University of Tehran

Faculty of Electrical and Computer Engineering

Machine Learning

Dr. A. Dehaqani, Dr. Tavassolipour

Homework #1

Name	Hesam Asadollahzadeh
Student No.	810198346

Table of Contents

Question #1- Bayes Optimal Classifier and Cauchy Distribution

a)

$$P(x|\omega_i) = \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_i}{b}\right)^2}, i = 1,2, \qquad a_2 > a_1$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$P(\omega_1|x) = P(\omega_2|x) \Leftrightarrow P(x|\omega_1)P(\omega_1) = P(x|\omega_2)P(\omega_2)$$

$$\Leftrightarrow \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}P(\omega_1) = \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}P(\omega_2) \Leftrightarrow$$

$$\left(\frac{x-a_1}{b}\right)^2 = \left(\frac{x-a_2}{b}\right)^2 \Leftrightarrow x^2 - 2a_1x + a_1^2 = x^2 - 2a_2x + a_2^2$$

$$\Leftrightarrow 2(a_2 - a_1)x = a_2^2 - a_1^2 = (a_2 - a_1)(a_2 + a_1) \Leftrightarrow$$

$$x = \frac{a_1 + a_2}{2}$$

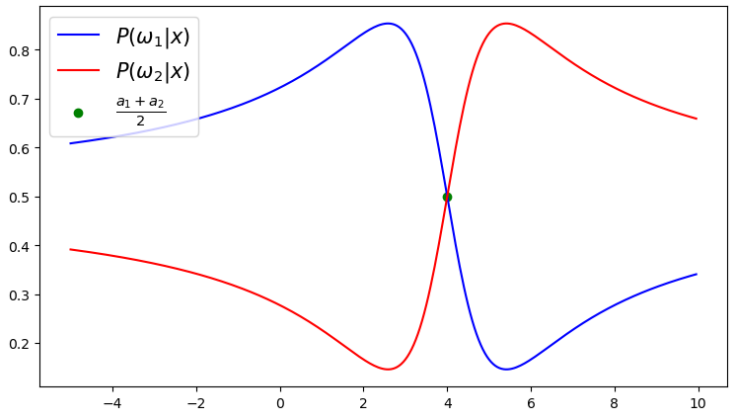


Figure 1. for ($a_1 = 3, a_2 = 5, b = 1$)

b)

$$P(Error) = \int_{R_1} P(\omega_2|x)dx + \int_{R_2} P(\omega_1|x)dx =$$

$$\int_{-\infty}^{\frac{a_1+a_2}{2}} P(\omega_2|x)dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} P(\omega_1|x)dx \xrightarrow{P(\omega_1)=P(\omega_2)=0.5}$$

$$\frac{1}{2}\left(\int_{-\infty}^{\frac{a_1+a_2}{2}} P(x|\omega_2)dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} P(x|\omega_1)dx\right)$$

$$= \frac{1}{2\pi b}\left(\int_{-\infty}^{\frac{a_1+a_2}{2}} \frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}dx + \int_{\frac{a_1+a_2}{2}}^{+\infty} \frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}dx\right)$$

$$= \frac{1}{2\pi}\left(\tan^{-1}\frac{x-a_2}{b}\Big|_{-\infty}^{\frac{a_1+a_2}{2}} + \tan^{-1}\frac{x-a_1}{b}\Big|_{\frac{a_1+a_2}{2}}^{+\infty}\right) =$$

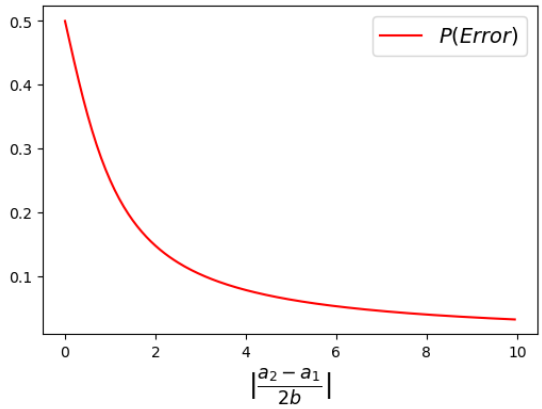
$$\frac{1}{2\pi}\left(\left(\tan^{-1}\frac{a_1-a_2}{2b} + \frac{\pi}{2}\right) + \left(\frac{\pi}{2} - \tan^{-1}\frac{a_2-a_1}{2b}\right)\right) =$$

$$\frac{1}{2\pi}\left(\pi + \tan^{-1}\frac{a_1-a_2}{2b} - \tan^{-1}\frac{a_2-a_1}{2b}\right) \rightarrow$$

Arctan is an odd function: $\tan^{-1}\frac{a_1-a_2}{2b} = -\tan^{-1}\frac{a_2-a_1}{2b} \Rightarrow$

$$= \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\frac{a_2-a_1}{2b} \xrightarrow{a_2>a_1} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2-a_1}{2b}\right|$$

c) First, we plot the error function with respect to $\frac{a_2-a_1}{2b}$:



Thus, we have:

$$P_{max}(Error) = \frac{1}{2},\ \ where\ \left|\frac{a_2-a_1}{2b}\right| = 0$$

$$\rightarrow \text{either } a_1 = a_2\ (1)\ \ \text{or}\ \ b \rightarrow \infty\ (2)$$

To conclude, the maximum value of the probability of error occurs when either the two distributions are the same (first case) or when both distributions are flat (second case).

d) Designing an optimal bayes classifier:

$$\omega_i = \begin{cases} \omega_1 & \text{if } \dfrac{P(x|\omega_1)}{P(x|\omega_2)} > 1 \\ \omega_2 & \text{otherwise} \end{cases}$$

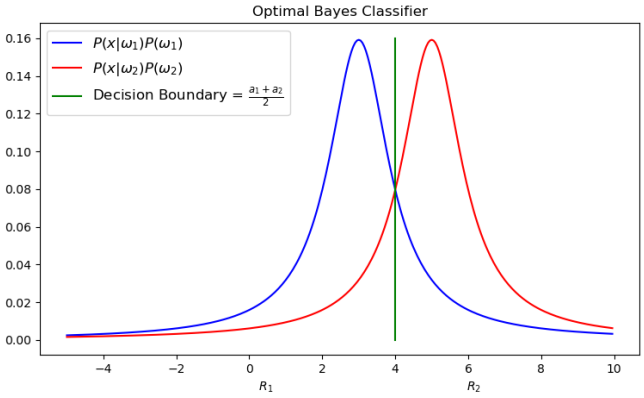
The decision boundary is:

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} = 1 \rightarrow P(x|\omega_1) = P(x|\omega_2) \rightarrow$$

$$\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} = \frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2} \rightarrow |x-a_1| = |x-a_2|$$

$$\begin{cases} a_1 \neq a_2 \rightarrow x_{decision\ boundary} = \dfrac{a_1+a_2}{2} \\ a_1 = a_2 \rightarrow \forall x: x \text{ is decision boundary} \end{cases}$$

$$P(Error) = \begin{cases} \dfrac{1}{2} - \dfrac{1}{\pi}\tan^{-1}\left|\dfrac{a_2-a_1}{2b}\right| & \text{if } a_1 \neq a_2 \\ \dfrac{1}{2} & \text{otherwise} \end{cases}$$



e) Designing a minimum-risk classifier:

$$\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$$

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{\lambda_{11}-\lambda_{12}}{\lambda_{22}-\lambda_{21}} = \frac{1}{2}$$

$$\omega_i = \begin{cases} \omega_1 & \text{if } \dfrac{P(x|\omega_1)}{P(x|\omega_2)} > \dfrac{1}{2} \\ \omega_2 & \text{otherwise} \end{cases}$$

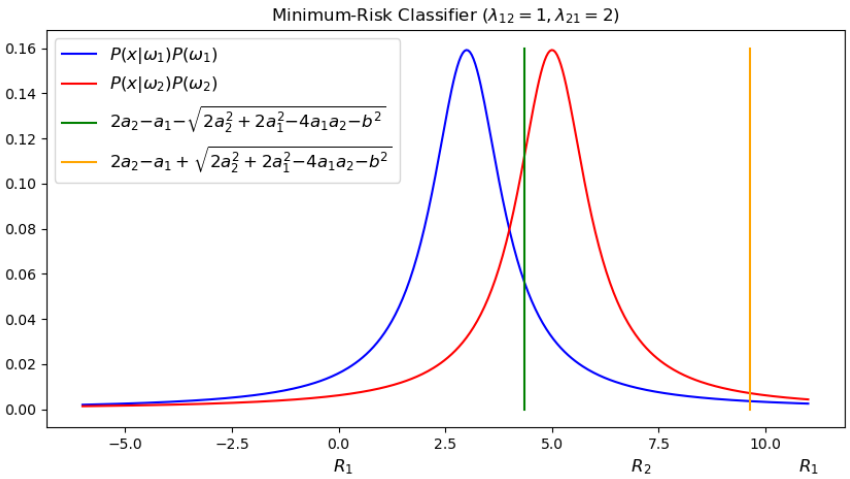
The decision boundary is:

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} = \frac{1}{2} \rightarrow P(x|\omega_1) = \frac{P(x|\omega_2)}{2} \rightarrow$$

$$\frac{1}{\pi b}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2} = \frac{1}{2\pi b}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2} \rightarrow \left(\frac{x-a_1}{b}\right)^2 = 1 + 2\left(\frac{x-a_2}{b}\right)^2$$

$$x^2 - 2(2a_2 - a_1)x + (2a_2^2 - a_1^2 + b^2) = 0 \rightarrow$$

$$x_1, x_2 = 2a_2 - a_1 \pm \sqrt{2a_2^2 + 2a_1^2 - 4a_1a_2 - b^2}$$



If we choose a value for λ_{21} that is greater than λ_{12}, the risk associated with selecting ω_2 and misclassifying ω_1 is elevated. Consequently, the classifier becomes more cautious about assigning data points to R1 (the decision region for class 1), resulting in a reduction in the size of R2 as compared to the previous version.

$$\boldsymbol{P(Error)} = \int_{R_1} P(\omega_2|x)dx + \int_{R_2} P(\omega_1|x)dx =$$

$$\int_{-\infty}^{x_1} P(x|\omega_2)P(\omega_2)dx + \int_{x_1}^{x_2} P(x|\omega_1)P(\omega_1)dx + \int_{x_2}^{+\infty} P(x|\omega_2)P(\omega_2)dx =$$

$$\frac{1}{2}\left(\int_{-\infty}^{x_1} P(x|\omega_2)dx + \int_{x_1}^{x_2} P(x|\omega_1)dx + \int_{x_2}^{+\infty} P(x|\omega_2)dx\right) =$$

$$\frac{1}{2}\left(\int_{-\infty}^{x_1}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}dx+\int_{x_1}^{x_2}\frac{1}{1+\left(\frac{x-a_1}{b}\right)^2}dx+\int_{x_2}^{+\infty}\frac{1}{1+\left(\frac{x-a_2}{b}\right)^2}dx\right)=$$

$$\frac{1}{2\pi}\left(\tan^{-1}\frac{x-a_2}{b}\Big|_{-\infty}^{x_1}+\tan^{-1}\frac{x-a_1}{b}\Big|_{x_1}^{x_2}+\tan^{-1}\frac{x-a_2}{b}\Big|_{x_2}^{+\infty}\right)=$$

$$\frac{1}{2\pi}\left(\left(\tan^{-1}\frac{x_1-a_2}{b}+\frac{\pi}{2}\right)+\left(\tan^{-1}\frac{x_2-a_1}{b}-\tan^{-1}\frac{x_1-a_1}{b}\right)\right.$$

$$\left.+\left(\frac{\pi}{2}-\tan^{-1}\frac{x_2-a_2}{b}\right)\right)=$$

$$\frac{1}{2}+\frac{1}{2\pi}\left(\tan^{-1}\frac{x_1-a_2}{b}+\tan^{-1}\frac{x_2-a_1}{b}-\tan^{-1}\frac{x_1-a_1}{b}-\tan^{-1}\frac{x_2-a_2}{b}\right)$$

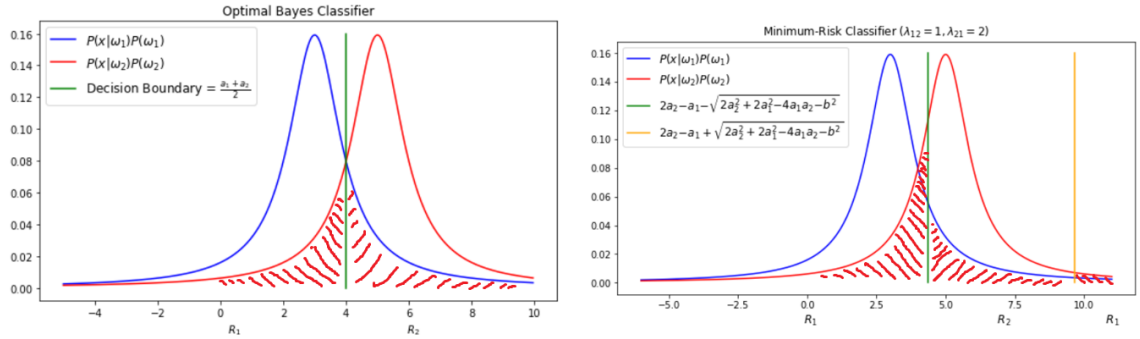


Figure 2. Error Region of Optimal Bayes and Minimum-Risk Classifiers

Question #2- Bayes Optimal Classifier and Rayleigh Distribution

$$P(x|\omega_i)=\begin{cases}\frac{x}{\sigma_i^2}\exp\left(-\frac{x^2}{2\sigma_i^2}\right), & x\geq 0\\ 0, & x<0\end{cases}\rightarrow$$

Decision Boundary: $P(\omega_1|x)=P(\omega_2|x)\rightarrow$

$$\frac{P(x|\omega_1)P(\omega_1)}{P(x)}=\frac{P(x|\omega_2)P(\omega_2)}{P(x)}\Rightarrow$$

$$\frac{x}{\sigma_1^2}\exp\left(-\frac{x^2}{2\sigma_1^2}\right)=\frac{x}{\sigma_2^2}\exp\left(-\frac{x^2}{2\sigma_2^2}\right)\overset{x\neq 0}{\longrightarrow}$$

$$\log\left(\frac{\sigma_2^2}{\sigma_1^2}\right)=\frac{x^2}{2\sigma_1^2}-\frac{x^2}{2\sigma_2^2}\overset{x>0}{\longrightarrow}\boldsymbol{x=2\sqrt{\frac{\log\frac{\sigma_2}{\sigma_1}}{\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}}}}$$

$$\omega_i = \begin{cases} \omega_1, & x > 2\sqrt{\dfrac{\log\frac{\sigma_2}{\sigma_1}}{\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}}} \\ \omega_2, & otherwise \end{cases}$$

Question #3- Bayes Optimal Decision Boundary on 2D data

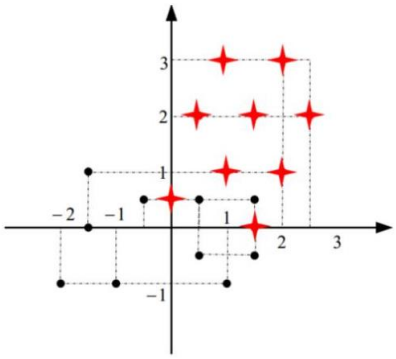


Figure 3. Train Data

a, b) For red class we'll have:

$$\hat{\mu}_{red} = \begin{bmatrix}\hat{\mu}_y \\ \hat{\mu}_x\end{bmatrix} = \frac{1}{9}\left(\begin{bmatrix}1\\ \frac{1}{2}\\ 0\end{bmatrix} + \begin{bmatrix}2\\ 1\\ \frac{1}{2}\end{bmatrix} + \begin{bmatrix}1\\ 1\end{bmatrix} + \begin{bmatrix}3\\ 1\end{bmatrix} + \begin{bmatrix}0\\ 3\\ \frac{1}{2}\end{bmatrix} + \begin{bmatrix}2\\ 3\\ \frac{1}{2}\end{bmatrix} + \begin{bmatrix}1\\ 2\end{bmatrix} + \begin{bmatrix}3\\ 2\end{bmatrix} + \begin{bmatrix}2\\ 5\\ \frac{1}{2}\end{bmatrix}\right)$$

$$= \begin{bmatrix}1.61\\ 1.33\end{bmatrix} \rightarrow \begin{bmatrix}\hat{\mu}_x \\ \hat{\mu}_y\end{bmatrix} = \begin{bmatrix}\hat{\mu}_{x_1} \\ \hat{\mu}_{x_2}\end{bmatrix} = \begin{bmatrix}1.33\\ 1.61\end{bmatrix}$$

$$\hat{\Sigma}_{red} = \begin{bmatrix}0.63 & 0.21\\ 0.21 & 1.11\end{bmatrix}$$

For black dots we'll have:

$$\hat{\mu}_{black} = \begin{bmatrix}\hat{\mu}_y \\ \hat{\mu}_x\end{bmatrix}$$

$$= \frac{1}{10}\left(\begin{bmatrix}-1\\ -2\end{bmatrix} + \begin{bmatrix}0\\ -\frac{3}{2}\end{bmatrix} + \begin{bmatrix}1\\ -\frac{3}{2}\end{bmatrix} + \begin{bmatrix}-1\\ -1\end{bmatrix} + \begin{bmatrix}\frac{1}{2}\\ -\frac{1}{2}\end{bmatrix} + \begin{bmatrix}-\frac{1}{2}\\ \frac{1}{2}\end{bmatrix} + \begin{bmatrix}\frac{1}{2}\\ \frac{1}{2}\end{bmatrix} + \begin{bmatrix}-1\\ 1\end{bmatrix}\right.$$

$$\left. + \begin{bmatrix}-\frac{1}{2}\\ \frac{3}{2}\end{bmatrix} + \begin{bmatrix}\frac{1}{2}\\ \frac{3}{2}\end{bmatrix}\right) = \begin{bmatrix}-0.15\\ -0.15\end{bmatrix} = \begin{bmatrix}\hat{\mu}_x \\ \hat{\mu}_y\end{bmatrix} = \begin{bmatrix}\hat{\mu}_{x_1} \\ \hat{\mu}_{x_2}\end{bmatrix}$$

$$\hat{\Sigma}_{black} = \begin{bmatrix} 1.73 & 0.003 \\ 0.003 & 0.56 \end{bmatrix}$$

c) The Gaussian discriminant functions are defined as below:

$$g_i(x) = P(x|\omega_i) = x^T W_i x + w_i^T x + w_{i_0}$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

$$w_{i_0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i).$$

For red class we'll have:

$$W_{red} = \begin{bmatrix} -0.85 & 0.16 \\ 0.16 & -0.48 \end{bmatrix}, w_{red} = \begin{bmatrix} 1.76 \\ 1.12 \end{bmatrix}, w_{red_0} = -2.55$$

$$g_{red}(x) = (-0.85x_1^2 + 0.32x_1x_2 - 0.48x_2^2) + (1.76x_1 + 1.12x_2) - 2.55$$

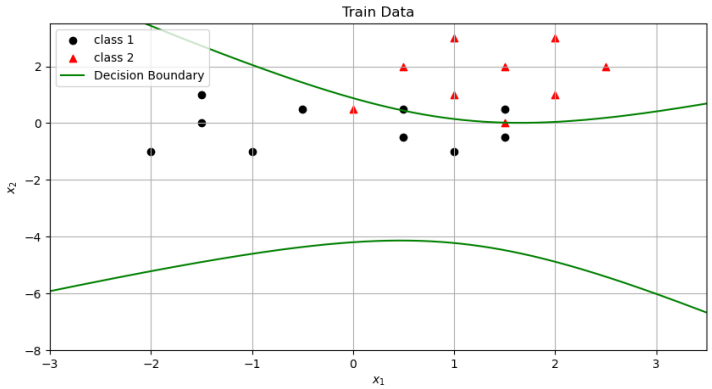
For black class we'll have:

$$W_{black} = \begin{bmatrix} -0.29 & 0.001 \\ 0.001 & -0.9 \end{bmatrix}, w_{black} = \begin{bmatrix} -0.09 \\ -0.27 \end{bmatrix}, w_{black_0} = -0.7$$

$$g_{black}(x) = (-0.29x_1^2 - 0.003x_1x_2 - 0.9x_2^2) + (-0.09x_1 - 0.27x_2) - 0.7$$

Using discriminant functions, the decision boundary is:

$$g_{red}(x) - g_{black}(x) = 0 \rightarrow$$
$$-0.56x_1^2 + 0.32x_1x_2 + 0.42x_2^2 + 1.85x_1 + 1.39x_2 - 1.55 = 0$$



As it's shown in the figure, there are 3 misclassified data points. Thus, the empirical error is:

$$P(Error)_{emp} = \frac{4}{19} \approx 21\%$$

d) Adding risk to the problem definition we'll get:

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|x)$$

$$\text{Given } \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 2a \\ a & 0 \end{pmatrix} \quad a > 0:$$

$$R(\alpha_{red}|x) = a \times P(\omega_2|x)$$

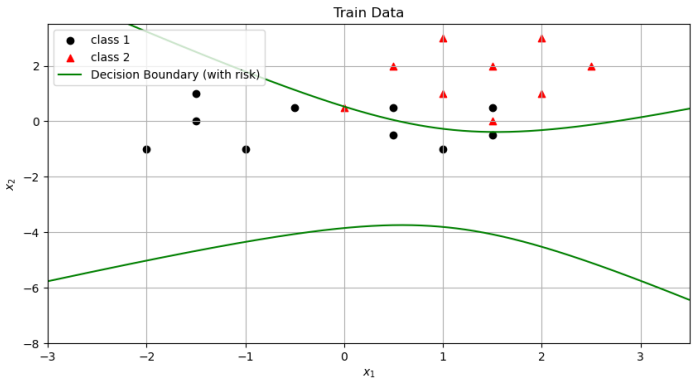
$$R(\alpha_{black}|x) = 2a \times P(\omega_1|x)$$

Thus, the decision boundary would be derived as follows:

$$\log\big(R(\alpha_{red}|x)\big) - \log\big(R(\alpha_{black}|x)\big) = 0 \rightarrow$$

$$g_{red}(x) - g_{black}(x) + \log 2 = 0 \Rightarrow$$

$$-0.56x_1^2 + 0.32x_1x_2 + 0.42x_2^2 + 1.85x_1 + 1.39x_2 - 0.86 = 0$$



Again, choosing $\lambda_{12} > \lambda_{21}$ leads to increasing the risk of choosing $black$ and misclassifying red. Thus, the classifier is now more conservative about decision region of red class and decision region of the black class is now smaller than the previous version.

e) Choosing $\lambda_{12} = 2\lambda_{21}$ & $P(\omega_1) = P(\omega_2)$ is equivalent to choosing $\lambda_{12} = \lambda_{21}$ & $P(\omega_2) = 2P(\omega_1)$. Thus, the decision boundary of this part would be just equal to the previous part. Empirical error is:

$$P(Error)_{emp} = \frac{3}{19} \approx 16\%$$

Question #4- Parameter Estimation Using MLE & MAP

a) Log-likelihood and MLE Estimation:

$$P(X) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad D = \{X_1, X_2, \dots, X_n\}$$

$$\hat{\lambda}_{MLE} = \arg\max_{\lambda} P(D;\ \lambda)$$

$$P(D;\ \lambda) = \prod_{i=1}^{n} P(X_i; \lambda)$$

$$LL(\lambda) = \log\lambda \times \Sigma_{i=1}^{n} x_i - n\lambda - \Sigma_{i=1}^{n} \log x_i!$$

$$\frac{\partial LL(\lambda)}{\partial \lambda} = 0 \rightarrow \frac{1}{\lambda}\Sigma_{i=1}^{n} x_i - n = 0 \Rightarrow$$

$$\hat{\lambda}_{MLE} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$$

b) Posterior distribution is:

$$P(\lambda) = c\lambda^{\alpha-1}e^{-\beta\lambda}$$

$$P(\lambda|D) \sim P(D|\lambda)P(\lambda) = P(X_1 \dots X_n|\lambda)P(\lambda) \xrightarrow{given\ i.i.d\ samples}$$

$$P(\lambda|D) \sim \prod_{i=1}^{n} P(X_i|\lambda)P(\lambda) = \left(\prod_{i=1}^{n} \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}\right) c\lambda^{\alpha-1}e^{-\beta\lambda} =$$

$$c'\lambda^{\Sigma_{i=1}^{n} x_i + \alpha - 1}e^{-(\beta+n)\lambda} = \text{Gamma}(\lambda|\Sigma_{i=1}^{n} x_i + \alpha, \beta + n)$$

c) Yes, this is correct. The result implies that both the prior and posterior distributions are Gamma distributions, but with different parameters. This means that the posterior distribution can be used as a reproducing density, which is a density that has the same functional form as the prior distribution. In addition, since the prior and posterior distributions belong to the same family of distributions, we can say that the prior is a conjugate prior for the likelihood function. This is useful because it allows us to easily update our beliefs about the parameter λ based on new data, by simply multiplying the prior distribution by the likelihood function and normalizing the result. Overall, the fact that the prior and posterior distributions are both Gamma distributions with different

parameters is a powerful result that enables us to make more accurate and efficient Bayesian inference in a wide range of applications.

d) MAP estimator should be derived as follows:

$$\hat{\lambda}_{MAP} = \arg\max_{\lambda} P(\lambda|D)P(\lambda) \rightarrow$$

$$\hat{\lambda}_{MAP} = \frac{\Sigma_{i=1}^{n} x_i + \alpha - 1}{n + \beta}$$

e) Let's rewrite $\hat{\lambda}_{MAP}$ as follows:

$$\hat{\lambda}_{MAP} = \frac{n}{n+\beta}\frac{\Sigma_{i=1}^{n} x_i}{n} + \frac{\alpha - 1}{n+\beta} = \frac{n}{n+\beta}\hat{\lambda}_{MLE} + \frac{\alpha - 1}{n+\beta}$$

$$\lim_{n\to\infty} \hat{\lambda}_{MAP} = \lim_{n\to\infty} \frac{n}{n+\beta}\hat{\lambda}_{MLE} + \frac{\alpha - 1}{n+\beta}$$

We know that:

$$\lim_{n\to\infty} \frac{n}{n+\beta} = 1, \lim_{n\to\infty} \frac{\alpha - 1}{n+\beta} = 0 \rightarrow$$

$$\lim_{n\to\infty} \hat{\lambda}_{MAP} = \lim_{n\to\infty} \frac{n}{n+\beta}\hat{\lambda}_{MLE} + \frac{\alpha - 1}{n+\beta} = \hat{\lambda}_{MLE}$$

Thus, the behavior of Maximum a Posteriori (MAP) estimation and Maximum Likelihood Estimation (MLE) can be studied in the context of increasing the number of data points. As the number of data points goes to infinity, the MAP estimation tends **asymptotically** to the MLE estimation. This happens because, in the presence of an infinite amount of data, the impact of the prior distribution on the estimation becomes **negligible**. As a result, the MAP estimator, which takes into account both the prior and the likelihood, converges to the MLE estimator, which only considers the likelihood.

f) In statistical inference and machine learning, the choice between Maximum a Posteriori (MAP) estimation and Maximum Likelihood Estimation (MLE) depends on the **size of the dataset** and the quality of the **prior knowledge**. For data sets with a limited number of observations, the MAP estimator tends to be more accurate than the MLE estimator since it incorporates the prior distribution of the parameters. The prior distribution provides a way to encode any prior knowledge (including domain knowledge)

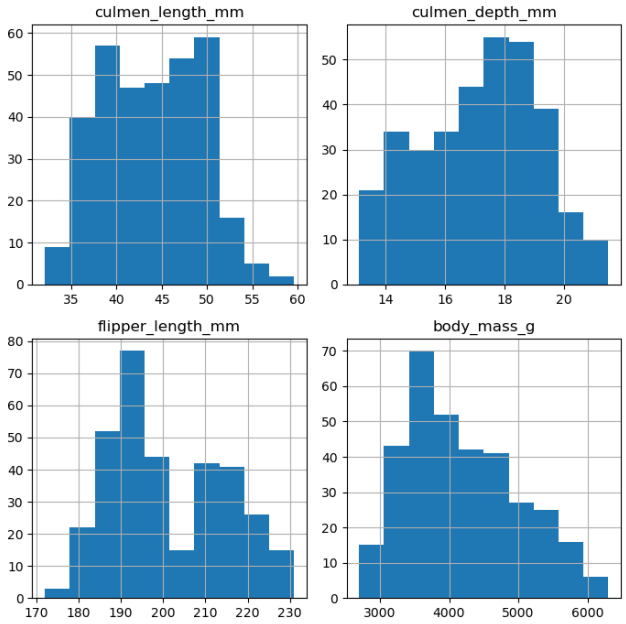
that we may have about the parameters, which can help in achieving better estimates. In this scenario, the MAP estimator is a more appropriate choice, as it utilizes both the data and prior knowledge to provide the estimate.

However, if we have a large data set or our prior knowledge of the distribution is limited, then it's better to use MLE estimation. In this case, the likelihood function provides more information about the parameters than the prior, and the impact of the prior becomes negligible. The MLE estimator only considers the likelihood of the data and is a more suitable choice for such scenarios where the prior knowledge is limited or unreliable.

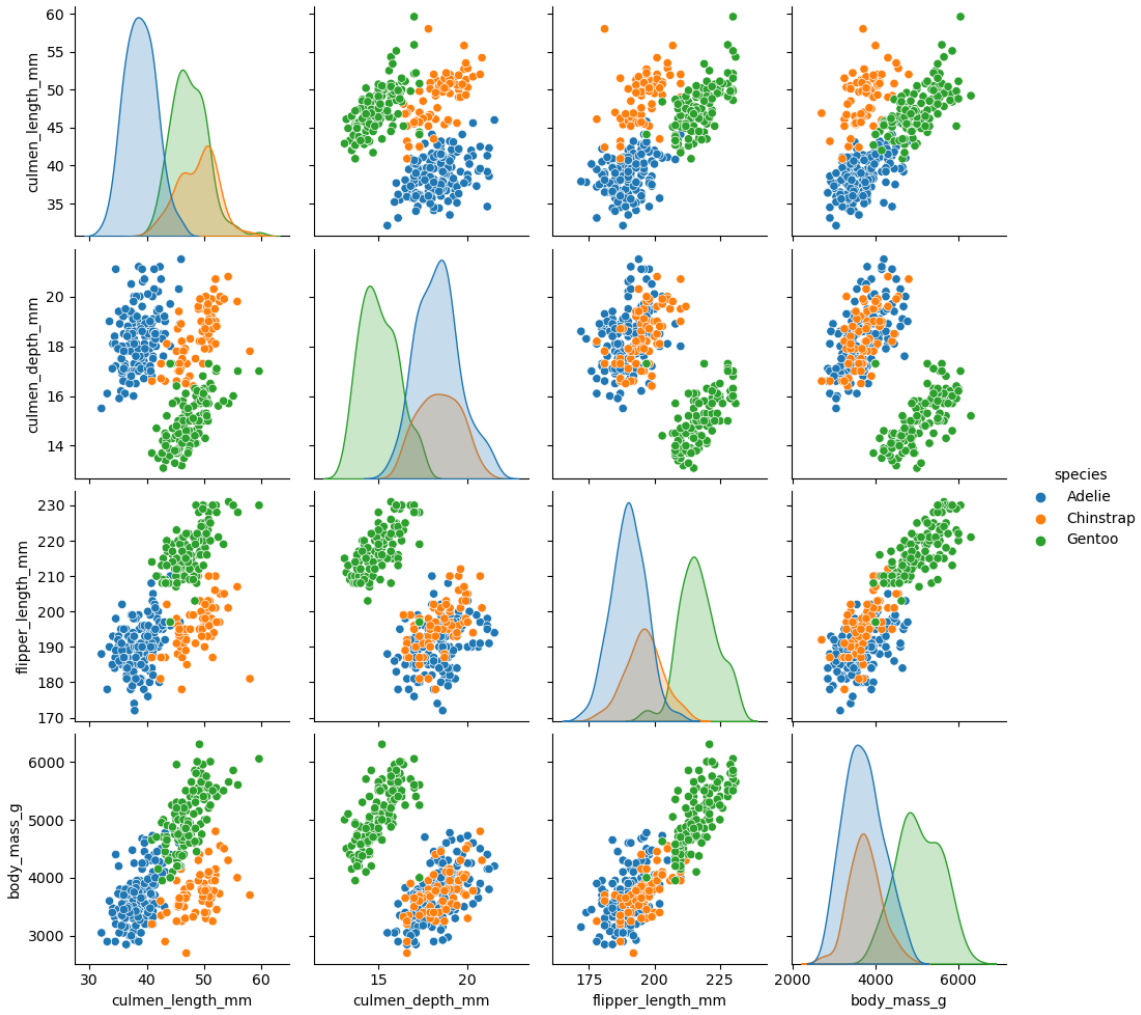
It's important to note that the choice between MAP and MLE estimation ultimately depends on the specific problem and the available data. A well-informed prior can improve the accuracy of MAP estimation, but an incorrect prior can lead to biased results. In contrast, the MLE estimator provides unbiased estimates but may suffer from high variance when the sample size is small. Therefore, the choice of estimator should be made carefully, taking into account the properties of the dataset and the prior information available.

Question #5- Naïve Bayes Classifier on Penguins Dataset

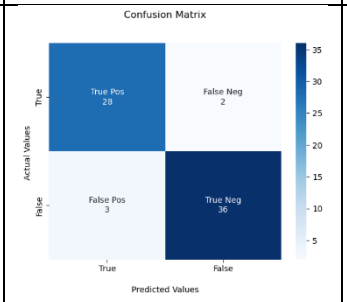
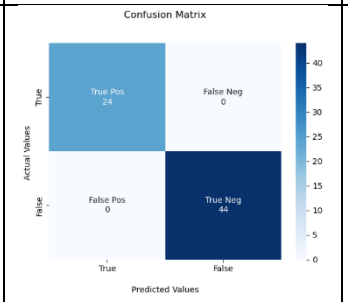
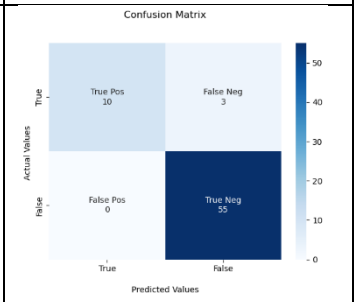
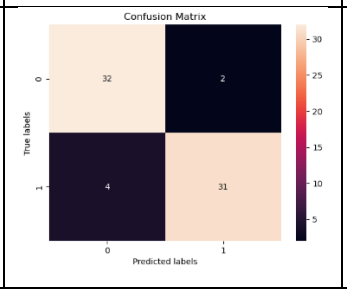
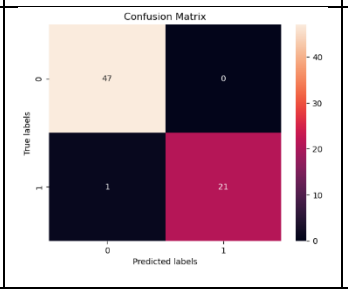
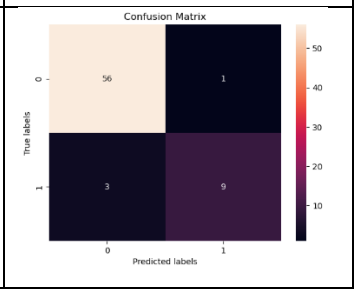
- Preprocessing: The line of code **df = df.fillna(df.quantile(0.5))** fills missing values in a pandas DataFrame df with the median value of the column.
- The presence of missing data can cause problems in data analysis, so it's important to handle missing values in a meaningful way. Filling missing values with the median value can be a simple and effective method of handling missing data.
- Filling missing values with the median value helps to maintain the overall distribution of the data. If the data has a normal distribution, then using the mean value to fill in the missing data would be appropriate. However, if the data is skewed, then using the median value would be a better choice. The figure below clearly depicts the skewed distributions of most of the features, indicating that filling missing values with the median would be a more effective approach.



Pairplot of features:



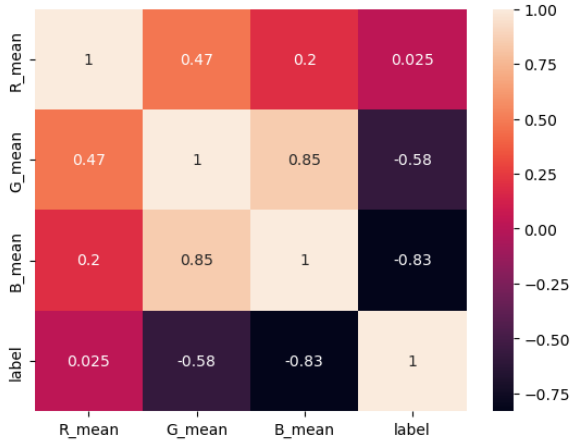
Comparing self-implemented Gaussian Naïve Bayes Classifier with Scikit-Learn's Classifier:

	Adelie vs All	Gentoo vs All	Chinstrap vs All
Self			
Sklearn			

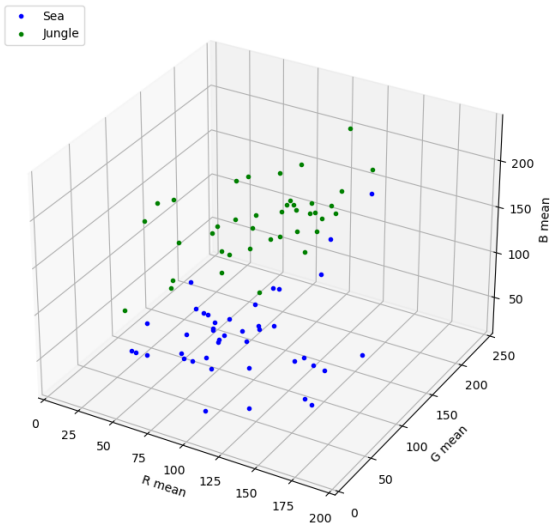
As depicted in the table, it is evident that the outcomes produced by the self-implemented classifier and scikit-learn's classifier exhibit a remarkable degree of similarity.

Question #6- Image Classification

To accurately categorize images according to their overall color as either green or blue, we adopt a well-defined procedure. We commence by computing the per-channel image mean for each image. Next, we compare the mean values of the green and blue channels for each image. If the green channel's mean value is greater than that of the blue channel, we classify the image as Jungle. Conversely, if the blue channel's mean value is higher than that of the green channel, we classify the image as Sea. This method enables us to efficiently and precisely differentiate between images with dominant green or blue hues.

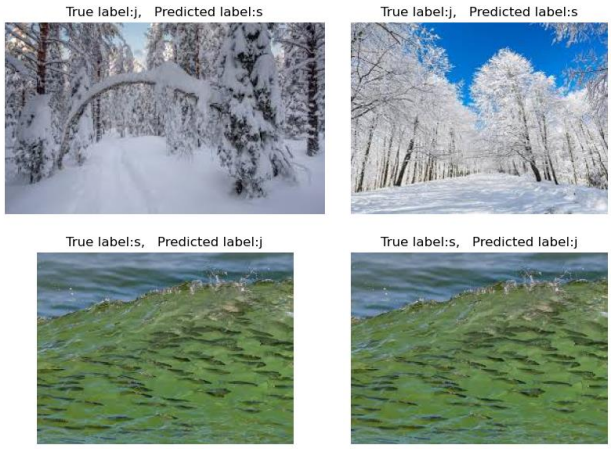


As expected, blue channel and green channel have high correlations with label.



Distribution of Data (Pay attention to high separability along blue channel)

Misclassified data:



Using the described method, only three data points were misclassified. These misclassifications can be attributed to the presence of a blue hue in the jungle images caused by the snow and perspective of the image, as well as a green hue in the sea image, which resulted in its misclassification into the jungle category.