PRINCIPLES OF MACHINE LEARNING

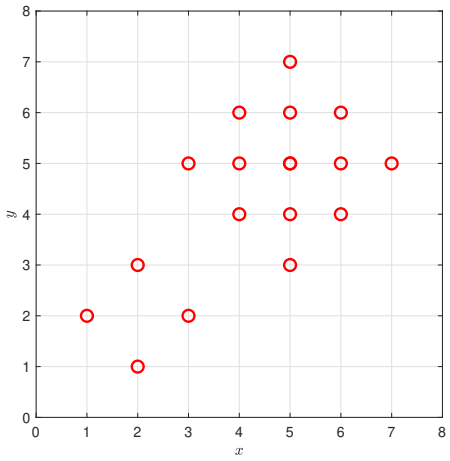
UNSUPERVISED LEARNING

ACADEMIC YEAR 2021/2022

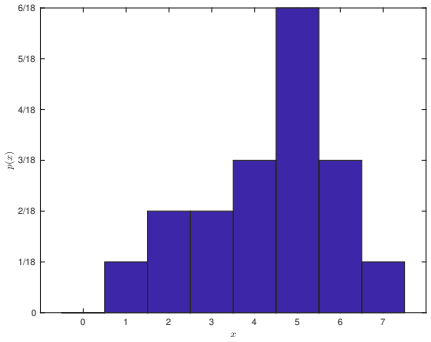
QUEEN MARY UNIVERSITY OF LONDON

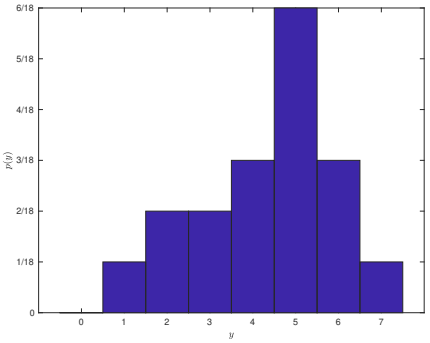
SOLUTIONS

EXERCISE ♯1 (SOL): Let's plot the dataset first:

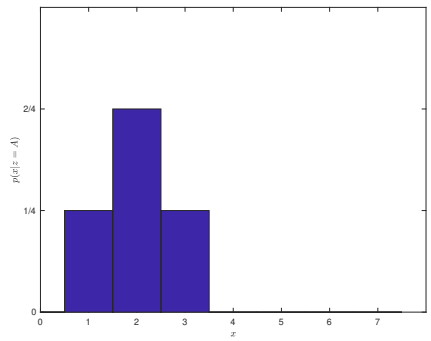


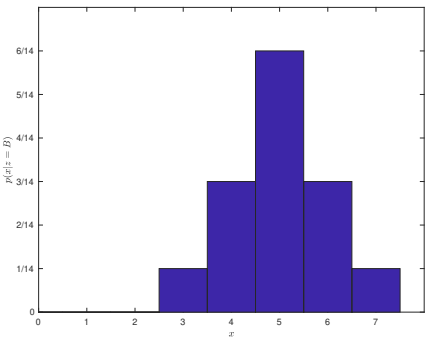
The histograms for the marginal densities $p(x)$ and $p(y)$ are:





The histograms for the marginal densities $p(x|z = A)$ and $p(x|z = B)$ are:



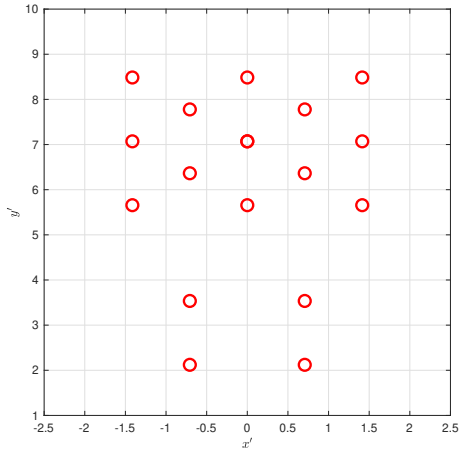


The mean $\boldsymbol{\mu}$ of $p(x, y)$ can be calculated by averaging x and y. The result is $\boldsymbol{\mu} = [4.3, 4.3]^T$. We shouldn't expect $\boldsymbol{\Sigma}$ to be diagonal: the plot shows that x and y are not independent, specifically, higher values of x are associated to higher values of y.

As previously mentioned, $\boldsymbol{\mu} = [\mu_x, \mu_y]^T$, where μ_x and μ_y are the means of the marginal densities $p(x)$ and $p(y)$. The probability density $p(x, y)$ cannot be expressed as the product of $p(x)$ and $p(y)$, as a and y are not independent.

If we treat the samples where $z = A$ and $z = B$ separately, we can build two new probability densities, namely $p(x, y|z = A)$ and $p(x, y|z = B)$. Their means are $\boldsymbol{\mu}_A = [2, 2]^T$ and $\boldsymbol{\mu}_B = [5, 5]^T$. Note that $\boldsymbol{\mu} = (4 \times \boldsymbol{\mu}_A + 14 \times \boldsymbol{\mu}_B)/18$, where 4 is the number of A samples, 14 is the number of B samples and 18 is the total number of samples in the dataset.

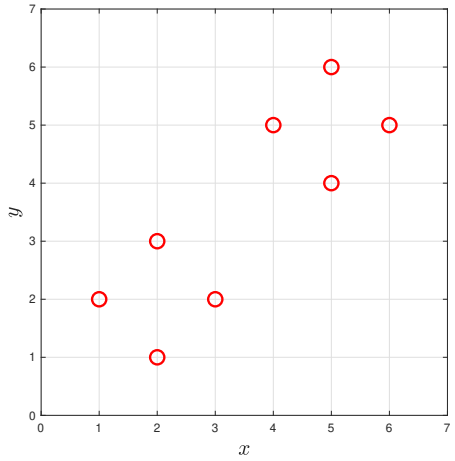
After applying PCA we obtain:



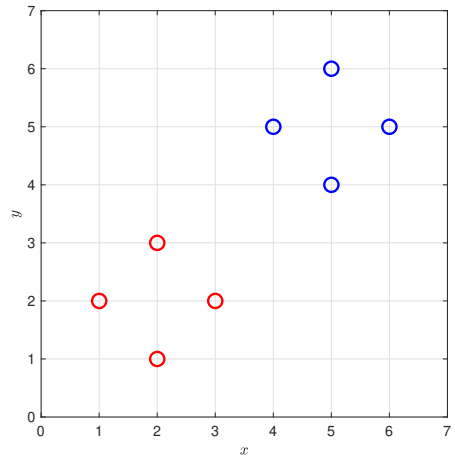
Note that PCA rotates the dataset so that the directions along which data spreads are aligned with one of the new axes. We can obtain the transformation visually or use Python (Matlab, R...).

In this case, we would expect the covariance matrix of the new probability distribution $p(x', y')$ to be diagonal, as the new attributes (*components*) are independent. Independence also means that $p(x', y') = p(x')p(y')$.

EXERCISE ♯2 (SOL): The first step is always to plot the dataset:



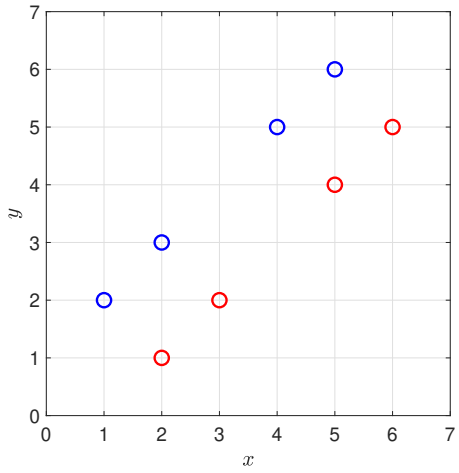
If the initial values of the prototypes are $\boldsymbol{\mu}_1 = [0,0]^T$ and $\boldsymbol{\mu}_2 = [7,7]^T$ the 2 identified clusters will be as follows:



The final prototypes will be in the locations $\boldsymbol{\mu}_1 = [2,2]^T$ and $\boldsymbol{\mu}_2 = [5,5]^T$. The intra-cluster sample scatter can be calculated by adding the distances between any two samples within a cluster or by adding the distances between the samples and the cluster center. Let's do it both ways for the first cluster:

$$\begin{aligned} I_1 &= \left((1^2+1^2)+(2^2+0^2)+(1^2+1^2)+(1^2+1^2)+(0^2+2^2)+(1^2+1^2)\right) = \\ &= (2+4+2+2+4+2) = 16 \text{ (distances between two pair of samples)} \\ I_1 &= 4 \times \left((1^2+0^2)+(1^2+0^2)+(1^2+0^2)+(1^2+0^2)\right) = 16 \text{ (distances to cluster centers)} \end{aligned}$$

Cluster 2 has the same intra-cluster sample scatter, $I_2 = 16$. Therefore, the overall quality is $I_1 + I_2 = 32$. If the initial values of the prototypes are $\boldsymbol{\mu}_1 = [1,6]^T$ and $\boldsymbol{\mu}_2 = [6,1]^T$ the 2 identified clusters will be:



The final prototypes will be in the locations $\boldsymbol{\mu}_1 = [4,3]^T$ and $\boldsymbol{\mu}_2 = [3,4]^T$. The intra-cluster sample scatter will be the same for both. Let's calculate it for the first cluster based on the distances to the cluster centre:

$$I_1 = 4 \times \left((1^2+1^2)+(2^2+2^2)+(1^2+1^2)+(2^2+2^2)\right) = 4 \times (2+8+2+8) = 80$$

The quality of this clustering arrangement is $I_1 + I_2 = 160$.

k-means might provide different final solutions depending on the initial location of prototypes. k-mean solutions can be seen as local minima where the algorithm gets stuck. In this case, the first clustering arrangement happens to also be the global minimum.