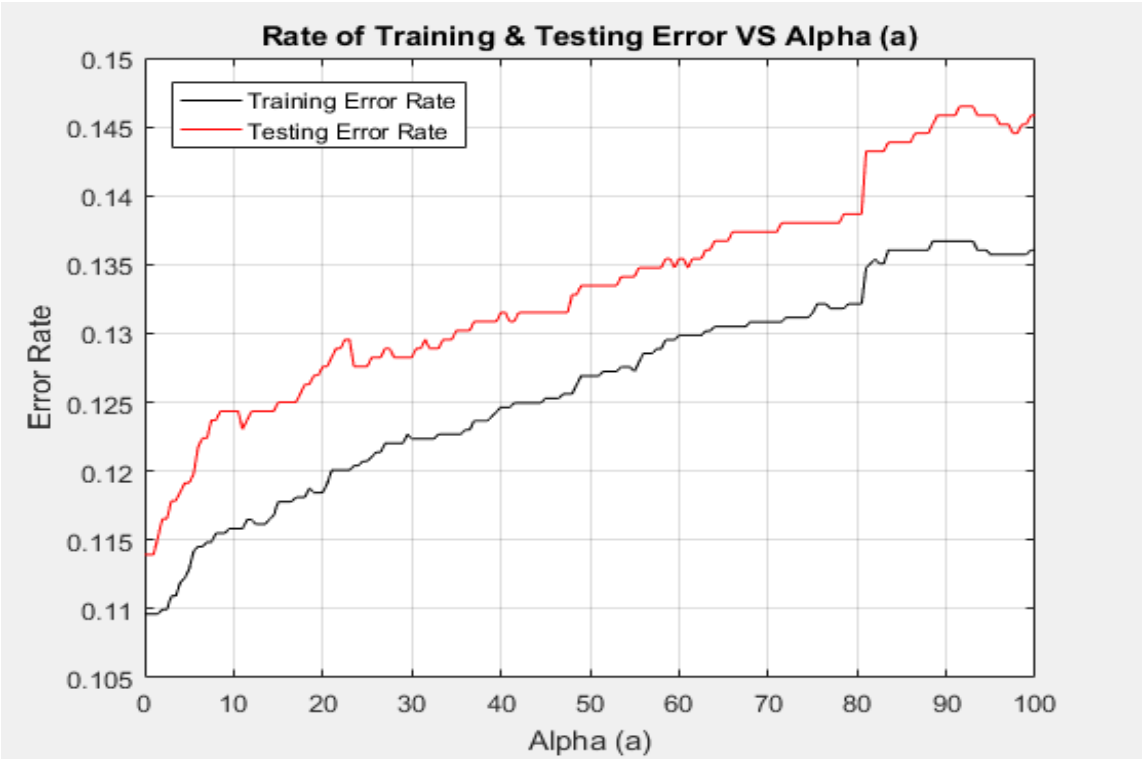
EE5907 CA1 Programming Assignment

Name: Zhu Qi
Matric Number: A0224460N

Contents

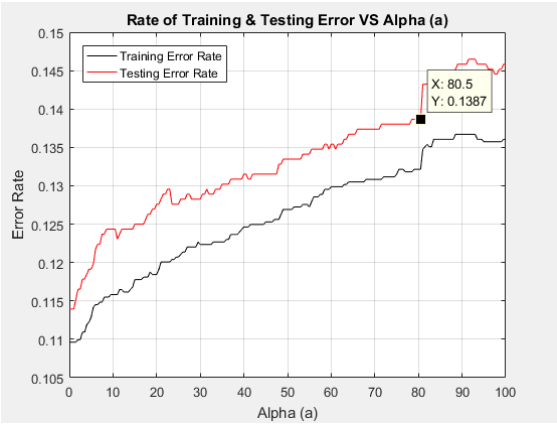
Q1. Beta-binomial Naïve Bayes

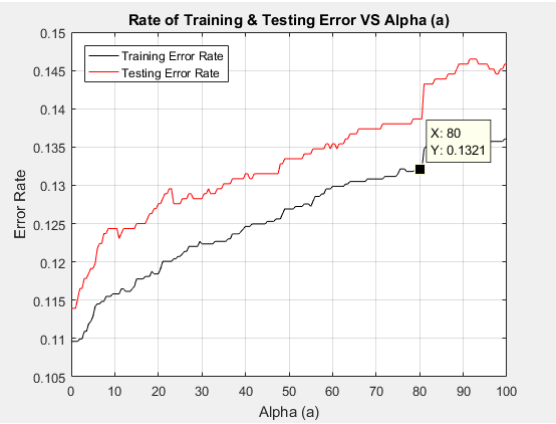
1 Plots of training and test error rates versus α.



2 What do you observe about the training and test errors as α change?

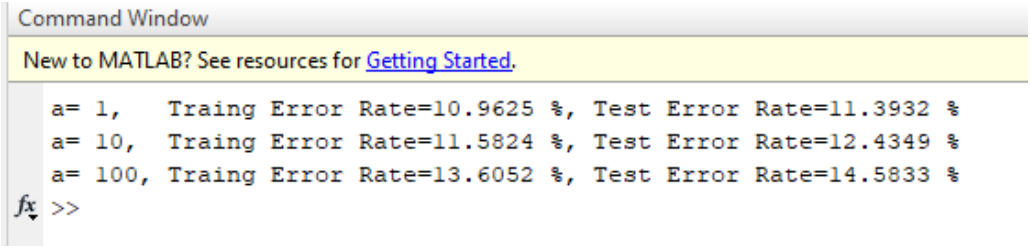
From the observation shown above, both the training and test error rates will *increase proportionally* to α at a similar trend.
Generally speaking, the testing error rate is higher than the training error rate. Both the training and test error rates fluctuate while increasing and *increased sharply at* the beginning from *α = 0 to α = 10*, as well as at *α= 80* (the screenshots fo the coordinates are showing in the figures below). Therefore we can know if we have sufficient training, or whether we have chosen a suitable value of α t to reduce the error rate by observing these trends.





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

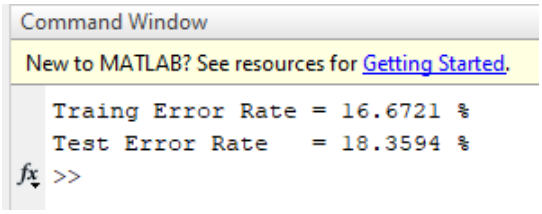
α	Training Error Rate	Testing Error Rate
1	0.109625	0.113932
10	0.115824	0.124349
100	0.136052	0.145833



Q2. Gaussian Naïve Bayes

1 Training and testing error rates for the log-transformed data.

	Training Error Rate	Testing Error Rate
Log Transformed Data	0.166721	0.183594

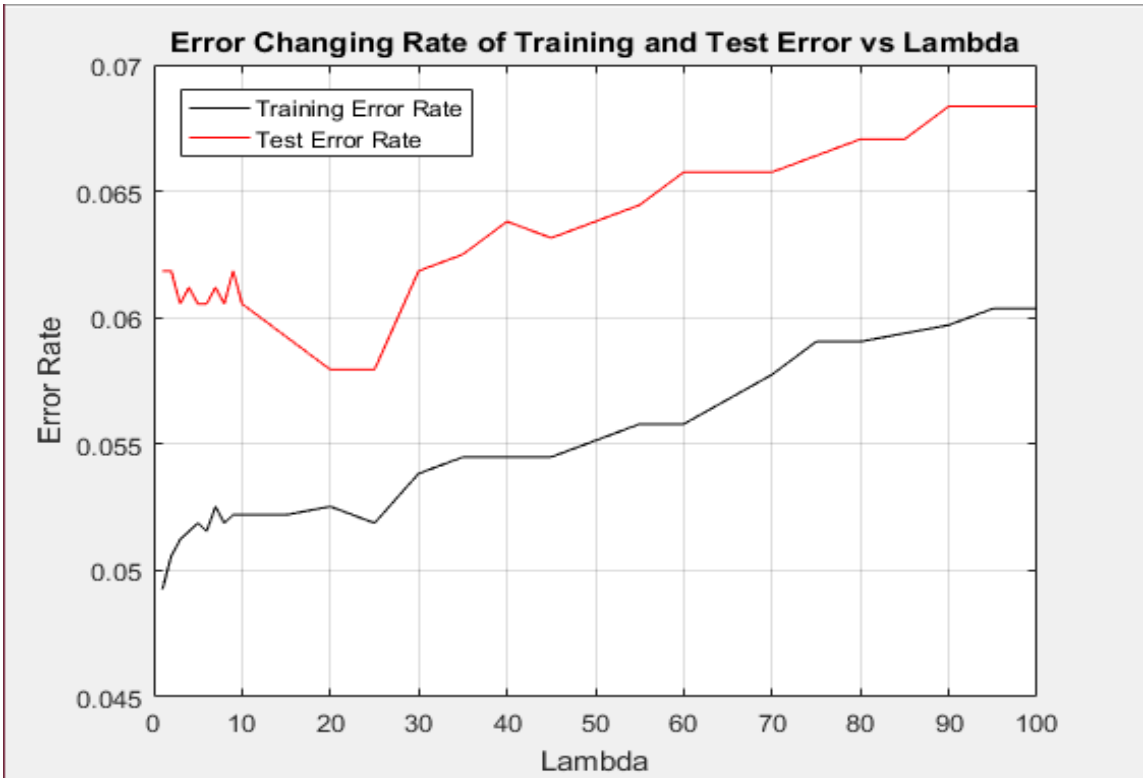


From the observation of the error rate for the log-transformed data using the Gaussian Naïve Bayes classifier (training error rate = 16.67% and testing error rate = 18.36%) is higher than the result error rates from the beta-binomial Naïve Bayes classifier in the worst scenario (when α=100, training error rate = 13.60% and testing error rate = 14.58%).

Therefore, by comparing questions one and two, we know that Beta-Binomial Naïve Bayes Classifier is better for this case.

Q3. Logistic Regression

1 The plot of training and test error rates versus λ



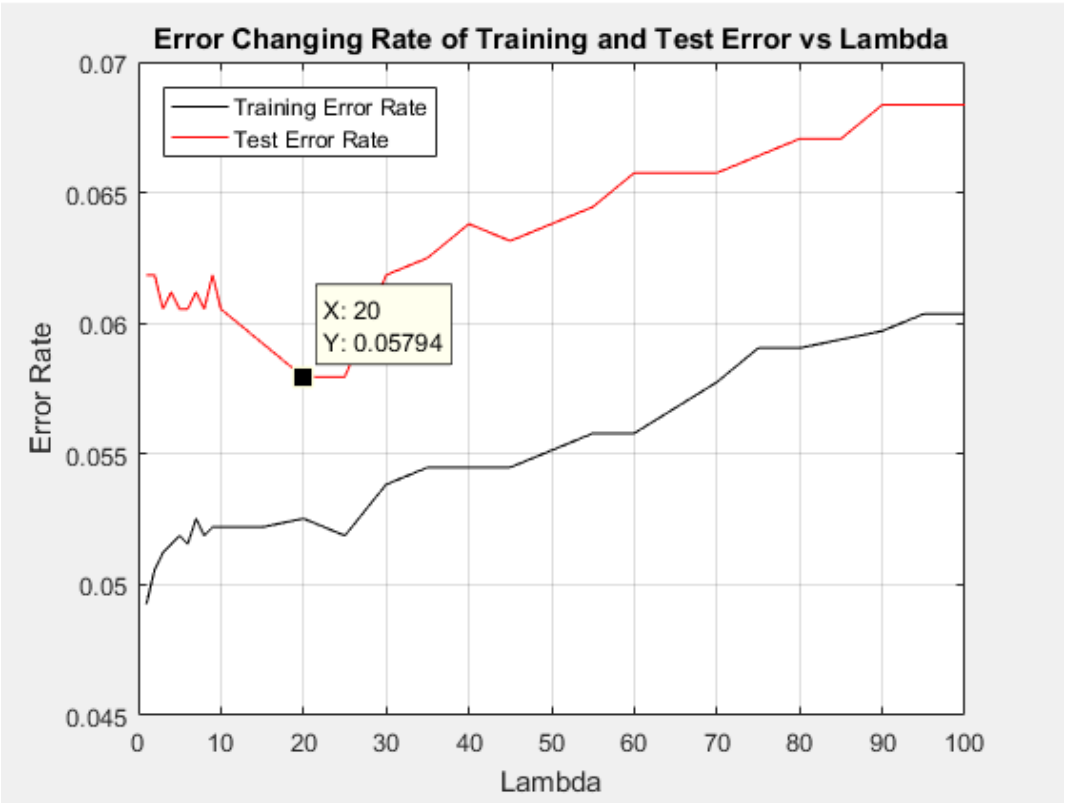
2 What do you observe about the training and test errors as λ change?

From the observation of the plot above, generally speaking, the error rate of test data is higher than the training data and both of them are remaining less than 7.00%.

While **λ** is increasing, the difference between Training and testing error rate forms 2 trends:

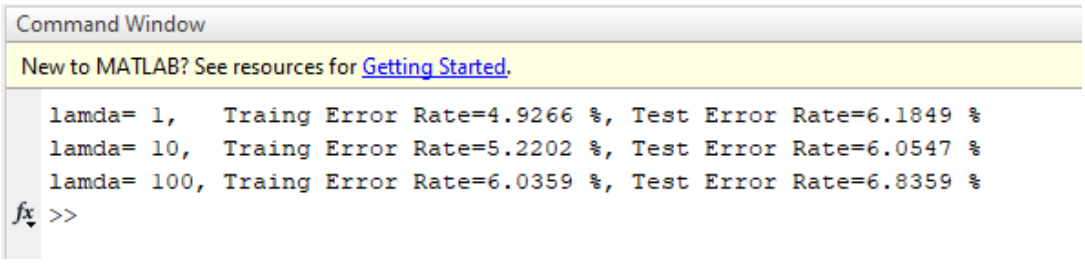
1) When 0 ≤ λ ≤20 (coordinate is indicated in the plot below), the difference between training and the test error rate is gradually decreasing, the training error rate is increased proportionally to λ with fluctuation, while the testing error rate has a huge spike of fluctuation λ ≤10, and then decreased proportionally to λ.

2) When λ > 20, the difference between training and the test error rate is roughly remaining as a constant, and both training and test error rates are increasing proportionally with **λ** with some small fluctuations.



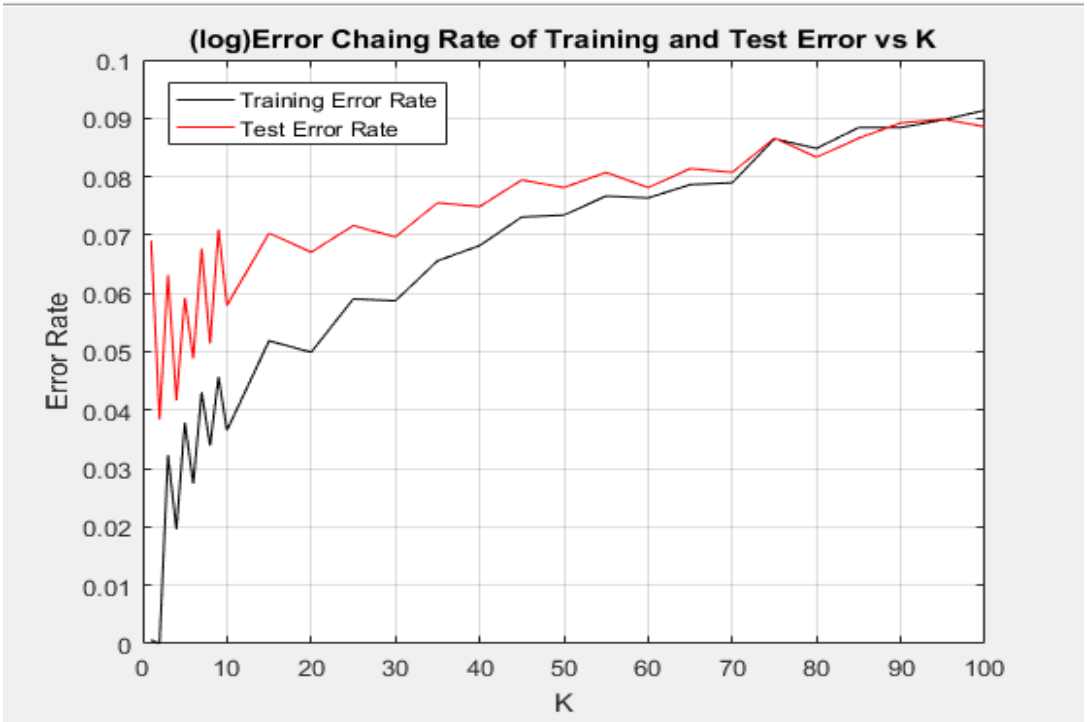
3 Training and testing error rates for λ = 1, 10, and 100.

λ	Training Error Rate	Testing Error Rate
1	0.049266	0.061849
10	0.052202	0.060547
100	0.060359	0.068359



Q4. K-Nearest Neighbours

1 The plot of training and test error rates versus K



2 What do you observe about the training and test errors as K change?

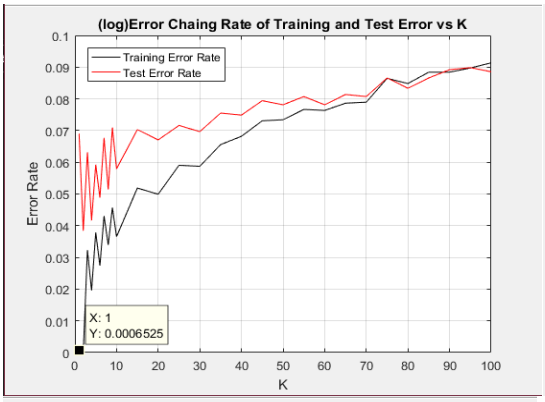
As K increased from 1 to 100, the relationship between error rates and K can be categorized into 2 parts:

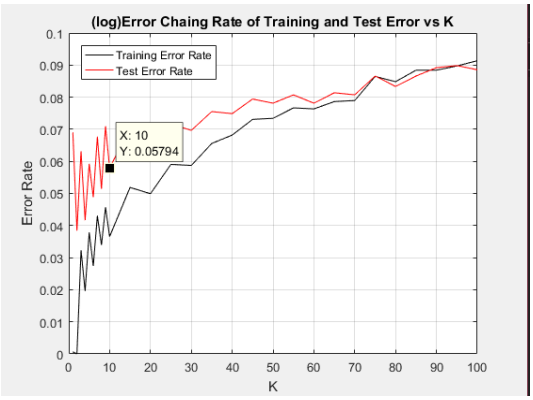
1) When 0 ≤ K ≤ 75, Test Error Rate is always higher than the Training Error Rate, and they are gradually getting closer to each other while K increasing:
   When 0 ≤ K ≤10, both Training & Test error rates increased with huge spikes and fluctuations.

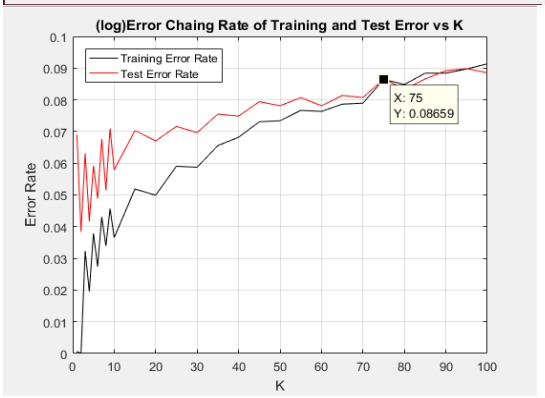
   a. At K=1, the Training Error Rate is very low (1, 0.0006525) but not equal to 0. Because some samples have the same features distance=0 but from different classes.

   b. When K >10, the waveform became smoother and the difference between the error rate of Training and Test error reduced gradually until they cross with each other at K =75.

2) When k > 75, the waveform is smoother and the error rate of Test data is smaller than the Training data.

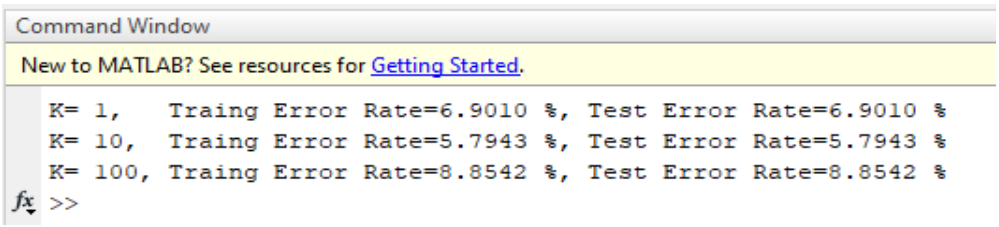






3 Training and testing error rates for K = 1, 10, and 100.

K	Training Error Rate	Testing Error Rate
1	0.000653	0.069010
10	0.036542	0.057943
100	0.091354	0.088542



Q5. Survey

Time Spent

I am a part-time student, and I have spent 4 working-days on this assignment, spent *16 hours* (around 4 hours per day) in total. Before doing the assignment, I also spent a few days revising the notes since I was not familiar with the formula and terms.

Feedback

This is my 1st semester in NUS, and I have only taken 2 modules, EE5907 & EE5902, where I can tell the difference between the teaching mode between 2 professors with no doubt. Frankly speaking, Prof. Thomas gave a clear explanation of the topics he covered, and he always encourages his students to ask questions and gave feedback during his class. I enjoyed his way of teaching and believe it could lead me to a better understanding of pattern recognition.