CSCI-SHU 360 Machine Learning
Solution to homework 1

Yufeng Xu yx3038@nyu.edu

February 7, 2024

1 Simpson's Paradox

1.1

We define "you" as the 1^{st} player, "friend" as the 2^{nd} player. Therefore,

$$W_i := \text{the } i^{th} \text{ player wins}$$
$$U_j := \text{the player uses the } j^{th} \text{ machine}$$

With maximum likelihood estimation, we have

$$p(W_1|U_1) = \frac{40}{40+60} = \frac{2}{5}$$
$$p(W_2|U_1) = \frac{30}{30+70} = \frac{3}{10}$$
$$p(W_1|U_2) = \frac{210}{210+830} = \frac{21}{104}$$
$$p(W_2|U_2) = \frac{14}{14+70} = \frac{1}{6}$$

where $p(W_1|U_1) = \frac{2}{5} > \frac{3}{10} = p(W_2|U_1)$, $p(W_1|U_2) = \frac{21}{104} > \frac{1}{6} = p(W_2|U_2)$. Hence, "You" is more likely to win on both machine 1 and 2 than "friend".

1.2

$$p(W_1) = \frac{40+210}{40+60+210+830} = \frac{250}{1140} = \frac{25}{114}$$
$$p(W_2) = \frac{30+14}{30+70+14+70} = \frac{44}{184} = \frac{11}{46}$$

where $p(W_1) = \frac{25}{114} < \frac{11}{46} = p(W_2)$. Hence, overall "friend" is more likely to win than "you".

1.3

$$\begin{aligned} p(W_1) &= p(W_1, U_1) + p(W_1, U_2) \\ &= p(W_1|U_1)p(U_1) + p(W_1|U_2)p(U_2) \\ &= \frac{2}{5} \cdot \frac{40+60}{40+60+210+830} + \frac{21}{104} \cdot \frac{210+830}{40+60+210+830} \\ &= \frac{4}{114} + \frac{21}{114} = \frac{25}{114} \end{aligned}$$

$$
\begin{aligned}
p(W_2) &= p(W_2, U_1) + p(W_2, U_2) \\
&= p(W_2|U_1)p(U_1) + p(W_2|U_2)p(U_2) \\
&= \frac{3}{10} \cdot \frac{30+70}{30+70+14+70} + \frac{1}{6} \cdot \frac{14+70}{30+70+14+70} \\
&= \frac{30}{184} + \frac{14}{184} = \frac{44}{184} = \frac{11}{46}
\end{aligned}
$$

where $p(W_1|U_1) > p(W_2|U_1)$, $p(W_1|U_2) > p(W_2|U_2)$, but $p(W_1) < p(W_2)$.
This is because $p(W_1|U_1) > p(W_2|U_1) > p(W_1|U_2) > p(W_2|U_2)$, but the 1^{st} player("you") plays the 2^{nd} machine much more than the 1^{st} one, i.e., $p(U_1) << p(U_2)$. As a result, your probability of winning is close to $p(W_1|U_2)$ and becomes smaller than $p(W_2)$.

2 Matrix as Operations

2.1

We know that

$$
\begin{aligned}
Wa_1 &= \begin{pmatrix} W_{11} W_{12} \\ W_{21} W_{22} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} W_{11} + W_{12} \\ W_{21} + W_{22} \end{pmatrix} = b_1 = \begin{pmatrix} -0.8 \\ 1.6 \end{pmatrix} \\
Wa_2 &= \begin{pmatrix} W_{11} W_{12} \\ W_{21} W_{22} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} W_{11} - W_{12} \\ W_{21} - W_{22} \end{pmatrix} = b_2 = \begin{pmatrix} 2.6 \\ -0.2 \end{pmatrix}
\end{aligned}
$$

Hence

$$
\begin{cases}
W_{11} = 0.9 \\
W_{12} = -1.7 \\
W_{21} = 0.7 \\
W_{22} = 0.9
\end{cases}
$$

Hence

$$
W = \begin{pmatrix} 0.9 & -1.7 \\ 0.7 & 0.9 \end{pmatrix}
$$

2.2

$$
V = \begin{pmatrix} \cos(-\alpha) & -\sin(-\alpha) \\ \sin(-\alpha) & \cos(-\alpha) \end{pmatrix} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}
$$

Because $\tan(\alpha) = 3$, suppose $\alpha < 180°$, then $\cos(\alpha) = \frac{1}{\sqrt{10}}$, $\sin(\alpha) = \frac{3}{\sqrt{10}}$. Hence

$$
V = \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix}
$$

Next, we know

$$
\Sigma x = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{11} x + \Sigma_{12} y \\ \Sigma_{21} x + \Sigma_{22} y \end{pmatrix} = \begin{pmatrix} x \\ 2y \end{pmatrix}
$$

$$
\text{for } \begin{pmatrix} x \\ 2y \end{pmatrix} \in \mathbb{R}^2
$$

Hence,

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$$

Next, because U rotates counter-clockwise by β degrees,

$$U = \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}$$

Because $\tan(\beta) = \frac{1}{3}$, $\cos(\beta) = \frac{3}{\sqrt{10}}$, $\sin(\beta) = \frac{1}{\sqrt{10}}$. Hence

$$U = \begin{pmatrix} \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{pmatrix}$$

Therefore,

$$\begin{aligned} U\Sigma V &= \begin{pmatrix} \frac{3}{\sqrt{10}} & -\frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{10}} & \frac{3}{\sqrt{10}} \\ -\frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ -3 & 1 \end{pmatrix} \\ &= \frac{1}{10} \begin{pmatrix} 3 & -2 \\ 1 & 6 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ -3 & 1 \end{pmatrix} = \frac{1}{10} \begin{pmatrix} 9 & 7 \\ -17 & 9 \end{pmatrix} = \begin{pmatrix} 0.9 & 0.7 \\ -1.7 & 0.9 \end{pmatrix} = W^T \end{aligned}$$

2.3

$$W^T W = \begin{pmatrix} 0.9 & 0.7 \\ -1.7 & 0.9 \end{pmatrix} \begin{pmatrix} 0.9 & -1.7 \\ 0.7 & 0.9 \end{pmatrix} = \begin{pmatrix} 1.3 & -0.9 \\ -0.9 & 3.7 \end{pmatrix}$$

Hence $\det(W^T W - \lambda I) = (1.3-\lambda)(3.7-\lambda) - 0.9^2 = \lambda^2 - 5\lambda + 4 = (\lambda - 1)(\lambda - 4)$. Therefore, the eigenvalues and eigenvectors of $W^T W$ are

$$\begin{cases} \lambda_1 = 4, u_1 = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 & -3 \end{pmatrix}^T \\ \lambda_2 = 1, u_2 = \frac{1}{\sqrt{10}} \begin{pmatrix} 3 & 1 \end{pmatrix}^T \end{cases}$$

From 2.2 we know that $W^T = U\Sigma V$, where U, V are orthogonal and Σ is symmetric. Hence

$$W = (U\Sigma V)^T = V^T \Sigma^T U^T = V^{-1} \Sigma U^{-1}$$

Therefore, if we transform the unit circle by W, it is equivalent to rotate the unit cricle by β degrees clockwise (the result is still a unit circle), scaling the y-axis by 2 and keeping x-axis unchanged (result is an ellipse), and rotating the ellipse by α degrees anti-clockwise.

In the end, we will get an inclined ellipse. Moreover, suppose that the semi-major axis of this ellipse is of length a, semi-minor axis is of length b, we have $a^2 = 4 = \lambda_1, b^2 = 1 = \lambda_2$; the major axis is in the same direction as u_1, whil the minor axis is in the same direction as u_2.

This is bacause

$$W^T W = U\Sigma V V^{-1} \Sigma U^{-1} = U\Sigma^2 U^{-1} = \begin{pmatrix} u_2 & -u_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} u_2 & -u_1 \end{pmatrix}^{-1}$$

Hence, if we transform the unit circle with $W^T W$, it is equivalent to scaling the y-axis by 2 twice, therefore $a^2 = 2^2 = 4 = \lambda_1$, $b^1 = 1^2 = 1 = \lambda_1$. The matrix $U = \begin{pmatrix} u_2 & -u_1 \end{pmatrix}$, which is equivalent to mapping the standard coordinate system to the space spanned by u_2 and u_1, which explains why the long-axis and short-axis are in the direction of u_2 and u_1.

2.4

$$\det(W) = \det \begin{pmatrix} 0.9 & -1.7 \\ 0.7 & 0.9 \end{pmatrix} = 0.9 \cdot 0.9 - (-1.7) \cdot 0.7 = 2$$

The area of the ellipse is $2 \cdot 1 \cdot \pi = 2\pi = \det(W) \cdot$ area of a unit circle. Therefore,

$$\frac{\text{the area of the shape transformed by } W}{\text{the area of the original shape}} = \det(W)$$

This is because W can be diagonalized as follows:

$$W = U\Sigma V = \begin{pmatrix} u_1 & u_2 \dots u_n \end{pmatrix}^{-1} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{pmatrix} u_1 & u_2 \dots u_n \end{pmatrix}$$

where $\lambda_1 \dots \lambda_n$ are eigenvalues of W, $u_1 \dots u_n$ are its orthonormal eigenvectors. Therefore, transformations of a shape with U and V do not influence the area of it, where as Σ multiplies its area by $\prod_{i=1}^{n} |\lambda_i| = |\det(W)|$. Therefore, transforming a shape with AB will multiply its area by $|\det(AB)|$. On the other hand, transforming a shape with AB is equivalent to transforming it with B, then transforming it with A, which multiplies its area by $|\det(A)||\det(B)|$. These two understandings are equivalent, therefore $\det(AB) = \det(A)\det(B)$.

3 Some Practices

3.1

Because

$$Var(X^3) = \mathbb{E}[(X^3)^2] - \big(\mathbb{E}[X^3]\big)^2 = \mathbb{E}[X^6] - \big(\mathbb{E}[X^3]\big)^2 \geq 0$$

Therefore

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

3.2

Because X is a discrete random variable, we know that

$$\mathbb{E}(X^6) = \sum_x x^6 p(X = x)$$
$$\mathbb{E}(X^3) = \sum_x x^3 p(X = x)$$

By Cauchy-Schwarz inequality, we know

$$\begin{aligned} \mathbb{E}(X^6) = \sum_x x^6 p(X = x) &= \left(\sum_x x^6 p(X = x) \right) \left(\sum_x p(X = x) \right) \\ &= \left(\sum_x \left(x^3 \sqrt{p(X = x)} \right)^2 \right) \left(\sum_x \left(\sqrt{p(X = x)} \right)^2 \right) \\ &\geq \left(\sum_x x^3 \sqrt{p(X = x)} \sqrt{p(X = x)} \right)^2 = \mathbb{E}(X^3)^2 \end{aligned}$$

3.3

Because A, B are PSD matrices, we know

$$\forall z \in \mathbb{R}^n, z^T A z \geq 0, z^T B z \geq 0$$

Then for $\forall \lambda \in [0, 1]$,

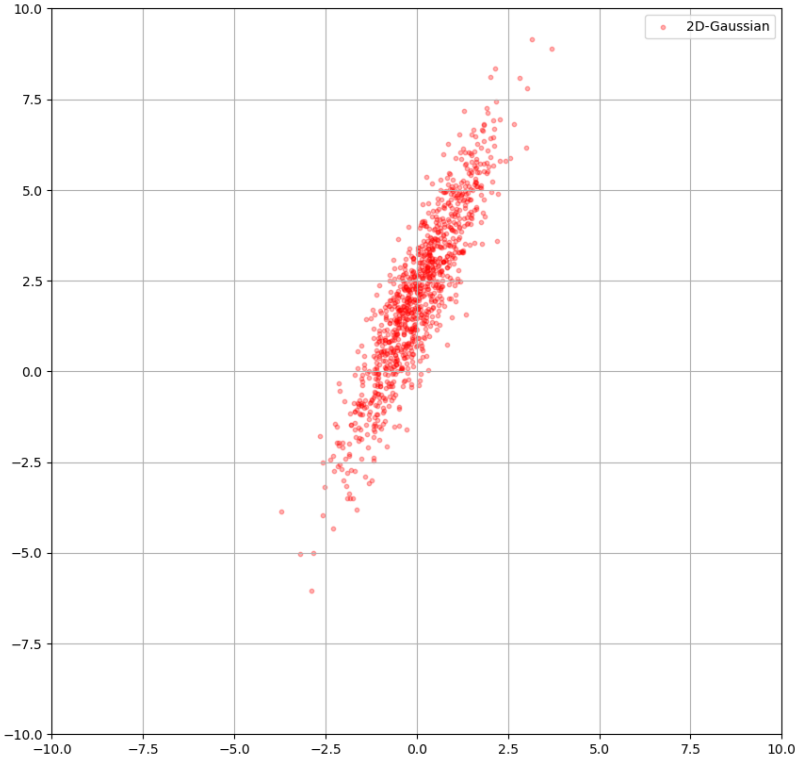
$$\begin{aligned} z^T(\lambda A + (1-\lambda)B)z &= \sum_{i,j} (\lambda A + (1-\lambda)B)_{i,j} z_i z_j \\ &= \sum_{i,j} \left(\lambda A_{i,j} + (1-\lambda)B_{i,j}\right) z_i z_j \\ &= \lambda \sum_{i,j} A_{i,j} z_i z_j + (1-\lambda) \sum_{i,j} B_{i,j} z_i z_j \end{aligned}$$

where $\lambda, 1-\lambda, \sum_{i,j} A_{i,j} z_i z_j, \sum_{i,j} B_{i,j} z_i z_j \geq 0$, hence LHS$\geq 0$, $\lambda A + (1-\lambda)B$ is also a PSD matrix.

4 Density Estimation of Multivariate Gaussian

4.1

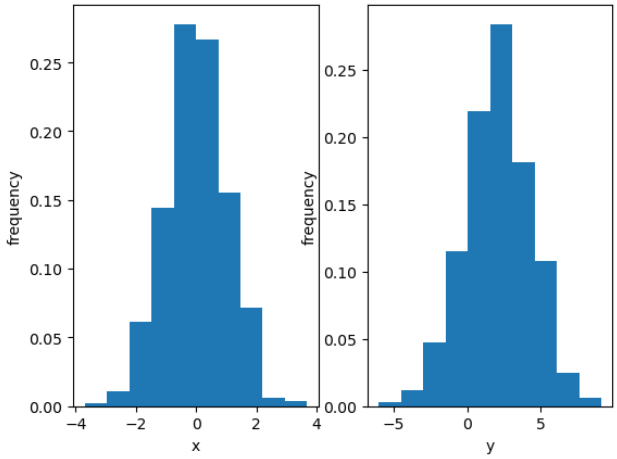
Figure 1: the 1000 points in X sampled from multivariate Gaussian distribution



Below is the image of the 1000 points sample from Gaussian distribution. With numpy, we obtained the mean and covariance of X, which are $\begin{pmatrix} 0.01909265 & 2.06052385 \end{pmatrix}$ and $\begin{pmatrix} 1.03589238 & 2.06244165 \\ 2.06244165 & 5.08839176 \end{pmatrix}$, respectively.

4.2

Figure 2: the histograms of x-coordinates (left) and y-coordinates (right) of X



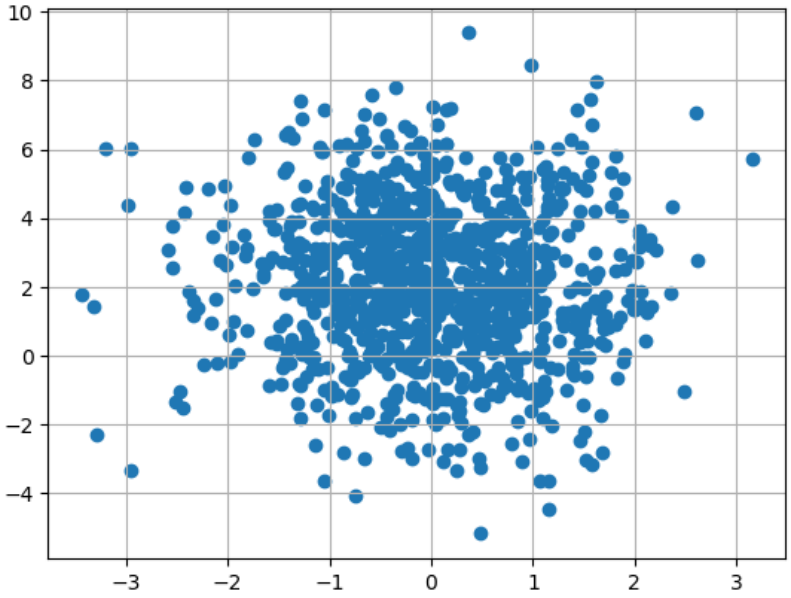
4.3

From 4.2, we can confidently conjecture that both x and y-coordinates of X follow Gaussian distribution. By calculation with numpy, we got $\mu_x = 0.019092651459051667$, $\sigma_x^2 = 1.0348564864741474$, $\sigma_x = 1.0172789619736307$; $\mu_y = 2.0605238541899022$, $\sigma_y^2 = 5.083303371449156$, and $\sigma_y = 2.2546182318630255$.

4.4

After generating 1000 samples $x \sim N(\mu_x, \sigma_x^2)$ and 1000 samples $y \sim N(\mu_y, \sigma_y^2)$, we plotted a graph of coordinates (x, y) as follows:

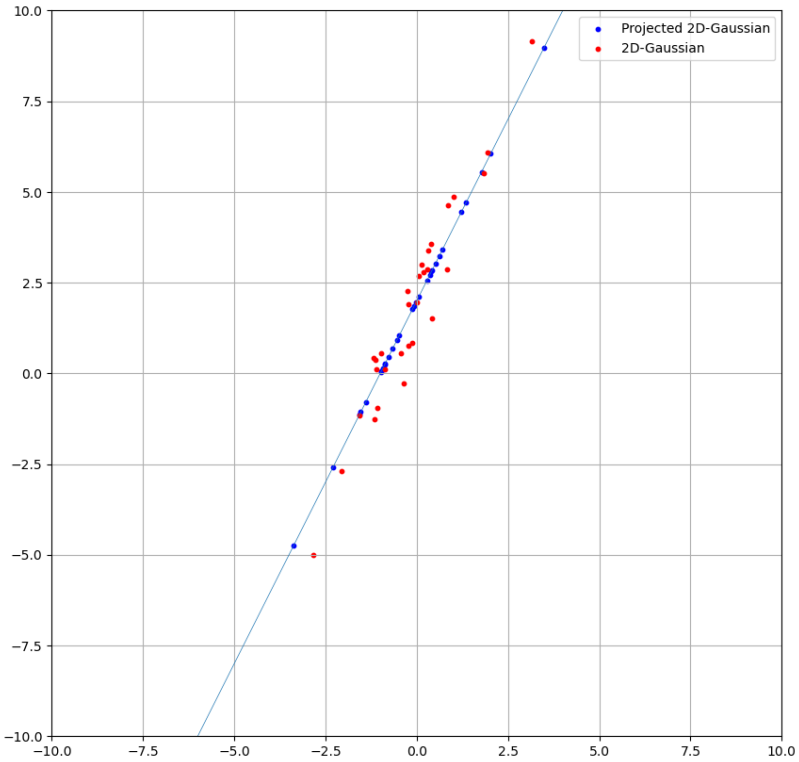
Figure 3: the distribution of (x, y) with $x \sim N(\mu_x, \sigma_x^2)$, $y \sim N(\mu_y, \sigma_y^2)$



Compared to the graph in 4.1, where the distribution of (x, y) are basically within an inclined ellipse, the distribution in this graph is mostly within an **horizontally placed** ellipse within (μ_x, μ_y) as its center, σ_x and σ_y as its semi-major axis and semi-minor axis. This is because x and y are independently sampled in this problem, whereas the x and y-coordinates of X are not indenpendent.

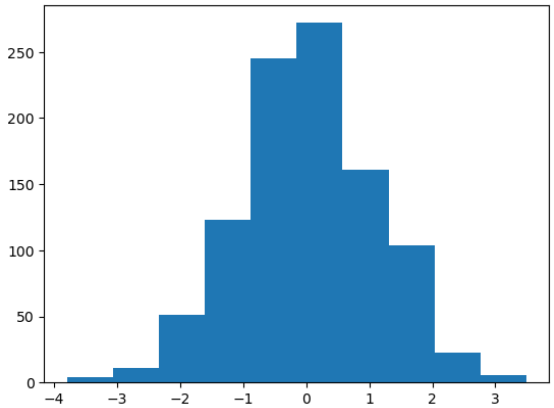
4.5

Figure 4: 30 samples from X, $y = 2x + 2$, and the projection of the samples on $y = 2x + 2$



4.6

Figure 5: the histogram of the x-coordinates of the projected samples



From the graph above, we can tell the x-coordinates of the projected points also follow Gaussian distribution. After calculating with numpy, we obtained $\mu_{x_{proj}} = 0.028028071967771222$, $\sigma^2_{x_{proj}} = 1.1843834728271618$.

5 KNN for Iris flowers classification

5.1

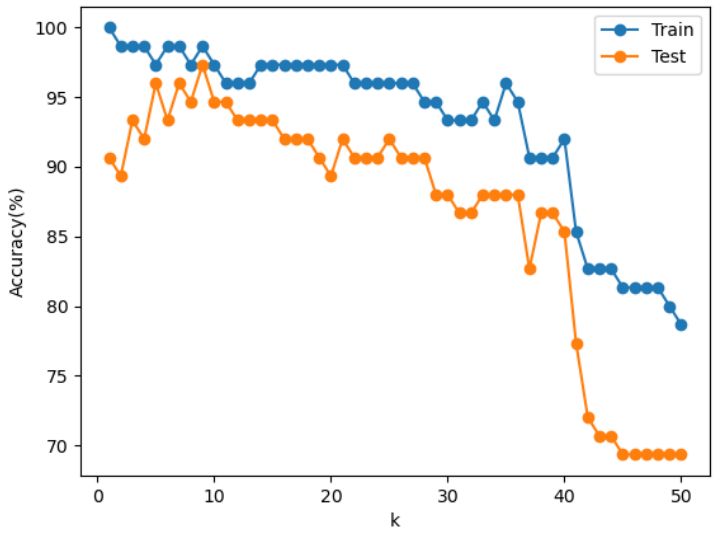
After counting the number of samples in each class, it turns out there are 50 samples in setosa, versicolor, and virginica class.

5.2

After applying KNeighborsClassifier with k = 1 to the data, the accuracy turns out be 100%. However, this result makes no sense because the top-1 closest point to a data point is just itself, which is meaningless for our classification task.

5.3

Figure 6: train and test accuracy with different k's



According to graph above, it turns out the optimal k is 9, because when $k = 9$ both train and test accuracy are high, and the gap between them is small.

5.4

After putting the given data (sepal width = 5.0, petal width = 4.1, sepal length = 3.8, petal length = 1.2) into my prediction model (KNN with k = 9), the predicted class turns out to be class 1, which is versicolor.

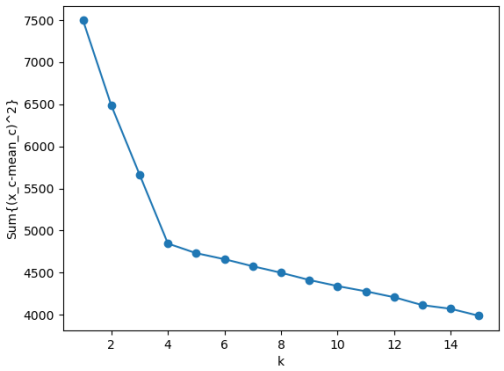
6 K Means clustering

6.1

From the graph below, we can tell the elbow point appears when k=4. At this point, the samples within the same class have high similarity, while k is not too large, which guarantees the model's robustness. Therefore

k=4 clusters should be used for this data.

Figure 7: the sum of within-class inertia for different k's

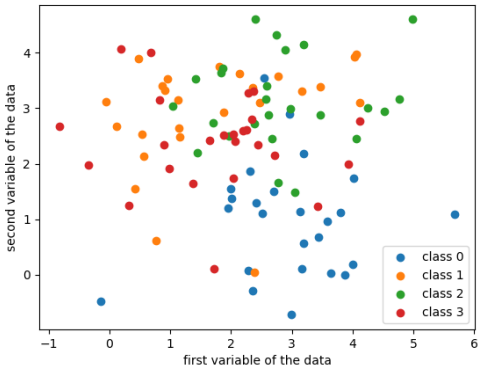


6.2

After rerunning k-means with $k = 4$, we found out there are 25 samples in each class. The value of inertia is 4844.925817623823.

6.3

Figure 8: the plot of the data mapped to its first two dimensions. The color of each data indicates the class it belongs to.



From the graph below, we can see the number of samples in each class is appropriate. However, the data points do not look well-classified, which is likely due to our choice of dimensions when doing the visualization. Maybe by choosing two different dimensions will make this plot look well-segmented. Overall, k=4 should be a good choice for number of centers because it demonstrates accuracy and robustness in 6.1.