EE5907: Programming Assignment CA1

By: Zhou Yidi(A0149755N)

Question 1: Beta-binomial Naïve Bayes

Objective:

-Fit a beta-binomial naïve Bayes classifier on the dataset.

-Dataset has to be binarized before hand.

-assume a prior Beta(α, α) on the feature distribution

-use posterior predictive for training and testing

-compute error rate on test data using $\alpha = \{0, 0.5, 1, 1.5, 2, \cdots, 100\}$

Simulation Result:

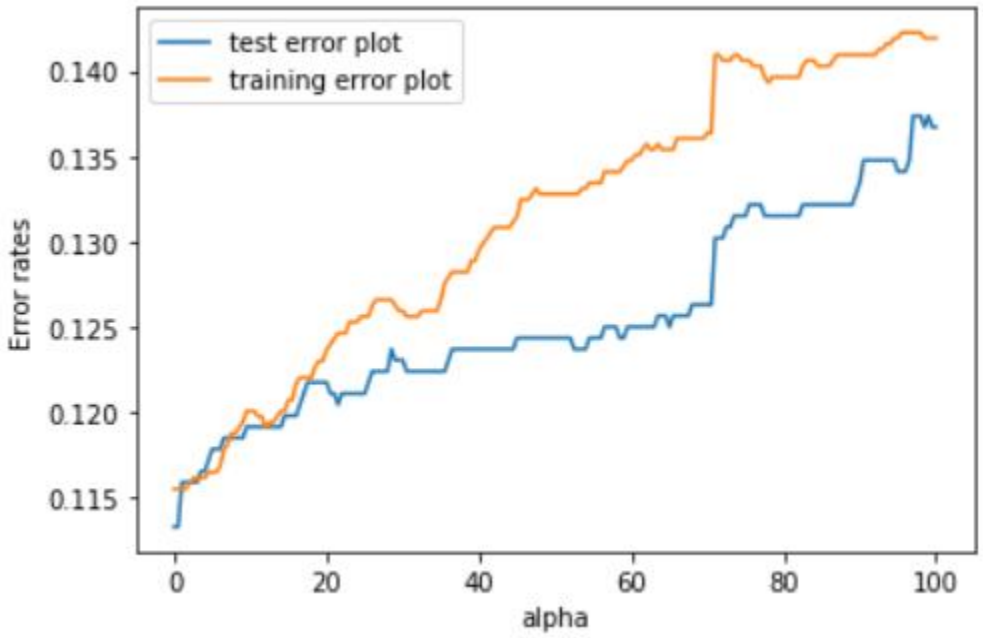

Figure 1: Plots of training and test error rates versus α

As α increases, the training and test error rates both generally show an increasing trend.

	Training Error Rate	Test Error Rate
α = 1	0.11549755301794451	0.11588541666666663
α = 10	0.1200652528548124	0.119140625
α = 100	0.1419249592169658	0.13671875

Question 2: Gaussian Naïve Bayes

Objective:

-Fit a Gaussian naïve Bayes classifier on the dataset.

-Dataset has to be log-transformed before hand.

-use maximum likelihood to estimate class conditional mean and variance of each feature and use ML estimates as plug-in estimator for testing

-compute error rate on test data

Simulation Result:

The test error rate is 0.16015625.

The training error rate is 0.16574225122349107.

Question 3: Logistic Regression

Objective:

-fit a logistic regression model with l2 regularization

-Dataset has to be log-transformed before hand.

-include bias term and l2 regularization should not not apply to the bias term

-compute error rate on test data using $\lambda = \{1, 2, \cdots, 9, 10, 15, 20, \cdots, 95, 100\}$

Simulation Result:

Figure 2: Plots of training and test error rates versus λ

As λ increases, the training and test error rates both generally show an increasing trend. However for small values of λ, when λ increases, the training and test error rates both decrease. This could be due to the model being well regularized for small values of λ, so generalization is better. Beyond the small values of λ, weights are penalized too much and move them away from the optimum, leading to worse performance.

	Training Error Rate	Test Error Rate
λ = 1	0.04893964110929849	0.05859375
λ = 10	0.04926590538336051	0.060546875
λ = 100	0.06264274061990216	0.068359375

Question 4: K Nearest Neighbour

Objective:

-implement a KNN classifier

-Dataset has to be log-transformed before hand.

-use Euclidean distance to measure distance between neighbors

-compute error rate on test data using K = {1, 2, · · · , 9, 10, 15, 20, · · · , 95, 100}

Simulation Result:

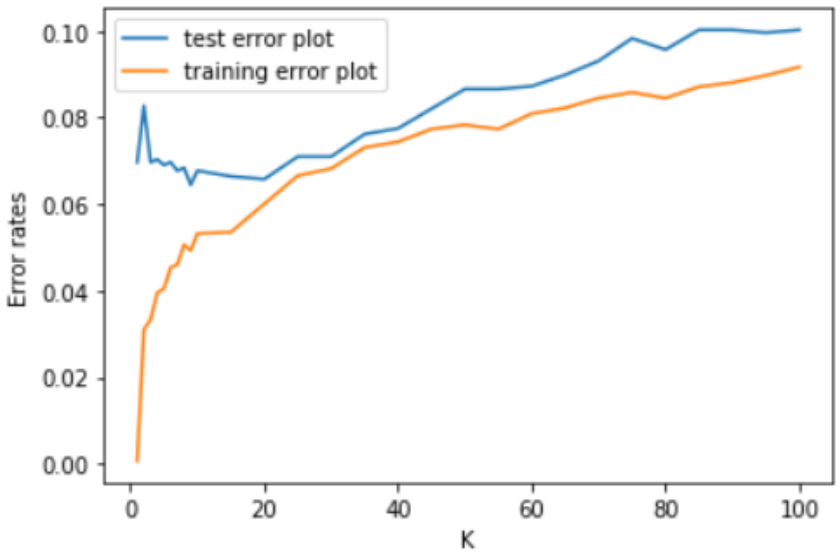

Figure 3: Plots of training and test error rates versus K

As K increases, the training and test error rates both generally show an increasing trend. For small values of k, the curve is jagged. This is due to every alternate odd values of k have lower error rates than their adjacent even values of k. There are higher chances of a tie occurring in even values of k, resulting in predicting the same class(class 0) every time. This will lead to more errors.

In addition for small values of K, we can see that the training error rate and test error rate has a huge gap, training error rate is very much smaller than the test error rate, which shows symptoms of overfitting.

	Training Error Rate	Test Error Rate
K = 1	0.0006525285481240317	0.06966145833333337
K = 10	0.053181076672104366	0.06770833333333337
K = 100	0.09168026101141924	0.10026041666666663

Question 5: Survey

I spend about around 3 weeks and a little bit more on this assignment. This assignment trains me well to be a good foundational data scientist, it opens me up to the world of data science. I am very happy to be able to code out the 4 questions, and have a feel of the numpy library. I realise that numpy library is very powerful and if used well, it can actually shorten the run time of the code a lot.