Exercise 5 - Big Data Analytics

Thi Tra My Nguyen-230005

January 2, 2021

1 Maximum Likelihood Estimation

Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$\begin{aligned} L(\theta) = P(X=3) \times P(X=0) \times P(X=2) \times P(X=1) \times P(X=3) \\ \times P(X=2) \times P(X=1) \times P(X=0) \times P(X=2) \times P(X=1) \end{aligned}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \Big(\frac{2\theta}{3}\Big)^2 \Big(\frac{\theta}{3}\Big)^3 \Big(\frac{2(1-\theta)}{3}\Big)^3 \Big(\frac{1-\theta}{3}\Big)^2$$

Let us look at the log likelihood function.

$$\begin{aligned} l(\theta) = \log L(\theta) &= \sum_{i=1}^{n} \log P(X_i|\theta) \\ &= 2\Big(\log\frac{2}{3}\Big) + 3\Big(\log\frac{1}{3} + \log\theta\Big) + 3\Big(\log\frac{2}{3} + \log(1-\theta)\Big) + 2\Big(\log\frac{1}{3} + \log(1-\theta)\Big) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

The derivative of $l(\theta)$:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

and the solution gives us the maximum likelihood estimation which is 0.5.

2 Spectral Clustering

2.1

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$L = D - A = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

2.2

Let k be the number of connected components of G and $S_1, S_2, S_3, ..., S_k$ be the connected components of G.

We will prove that the number eigenvalues O is at least the number of connected components. In other word the number of connected components is at most the multiplicity the eigenvalue 0 of the Laplacian matrix.

Towards this, we define the following set of k vectors $u_1, u_2, u_3, ..., u_k$ where u_i is defined as follows

$$u_i(j) = \begin{cases} 1 & \text{if } j \in S_i \\ 0 & \text{otherwise} \end{cases}$$

Then, we have

$$\begin{cases} < u_i, u_j >= 0 & \text{for all } i \neq j \\ Lu_i = 0 & \text{for all } i \in \{1, 2, ..., k\} \end{cases}$$

There are k mutually orthogonal vectors that are all eigenvectors of L corresponding to the eigenvalue 0.

Thus, the multiplicity of the eigenvalue 0 of L is at least k, which equals the number of connected components.

2.3

2.3.1

We will prove that the multiplicity of eigenvalue 0 of L is at most k.

We have

$$u^T L u = \sum_{(i;j) \in E} (u_i - u_j)^2 \tag{1}$$

Then $u^T L u = 0$ if and only if $u_i = u_j$ for all $(i; j) \in E$. That is, u should have equal value for all vertices in a connected component.

Thus, if $Lu = 0$, then $u^T L u = 0$, which implies that there exists scalars $\alpha_1, \alpha_2, \alpha_3, ..., \alpha_k$ such that $u = \sum_{i=1}^{k} \alpha_i u_i$.

This implies that every eigenvector of L corresponding to the eigenvalue 0 is contained in the subspace spanned by $\{u_1, u_2, ..., u_k\}$.

Thus, there are at most k linear independent eigenvectors of L corresponding to the eigenvalue 0.

$\longrightarrow$ The multiplicity of eigenvalues 0 of L is at most the connected components.

Combine with the result of the previous part, we can conclude that the number of connected components is equal to the multiplicity of eigenvalue 0.

2.3.2 How to make use of that for clustering the nodes

The implementation of spectral clustering: 3 main steps

1. We form a graph between data points. The edges of the graph capture the similarities between the points.
2. Compute the first k eigenvectors of its Laplacian matrix to define a feature vector for each object.
3. Run k-means on these features to separate objects into k-classes.

So how to choose k?

$\longrightarrow$ When the similarity graph is not fully connected, the multiplicity of the eigenvalue 0 gives us an estimation of k.

3 Principal Component Analysis

3.1

Perform PCA on the dataset

$$A = \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 3 & 2 \\ 4 & 3 \\ 5 & 5 \end{bmatrix}$$

$$A_{mean} = \begin{bmatrix} 3,4 & 3 \end{bmatrix}$$

Center the data

$$A^* = \begin{bmatrix} -1,4 & -1 \\ -0,4 & 0 \\ -0,4 & -1 \\ 0,6 & 0 \\ 1,6 & 2 \end{bmatrix}$$

Covariance matrix

$$C = \begin{bmatrix} var(X) & cov(X,Y) \\ cov(Y,X) & var(Y) \end{bmatrix} = \begin{bmatrix} 1,3 & 2,5 \\ 1,25 & 1,5 \end{bmatrix}$$

Eigenvalues and eigenvectors of the covariance matrix

The characteristics polynomial of C

$$\begin{aligned} f(\lambda) &= det(C - \lambda I) \\ &= det\Big(\begin{bmatrix} 1,3-\lambda & 1,25 \\ 1,25 & 1,5-\lambda \end{bmatrix}\Big) \\ &= \Big(\lambda - \frac{2,8+\sqrt{6,29}}{2}\Big)\Big(\lambda - \frac{2,8-\sqrt{6,29}}{2}\Big) \end{aligned}$$

The roots of the polynomial are

$$\lambda_1 = \frac{2,8+\sqrt{6,29}}{2}$$
$$\lambda_2 = \frac{2,8-\sqrt{6,29}}{2}$$

The eigenvectors associated to $\lambda_1 = \frac{2,8+\sqrt{6,29}}{2}$ are the vector $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ that solve the equation

$$\begin{bmatrix} 1,3-\lambda_1 & 1,25 \\ 1,25 & 1,5-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow u = \begin{bmatrix} -0,6783269 \\ -0,73476024 \end{bmatrix}$$

The eigenvectors associated to $\lambda_2 = \frac{2,8-\sqrt{6,29}}{2}$ are $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ that solve the equation

$$\begin{bmatrix} 1,3-\lambda_2 & 1,25 \\ 1,25 & 1,5-\lambda_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow v = \begin{bmatrix} -0,73476024 \\ 0,6783269 \end{bmatrix}$$

Find the transformation matrix V based on the selection of PC_s

The eigenvector with the highest eigenvalue is the principal component. The less significant component can be ignore, so as to reduce the dimensions of the dataset

$$\longrightarrow V = \begin{bmatrix} -0,6783269 \\ -0,73476024 \end{bmatrix}$$

Derive the new dataset by taking $Y = AV$

$$Y = \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 3 & 2 \\ 4 & 3 \\ 5 & 5 \end{bmatrix} \begin{bmatrix} -0,6783269 \\ -0,73476024 \end{bmatrix} = \begin{bmatrix} -2,826174 \\ -4,23926 \\ -3,05045 \\ -4,917588 \\ -7,0654357 \end{bmatrix}$$

3.2

Naive dimensionality reduction

$$B = \begin{bmatrix} 2 \\ 3 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

$$\mu_B = 3,4$$

$$\sigma_B^2 = \frac{1}{5} \sum_{i=1}^{5} (x_i - 3,4)^2 = 1,04$$

PCA reduction

$$Y = \begin{bmatrix} -2,826174 \\ -4,23926 \\ -3,05045 \\ -4,917588 \\ -7,0654357 \end{bmatrix}$$

$$\mu_Y = -4,51059$$

$$\sigma_Y^2 = \frac{1}{5}\sum_{i=1}^{5}[y_i - (-4,51059)]^2 = 2,12319$$

What do you observe? What does this tell you about projection quality?

We can see that the variance of Y is much higher than the variance of B.

To tell about the projection quality, I will calculate the variance of each dimension in the original dataset by using the given formula:

$$var_A = \begin{bmatrix} 1,04 & 1,2 \end{bmatrix}$$

Total variance is $2,24$. I will devide individual variance by the total variance, I will see how much variance each variable explains

$\rightarrow$ variable 1 explains $46,43\%$ of the total variance.

PCA computes a new variable (principal component) and the total variance remains the same.

I will devide the σ_Y^2 by the total variance: $94,79\%$

$\rightarrow$ the new variable explains $94,79\%$ of the total variance.

Therefore, when we keep only this projection, information loss will be just $5,21\%$.

3.3

2-dimensional Possion distributed dataset

$$A = \begin{bmatrix} 41 & 46 \\ 49 & 52 \\ 52 & 64 \\ 47 & 53 \\ 64 & 54 \\ 45 & 47 \\ 45 & 56 \\ 37 & 63 \\ 46 & 64 \\ 53 & 53 \\ 42 & 58 \\ 47 & 59 \\ 52 & 55 \\ 45 & 60 \\ 42 & 63 \end{bmatrix}$$

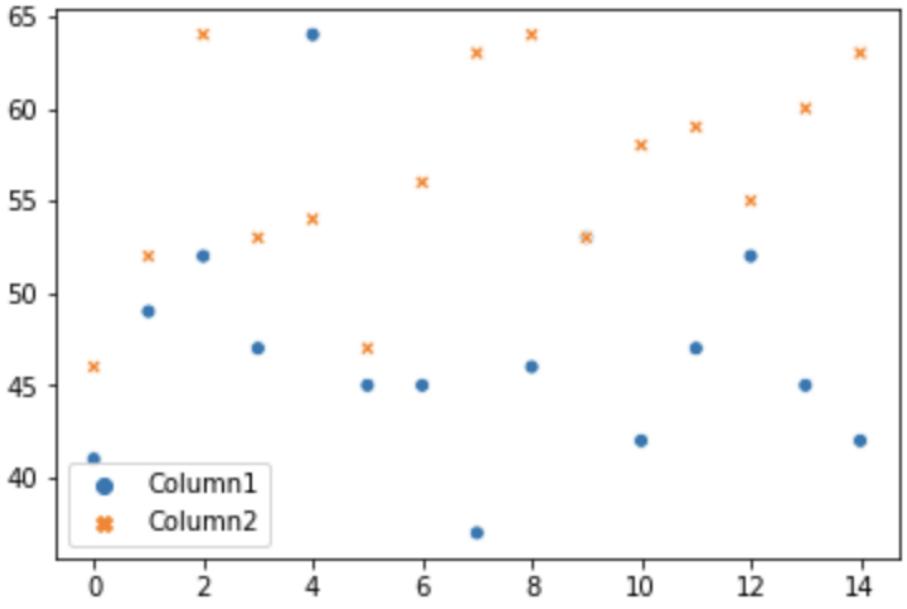

Figure 1: Scatter plot

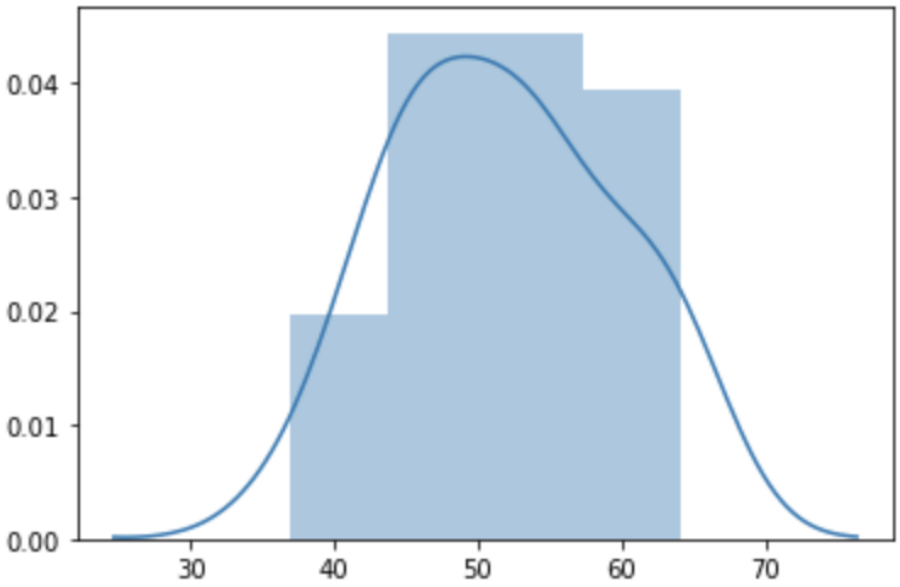

Figure 2: Distribution

$$varA = \begin{bmatrix} 38,515556 & 31,448889 \end{bmatrix}$$

$$\text{total variance} = 69,964445$$

After PCA we only keep *the first principal component* that has variance 41.32063110462275, which only explains $59,06\%$ total variance.

The reason that PCA fails with this dataset is that PCA is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to find hidden linear correlations between variables of the dataset. Therefore, when the data is not linear correlated, PCA is not enough.

We apply the Central Limit Theorem to find a linear decomposition by maximizing non-Gaussianity of the component.

Central Limit Theorem idea: any mixture of components will be more Gaussian than the components themselves.