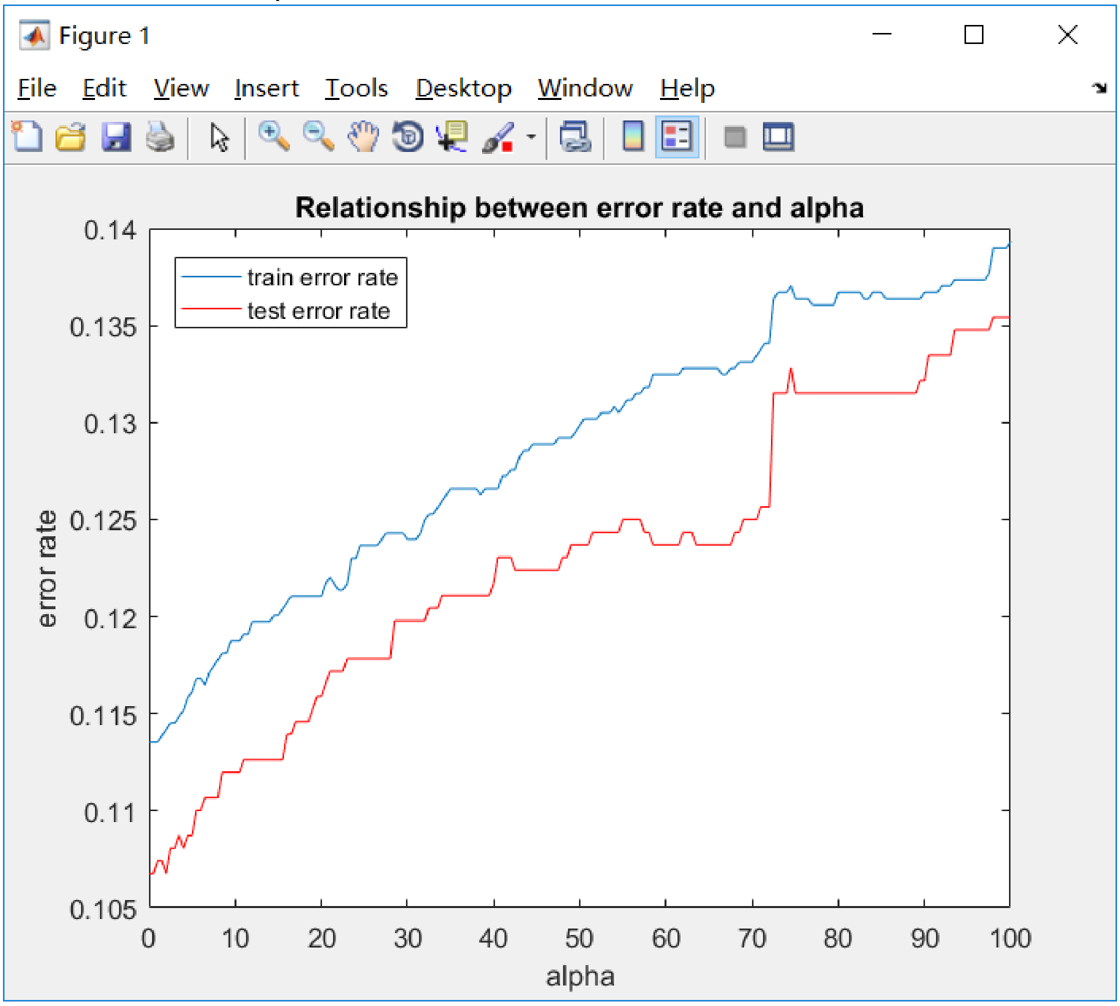
EE5907 Programming Assignment Report

Student name: LIANG AOLING Student Number: A0177210N

Q1. Beta-bernoulli Naive Bayes

(1). The below figure represents the plots of the training error rate and the test error rate versus α parameter.



(2). When α increases, the general trends of the training error rate and the test error rate all increase. But in some time the training error rates are the same when in two very close different α. The test error rates have the same situation. And when α is 70-80, there is a sudden increase in both curves.

(3). The training error rate for α=1 is 0.1135

The training error rate for α=10 is 0.1188

The training error rate for α=100 is 0.1393

The test error rate for α=1 is 0.1074

The test error rate for α=10 is 0.1120

The test error rate for α=100 is 0.1354

Q2. Gaussian Naive Bayes

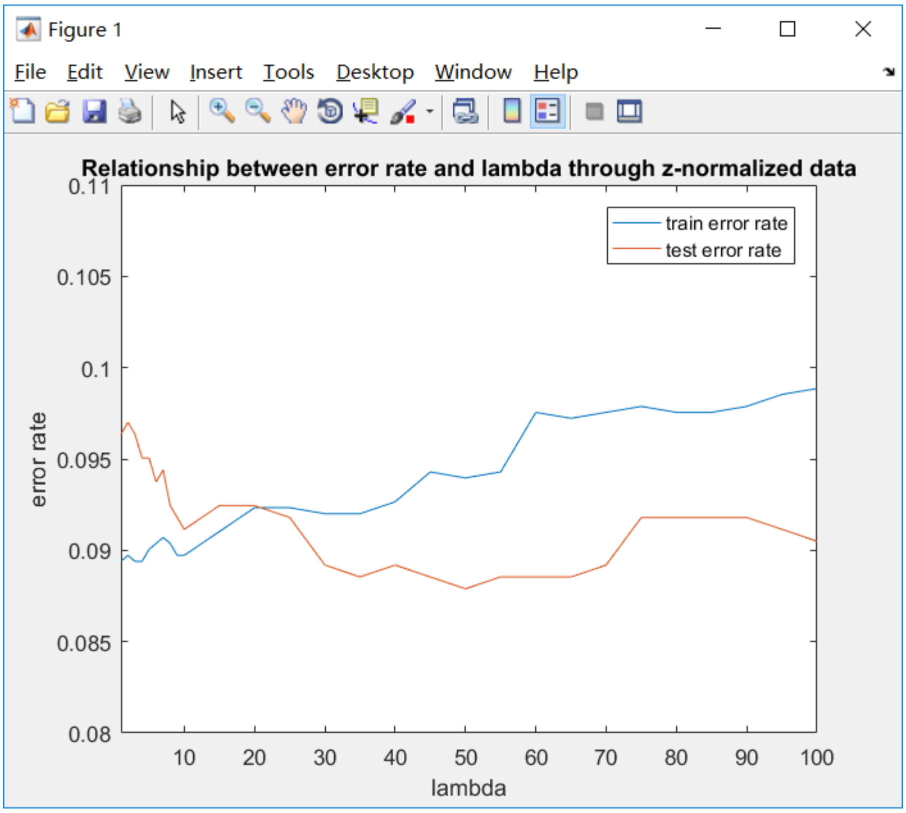
For z-normalized data, the training error rate is 0.1886 and the test error rate is 0.1628.

For log-transformed data, the training error rate is 0.1726 and the test error rate is 0.1595.

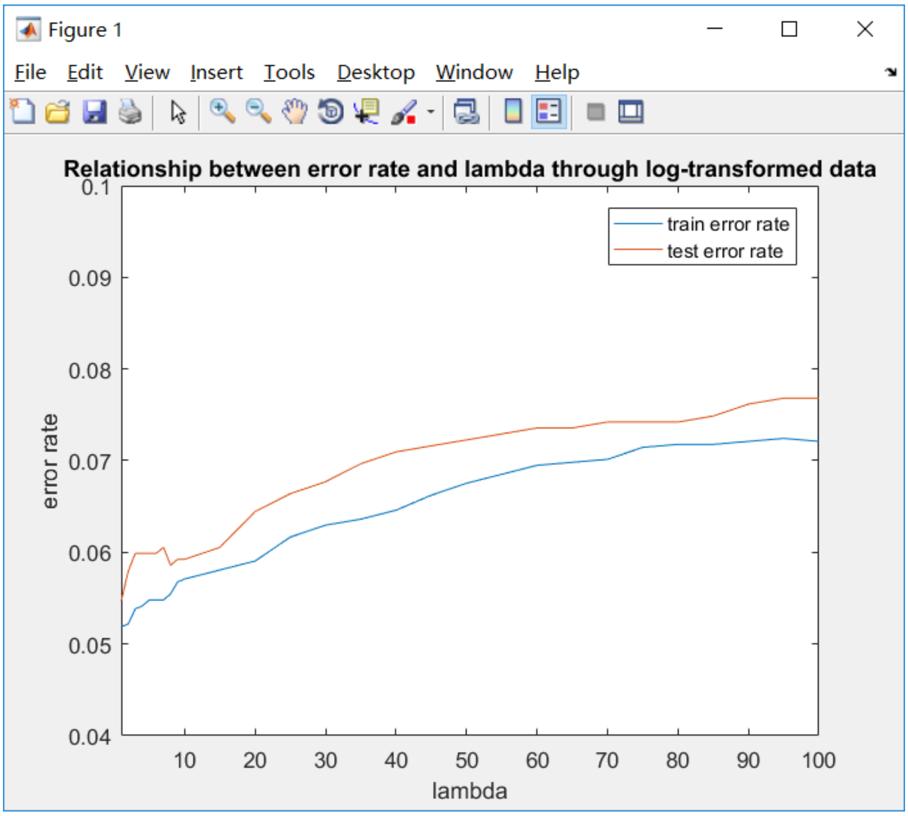
Q3. Logistic regression

(1). The plot of training and test error rates versus is λ as below:

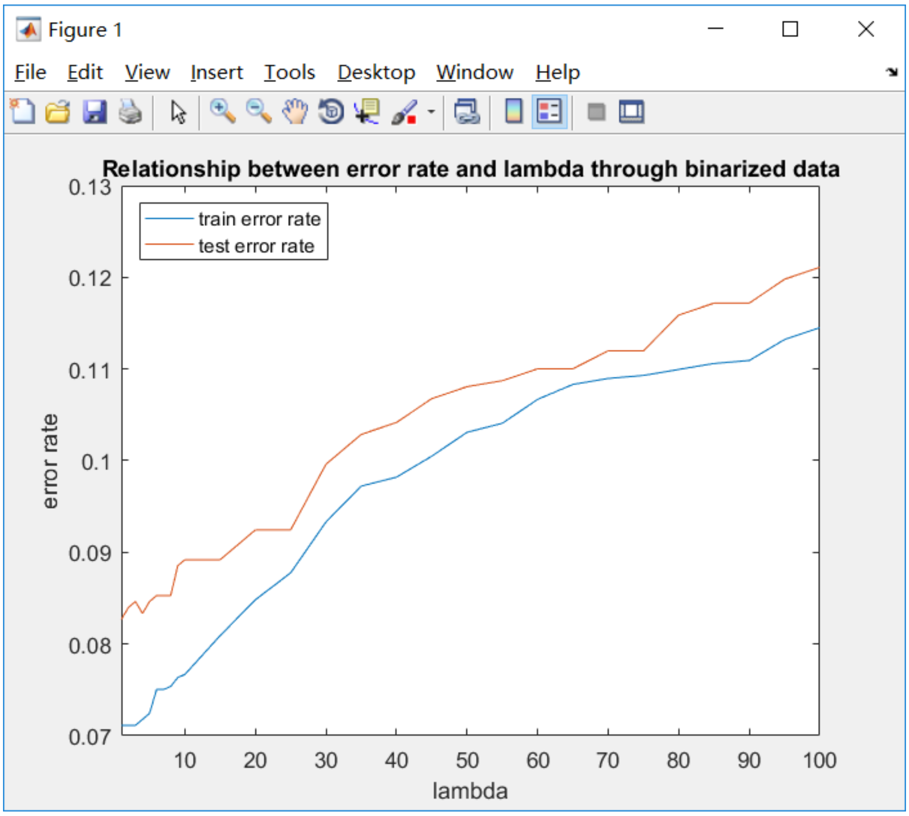
For z-normalized data:



For log-transformed data:



For binarized data:



(2). For log-transformed strategy and binarization strategy, the error rates' trends are similar, where the test error rates are higher than the training error rates and when the lambda increases, the error rates all increase. For z-normalized data, when lambda increases, the training error rate increases while the test error rate decreases. And the error rates for the log-transformed strategy are the lowest so this strategy can be best fit the logistic-regression algorithm. The error rates for the binarization strategy have the biggest change, so this strategy may not very fit for this algorithm. As for the z-normalized strategy, the test error rate has a decreasing trend and two error rates are not very high, so it can also fit for the algorithm.

(3). The training and testing error rates for λ= 1, 10 and 100 are as below:

z-normalized data:

The training error rate for λ=1 is 0.0894
The training error rate for λ=10 is 0.0897
The training error rate for λ=100 is 0.0989
The test error rate for λ=1 is 0.0964
The test error rate for λ=10 is 0.0911
The test error rate for λ=100 is 0.0905

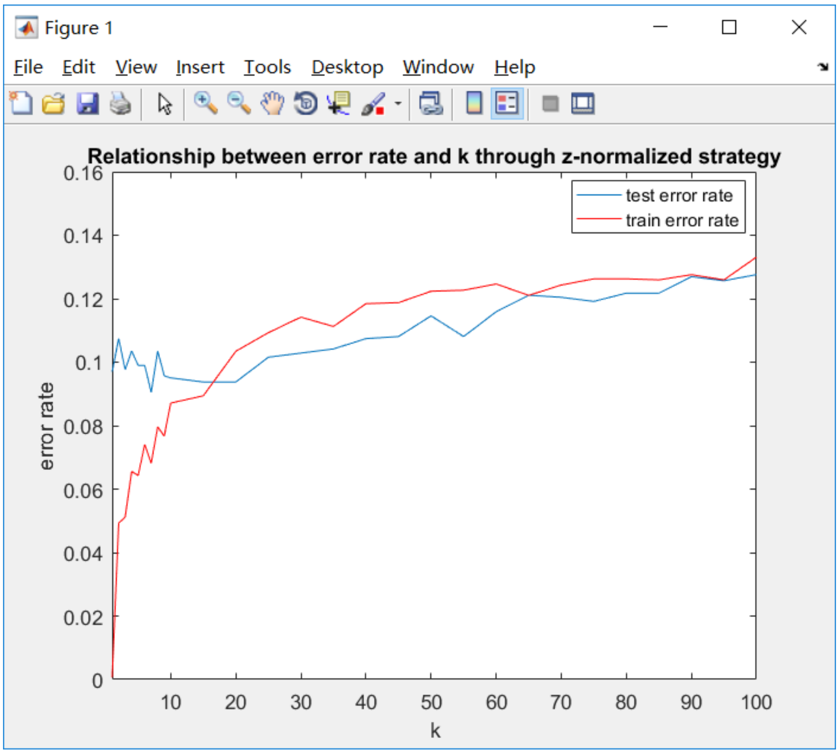
Log-transformed data:

The training error rate for λ=1 is 0.0519
The training error rate for λ=10 is 0.0571
The training error rate for λ=100 is 0.0721
The test error rate for λ=1 is 0.0547

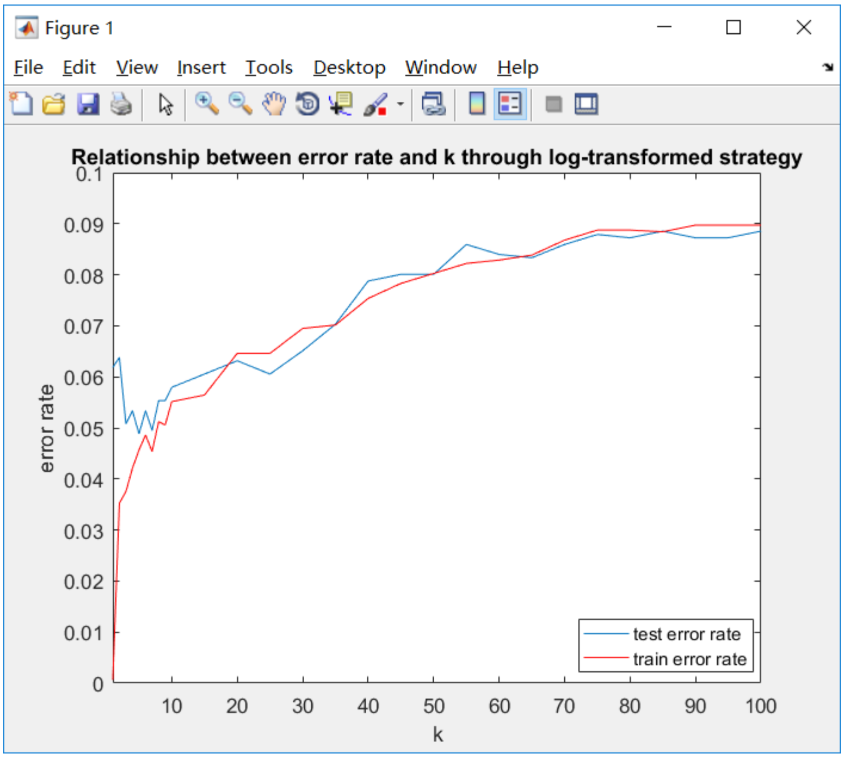
The test error rate for λ=10 is 0.0592
The test error rate for λ=100 is 0.0768
binary data:
The training error rate for λ=1 is 0.0711
The training error rate for λ=10 is 0.0767
The training error rate for λ=100 is 0.01145
The test error rate for λ=1 is 0.0827
The test error rate for λ=10 is 0.0892
The test error rate for λ=100 is 0.1211

Q4. K-Nearest Neighbors

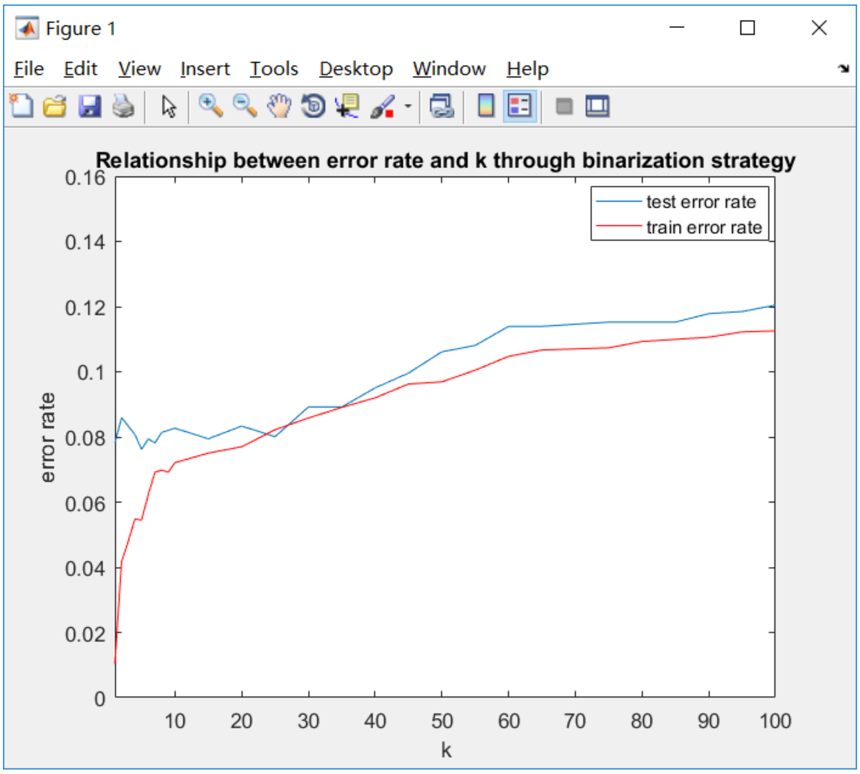
(1). The plot of training and test error rates versus K is as below:
For z-normalized data**:**



For log-transformed data:



For binarized data:



(2).

a. When K increases, the two error rates' trend is increasing.

b. When k = 1, not only the distance from Xtrain(i,j) to Xtrain(i,j) (--the same i and j) is equal to 0, but also the distances from some other points in other rows of Xtrain data to the previous Xtrain(i,j) can be equal to 0. And when I use the sort function to sort the distance and its relative label number, the val number will be also sort from small to large. So the first rows in val matrices are not all equal to i such that the training error is not 0, whose relative actual ytrain data may not be equal to its original ytrain data so that it can be count as error. We can see from the below case tables:

12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
12	13	14	15	16	17	18	19	2	21	22	23	24	25	26	27	2
1158	66	3061	2969	128	1716	1070	2085	20	1298	2956	126	1435	1277	1070	2897	68
1330	880	2930	1055	2316	2649	1521	2250	87	2735	1877	334	2600	1590	1643	197	275

384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400
384	385	386	387	388	389	390	69	392	393	34	395	396	397	398	399	400
1372	2532	1343	186	1460	2669	845	391	1305	2850	394	274	1383	702	143	1463	2645

(3) The log-transform strategy has lowest error rate when k is very large. But the three processing strategies have the similar error rate trends. When k approaches 0, the training error rates all approach to 0. When k increases, the test error rate does not have very big change while the training error increases more. In some k, the training error rate can exceed the test error rate.

(4) The training and testing error rates for K = 1, 10 and 100 are as below:

z-normalized data:

The training error rate for k=1 is 6.5253*e-4
The training error rate for k=10 is 0.0871
The training error rate for k=100 is 0.1331
The test error rate for k=1 is 0.0970
The test error rate for k=10 is 0.0951
The test error rate for k=100 is 0.1276

Log-transformed data:

The training error rate for k=1 is 6.5253*e-4
The training error rate for k=10 is 0.0551
The training error rate for k=100 is 0.0897
The test error rate for k=1 is 0.0618
The test error rate for k=10 is 0.0579
The test error rate for k=100 is 0.0885

binarized data:

The training error rate for k=1 is 0.0104
The training error rate for k=10 is 0.0721
The training error rate for k=100 is 0.1126
The test error rate for k=1 is 0.0781
The test error rate for k=10 is 0.0827
The test error rate for k=100 is 0.1204

Q5. Survey

I spent nearly one week in total on this assignment, which includes three days' understanding for the different algorithms and project topics, three days' programming and optimization, one day's management of report and codes.