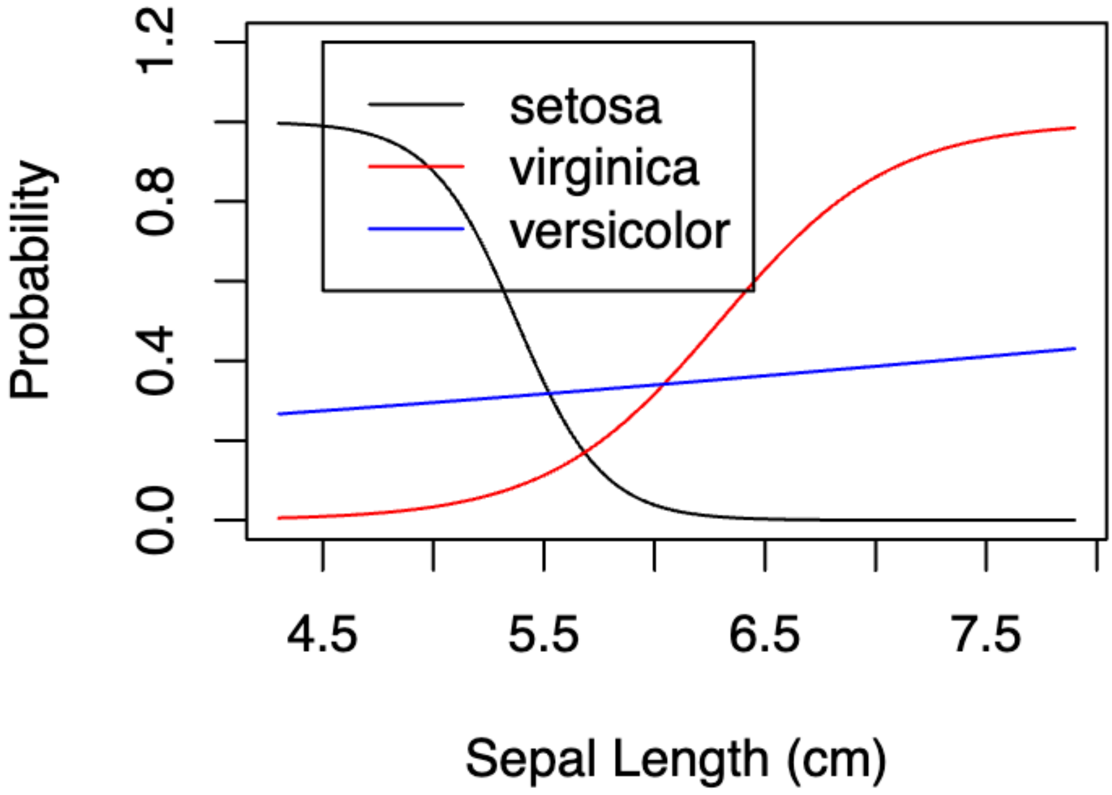
REG Assignment 2

- Sampad Kumar Kar
- MCS202215
- sampadk04@cmi.ac.in

P1



P1.1

If the `Sepal.Length`` is `6.5 cm`, then based on the probability scores as obtained by the Logistic Regression model, `virginica` is the most likely class, with a probability score of around `0.6`.

P1.2

Yes, `Sepal.Length` does indeed have an effect on the identification of the class `versicolor`. Based on the probability plots, we can observe that if the `Sepal.Length` lies approximately between `5.5` to `6`, the model predicts the class `versicolor` with a probability score of around `0.4`.

P2

The linear discriminant rule for class $k \in \{1, 2\}$ is presented as:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^{-1} \Sigma^{-1} \mu_k + \log(\pi_k)$$

We are also given the following summary statistics from two (independent) datasets:

$\overline{X}_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$, $\overline{X}_2 = \begin{pmatrix} 5 \\ 7 \end{pmatrix}$ and $S = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$.

Also, $S^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$.

Also, $n_1 = 20$ and $n_2 = 40$, which means that $\pi_1 = \frac{n_1}{n_1+n_2} = \frac{1}{3}$ and $\pi_2 = \frac{n_2}{n_1+n_2} = \frac{2}{3}$.

P2.(i)

We compute the linear discriminant function and use that to classify the data point $x = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ as either π_1 or π_2.

In [ ]:
```
# define the function to calculate the delta

lda <- function(x, mu, sigma_inv, pi)
{
    delta <- t(x)%*%sigma_inv%*%mu - (0.5)%*%t(mu)%*%sigma_inv%*%mu + log
    return(delta)
}
```

In [ ]:
```
# setting up the parameters and variables

x <- c(1,4)

mu_1 <- c(2,3)
mu_2 <- c(5,7)

sigma_inv <- matrix(c(2,-1,-1,1), ncol=2, nrow=2, byrow=TRUE)

pi_1 <- 1/3
pi_2 <- 2/3
```

In [ ]:
```
# calculate delta for class 1 and class 2

delta_1 = lda(x, mu_1, sigma_inv, pi_1)
delta_2 = lda(x, mu_2, sigma_inv, pi_2)

# print out the results
print(delta_1)
```

```
print(delta_2)
```

```
         [,1]
[1,] 1.401388
          [,1]
[1,] -3.905465
          [,1]
[1,] -3.905465
```

So, we obtain the following results:

- $\delta_1(x) = 1.401388$
- $\delta_2(x) = -3.905465$

Since, $\delta_1(x) > \delta_2(x)$, we can classify x as π_1.

P2.(ii)

Under the following assumptions, we can expect our methods to be reliable:

- The data is normally distributed.
- The covariance matrices are equal across all classes.
- The observations are independent of each other.