Pattern Recognition CA1

An email classification dataset of 57 features and 2 classes, spam and non-spam, will be classified in this project using different algorithms. Three basic preprocessing techniques are applied to the testing and training datasets; a) z-normalization, b) log transformation and c) binarization. An additional plot of sensitivity versus specificity is shown for each classification approach where sensitivity and specificity are defined as follows:

Sensistivity = True Positive/True Positive + False Negative

Specificity = True Negative/True Negative + False Positive

Preprocessing of the data revealed that 29 features from testing dataset belonging to class spam and 34 features from testing dataset belonging to class non-spam had a higher variance as compared to training dataset features. Thus, overall, variance among the features of the testing dataset is higher.

Q1 Beta-Bernoulli Naïve Bayes

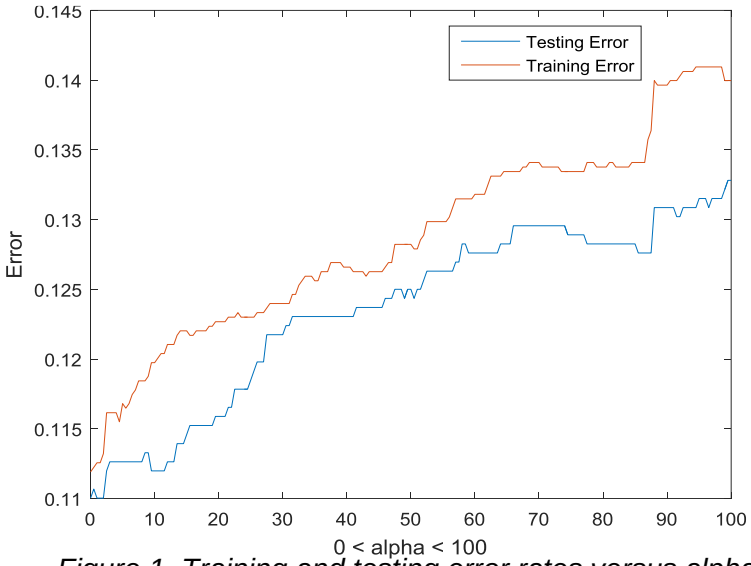

Figure 1. Training and testing error rates versus alpha

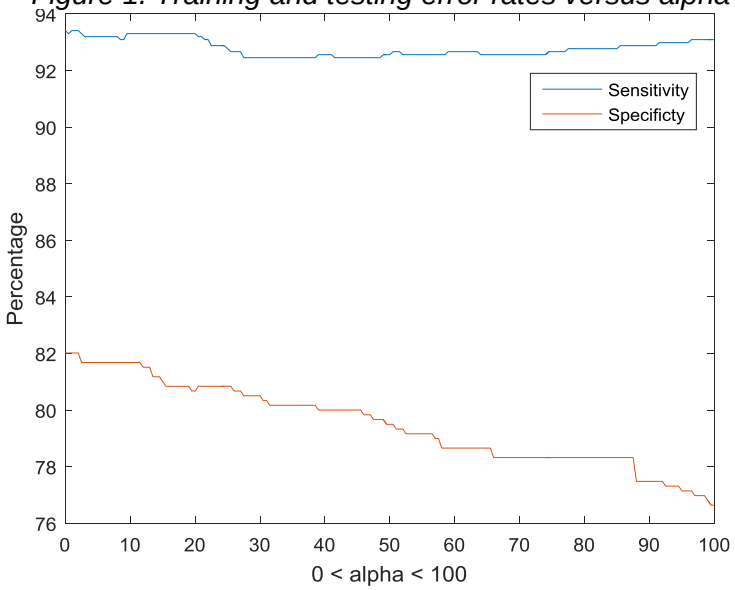

Figure 2. Sensitivity and specificity versus alpha

***Observations*:** A Beta-Bernoulli Naïve Bayes is implemented to classify binarized data. The hyperparameters for our prior distribution Beta, α and β, are assumed to be equal varying from 0 to 100. Through this assumption, for $0 < \alpha < 1$, the prior distribution has a concave shape with higher variance and lower density function, which changes to uniform distribution when $\alpha = 1$. As α further increases, the shape of this prior becomes a convex shape which is narrower with lower variance and higher density function. Thus, as α increases, the posterior distribution depends more on the hyperparameters of the prior rather than the observed counts. This overfitting results in higher misclassifications, due to which error rate is observed to increase as α increases. This can also be seen from the sensitivity vs specificity graph which shows that as α increases, Posterior P(y==1|x,D) overfits classifying ~93% of spam correctly while decreasing the number of correct classification to ~76.64% for non-spam data. Another observation is that training error is higher than testing error for this regression model. This may be because N (total number of samples) is higher for training dataset as compared testing which leans towards maximum likelihood estimation, leading to higher misclassification.

Table 1. Training, testing error rates, sensitivity and specificty for alpha = 1,10,100

α		1	10	100
Binarized Data	Testing	0.1119	0.1197	0.1328
	Training	0.11	0.112	0.14
	Sensitivity	93.41%	93.3%	93.09%
	Specificity	82.02%	81.68%	76.64%

Q2 Gaussian Naïve Bayes

Table 2. Training, testing error rates, sensitivity and specificty for gaussian distribution

Z-normalized data	Testing	0.1888
	Training	0.1759
	Sensitivity	72.5%
	Specificity	94.8%
Log-transformed data	Testing	0.1810
	Training	0.1635
	Sensitivity	74.6%
	Specificity	93.4%

Observation: In this section a Gaussian distribution is assumed on the z-normalized and log-transformed data, whereas maximum likelihood estimate is used to determine the mean and variance parameters for the Gaussian distribution. It is observed that the testing error is higher than the training error, while taking the logarithm of data yields slightly improved error rates as compared to normalized data. However, these error rates are higher as compared to assuming a Beta-Bernoulli distribution. This reflects that estimating parameters using maximum likelihood over fits the distribution, resulting in higher misclassifications. The sensitivity and specificity values show that only ~73% of spam data is correctly classified where as ~93% of non-spam data is correctly classified. This shows that this distribution represents the non-spam data better than spam data.

Q3 Logistic Regression

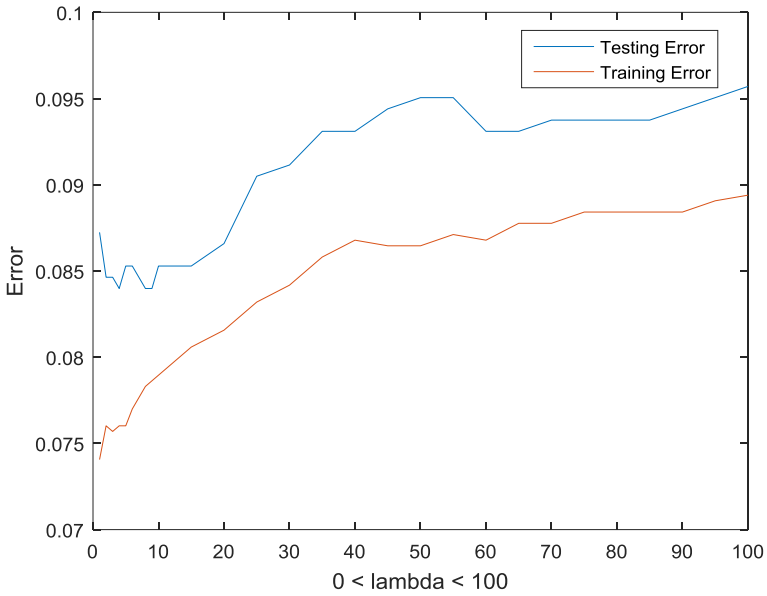

Figure 3. Training and testing error for normalized data versus lambda

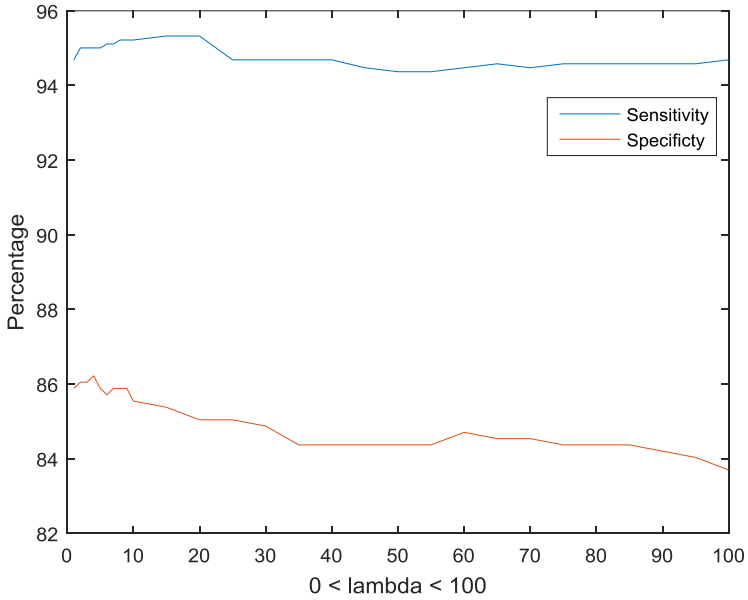

Figure 4. Sensitivity and specificity versus lambda

Observations: Testing error is consistently higher than training error. Among all three pre-processing techniques, the minimum error rate for both training and testing data is achieved when data is log transformed, where the error lies in a range of 0.06 to 0.07. This is verified by the sensistivity and specificity plots which indicate ~95% and ~89% correct classification of spam and non-spam data, respectively. Testing and training error is higher for both z-normalized data and binarized data. This may be because due to these pre-processing procedures, some information about the features might be lost.

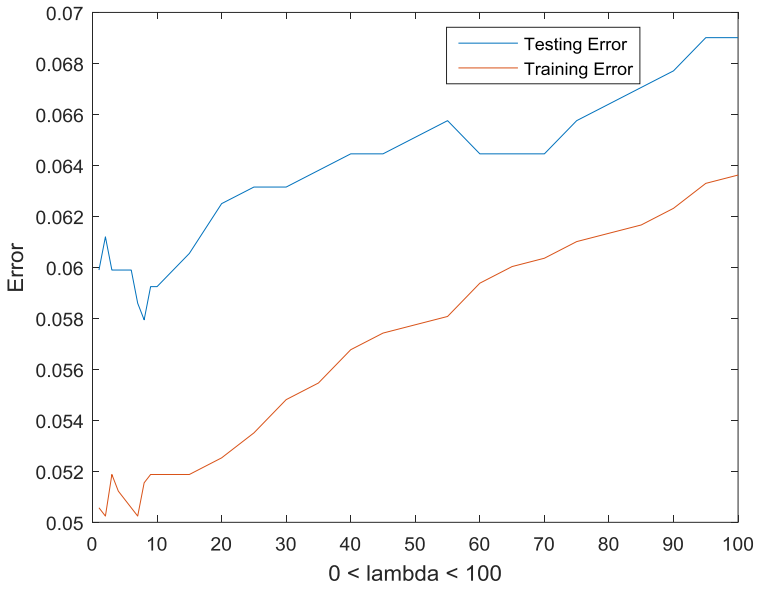

Figure 5. Training and testing error for log-transformed data versus lambda

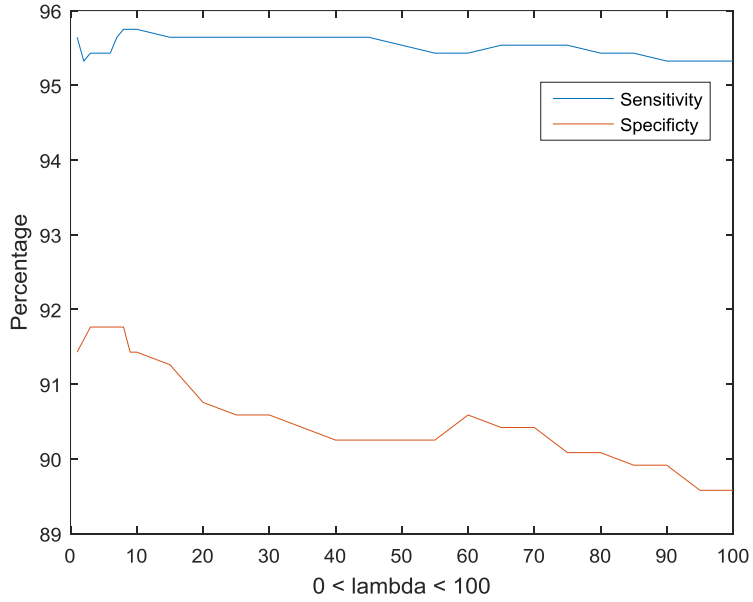

Figure 6. Sensitivity and specificity versus lambda

For binarized data, the sensitivity and specificity plot indicates that as the regularization term lambda increases, mis-classification of non-spam email increases. This may reflect that lambda has simplified the decision boundary to such an extent that misclassifications of non-spam data has increased. For normalized data however, spam is consistently classified correctly about 95% while non-spam has a correct classification rate of ~84%. The overall trend of training and testing error rates shows an increase as the regularization term lambda grows. This can be due to the simplification of decision boundary with an increase in lambda. This also suggests that spam and non-spam data are not completely seperable by simple decision boundaries.

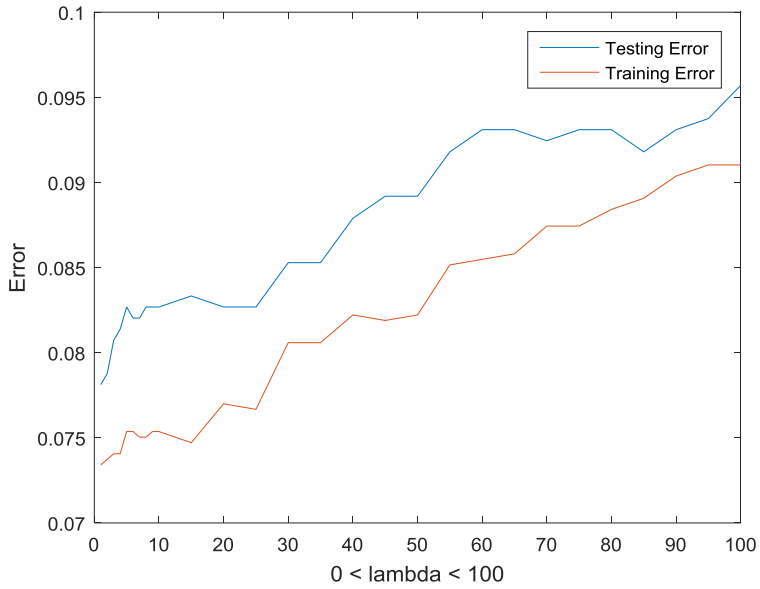

Figure 7. Training and testing error for binarized data versus lambda

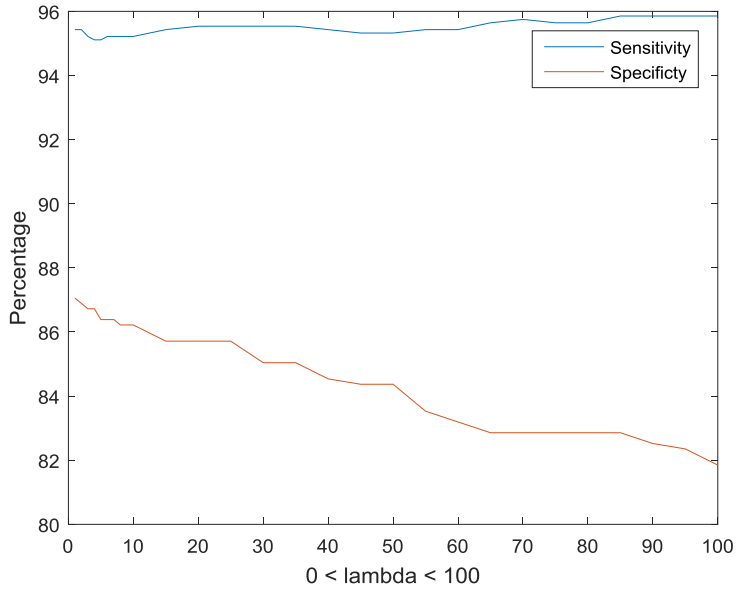

Figure 8. Sensitivity and specificity versus lambda

Table 3. Training, testing error rates, sensitivity and specificty for lambda = 1,10,100

λ		1	10	100
z-normalized Data	Testing	0.08724	0.08529	0.0957
	Training	0.07406	0.07896	0.0894
	Sensitivity	94.69%	95.22%	94.69%
	Specificity	85.88%	85.55%	83.7%
Log transformed Data	Testing	0.0599	0.05924	0.06901
	Training	0.05057	0.05188	0.06362
	Sensitivity	95.64%	95.75%	95.32%
	Specificity	91.43%	91.43%	89.58%

Binarized Data	Testing	0.07813	0.08268	0.0957
	Training	0.07341	0.07537	0.09103
	Sensitivity	95.43%	95.22%	95.86%
	Specificity	87.06%	86.22%	81.85%

Q4 K-Nearest Neighbor

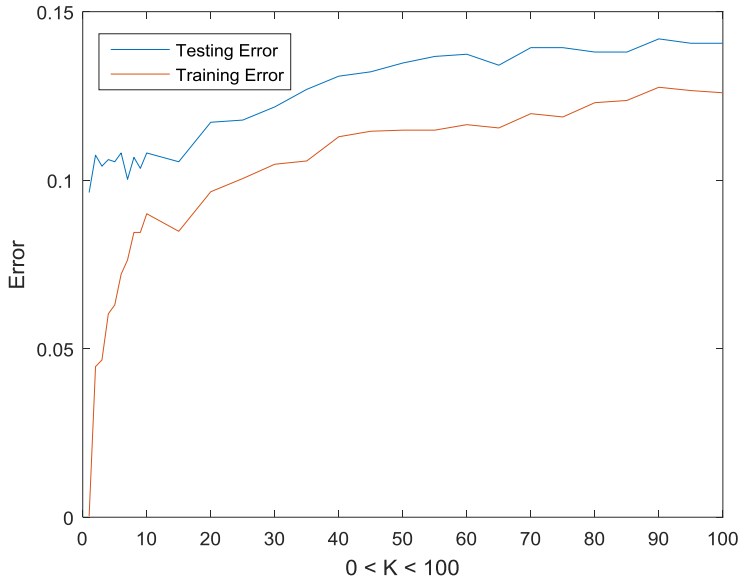

Figure 9. Training and testing error for normalized data versus K

Observations: For the z-normalized and log transformed data, the error is seen to grow with an increasing value of K. For binarized data however, this error is observed to fluctuate initially and then remain relatively consistent as K increases. This shows that the loss of information in features for binarized data acts to our advantage such that simple decision boundaries (increasing K) can correctly classify a higher number of data samples as compared to complex decision boundaries (smaller value of K). On the other hand, vice versa is true for z-normalized and log-transformed data where complex decision boundaries (lower K) result in better classification. The sensitivity and specificity graphs show that ~95% spam data is correctly classified for z-normalized and log transformed data, while the correct classification for non-spam continues to decrease, with an increase in K, to about ~72% and ~82% respectively. On the other hand, the sensitivity and specificity plots for binarized data shows that initially correct classification for non-spam is higher than spam. As K increases, correct classification for spam increases while it decreases for non-spam email samples. At around K=~85, the two plots intersect each other after which correct classification for spam is higher than non-spam.
This confirms that with lower values of K, decision boundary would be overfitting which favours classification of z-norrmalized and log-transformed data. As K increases and the decision boundary simplifies, the outcome reverses and correct classification increases for binarized data.

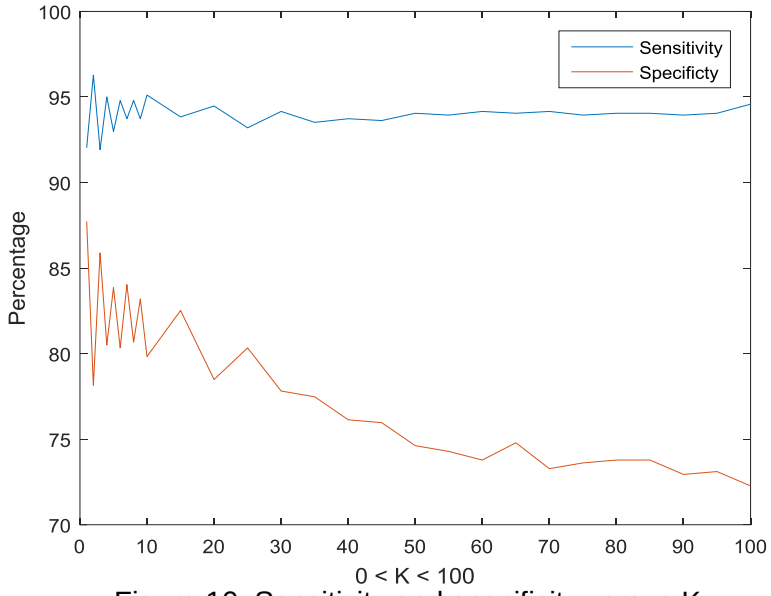

Figure 10. Sensitivity and specificity versus K

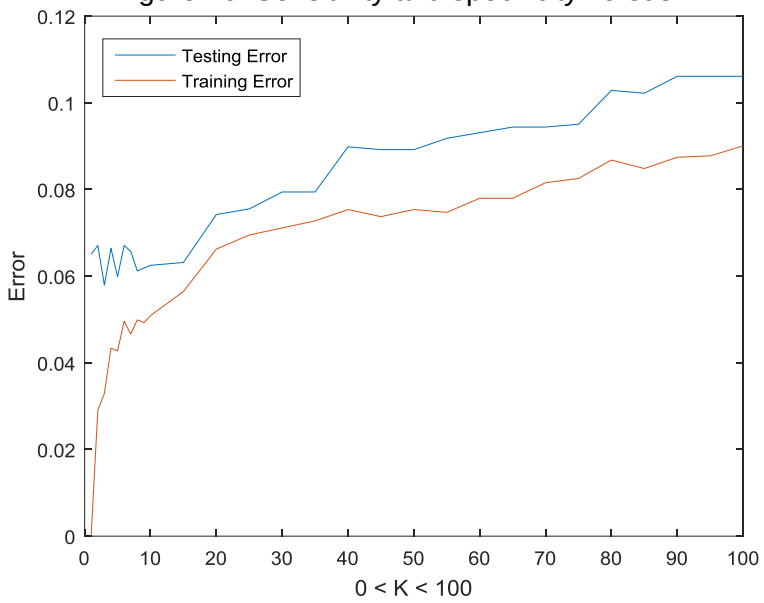

Figure 11. Training and testing error for log transformed data versus K

Another observation shows that at K=1, unlike an ideal case scenario, the training error rate is not zero. This is because there is an instance where the data has same features but exists in both classes, due to which a misclassification error is shown. For binarized data, this error at K=1 is even higher because there are multiple instances which have the same binary features and belong to both classes spam and non-spam.

Final Remarks: The minimum classification error achieved was through l2 regularied logistic regression applied to log transformed data. The error rate was ~0.06 where ~95% of spam and ~90% of non-spam emails were correctly classified. On the other hand, maximum error rate was observed by using Gaussian naïve bayes where the error was ~0.18. Spam was correctly classified ~75% and non-spam was correctly classified ~95%.

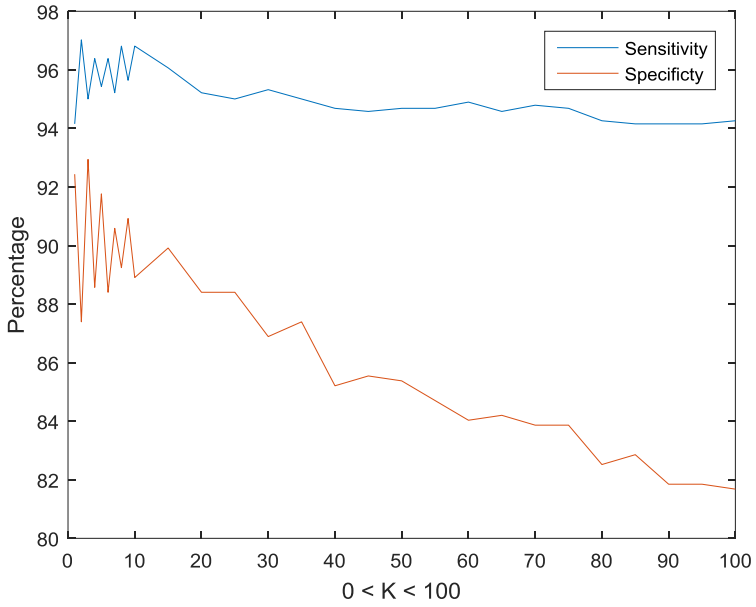

Figure 12. Sensitivity and specificity versus K

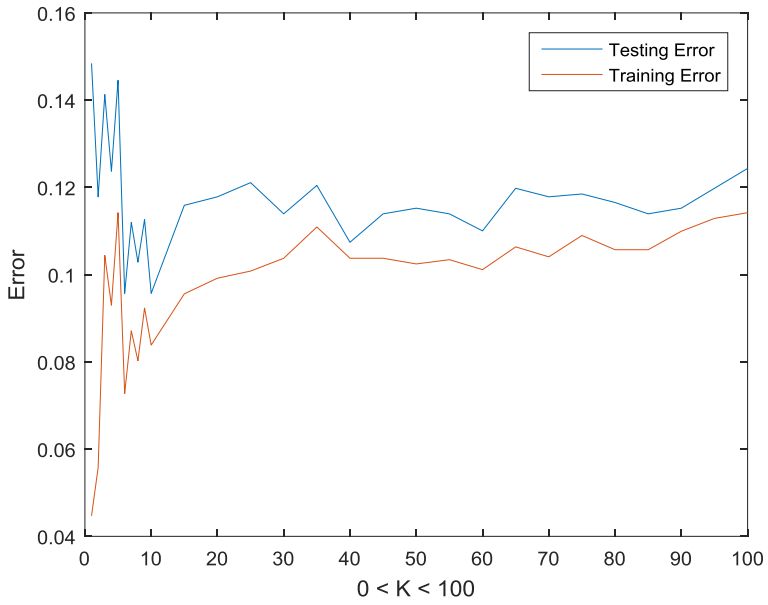

Figure 13. Training and testing error for binarized data versus K

This goes on to prove that assuming maximum likelihood estimates can result in an overfitting distribution. Among the preprocessing techniques, it is observed that log-transformed data yields lower error rates for both logistic regression and KNN algorithm. Whereas z-normalized data and binarized data result in similar error rates. This might suggest the fact that these pre-processing techniques result in a certain level of loss of information from the features which is relatively preserved by taking the log transformation. Finally, majority of the algorithms classify spam email correctly as compared to non-spam emails. This may be because features of non-spam email have higher variance.

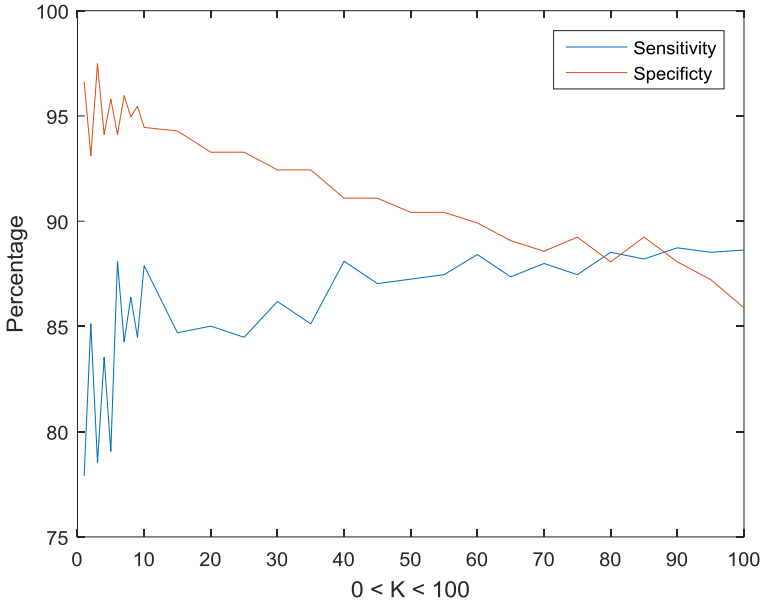

Figure 14. Sensitivity and specificity versus K

Table 4. Training, testing error rates, sensitivity and specificty for lK= 1,10,100

K		1	10	100
z-normalized Data	Testing	0.09635	0.1081	0.1406
	Training	0.0003263	0.09005	0.1259
	Sensitivity	92.03%	95.11%	94.58%
	Specificity	87.73%	79.83%	72.27%
Log transformed Data	Testing	0.0651	0.0625	0.1061
	Training	0.0003263	0.0509	0.09005
	Sensitivity	94.16%	96.81%	94.26%
	Specificity	92.44%	88.91%	81.68%
Binarized Data	Testing	0.1484	0.0957	0.1243
	Training	0.0447	0.08385	0.1142
	Sensitivity	96.64%	94.45%	85.88%
	Specificity	77.9%	87.89%	88.63%

Additional Preprocessing Strategy: Binarization of Normalized Data

Binarization is performed by a simple thresholding technique around zero which might render a feature completely useless if the feature lies above or below zero. Another approach to binarized data is tested where the normalized data is binarized, in effect binarizing each feature with its mean as the thresholding value. Using this approach, results for Beta-Bernoulli Naïve Bayes, Logistic Regression and KNN were re-evaluated. The graphs below show that using this approach, error rate for the Beta-Bernoulli and KNN algorithm is lower than normal binarization. However, the error rate for logistic regression is slightly higher than before. A plausible explanation is that higher regularization term simplifies the decision boundary assuming that data is seperable. However, this pre-processing technique sustains the variance in the data (since data is binarized around its mean) which may result in even higher level of misclassification which was suppressed (to a certain extent) when all data was just binarized about zero.

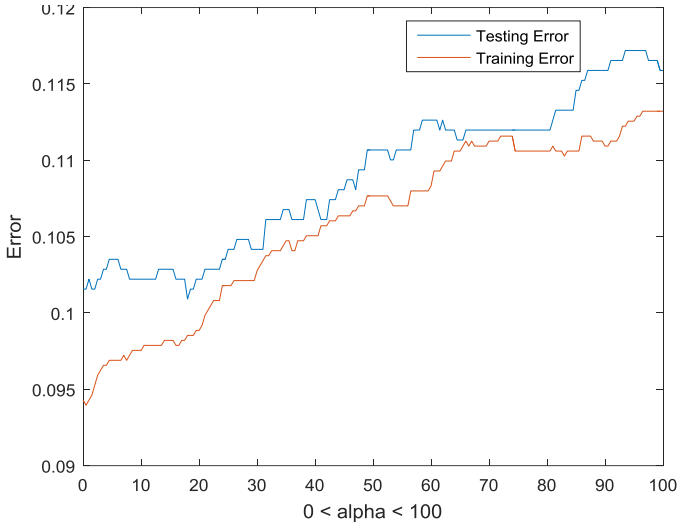

Figure 15. Training and testing error for Beta-Bernoulli distribution

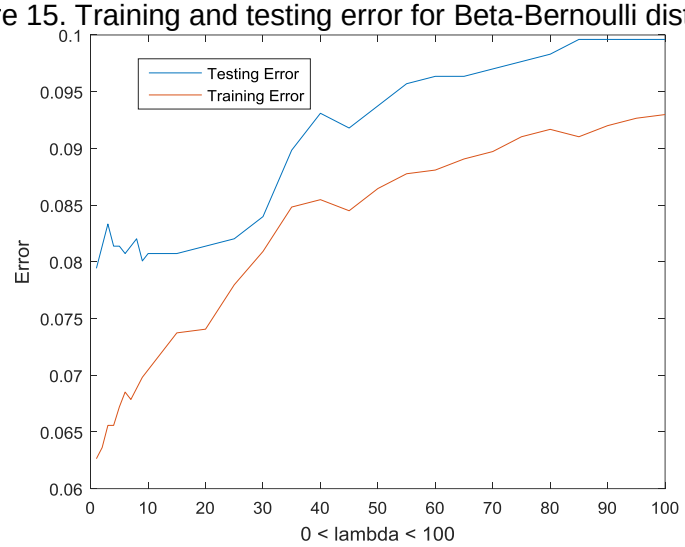

Figure 16. Training and testing error for Logistic Regression

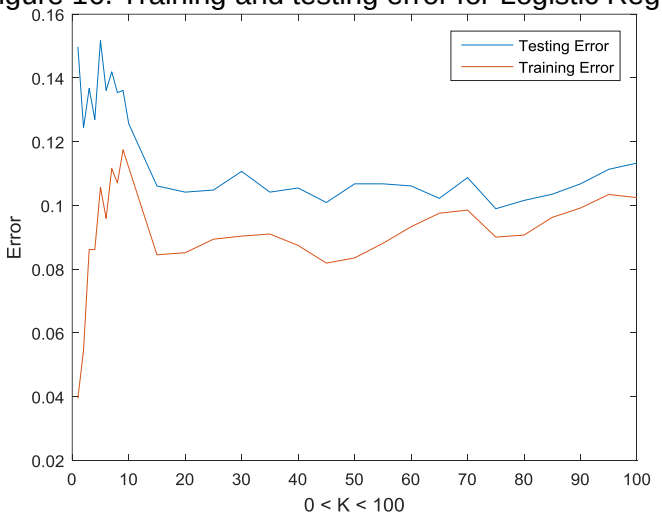

Figure 17. Training and testing error for K-Nearest Neighbor Search

Survey

I worked for almost about 2 days for 2~3 weeks for the assignment.