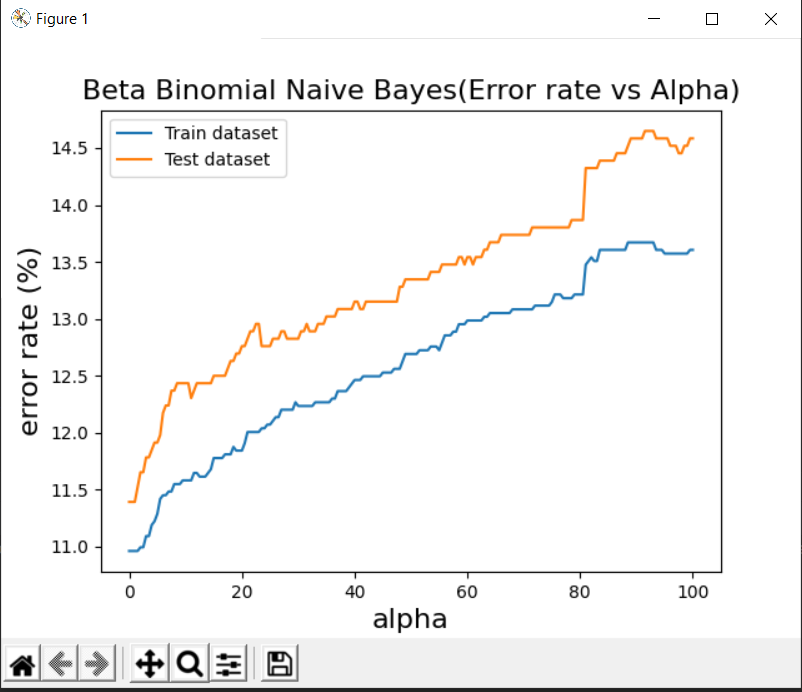


|  |  |
| --- | --- |
| **Module** | EE5907 |
|  |  |
| **Name** | Zavier Ong Jin Jie |
|  |  |
| **Matriculation No.** | A0138993L |

# Question 1. Beta-Binomial Naïve Bayes

**Plots of Training and test error rates versus α**



**Observation about training and test errors as α change**

We can observe that:

1. Error rates for both datasets increases as alpha increases.
2. Trend for both train and test datasets are similar
3. Test dataset always has a higher error rate compared to train dataset
4. There is a spike in error rate when α ≈ 80, this would suggest a poor fit of the emails due to the change in prior distribution.

**Training and testing error rates for α = 1, 10 and 100**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 10 | 100 |
| Training | 0.10962 | 0.11289 | 0.12692 |
| Test | 0.11393 | 0.11914 | 0.13346 |

Note: These error rates are rounded up at 5dp for the ease of viewing. The actual value of the errors rates is displayed when running the code.

# Question 2. Gaussian Naïve Bayes

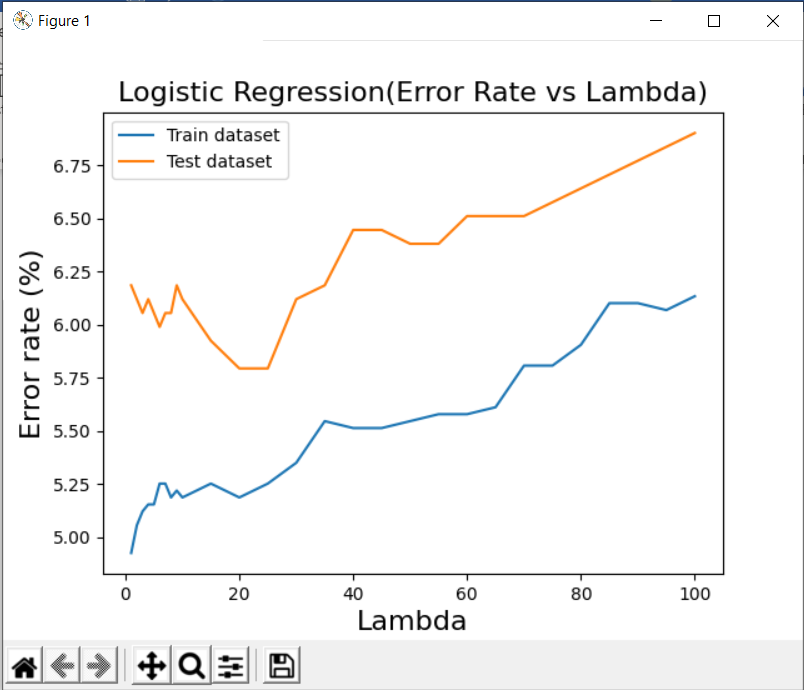
**Training and testing error rates for log-transformed data**

|  |  |
| --- | --- |
|  | Error Rate |
| Training | 0.16672 |
| Test | 0.18359 |

Note: These error rates are rounded up at 5dp for the ease of viewing. The actual value of the errors rates is displayed when running the code.

# Question 3. Logistic Regression

**Plots of training and test error rates versus λ**



**Observation about training and test errors as λ change**

We can observe that:

1. Error rate for both training and test sets are generally increasing as λ increases
2. For smaller values of λ (1-10), the error in the training data is about 1% lesser as compared to the test set.
   1. This behavior is to be expected as when λ is small, the model is too complex and overfits to the training data and therefore, leading to higher error rates in the test data.
3. When λ is more than 10, the test error decreases to about 5.75% before steadily increasing again.
   1. This behavior is also to be expected as when λ is too large, the model is too simple, resulting in a possible underfitting of the training data. Therefore, the model is not useful enough to make accurate predictions in the test data.
4. At λ ≈ 25 is where the test data has the smallest error rate. Which suggest setting the regularization parameter λ to 25 would produce the most accurate model.
5. After trying multiple values for the error threshold to determine convergence (between 10^-3 to 10^-8), it did not affect the final result in a significant way. This could be because the rate of convergence is much faster when using the newtons method.

**Training and test error rates for λ = 1, 10 and 100**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 10 | 100 |
| Training | 0.04927 | 0.05188 | 0.06134 |
| Test | 0.06185 | 0.06120 | 0.06901 |

Note: These error rates are rounded up at 5dp for the ease of viewing. The actual value of the errors rates is displayed when running the code.

# Question 4. K-Nearest Neighbors

**Plots of training and test error versus K**



Figure . Equal count classified to not spam

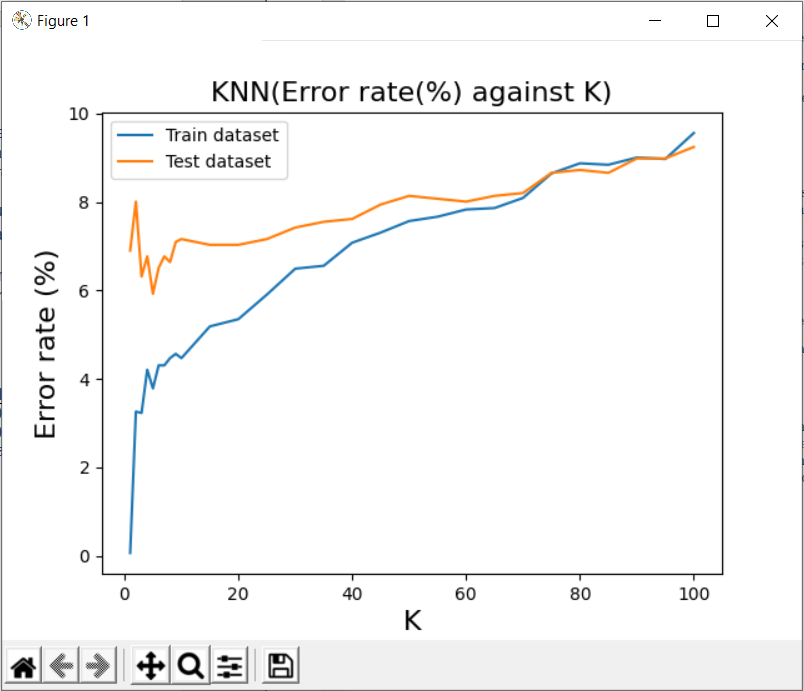


Figure . Equal count classified to spam

**Observation about training and test errors as K change**

In this question, I ran the code twice with slight differences, when K is even, there is a chance that the number of spam mails and not spam mails are equivalent to each other. Therefore, an assumption is made to either classify them as spam or not spam, which explains why Figure 1 and 2 has slight differences in error rates.

We can observe that:

1. While the error rates in both training and test datasets are increasing, the test data increases gradually while the training dataset behaves like a function of log.
2. When K = 1, the first nearest neighbor for any email in the training data is itself, which explains the near 0 error rate on the training data. On the other hand, the first nearest neighbor for any email in the test data is not itself, explaining the much higher error rate.
3. At small values of K, the model is being overfitted to the training data, explaining the difference in error rates between the training and test dataset.
4. At larger values of K, both error rates tend to converge to the same error rate. It is also useful to note that if K increases, the decision surface becomes smoother. However, if K = N, it would become underfitting and everything would become one class which is the majority class in the dataset.
5. In this case where the result of the mail is binary (spam or not spam), if may be a better idea to not calculate even K values when K is small so as to avoid wrongly classifying them when spam count is equivalent to not spam count. (E.g. K=2, spam count = 1, not spam count =1)

**Training and test error rates for K = 1, 10 and 100**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 10 | 100 |
| Training | 0.00065 | 0.05024 | 0.09168 |
| Test | 0.06901 | 0.07357 | 0.09180 |

Table . Equal count classified as not spam

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 10 | 100 |
| Training | 0.00065 | 0.04470 | 0.09560 |
| Test | 0.06901 | 0.07161 | 0.09245 |

Table . Equal count classified as spam

Note: These error rates are rounded up at 5dp for the ease of viewing. The actual value of the errors rates is displayed when running the code.

# Question 5: Survey

I spent 4 weekends to complete the assignment (1 weekend for 1 question), which amounts to about 48 hours.

I think that this assignment is helpful as I can practice and try to apply the concepts taught during lecture. Though it was difficult, I enjoyed testing different possibilities and trying different parameters to see how it affected the results.