EE5907 Pattern Recognition
programming assignment 1

Name	Chen Yuxiang
Matric Number	A0206691A

September 21, 2019

This report is to utilize different classifiers (i.e, Beta-binomial Naive Bayes, Gaussian Naive Bayes, Logic Regression and K-Nearest Neighbours) to classify span e-mails and good emails on the same train data and testing data.

1 Beta-binomial Naive Bayes

1.1 Conditions

data pre-process: binarization
class label prior $\lambda : \lambda_{ML}$
prior Beta$(\alpha, \alpha) : \alpha = \{0.5, 1, 1.5, 2..., 100\}$
use λ_{ML}as a plug-in estimator for testing
for the features, use Bayes(i.e, posterior predictive) training and testing

1.2 Algorithm

$$posterior : p(y|x) \propto p(y)p(x|y, D) \tag{1}$$

$$prior : p(y = 1) = \lambda_{ML} = \frac{N_1}{N}, p(y = 0) = 1 - \lambda_{ML} = 1 - \frac{N_1}{N} \tag{2}$$

$$p(x|y = c, D) = \prod_{j=1}^{d} p(x_j|x_{i\in j,c}, y = c) \tag{3}$$

$$p(x_j|x_{i\in j,c}, y = c) = \frac{N_{j,c} + \alpha}{N_c + 2\alpha} \tag{4}$$

$$p(\tilde{y} = 1|x) \propto \log(\frac{N_1}{N}) + \sum_{j=1}^{d} \log(\frac{N_{j,1} + \alpha}{N_1 + 2\alpha}) \tag{5}$$

$$p(\tilde{y} = 0|x) \propto \log(1 - \frac{N_1}{N}) + \sum_{j=1}^{d} \log(\frac{N_{j,0} + \alpha}{N - N_1 + 2\alpha}) \tag{6}$$

$$if : p(\tilde{y} = 1|x) > p(\tilde{y} = 0|x), \tilde{y} = 1$$

$$else : \tilde{y} = 0 \tag{7}$$

1.3 results

(1)plots of training and test error rates versus α (figure 1).

Figure 1: training and test error rates versus α for Beta-binomial Naive Bayes classifier

(2)what do you observe about the training and test errors as α change?
As α increases, both the training error rate and the test error rate rise accordingly, regardless of some local decrease.
Besides, the training error rate is always better than the test error rate for each α.

(3)training and test error rates for $\alpha = 1, 10, 100$.

α	training error rate	testing error rate
1	0.1141925	0.12369792
10	0.11745514	0.12695312
100	0.13637847	0.14648438

Table 1: training and test error rates for $\alpha = 1, 10, 100$

2 Gaussian Naive Bayes

2.1 Conditoins

data pre-process: log transform
class label prior λ : λ_{ML}
use λ_{ML}as a plug-in estimator for testing
use ML to estimate the class conditional mean and variance of each feature
include bias term in the logistic regression
$l_2$regularization should not apply to the bias term

2.2 Algorithms

$$posterior : p(y|x) \propto p(y|\lambda_{ML})p(x|y, \eta_{ML}) \tag{8}$$

$$prior : p(y=1) = \lambda_{ML} = \frac{N_1}{N}, p(y=0) = 1 - \lambda_{ML} = 1 - \frac{N_1}{N} \tag{9}$$

$$p(x|y=c, \eta_{ML}) = \prod_{j=1}^{d} p(x_j|\eta_{jc}^{ML}, y=c) \tag{10}$$

$$p(x_j|\eta_{jc}^{ML}, y=c) = \frac{1}{\sqrt{2\sigma_{jc}^2}} e^{-\frac{(x_j-\mu_{jc})^2}{2\sigma_{jc}^2}} \tag{11}$$

$$p(\tilde{y}=1|x) \propto \log(\frac{N_1}{N}) + \sum_{j=1}^{d} \log(\frac{1}{\sqrt{2\sigma_{j1}^2}} e^{-\frac{(x_j-\mu_{j1})^2}{2\sigma_{j1}^2}}) \tag{12}$$

$$p(\tilde{y}=0|x) \propto \log(1 - \frac{N_1}{N}) + \sum_{j=1}^{d} \log(\frac{1}{\sqrt{2\sigma_{j0}^2}} e^{-\frac{(x_j-\mu_{j0})^2}{2\sigma_{j0}^2}}) \tag{13}$$

$$if : p(\tilde{y}=1|x) > p(\tilde{y}=0|x), \tilde{y} = 1$$
$$else : \tilde{y} = 0 \tag{14}$$

2.3 Results

Training and test error rates for the log transformed data:

training error rate	testing error rate
0.20424144	0.1953125

Table 2: Training and test error rates for the log transformed data

3 Logistic Regression

3.1 Conditions

data pre-process: log transform

l_2 regularization for logistic regression model

regularization parameter value: $\lambda = \{1, 2, ..., 9, 10, 15, 20, ..., 95, 100\}$

initialize $\vec{\omega} = \vec{0}$

3.2 Algorithm

$$posterior : p(y_i = 1|x_i) = \mu_i = \frac{1}{1 + e^{-w^T x_i}} \quad (15)$$

$$p(y_i = 0|x_i) = 1 - \mu_i = \frac{1}{1 + e^{w^T x_i}} \quad (16)$$

$$\mu = (\mu_1, \mu_2, ..., \mu_N) \quad (17)$$

$$g = X^T(\mu - y) \quad (18)$$

$$g_{reg} = g + \lambda \begin{bmatrix} 0 \\ w_{d\times 1}^T \end{bmatrix} \quad (19)$$

$$H = X^T S X \quad (20)$$

$$S = diag(\mu_i(1 - \mu_i)) \quad (21)$$

$$H_{reg} = H + \lambda \begin{bmatrix} 0 & 0 \\ 0 & I_{d\times d} \end{bmatrix} \quad (22)$$

$$d = -H^{-1} g \quad (23)$$

$$w[n+1] = w[n] + \eta d, \eta = 1 \quad (24)$$

$$if : p(\tilde{y} = 1|x) > p(\tilde{y} = 0|x), \tilde{y} = 1$$

$$else : \tilde{y} = 0 \quad (25)$$

3.3 Results

(1)plots of training and test error rates versus λ (figure 2).

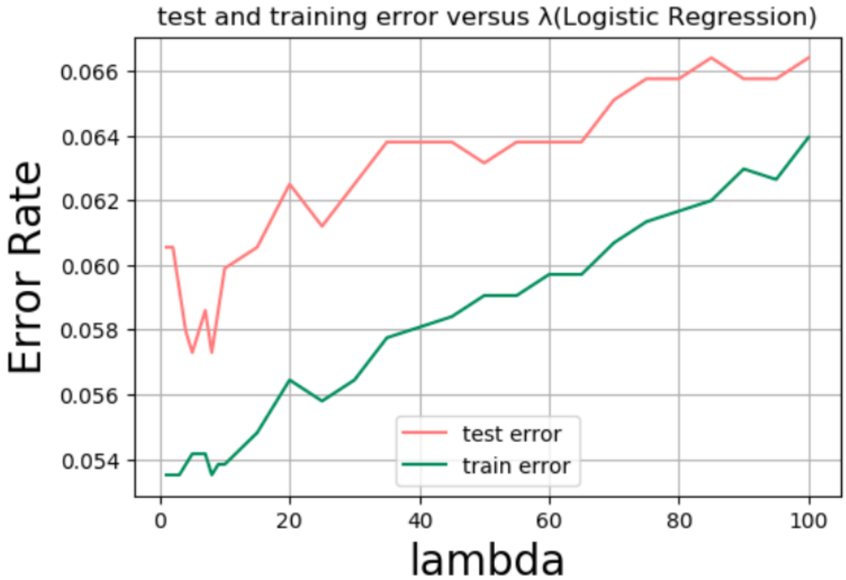

Figure 2: plots of training and test error rates versus λ for logistic regression classfier

(2)what do you observe about the training and test errors as λ change?
Similarly as in Beta-binomial classifier, the overall trend for both error rates is increasing as λ increases, regardless of some local decrease. And train error rate is always better than test error rate, and the test error rate fluctuates more severly.
However, the test error rate has a decrease at the beginning, and for $\lambda = 8$ test error rate reaches the lowest value,which indicates that l_2 regularization does some effort to restrain overfitting. But after that, test error still goes up, which is influenced by excessive regularization.

(3)training and testing error rates for $\lambda = 1, 10, 100$.

λ	training error rate	test error rate
1	0.05350734094616638	0.060546875
10	0.0538336052202284	0.05989583333333337
100	0.06394779771615011	0.06640625

Table 3: training and testing error rates for $\lambda = 1, 10, 100$

4 K-Nearest Neighbors

4.1 Conditions

data pre-process: log transform

distance:Euclidean distance

$K = \{1, 2, ..., 9, 10, 15, 20, ..., 95, 100\}$

4.2 Algorithm

$$dist_i = \{\|x_i - x_j\|_2 \big| j = 1, 2, ..., N\} = \big(\sum_{j=1}^{d} |a_j - b_j|^2\big)^{\frac{1}{2}}, i = 1, 2, ..., N \tag{26}$$

$$index = argsort(dist) \tag{27}$$

$$p(\tilde{y}_i = 1|x) = \frac{k_c}{K} \tag{28}$$

$$k_c = sum(y[index[0, 0 : k]] == 1) \tag{29}$$

$$if : p(\tilde{y} = 1|x) > p(\tilde{y} = 0|x), \tilde{y} = 1$$

$$else : \tilde{y} = 0 \tag{30}$$

4.3 Results

(1)plots od training and test error rates versus K (figure 3).

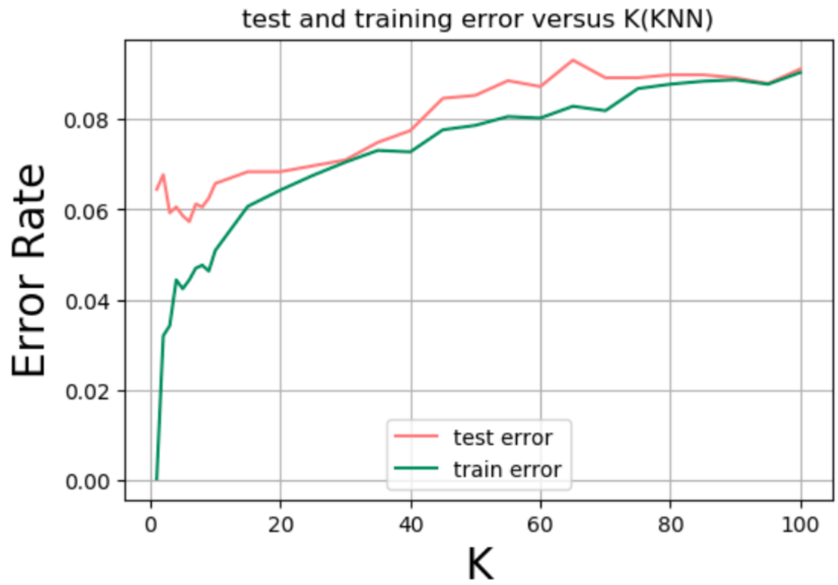

Figure 3: plots od training and test error rates versus K for KNN classifier

(2)what do you observe about the training and test error as K change?
With the rise of K, similarly as the former two plots, both error rates increase. For the train error rate, it is almost zero when k=1 and then increases dramatically to around 0.03. For test error rate, similar with that of logistic regression classifier, there is a decrease at the beginning and then the error rate goes up since $K = 6$ reaches the minimum error rate. At $K = 30$ and after $K > 90$, test error rate is almost equal to training error rate. When K is small, the fluctuation of teat error rate is severe, which indicates that small value of K influences error rate significantly. Besides, the training error rate is always better than test error rate.

(3)training and test error rates for $K = 1, 10, 100$.

K	training error rate	test error rate
1	0.00032626427406201586	0.064453125
10	0.05089722675367048	0.06575520833333337
100	0.09037520391517129	0.09114583333333337

Table 4: training and test error rates for $K = 1, 10, 100$

5 Survey

I spent about 40 hours in this assignment, about 10 hours for each question averagely. And the report took about 5 hours.