CSCI-SHU 360 Machine Learning
Solution to homework 1

Josiah Li yl11912@nyu.edu

February 15, 2025

1 Simpson's Paradox

1.1

We use A to indicate that I win, and B to indicate that my friend wins.
We use M_i to indicate the player plays on the ith machine. We have

$$P(A|M_1) = \frac{2}{5}, P(A|M_2) = \frac{21}{104}$$
$$P(B|M_1) = \frac{3}{10}, P(B|M_2) = \frac{1}{6}$$

Thus, on both machine 1 and machine 2, I am more likely to win.

1.2

$$P(A) = \frac{40+210}{40+60+210+830} = \frac{25}{114}$$
$$P(B) = \frac{30+14}{30+70+14+70} = \frac{11}{46}$$

Thus, my friend is more likely to win.

1.3

$$P(A) = P(A|M_1)P(M_1) + P(A|M_2)P(M_2)$$
$$P(B) = P(B|M_1)P(M_1) + P(B|M_2)P(M_2)$$

Thus, when the $P(M_i)$ increses, the total winning probability of player will get closer to the winning rate of machine i.
Thus, as I played much more rows on machine 2, and the overall winning rate on machine 2 is smaller than machin 1, my winning rate will get closer to the smaller winning rate(which is the winning rate of machine 2), and my friend's winning rate will be closer to the higher winning rate(which is the winning rate of machine 1), and that lead to the result that my winning rate is high on both machine, but the overall winning rate is lower than my friend.

2 Matrix as Operations

2.1

$$a_1 = (1, 0), b_1 = (2, 0)$$
$$a_2 = (0, 1), b_2 = (3, 2)$$

Suppose $W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$ We have

$$b_1 = Wa_1 = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} W_{11} \\ W_{21} \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$
$$b_2 = Wa_2 = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} W_{12} \\ W_{22} \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

Thus, we have

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

2.2

From $\tan(\alpha) = 2$, we have $\sin(\alpha) = \frac{2\sqrt{5}}{5}$ and $\cos(\alpha) = \frac{\sqrt{5}}{5}$(Suppose $\alpha \in (0, \pi)$)
Moreover, we have the formula of the rotation matrix:

$$R = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

When θ is the clockwise rotation angle.
Thus,

$$V = \begin{bmatrix} \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \\ -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{bmatrix}$$

As Σ scales the x-axis by a factor of 4 in two-dimensional space, the matrix is:

$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$

By using the same approach, we can get matrix U. $\tan(\beta) = \frac{1}{2}, \sin(\beta) = \frac{\sqrt{5}}{5}, \cos(\beta) = \frac{2\sqrt{5}}{5}$,
And

$$U = \begin{bmatrix} \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix}$$

Multiplying three matrices together, we have:

$$\begin{aligned} U\Sigma V &= \begin{bmatrix} \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \\ -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{bmatrix} \\ &= \begin{bmatrix} \frac{2\sqrt{5}}{5} & -\frac{\sqrt{5}}{5} \\ \frac{\sqrt{5}}{5} & \frac{2\sqrt{5}}{5} \end{bmatrix} \begin{bmatrix} \frac{4\sqrt{5}}{5} & \frac{8\sqrt{5}}{5} \\ -\frac{2\sqrt{5}}{5} & \frac{\sqrt{5}}{5} \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix} \\ &= W \end{aligned}$$

Thus, we have

$$U\Sigma V = W$$

2.3

We have

$$W = \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix}, W^T = \begin{bmatrix} 2 & 0 \\ 3 & 2 \end{bmatrix}$$

Thus

$$W^TW = \begin{bmatrix} 2 & 0 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 13 \end{bmatrix}$$

$$\begin{aligned} det(W^TW - \lambda I) &= det\left(\begin{bmatrix} 4-\lambda & 6 \\ 6 & 13-\lambda \end{bmatrix}\right) \\ &= (13-\lambda)(4-\lambda) - 36 \\ &= (\lambda - 16)(\lambda - 1) \end{aligned}$$

Thus, the eigenvalues and eigenvectors of W^TW are:

$$\begin{cases} \lambda_1 = 1, u_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix} \\ \lambda_2 = 16, u_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \end{cases}$$

By applying the linear transformation W, the unit circle is rotated clockwise by α degrees, and then scales along x-axis by 4, finally, rotated counter-clockwise by β degrees. And this transformation results in a ellipse.

2.4

The determinant of W is $det(W) = 4$. As in the previous transformation, we scaled along x-axis by 4, the semi-major axis is 4, and the semi-minor axis is 1. Thus, the area of the ellipse is $1 \times 4 \times \pi = 4\pi$ As the origin unit circle has an area of π, we came up with the hypothesis that the determinant shows the change of the area of a transformed shape. Using a sentence to explain that $det(AB) = det(A)det(B)$. Linear transformation AB means to perform first A and B, and this means the total effect of the change of area will first be the effect of A, then multiplys the effect of B, which in total is $det(A)det(B)$.

3 Some Practices

3.1

As X is a discrete random variable, X^3 is also a discrete random variable. As $Var(X^3) = \mathbb{E}[X^6] - (\mathbb{E}[X^3])^2$ and $Var(X^3) \geq 0$, we can prove that:

$$(\mathbb{E}[X^3])^2 \leq \mathbb{E}[X^6]$$

3.2

$$(\mathbb{E}[X^3])^2 = (\sum_{i=1}^{n} p_i X_i^3)^2$$

$$\mathbb{E}[X^6] = \sum_{i=1}^{n} p_i X_i^6$$

We expand the square of expectation, which means:

$$(\mathbb{E}[X^3])^2 = (\sum_{i=1}^{n} p_i X_i^3)^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j X_i^3 X_j^3$$

We know that $2ab \leq a^2 + b^2$, and by letting $a = X_i^3$ and $b = X_j^3$, we can conclude that $X_i^3 X_j^3 \leq \frac{X_i^6 + X_j^6}{2}$ Thus, we can prove that:

$$\begin{aligned}(\mathbb{E}[X^3])^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j X_i^3 X_j^3 \\ &\leq \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{X_i^6 + X_j^6}{2} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{X_i^6}{2} + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{X_j^6}{2}\end{aligned}$$

As for i, we have $\sum_{i=1}^{n} p_i = 1$ (Which is also true for j), we can prove that:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{X_i^6}{2} + \sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j \frac{X_j^6}{2} &= \sum_{i=1}^{n} p_i \frac{X_i^6}{2} + \sum_{j=1}^{n} p_j \frac{X_j^6}{2} \\ &= \sum_{i=1}^{n} p_i X_i^6 \\ &= \mathbb{E}[X^6]\end{aligned}$$

Thus, we can prove:

$$(\mathbb{E}[X^3])^2 \leq \mathbb{E}[X^6]$$

3.3

For PSD matrices A we have $x^T A x \geq 0$ for every x. And for $0 \leq \lambda \leq 1$ we have $\lambda \geq 0, 1 - \lambda \geq 0$. As $x^T A x \geq 0, x^T B x \geq 0$ we can conclude that $x^T \lambda A x \geq 0, x^T (1-\lambda) B x \geq 0$ Thus we can prove:

$$\begin{aligned}x^T \lambda A x + x^T (1-\lambda) B x &\geq 0 \\ x^T (\lambda A x + (1-\lambda) B) x &\geq 0\end{aligned}$$

Thus we can prove that the matrix $\lambda A x + (1-\lambda) B)$ is also PSD for $0 \leq \lambda \leq 1$.

4 Density Estimation of Multivariate Gaussian

4.1

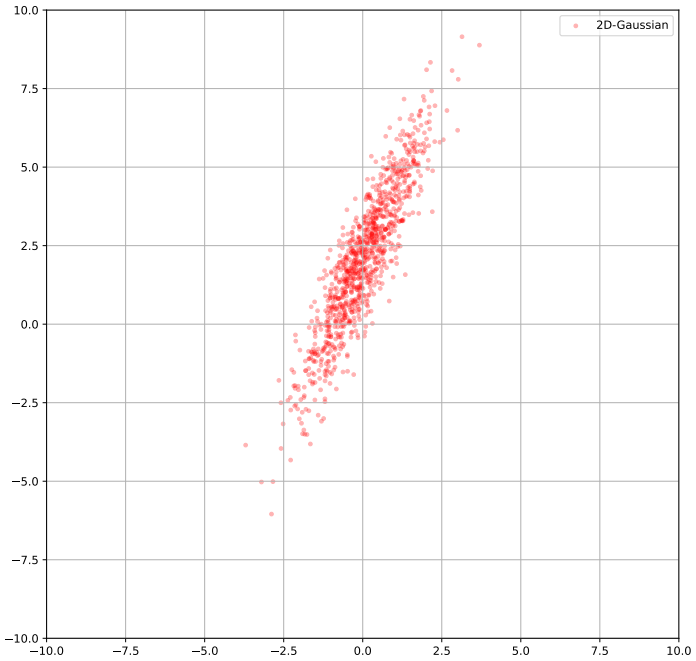
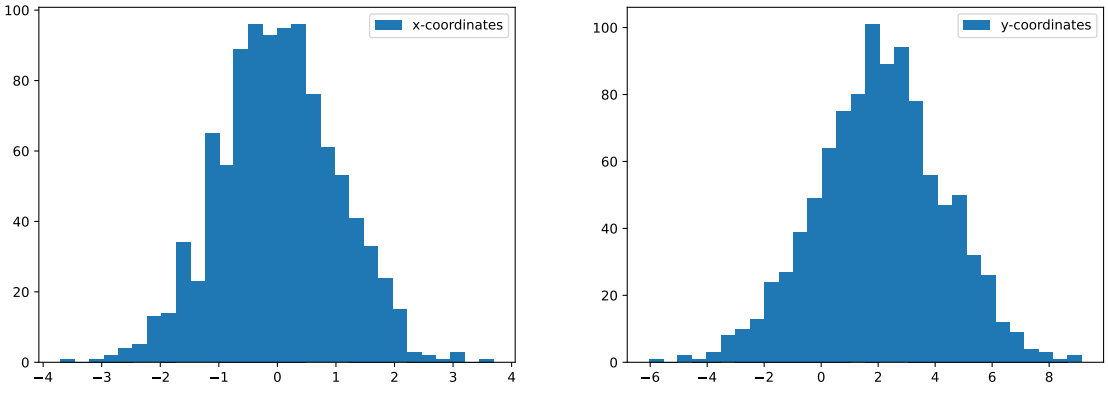


Figure 1: the 1000 points in X sampled from multivariate Gaussian distribution

By using numpy, we have the estimated mean and covariance. Which is

$$mean(X) = (0.01909265, 2.06052385)$$
$$cov(X) = \begin{bmatrix} 1.03589238 & 2.06244165 \\ 2.06244165 & 5.08839176 \end{bmatrix}$$

4.2



(a) x (b) y

Figure 2: Histogram of x and y coordinates of the 1000 points

4.3

By the histogram, we can assume the x and y coordinates follow some Gaussian distribution. By using numpy, we get the mean and the variance which is:

$$\mu_x = 0.019092651459051646, \sigma_x^2 = 1.0348564864741474$$
$$\mu_y = 2.0605238541899022, \sigma_y^2 = 5.083303371449154$$

4.4

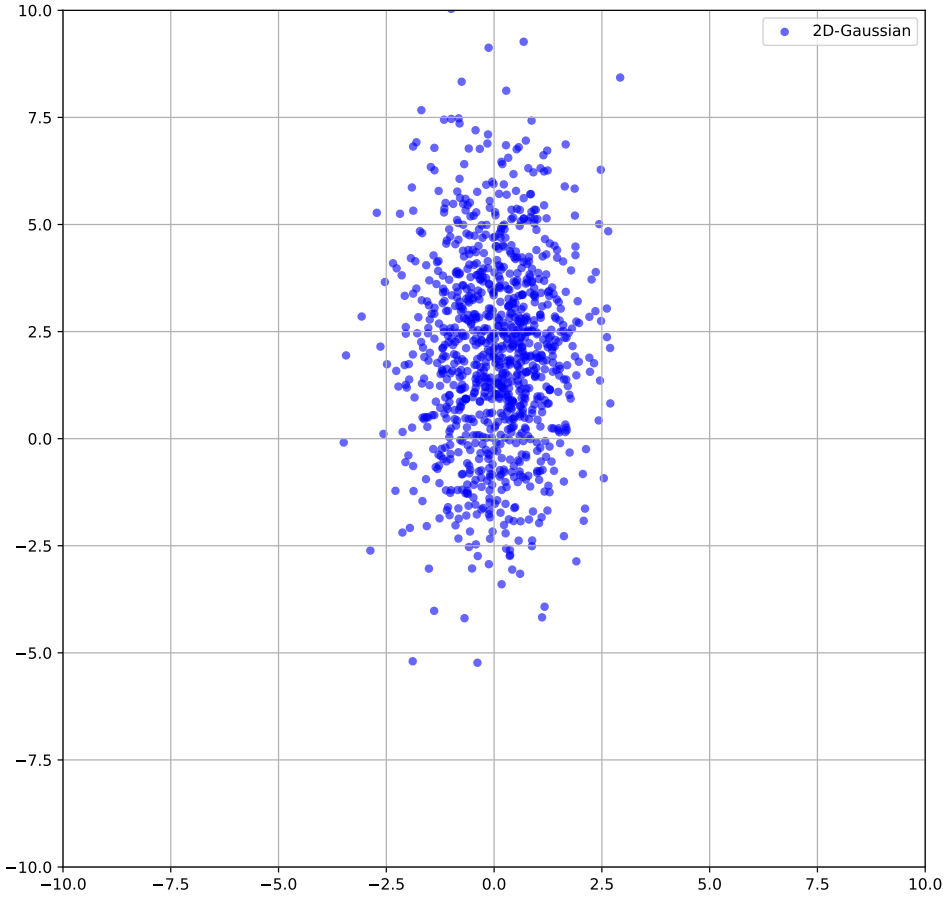


Figure 3: The 2D scatter plot of 1000 points

Difference: New 2D scatter graph is genearlly symmetrical along the y axis, while the original graph is not. The linear relationship between x and y (the covariance is not zero in the first case) lead to the difference.

4.5

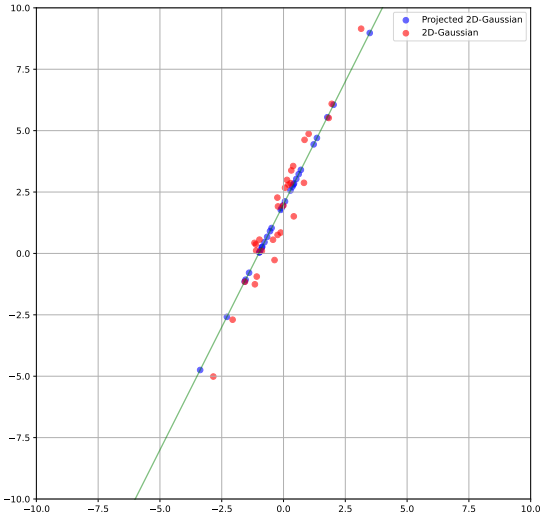


Figure 4: The project points

4.6

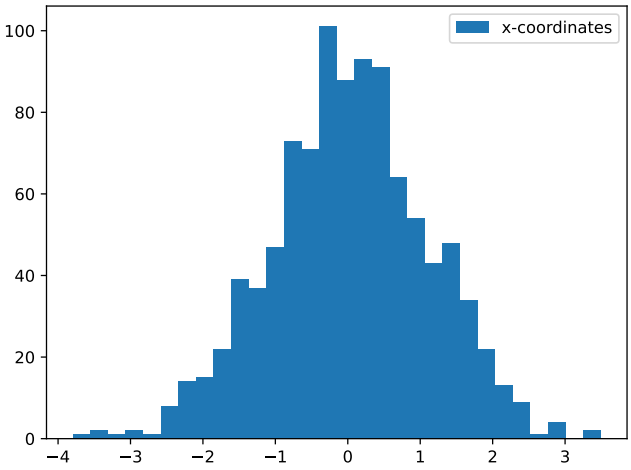


Figure 5: The histogram of the x-coordinates of the projected points

From the histogram, we can assume that the x-coordinates follow some Gaussian distribution. By using numpy, we can get the mean and the variance. Which is:

$$\mu = 0.028028071967771222, \sigma^2 = 1.1843834728271614$$

5 Matrix/Tensor Transpose

5.1

Check the corresponding ipynb file.

6 K Means clustering

6.1

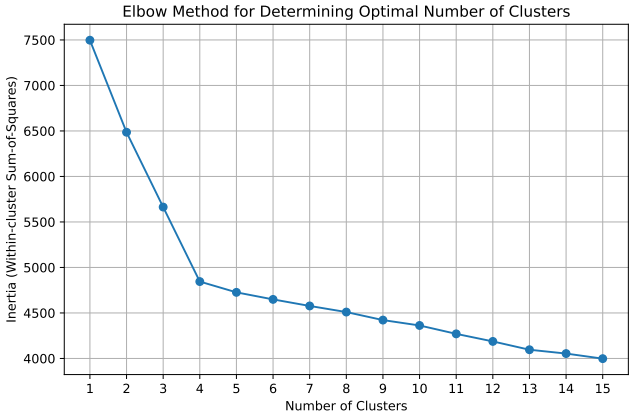


Figure 6: Inertia values corresponding to cluster number

By the figure, we can find that the "elbow" is 4, thus, 4 clusters should be used for this data.

6.2

By setting the cluster number to 4. We re-apply k-means and there are 25 observations placed in each cluster, and the value of inertia is 4844.925817623823.

6.3

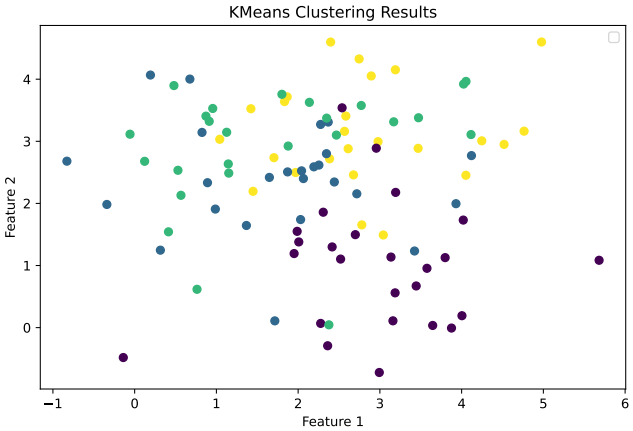


Figure 7: KMeans clustering visualization by using first two variables

Based on this plot, it might seemed the clustering is not that accurate. However, this is because we only selected the first two variables two do the visualization while the original data is actually much more complex than that. And we can be sure this is a good choice for the number of centers as this is the elbow point showed in 6.1.