EE5907 Pattern Recognition

CA1-SPAM EMAIL FILTERING

Report

Student name： **Luo Ke**

Student id： **A0177273X**

Date：2018/03/01

Q1. Beta-bernoulli Naïve Bayes

1. **Plots of training and test error rates versus α.**

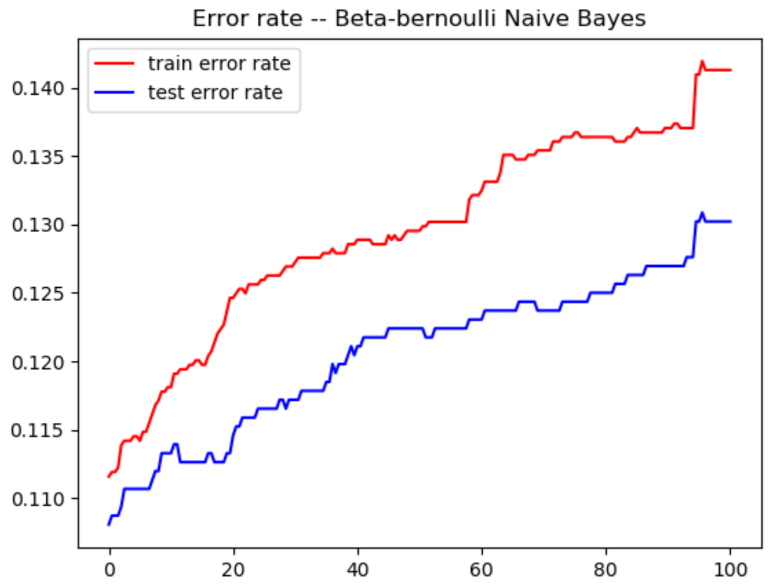

Figure 1
Training error rate and test error rate versus α parameter

2. **What do you observe about the training and test errors as α change?**
From figure 1, we can easily conclude that when α increases, the general trends both for training error and test error increase. But at some points, there are also decrease happen to both train and test error rate. In addition, as α changes, the train error rate is always higher than the test error rate.

3. **Training and testing error rates for α = 1, 10 and 100.**
(NOTE: Round the result to 4 digits)

α	**1**	**10**	**100**
Train error rate	0.1119	0.1181	0.1413
Test error rate	0.1087	0.1133	0.1302

Q2. Gaussian Naive Bayes

1. **Training and testing error rates for both z-normalized and log-transformed data.**

   (NOTE: Round the result to 4 digits)

Data processing	Training error rate	Testing error rate
z-normalization	0.1765	0.2025
log-transform	0.1612	0.1842

Q3. Logistic regression

1. **Plots of training and test error rates versus λ.**
   a. Binarization

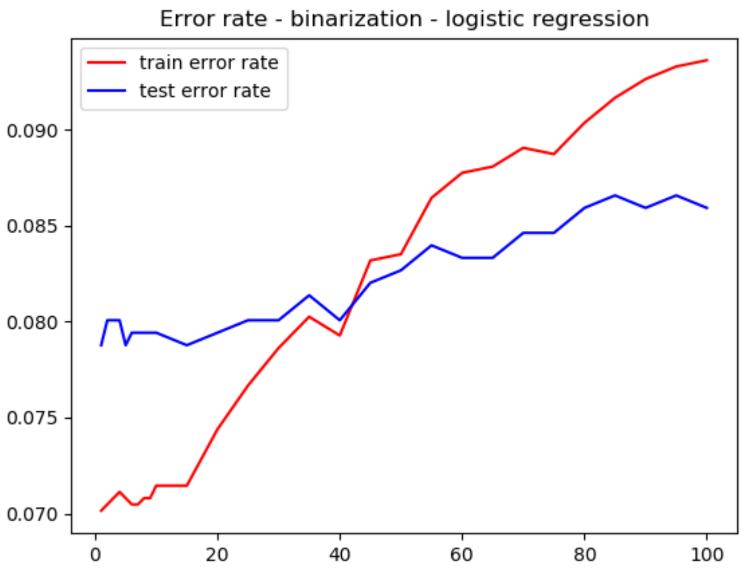

Figure 2
Training error rate and test error rate versus λ when apply binarization feature processing

b. Log-transform

Figure 3

Training error rate and test error rate versus λ when apply log-transform feature processing

c. Z-normalization

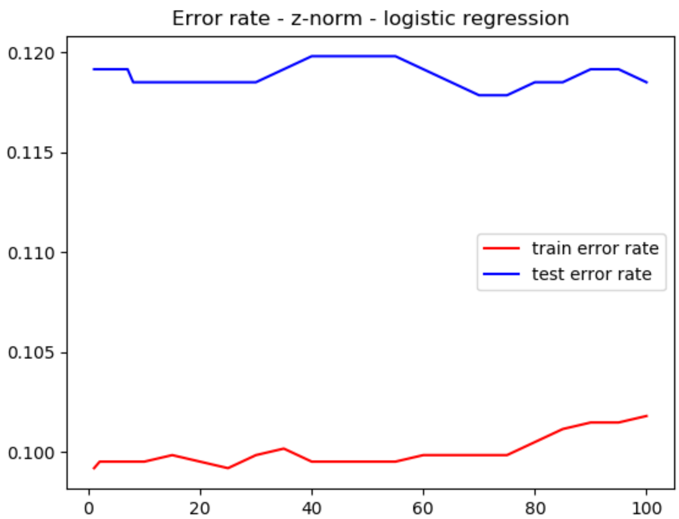

Figure 4

Training error rate and test error rate versus λ when apply z-normalization feature processing

2. **What do you observe about the training and test errors as λ change?**
Generally, as λ increase, both of the training error rate and test error rate increase. For binarization data processing, training error rate is less than test error rate when λ less than 40, and the situation reversed after λ is larger than 40. For log-transform data processing, test error rate is always higher than train error rate. And they both increase as λ increases. And for z-normalization data processing, test error rate is always higher than train error rate. But as train error rate has a softly increase as λ increases, test error rate is swing around 0.118 as λ increases.

3. **What do you observe about the error rates of the different preprocessing strategies?**
From the above three figures, we can see log-transform gives the lowest error rate both for test set and training set. When λ is less than 100, binarization has a lower error rate than z-normalization, but as λ keep increasing, the situation might be reversed. And the error rate of z-normalization is more stable than binarization.

4. **Training and testing error rates for λ = 1, 10 and 100.**
(NOTE: Round the result to 4 digits)

λ	**1**	**10**	**100**
Binarization_train	0.0701	0.0715	0.0936
Binarization_test	0.0788	0.0794	0.0859
z-normalization_train	0.0992	0.0995	0.1018
z-normalization_test	0.1191	0.1185	0.1185
log-transform_train	0.0600	0.0607	0.0633
log-transform_test	0.0671	0.0684	0.0703

Q4. K-Nearest Neighbors

1. Plots of training and test error rates versus K.

a. Binarization

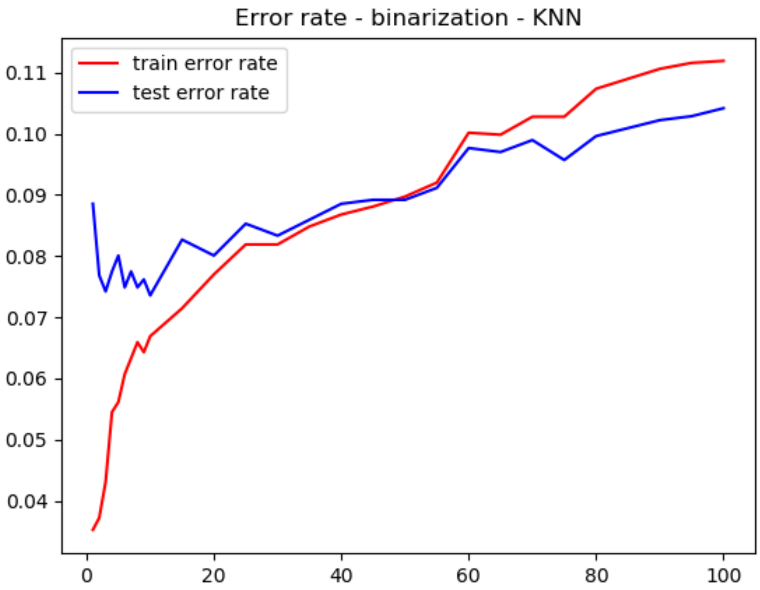

Figure 5

Training and test error rates versus K when apply binarization data processing

b. Log-transform

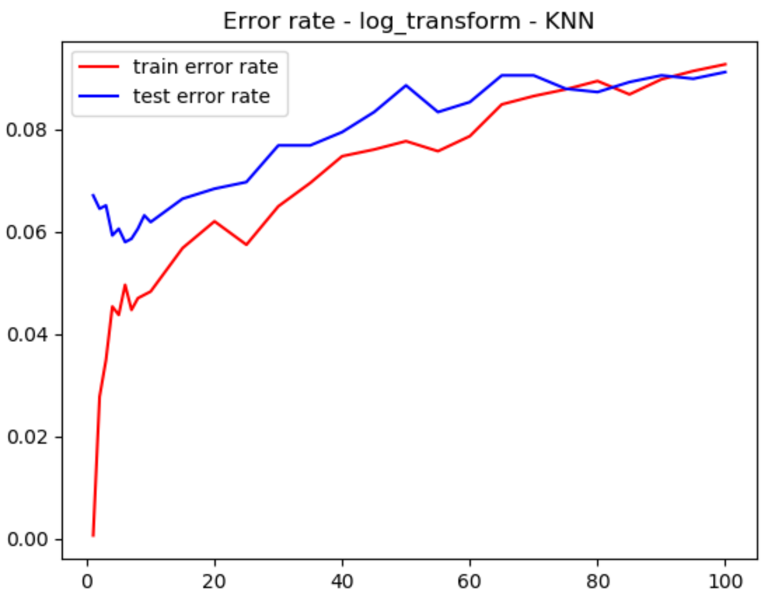

Figure 6

Training and test error rates versus K when apply log-transform data processing

c. Z-normalization

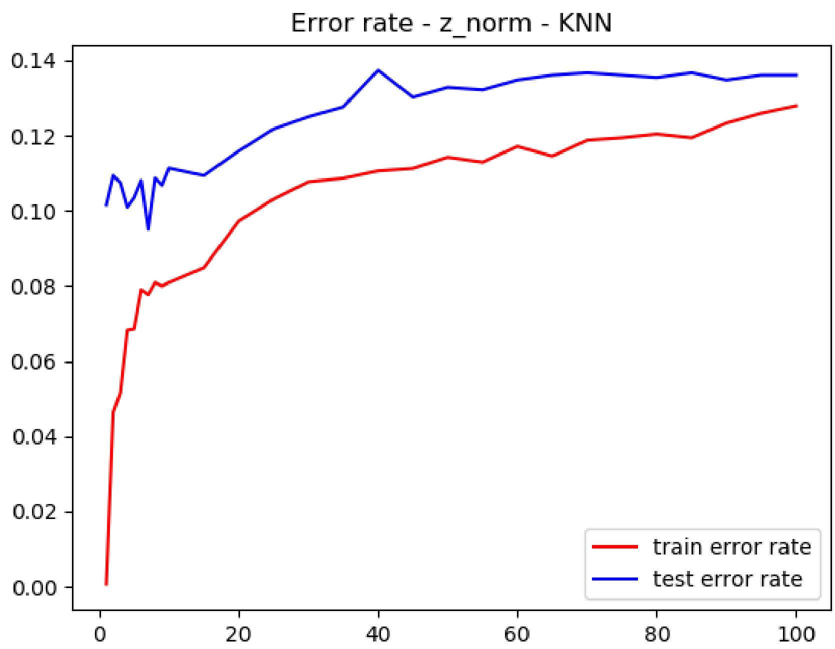

Figure 7
Training and test error rates versus K when apply z-normalization data processing

2. **What do you observe about the training and test errors as K change? Why is training error not 0 when K = 1?**
   a. Generally, as K increases, both training and test error rate increase. It is worth mention that when K =1, the training error rate is lowest, almost to 0. For binarization, the training error rate is less than test error rate when K is less than 50, and the situation versus when K is greater than 50. For log-transform, the training error rate is less than test error rate when K is less than 80. And when K is greater than 80, training and test error rate rise up alternately. For z-normalization, the training error rate is always less than test error rate.
   b. When K=1, only on nearest will be found. In this spam filtering case, the nearest mail to mail A which is labeled as 1 can be A or another mail B labeled as 0. In other words, the nearest mail may not be the mail itself, it can be other mails which is in a different category. When this situation happens, the error rate will not be one.

3. **What do you observe about the error rates of the different preprocessing strategies?**
   Log-transform has the best performance. Binarization has a little bit better performance than z-normalization. But as K goes up, the error rate for binarization increase faster than that of z-normalization.

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

   (NOTE: Round the result to 7 digits)

K	1	10	100
Binarization_train	0.0352365	0.0668842	0.1119086
Binarization_test	0.0885417	0.0735677	0.1041667
z-normalization_train	0.0006525	0.0809135	0.1278956
z-normalization_test	0.1015625	0.1113281	0.1360677
Log-transform_train	0.0006525	0.0482871	0.0926591
Log-transform_test	0.6705729	0.0618490	0.0911458

Q5. Survey

The total time I spend on this project is about: 50 hours.

Since I used too many for loop instead of matrix to do calculation, the code is time consuming especially in part 4.

And for part 3, when calculate the inverse of matrix h_reg in the training process with python, some bugs may happen.