EE 5907: Pattern Recognition

Project Report - CA1

Name: Liu Fusheng

ID: A0214203B

Q1: Beta-binomial Naive Bayes (24%)

(a) Plots of training and test error rates versus $\boldsymbol{\alpha}$.

For Beta-binomial Naive Bayes Classifier, we have

$$\begin{aligned}\log p(\tilde{y}=0|\tilde{x},D) &\propto \log p(\tilde{y}=0|\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 0,j},\tilde{y}=0)\\ &= \log \lambda^{ML} + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 0,j},\tilde{y}=0)\end{aligned}$$

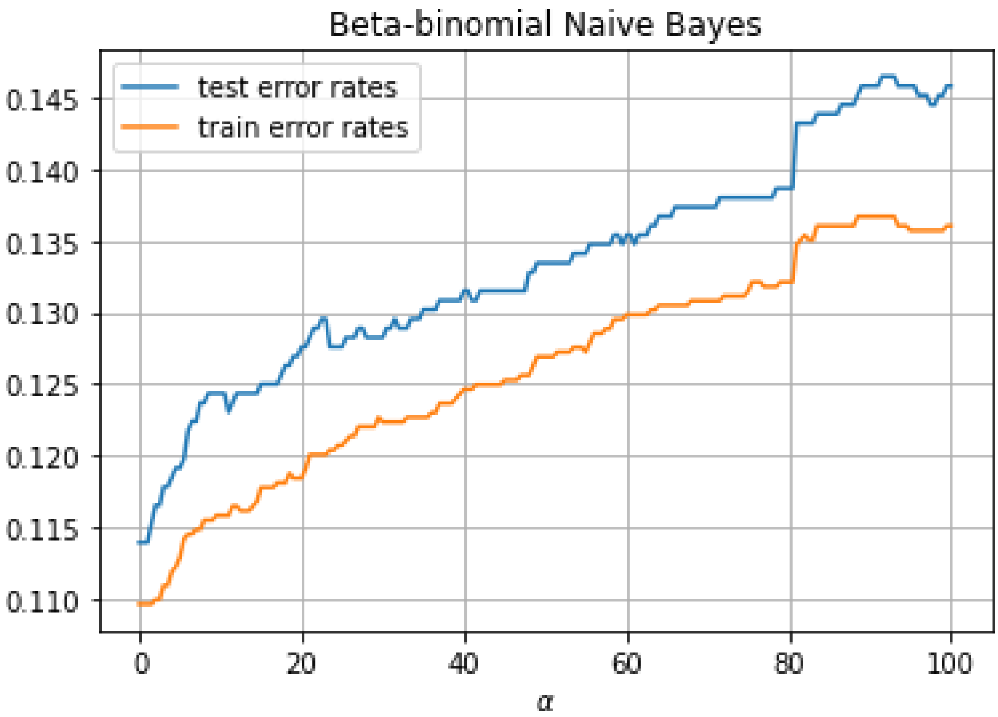
and

$$\begin{aligned}\log p(\tilde{y}=1|\tilde{x},D) &\propto \log p(\tilde{y}=1|\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 1,j},\tilde{y}=0)\\ &= \log(1-\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 1,j},\tilde{y}=1)\end{aligned}$$

with $p(\tilde{x}=1|D) = \frac{N_1+\alpha}{N+2\alpha}$, where $N_1 = \#\left\{\tilde{x}=1\right\}$ and $N = \#|D|$.

So for testing error, we only need to compare $\log p(\tilde{y}=0|\tilde{x},D)$ and $\log p(\tilde{y}=1|\tilde{x},D)$ for $\tilde{x},\tilde{y}$ in the testing set; and for training error, we only need to compare $\log p(\tilde{y}=0|\tilde{x},D)$ and $\log p(\tilde{y}=1|\tilde{x},D)$ for $\tilde{x},\tilde{y}$ in the training set.

The following image shows training and test error rates versus $\alpha = \{0, 0.5, 1, 1.5, 2, \cdots, 100\}$.



(b) What do you observe about the training and test errors as α change?

1. When α increases, both the training error and test error increase overall.

2. For any α, the training error is less than the test error.

3. The generalization error (test error - training error) is bounded and the bound is independent of α.

(c) Training and testing error rates for $\alpha = 1, 10$ and 100.

	Training error rates	Test error rates
$\alpha = 1$	0.10962479608482871	0.11393229166666667
$\alpha = 10$	0.11582381729200653	0.12434895833333333
$\alpha = 100$	0.13605220228384993	0.14583333333333334

Q2: Gaussian Naive Bayes (24%)

(a) Training and testing error rates for the log-transformed data.

For Gaussian Naive Bayes Classifier, we have

$$\begin{aligned}\log p(\tilde{y}=0|\tilde{x},D) &\propto \log p(\tilde{y}=0|\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 0,j},\tilde{y}=0)\\ &= \log \lambda^{ML} + \sum_{j=1}^{57} \log p(\tilde{x}_j|\mu_{j0},\sigma_{j0}^2)\\ &= \log \lambda^{ML} - \sum_{j=1}^{57} \log(2\pi\sigma_{j0}^2) - \frac{(\tilde{x}_j-\mu_{j0})^2}{2\sigma_{j0}^2}\end{aligned}$$

and

$$\begin{aligned}\log p(\tilde{y}=1|\tilde{x},D) &\propto \log p(\tilde{y}=1|\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x_j}|x_{i\in 1,j},\tilde{y}=0)\\ &= \log(1-\lambda^{ML}) + \sum_{j=1}^{57} \log p(\tilde{x}_j|\mu_{j1},{\sigma_{j1}}^2)\\ &= \log(1-\lambda^{ML}) - \sum_{j=1}^{57} \log(2\pi\sigma_{j1}^2) - \frac{(\tilde{x}_j-\mu_{j1})^2}{2\sigma_{j1}^2}\end{aligned}$$

with $\mu_{j0}, {\sigma_{j0}}^2, \mu_{j1}, {\sigma_{j1}}^2$ being the ML mean and ML variance of j-th feature, class $\{0,1\}$ separately. For testing error, we compare $\log p(\tilde{y}=0|\tilde{x},D)$ and $\log p(\tilde{y}=1|\tilde{x},D)$ for $\tilde{x},\tilde{y}$ in the testing set; and for training error, we compare $\log p(\tilde{y}=0|\tilde{x},D)$ and $\log p(\tilde{y}=1|\tilde{x},D)$ for $\tilde{x},\tilde{y}$ in the training set.

Training error rates	Test error rates
0.166721044045677	0.18359375

Q3: Logistic regression (24%)

(a) Plots of training and test error rates versus $\boldsymbol{\lambda}$.

For logistic regression on binary classification, we know

$$p(y=1|x) = \frac{1}{1+e^{-\mathbf{w}^\top x}}, \quad p(y=0|x) = \frac{1}{1+e^{\mathbf{w}^\top x}}$$

where $\mathbf{w} = \text{vec}\{b, w\}$, which is the whole parameter set that contains weight w and bias b. Note that we concatenate 1 to start of x_i, i.e., $\mathbf{w}, x_i \in \mathbb{R}^{58}$. But we do not want to regularize the bias term, so to optimize w, we minimize the negative log likelihoood with ℓ_2 regularization:

$$\begin{aligned} NLL_{reg}(\mathbf{w}) &= -\sum_{i=1}^{N} \log p(y_i|x_i, \mathbf{w}) + \frac{\lambda}{2} \|w\|_2^2 \\ &= -\sum_{i=1}^{N} [y_i \log \mu_i + (1-y_i)\log(1-\mu_i)] + \frac{\lambda}{2} \|w\|_2^2 \end{aligned}$$

The Newton's Method for Logistic Regression is given by

$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k - H_{reg}(\mathbf{w}_k)^{-1} g_{reg}(\mathbf{w}_k) \\ &= \mathbf{w}_k - \left(H(\mathbf{w}_k) + \lambda I\right)^{-1} \left(g(\mathbf{w}_k) + \lambda \mathbf{w}_k\right) \end{aligned}$$

Since we don't regularize the bias term, the iteration is modified as

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \left(H(\mathbf{w}_k) + \lambda \begin{pmatrix} 0 & 0 \\ 0 & I_D \end{pmatrix} \right)^{-1} \left(g(\mathbf{w}_k) + \lambda \begin{pmatrix} 0 \\ w_k \end{pmatrix} \right)$$

The following image shows training and test error rates versus $\lambda = \{1, 2, \cdots, 9, 10, 15, 20, \cdots, 100\}$.

(b) What do you observe about the training and test errors as $\boldsymbol{\lambda}$ change?

1. When λ increases, both the training error and test error fluctuated, but they are increasing asymptotically.
2. For any λ, the training error is less than the test error.
3. The generalization error (test error - training error) is bounded for all λ and the bound is independent to λ.

(c) Training and testing error rates for $\boldsymbol{\lambda = 1, 10}$ and 100.

	Training error rates	Testing error rates
$\lambda = 1$	0.04926590538336052	0.061848958333333336
$\lambda = 10$	0.05187601957585644	0.061197916666666664
$\lambda = 100$	0.06133768352365416	0.06901041666666667

Q4: K-Nearest Neighbors (24%)

(a) Plots of training and test error rates versus K.

For KNN classifier, we need to find K images among training set that are closest to each training (testing) image, and compare the corresponding probability by $\frac{\#\{y=0\}}{K}$ and $\frac{\#\{y=1\}}{K}$.

The following image shows training and test error rates versus $K = \{1, 2, \cdots, 9, 10, 15, 20, \cdots, 100\}$.

(b) What do you observe about the training and test errors as K change?

1. When K increases, both the training error and test error fluctuated, especially when K is small, but they are increasing asymptotically.

2. For small K, the training error is less than the test error, but when K is large enough, this may not hold.

3. The generalization error (test error - training error) is decreasing when K increasing.

(c) Training and testing error rates for $K = 1, 10$ and 100.

	Training error rates	Testing error rates
$K = 1$	0.0006525285481239804	0.06901041666666667
$K = 10$	0.04469820554649266	0.07161458333333333
$K = 100$	0.09559543230016314	0.09244791666666667

Q5: Survey (4%)

(a) Please give an estimate of how much time you spent on this assignment.

I spent around **30 hours** to finish this assignment.