SSY316, Hand-in 3

***Group**: 27*
***Student**: Jahanvi B Dinesh*
***Student**: Sumukh S Moudghalya*

Exercise 1

We need to estimate the probability $p(y = 1 \mid x)$ using the given logistic regression model:

$$p(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

Where:

- $\beta_0 = -6$
- $\beta_1 = 0.05$
- $\beta_2 = 1$
- $x_1 = 40$ (hours studied)
- $x_2 = 3.5$ (grade point average)

Substitute the values into the logistic regression equation:

$$p(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}$$

$$p(y = 1 \mid x) = \frac{e^{-6 + 0.05 \cdot 40 + 1 \cdot 3.5}}{1 + e^{-6 + 0.05 \cdot 40 + 1 \cdot 3.5}}$$

Calculate the numerator and denominator:

$$z = -6 + 0.05 \cdot 40 + 1 \cdot 3.5 = -6 + 2 + 3.5 = -0.5$$

The numerator is:

$$e^z = e^{-0.5} \approx 0.6065$$

The denominator is:

$$1 + e^z = 1 + 0.6065 \approx 1.6065$$

Finally, the probability:

$$p(y = 1 \mid x) = \frac{0.6065}{1.6065} \approx 0.3774$$

The probability that a student who studies for 40 hours and has a GPA of 3.5 receives a grade of 5 is approximately 37.74%.

Exercise 2

Question 1

Derive $\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$

The sigmoid function is defined as:

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

Differentiating $\sigma(a)$ with respect to a

Using the quotient rule:

$$\frac{d\sigma(a)}{da} = \frac{d}{da}\left(\frac{1}{1 + e^{-a}}\right)$$

Let $u = 1$ and $v = 1 + e^{-a}$. Then:

$$\frac{d\sigma(a)}{da} = \frac{u'v - uv'}{v^2}$$

- $u' = 0$
- $v' = \frac{d}{da}(1 + e^{-a}) = -e^{-a}$

Substitute into the formula:

$$\frac{d\sigma(a)}{da} = \frac{0 \cdot v - 1 \cdot (-e^{-a})}{(1 + e^{-a})^2} = \frac{e^{-a}}{(1 + e^{-a})^2}$$

Simplifying the result

Factor $\frac{e^{-a}}{1+e^{-a}}$ as $\sigma(a)$:

$$\sigma(a) = \frac{1}{1 + e^{-a}} \implies 1 - \sigma(a) = \frac{e^{-a}}{1 + e^{-a}}$$

So:

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$$

Question 2

Derive the gradient of the log-likelihood.

For logistic regression, the likelihood is:

$$\mathcal{L}(\beta) = \prod_{i=1}^{N} \sigma(x_i^T \beta)^{y_i} \cdot (1 - \sigma(x_i^T \beta))^{1-y_i}$$

The log-likelihood is:

$$\ell(\beta) = \sum_{i=1}^{N} \Big[y_i \log \sigma(x_i^T \beta) + (1 - y_i) \log(1 - \sigma(x_i^T \beta)) \Big]$$

Differentiating the log-likelihood

The gradient of $\ell(\beta)$ with respect to β is:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[\frac{y_i}{\sigma(x_i^T \beta)} \frac{\partial \sigma(x_i^T \beta)}{\partial \beta} - \frac{1 - y_i}{1 - \sigma(x_i^T \beta)} \frac{\partial \sigma(x_i^T \beta)}{\partial \beta} \right]$$

Using $\frac{\partial \sigma(x_i^T \beta)}{\partial \beta} = \sigma(x_i^T \beta)(1 - \sigma(x_i^T \beta)) x_i$, substitute:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \Big[y_i (1 - \sigma(x_i^T \beta)) x_i - (1 - y_i) \sigma(x_i^T \beta) x_i \Big]$$

Factor x_i and simplify:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \Big[y_i - \sigma(x_i^T \beta) \Big] x_i$$

Final Gradient Expression:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \Big(y_i - \sigma(x_i^T \beta) \Big) x_i$$

This is the gradient of the log-likelihood for logistic regression.

Exercise 3

Question 1

Consider a dataset with N samples divided into C classes. Let y be the class label, and $p(y = c) = \pi_c$ is the probability for class c. The likelihood of the dataset is:

$$L(\pi) = \prod_{n=1}^{N} \pi_{y_n}.$$

Taking the log-likelihood:

$$\log L(\pi) = \sum_{n=1}^{N} \log \pi_{y_n}.$$

Group terms by class and define N_c as the count of samples in class c:

$$\log L(\pi) = \sum_{c=1}^{C} N_c \log \pi_c.$$

To maximize the likelihood, use the constraint $\sum_{c=1}^{C} \pi_c = 1$. Apply Lagrange multipliers and solve to get:

$$\pi_c = \frac{N_c}{N}.$$

Question 2

Mean for each class c: Assuming Gaussian densities, the likelihood for class c is:

$$p(x|y=c,\theta) = \prod_{n:y_n=c} \mathcal{N}(x_n|\mu_c,\Sigma).$$

The log-likelihood simplifies to:

$$\log L(\mu_c) = -\frac{N_c}{2}\log|\Sigma| - \frac{1}{2}\sum_{n:y_n=c}(x_n-\mu_c)^T\Sigma^{-1}(x_n-\mu_c).$$

Taking derivative with respect to μ_c and setting it to zero yields:

$$\mu_c = \frac{1}{N_c}\sum_{n:y_n=c} x_n.$$

Shared covariance matrix Σ: Combine the log-likelihoods for all classes and solve for Σ:

$$\Sigma = \frac{1}{N}\sum_{c=1}^{C}\sum_{n:y_n=c}(x_n-\mu_c)(x_n-\mu_c)^T.$$

Rewrite using S_c:

$$\Sigma = \sum_{c=1}^{C}\frac{N_c}{N}S_c, \quad S_c = \frac{1}{N_c}\sum_{n:y_n=c}(x_n-\mu_c)(x_n-\mu_c)^T.$$

Question 3

Using the Gaussian assumption for each class with separate covariance matrices:

$$p(x|y=c,\theta) = \prod_{n:y_n=c} \mathcal{N}(x_n|\mu_c,\Sigma_c).$$

The log-likelihood is:

$$\log L(\Sigma_c) = -\frac{N_c}{2}\log|\Sigma_c| - \frac{1}{2}\sum_{n:y_n=c}(x_n-\mu_c)^T\Sigma_c^{-1}(x_n-\mu_c).$$

Taking derivative with respect to Σ_c and setting it to zero yields:

$$\Sigma_c = \frac{1}{N_c}\sum_{n:y_n=c}(x_n-\mu_c)(x_n-\mu_c)^T.$$

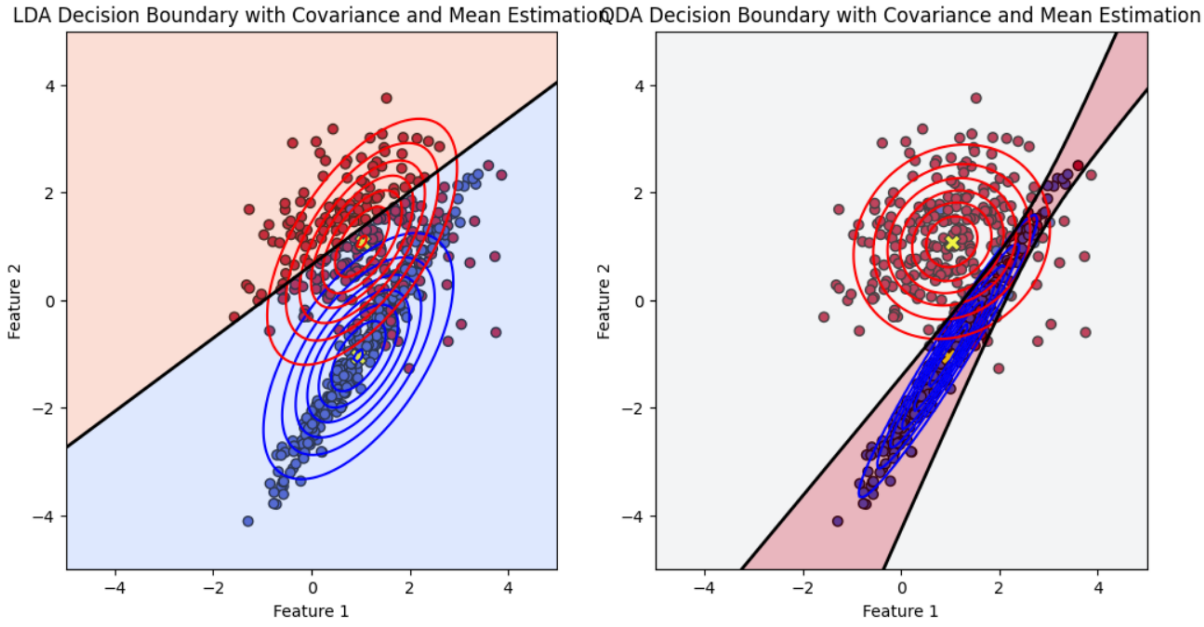
Question 4

Parameter	LDA	QDA
Mean of Class 0	$[0.95473542, -1.04313444]$	$[0.95473542, -1.04313444]$
Mean of Class 1	$[1.03126794, 1.08408911]$	$[1.03126794, 1.08408911]$
Shared Covariance Matrix	$\begin{bmatrix} 0.9554 & 0.6812 \\ 0.6812 & 1.3410 \end{bmatrix}$	Not applicable
Covariance Matrix of Class 0	Not applicable	$\begin{bmatrix} 0.8748 & 1.2292 \\ 1.2292 & 1.8290 \end{bmatrix}$
Covariance Matrix of Class 1	Not applicable	$\begin{bmatrix} 1.0367 & 0.1288 \\ 0.1288 & 0.8491 \end{bmatrix}$
Prior Probability (π)	0.502	0.502
Weights (w)	$[-1.6477276, 2.42332154]$	Not applicable
Bias Term (w_0)	-1.5785730634659674	Not applicable

Table 1: Parameters for Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

Question 5

- LDA accuracy: 0.692
- QDA accuracy: 0.076



Exercise 4

Question 1

To derive the negative log-likelihood (NLL) for the logistic regression model, we start with the likelihood of the binary outcome $y \in \{0, 1\}$:

$$P(y|X,\beta) = \prod_{i=1}^{N} p(y_i|x_i,\beta),$$

where

$$p(y_i|x_i,\beta) = \sigma(x_i^T\beta)^{y_i}(1-\sigma(x_i^T\beta))^{1-y_i},$$

and $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic (sigmoid) function.

Taking the log of the likelihood, we obtain the log-likelihood:

$$\text{Log-Likelihood: } \ell(\beta) = \sum_{i=1}^{N} \left[y_i \log(\sigma(x_i^T\beta)) + (1-y_i)\log(1-\sigma(x_i^T\beta)) \right].$$

The negative log-likelihood (NLL) is then:

$$\text{NLL: } L(\beta) = -\ell(\beta) = -\sum_{i=1}^{N} \left[y_i \log(\sigma(x_i^T\beta)) + (1-y_i)\log(1-\sigma(x_i^T\beta)) \right].$$

In simplified form, the NLL is:

$$L(\beta) = \sum_{i=1}^{N} \left[-y_i x_i^T\beta + \log(1+e^{x_i^T\beta}) \right].$$

Question 2

(a) Derive the Gradient of the NLL

The negative log-likelihood (NLL) is given as:

$$L(\beta) = \sum_{i=1}^{N} \left[-y_i x_i^T\beta + \log(1+e^{x_i^T\beta}) \right].$$

To derive the gradient, we differentiate $L(\beta)$ with respect to β.

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[-y_i x_i + \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} x_i \right].$$

Simplifying using $\sigma(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}}$, we have:

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[\sigma(x_i^T\beta)x_i - y_i x_i \right].$$

This can be written compactly as:

$$\frac{\partial L(\beta)}{\partial \beta} = X^T(\sigma(X\beta) - Y),$$

where:

- X is the $N \times d$ matrix of feature vectors,

- $\sigma(X\beta)$ is the N-dimensional vector of predictions,
- Y is the N-dimensional vector of true labels.

(b) Compute the Hessian Matrix

The Hessian is the second derivative of $L(\beta)$ with respect to β:

$$H(\beta) = \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T}.$$

From the gradient:

$$\frac{\partial L(\beta)}{\partial \beta} = X^T(\sigma(X\beta) - Y),$$

we compute the Hessian by differentiating again:

$$H(\beta) = X^T \text{diag}(\sigma(X\beta)(1 - \sigma(X\beta)))X,$$

where $\text{diag}(\sigma(X\beta)(1 - \sigma(X\beta)))$ is a diagonal matrix with entries $\sigma(x_i^T\beta)(1 - \sigma(x_i^T\beta))$.

(c) To Use the Hessian for Optimization (Newton-Raphson Algorithm)

The Newton-Raphson method updates β iteratively as follows:

$$\beta^{(t+1)} = \beta^{(t)} - H^{-1}(\beta^{(t)})\nabla L(\beta^{(t)}),$$

where:

- $\nabla L(\beta^{(t)})$ is the gradient of the NLL at $\beta^{(t)}$,
- $H(\beta^{(t)})$ is the Hessian at $\beta^{(t)}$.

This method converges faster than gradient descent due to the inclusion of second-order information. However, it requires computing and inverting the Hessian, which can be computationally expensive for high-dimensional data.

Question 3

Please find the code attached.

Question 4

Please find the code attached.

Question 5

Please find the code attached.

Question 6 , 8

6a, 8

Refer to Table 2

Table 2: Table for question Exercise 4 Question 6a

Metric	Newton Raphson	Scikit Learn
Accuracy	0.66	0.69
Precision	0.76	0.73
Recall	0.77	0.91
F1 Score	0.76	0.81

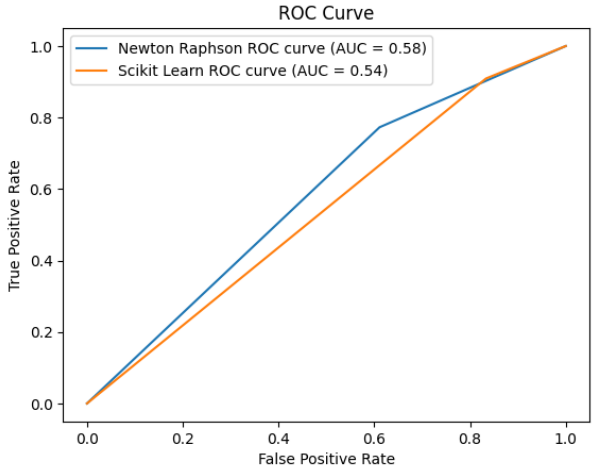


Figure 1: ROC Curve

6b

Refer to Figure 1: ROC Curve

Question 7

Nodes is the most significant predictor. The higher number of nodes implies a decrease in survivability. The older the age, the lower the chances of survival. Due to advances in treatment strategies, the more recent years indicate increased survivability.

Exercise 5

Question 1

We need to derive the Laplace approximation for Bayesian logistic regression. The likelihood is given as:

$$p(y_i|x_i, \beta) = \sigma(x_i^T\beta)^{y_i}(1 - \sigma(x_i^T\beta))^{1-y_i},$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic sigmoid function. The prior on β is Gaussian:

$$p(\beta) = N(\beta|0, I) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\beta^T\beta}.$$

The posterior distribution is:

$$p(\beta|X, Y) \propto p(Y|X, \beta)p(\beta).$$

Using the likelihood and prior, we have:

$$p(\beta|X,Y) \propto \prod_{i=1}^{N} \sigma(x_i^T\beta)^{y_i}(1-\sigma(x_i^T\beta))^{1-y_i} \cdot e^{-\frac{1}{2}\beta^T\beta}.$$

The log-posterior is:

$$\log p(\beta|X,Y) = \sum_{i=1}^{N} \Big[y_i \log \sigma(x_i^T\beta) + (1-y_i)\log(1-\sigma(x_i^T\beta))\Big] - \frac{1}{2}\beta^T\beta + \text{const.}$$

The negative log-posterior is:

$$-\log p(\beta|X,Y) = \underbrace{-\sum_{i=1}^{N} \Big[y_i \log \sigma(x_i^T\beta) + (1-y_i)\log(1-\sigma(x_i^T\beta))\Big]}_{\text{Negative log-likelihood (NLL)}} + \underbrace{\frac{1}{2}\beta^T\beta}_{\text{Prior term}} + \text{const.}$$

The mode of the posterior (Maximum A Posteriori, or MAP estimate) is the value of β that minimizes the negative log-posterior:

$$\hat{\beta} = \arg\min_{\beta} \left[L(\beta) + \frac{1}{2}\beta^T\beta\right],$$

where $L(\beta)$ is the negative log-likelihood.

The MAP estimate can be computed iteratively using the Newton-Raphson algorithm:

$$\beta^{(t+1)} = \beta^{(t)} - H^{-1}(\beta^{(t)})\nabla\tilde{L}(\beta^{(t)}),$$

where:

- $\tilde{L}(\beta) = L(\beta) + \frac{1}{2}\beta^T\beta$ is the negative log-posterior,
- $\nabla\tilde{L}(\beta)$ is the gradient of $\tilde{L}(\beta)$,
- $H = \nabla^2\tilde{L}(\beta)$ is the Hessian of $\tilde{L}(\beta)$.

The Laplace approximation models the posterior as a Gaussian distribution around the MAP estimate:

$$p(\beta|X,Y) \approx N(\beta|\hat{\beta}, \Sigma),$$

where: $\hat{\beta}$ is the MAP estimate, $\Sigma = H^{-1}$ is the covariance matrix, with H being the Hessian of the negative log-posterior at $\hat{\beta}$:

$$H = \nabla^2\tilde{L}(\hat{\beta}).$$

The Hessian of the negative log-posterior is:

$$H = X^T W X + I,$$

where:

- $W = \text{diag}(\sigma(X\hat{\beta})(1 - \sigma(X\hat{\beta})))$ is a diagonal matrix,
- I is the identity matrix (from the prior).

Question 2

Table 3: Table for question Exercise 5 Question 2

Metric	Laplace approximation
Accuracy	0.66
Precision	0.76
Recall	0.77
F1 Score	0.76