CP8318 Assignment 2
Student Name: Albina Cako
Student Number: 500390052

Question 1
Part 1:

The graph below shows the result obtained from the validation data of Dataset 1. Logistic Regression using Newton's Method was used to train the data. In red, we can observe the decision boundary.

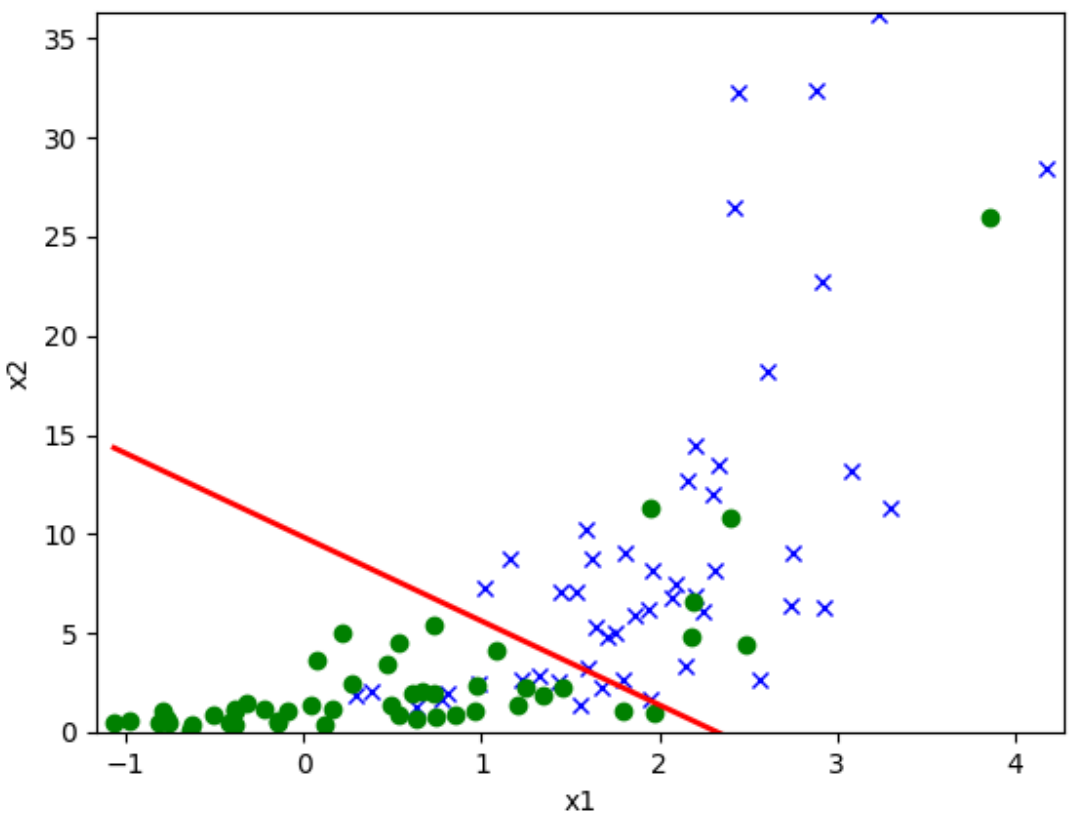

Figure 1. Plot of the validation data in dataset 1 using Logistic Regression.

Part 2.

The graph below shows the result of the results obtained from the validation data of Dataset 1. Gaussian Discriminant Analysis (GDA) was used to train the data. In red, we can observe the decision boundary.

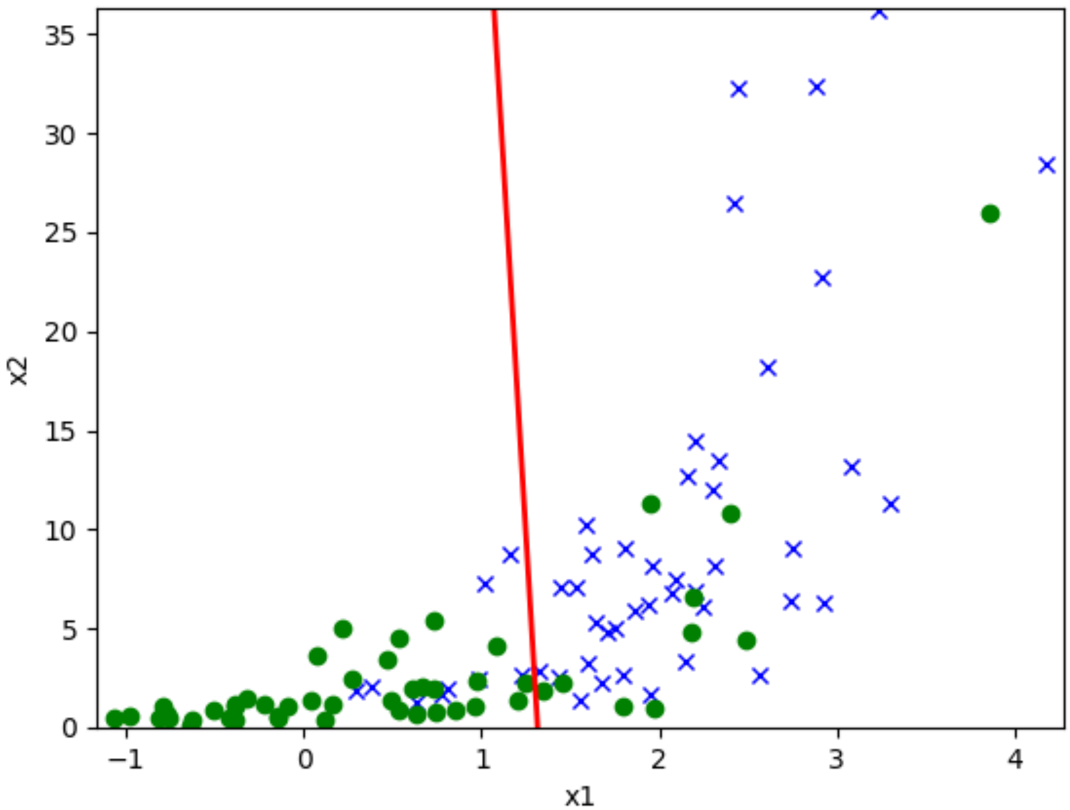

Figure 2. Plot of the validation data in dataset 1 using GDA.

Part 3.

Figure 1 and 2 show the decision boundary splitting the two groups of the validation data in dataset 1 using Logistic Regression and GDA to train the data, respectively. In the graph, the green circles represent the validation samples of the true labels equal to 1, while the blue stars represent the validation samples of the true labels equal to 0. The decision boundary is shown in red. As we observe in the graphs, the decision boundary on Figure 1 (Logistic regression) is a better fit, then the decision boundary on Figure 2 (GDA). This shows that Logistic regression has a higher accuracy and performs better in this dataset. This is due to the fact that the dataset is not Gaussian, thus Logistic regression performs better then GDA.

Part 4.

The graphs below shows the results obtained from the validation data of Dataset 2. Logistic Regression using Newton's Method and GDA were used to train the data, respectively. In red, we can observe the decision boundary.

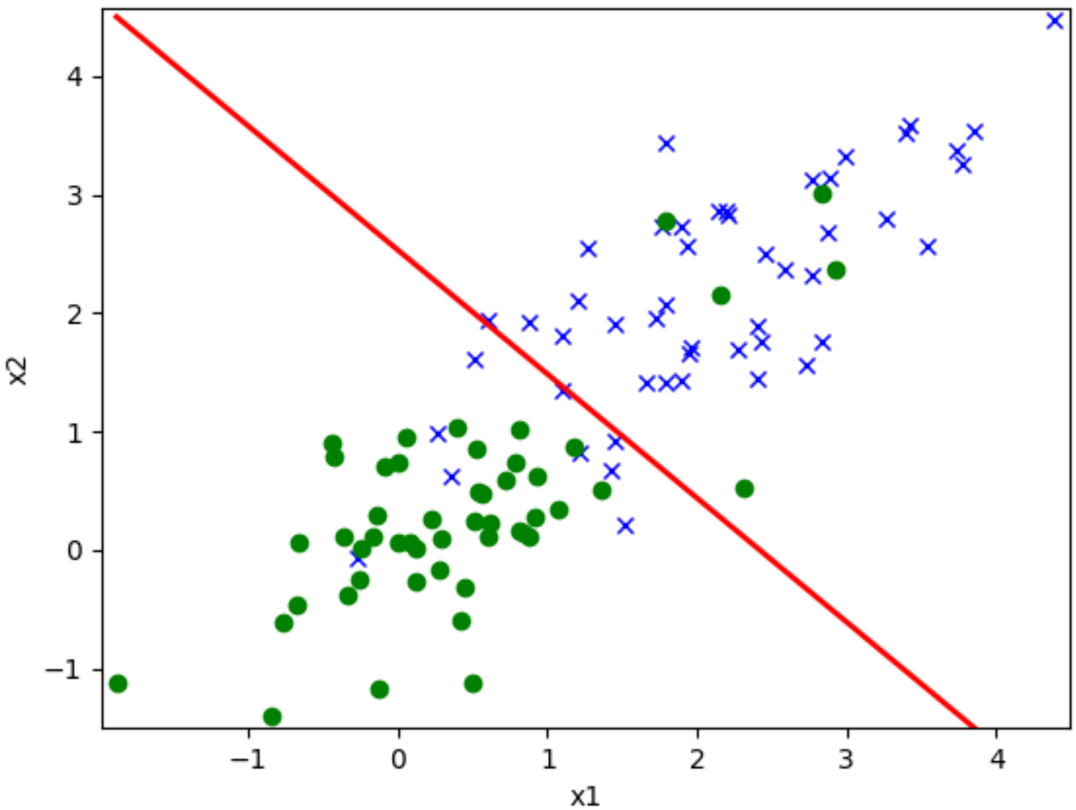

Figure 3. Plot of the validation data in dataset 2 using Logistic Regression.

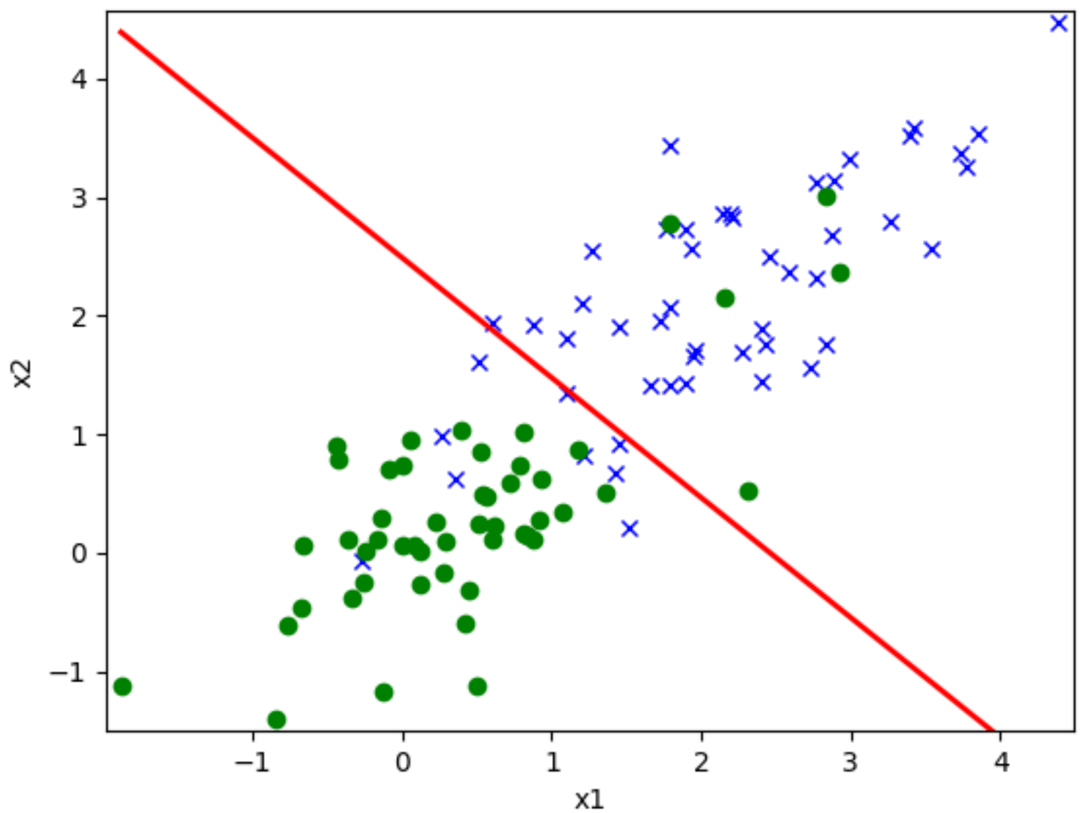

Figure 4. Plot of the validation data in dataset 2 using GDA.

Figure 3 and 4 show the decision boundary splitting the two groups of the validation dataset 2 using Logistic Regression and GDA to train the data, respectively. In the graphs, the green circles represent the validation samples of the true labels equal to 1, while the blue stars represent the validation samples of the true labels equal to 0. The decision boundary is shown in red. As we can see, in this dataset, the GDA and logistic regression perform very similarly. This is because Dataset 2 is more Gaussian then Dataset 1, thus GDA performs better in the Dataset 2.

GDA performs worse than logistic regression in Dataset 1. This is most likely due to the fact that the $x^{(i)}$ values are not Gaussian in Dataset 1. From lecture, we know that logistic regression works better than GDA when the dataset does not come from a multivariate Gaussian.

Part 5.

As we observe above, GDA performs better with a dataset being Gaussian. Thus, to improve the performance of the GDA, I would use logarithmic transformation on Dataset 1. I would transform the data by setting $x_2 := \log x_2$, which would make the $x^{(i)}$'s to become Gaussian. Thus, GDA would perform much better after this transformation, since it works better with Gaussian data.

Question 2.
Part 1.

Question 2

Part 1.

Given:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

If we use $a = \exp(\log a)$:

$$p(y;\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} = e^{\log\left(\frac{e^{-\lambda}\lambda^{y}}{y!}\right)}$$

$$= e^{\log e^{-\lambda} + \log \lambda^{y} - \log y!}$$

$$= e^{(-\lambda + y \log \lambda - \log y!)}$$

$$\boxed{p(y;\lambda) = \frac{1}{y!} e^{y \log \lambda - \lambda}}$$

which can be rewritten as: $\boxed{p(y;\lambda) = \frac{1}{y!} \exp(y \log \lambda - \lambda)}$

Therefore, Poisson distribution is in the exponential family.

-The standard form of the exponential family is:

$$p(y;\eta) = b(y) \exp(\eta T(y) - a(\eta))$$

Part 1 continued:

By comparing our equation with the standard form of the exponential family we get:

$$b(y) = \frac{1}{y!}$$

$$\eta = \log \lambda$$

$$T(y) = y$$

$$a(\eta) = \lambda = e^{\eta}$$

Part 2.

We need to prove that:

$g(\eta) = E[y;\eta] = e^{\eta} = e^{\theta^T x}$

For Poisson distribution we know that:

$E[y;\eta] = \mu = \lambda$

From part 1 we know that:

$\lambda = e^{\eta}$

Therefore $E[y;\eta] = \lambda = e^{\eta}$

$E[y;\eta] = e^{\eta}$ can be proved another way also:

$$E[y;\eta] = \frac{\partial a(\eta)}{\partial \eta}$$

$$= \frac{\partial e^{\eta}}{\partial \eta}$$ ⇒ from part 1 we know that $a(\eta) = e^{\eta}$

$$= e^{\eta}$$

Generalized linear models usually make the assumption that:

$\eta = \theta^T x$ (we set $\eta = \theta^T x$)

Therefore: $g(\eta) = E[y;\eta]$

$= \lambda$

$= e^{\eta}$

$= e^{\theta^T x}$

Part 3.

Below is the graph of the validation dataset using Poisson Regression for training of the model.

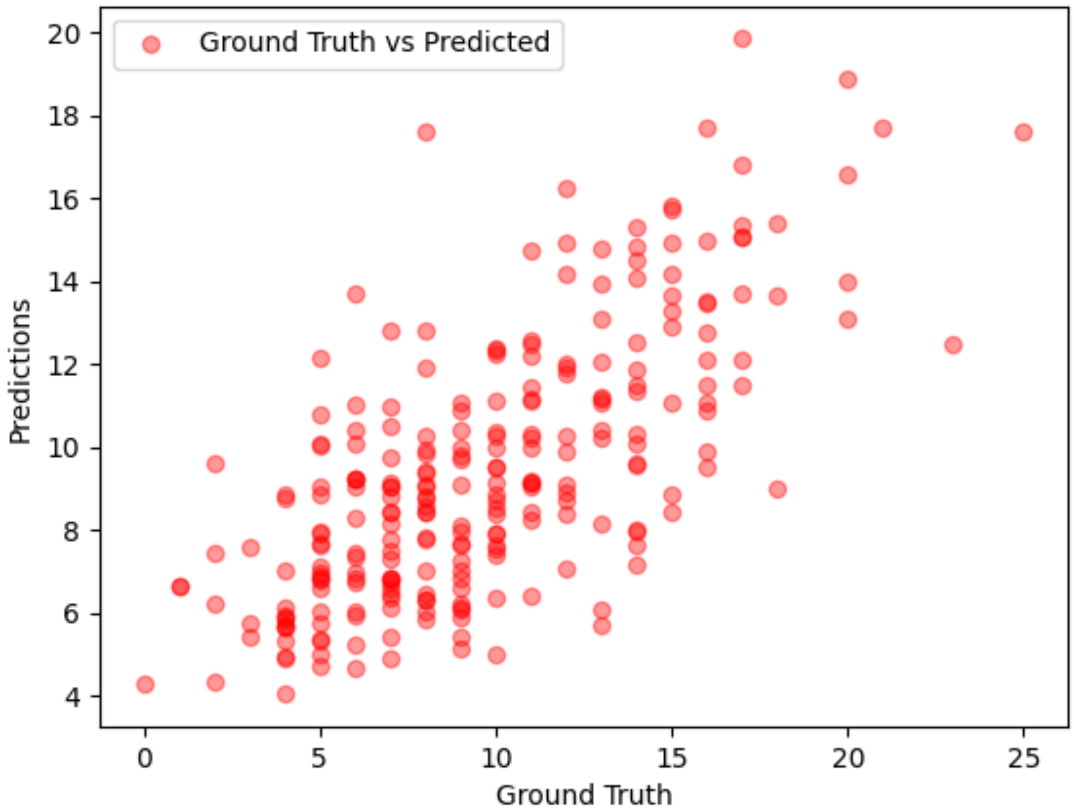

Figure 5. Ground Truth vs. Predictions plot on the validation dataset using Poisson Regression

Figure 5 represents the graph of the ground truth vs. the predictions of number of visitors per day. The x-axis, which is the ground truth or the true counts, represents the true number of visitors per day. The values in the x-axis are integers. The y-axis represents the predicted number of visitors per day and the values are real numbers. The graph shows the prediction of the visitors per day. For example, on the days where there are 0 true visitors (represented by ground truth), the graph shows the prediction of roughly 4 visitors per day. Another example, for 5 visitors, the model gives a predicted range of roughly 4-12 visitors per day. The model does not seem to perform at a very high accuracy, however it most true values it gives a prediction range where the true value lies.

Question 3.

For this question I have pasted the results from my code. For part 4 I have written the explanation of the difference between the performance of SVM and Naïve Bayes models.

Part 1.

The code printed: Size of dictionary 1722.

Part 2:

The code printed: Naive Bayes had an accuracy of 0.978494623655914 on the testing set.

As we can see, the naïve bayes model had a high accuracy, thus performed very well in the spam classification dataset.

Part 3:

The code printed: The top 5 indicative words for Naive Bayes are: ['claim', 'won', 'prize', 'tone', 'urgent!'].

Part 4:

The code printed: The optimal SVM radius was 0.1. The SVM model had an accuracy of 0.9695340501792115 on the testing set.

In this spam classification dataset, both Naïve Bayes and SVM performed with high accuracy of 0.978494623655914 and 0.9695340501792115, respectively. In machine learning, different models are run on different datasets, as each model has certain assumptions that need to be followed for the models to perform well. Naïve Bayes model performed better in this case, with a higher accuracy. This could be due to the fact that Naïve Bayes uses the assumption that features are independent, which this is the case for the messages on this dataset. SVM also had a high accuracy, however, it was outperformed by Naïve Bayes model.