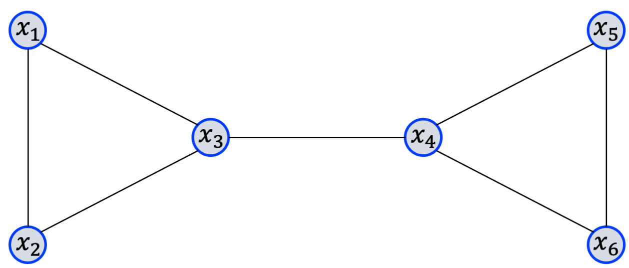
MTH793P **Advanced Machine Learning, Semester B, 2023/24**

Coursework 5

---

1 Evaluating clustering algorithms

In this problem we assume that we have a data set $\{x_1, \ldots, x_6\} \subset \mathbb{R}^2$ that looks as follows:



The Euclidean distances along all the edges segments is equal to 1.
We will consider to alternative clusterings on these data:

- $\mathcal{C} = \{C_1, C_2\}$, with $C_1 = \{x_1, x_2, x_3\}$ and $C_2 = \{x_4, x_5, x_6\}$.
- $\mathcal{C}' = \{C_1', C_2', C_3'\}$, with $C_1' = \{x_1, x_2\}$, $C_2' = \{x_3, x_4\}$, and $C_3' = \{x_5, x_6\}$.

1. Compute the Dunn-Index (DI) for $\mathcal{C}$ and $\mathcal{C}'$, using the single-linkage inter-cluster distance, and the diameter intra-cluster distance. Which clustering is better?
2. Compute the mean Silhouette Coefficient (SC) for $\mathcal{C}$ and $\mathcal{C}'$. Which one is better?
3. Suppose that we know that $\mathcal{C}'$ is the ground-truth for this dataset. Compute the Rand Index (RI) for $\mathcal{C}$.

2 Matrix Factorisation and SVD

1. Find vectors $u \in \mathbb{R}^2$ and $v \in \mathbb{R}^3$ such that the following identity is satisfied for all the given values:

$$uv^\top = \begin{pmatrix} 1 & 0 & ? \\ -2 & ? & 4 \end{pmatrix}.$$

What value do you obtain at the missing entry denoted by a question mark?
HINT: What should be the rank of the matrix?

2. Compute (by hand) the singular value decomposition of the matrix

$$X = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

Hint: (1) Find the eigenvalues of $XX^\top$ by computing the characteristic polynomial, and deduce the singular values of X. (2) Find the eigenvectors of both $XX^\top$ and $X^\top X$, and conclude what the U and V matrices in the SVD should be.

3. Given the matrix

$$X := \begin{pmatrix} -2 & 3 & 2 \\ 2 & 2 & 3 \end{pmatrix},$$

we are interested in finding the best approximation $\hat{X}$, with $\text{rank}(\hat{X}) = 1$. By 'best' we mean with respect to the Frobenius norm. Use SVD to find $\hat{X}$ (by hand).