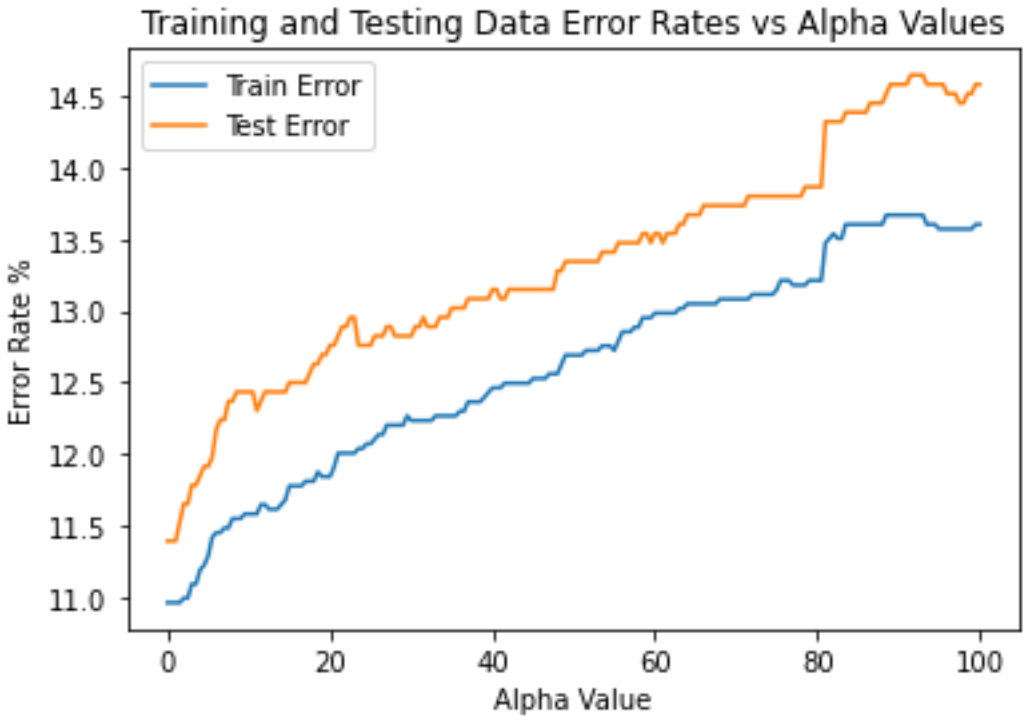
Project Report

Please refer to the source code within the Jupyter Notebook for more information.
All error rates are reported to 3 decimal places.

Q1. Beta-binomial Naïve Bayes

• Plots of training and test error rates versus α



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

Referring to the plot above, in general, the errors for both train and test increase as alpha increases. Specifically, both error rates increase steeply from alpha=1 to alpha=7 and face a sharp jump when alpha changes from 81 to 82.

For every value of alpha, the test error rates are always higher than the train error rates. The gap between train and test error is narrower when alpha is smaller than around 8, and widest when alpha is around 93.

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

- The training error for alpha=1 is: 10.962%
- The testing error for alpha=1 is: 11.393%
- The training error for alpha=10 is: 11.582%
- The testing error for alpha=10 is: 12.435%
- The training error for alpha=100 is: 13.605%
- The testing error for alpha=100 is: 14.583%

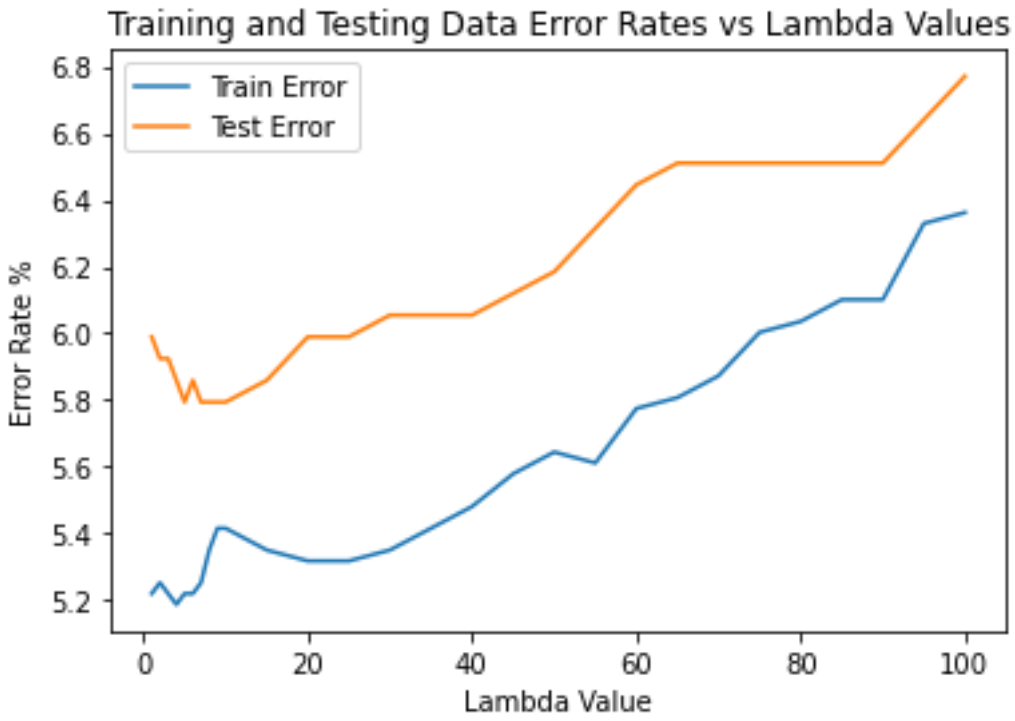
Q2. Gaussian Naive Bayes

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

- Training error: 16.672%
- Testing error: 18.359%

Q3. Logistic regression

• Plots of training and test error rates versus λ



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

Referring to the plot above, in general, the errors for both train and test increase as lambda increases. Specifically, at the beginning when lambda goes from 1 to 7, the train error rates fluctuate around 5.2% while the test error rates fall from around 6% to 5.80% (fall of 0.2%). Subsequently, when lambda increase from 7 to 10, the train error rates increase steeply by around 0.15% while the test error rates fluctuate steadily around 5.8%. As lambda increases from 10 to 20, train error rates fall by around 0.1% while test error rates rise by around 0.2%. Finally, for lambda >=20, both train and test errors rise as lambda increases.

For every value of lambda, the test error rates are always higher than the train error rates. The gap between train and test error is narrower when lambda is around 10, 90, 95, and wider when lambda is around 1, 30, 65.

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

- The training error for lambda=1 is: 5.220%
- The testing error for lambda=1 is: 5.990%
- The training error for lambda=10 is: 5.416%
- The testing error for lambda=10 is: 5.794%
- The training error for lambda=100 is: 6.362%
- The testing error for lambda=100 is: 6.771%

Q4. K-Nearest Neighbors

• Plots of training and test error rates versus K

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

Referring to the plot above, in general, the errors for both train and test increase as K increases. Specifically, at the beginning when K goes from 1 to 2, the train error rates increase very sharply from around 0% to 3.5% while the test error rates also increase fairly from about 6.9% to 8.4%. Subsequently, when K increases from 2 to 5, the train error rates increase steadily by around 0.2% while the test error rates decrease sharply by around 2.4%. As K increases from 5 to 10, train error rates continue its steady increase by around 0.7% while test error rates begin to rise by about 2.4%. As K increases from 10 to 20, train error rates continue its increase by around 0.5% while test error rates drop slightly by around 0.35%. Finally, for K >=20, both train and test error rates rise as K increases.

For K<=75, the test error rates are always higher than the train error rates. The gap between train and test error is widest when K is around 1 and the gap reduces as K increases. At K=60, the train and test error rates are very similar. When K>75, the

train error rates become slightly higher or slightly below but still very close to test error rates.

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

- The training error for K=1 is: 0.065%
- The testing error for K=1 is: 6.901%
- The training error for K=10 is: 5.024%
- The testing error for K=10 is: 7.357%
- The training error for K=100 is: 9.168%
- The testing error for K=100 is: 9.180%

Q5. Survey

I took around 30 hours to do this project.

I thought maybe it would be nice if we could be asked to try cross validation in Q4 to set K, like how we learnt in lectures.