EE5907
Pattern Recognition

CA1 Report

Liew Jia Min

A0161901N

Q1. Beta-binomial Naive Bayes

For Beta-binomial Naive Bayes Classifier, we assume that the features are conditionally independent given a class label and all the features are binary.

The naive bayes classifier can be defined by the class label prior and the feature likelihood and is estimated using the following equations:

$$p(\tilde{y} = 1|\tilde{x}, D) \propto \mathbf{log\,p}(y = 1|\lambda^{ML}) + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 1)$$

$$= log\lambda^{ML} + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 1)$$

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

$$= log(1 - \lambda^{ML}) + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 0)$$

We do not assume any prior on the class label. Hence the class label prior λ can be estimated using maximum likelihood estimation:

$$\lambda^{ML} = \frac{N_1}{N} = \frac{Number\ of\ \ training\ samples\ with\ label\ 1}{Total\ number\ of\ training\ samples}$$

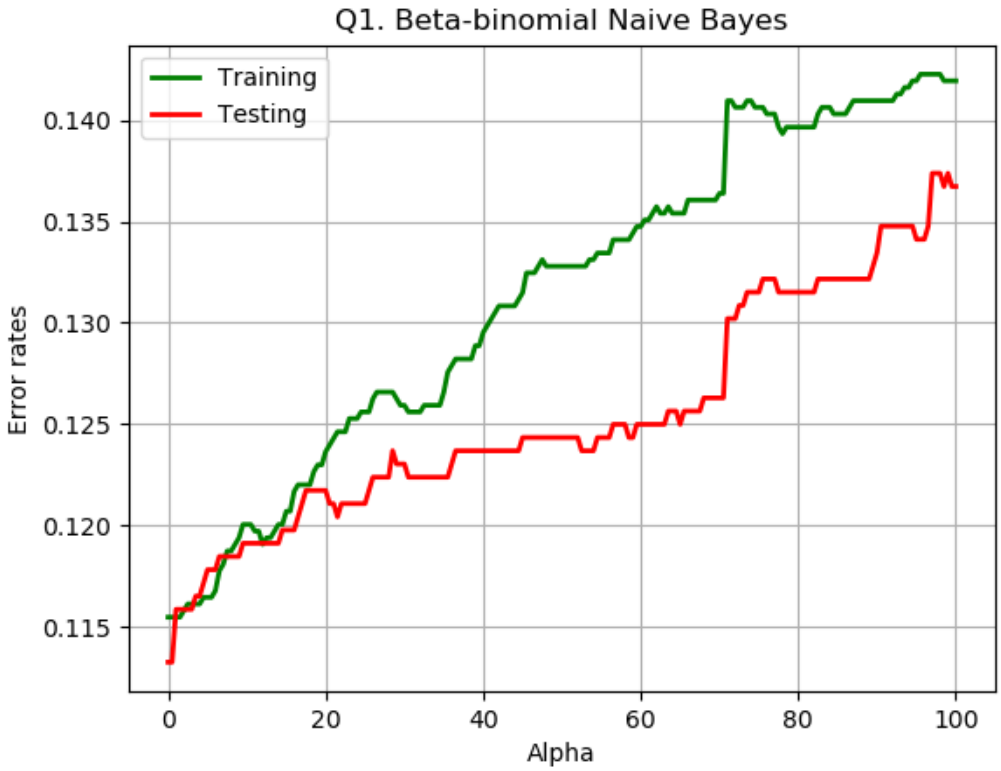
For the features, the posterior predictive distribution can be estimated by:

$$p(\tilde{x} = 1|D) = \frac{N_1 + a}{N + a + b} = \frac{N_1 + \alpha}{N + 2\alpha}$$

where N_1 = Number of data samples from label c where j^{th} feature = 1 and N = Number of data samples of label c. In this example we assume a prior Beta(α, α) on the feature distribution. Hence a = b = α.

To predict the target label of the email data, we fit the classifier on the binarized data and compare $p(\tilde{y} = 1|\tilde{x}, D)$ and $p(\tilde{y} = 0|\tilde{x}, D)$ and pick the label with the higher posterior predictive distribution. This is repeated for each value of α = {0, 0.5, 1, 1.5, 2, · · · , 100}.

a) Plots of training and test error rates vs α



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

- As α increases, the error rate for both training and testing data increases. As α increases, the posterior predictive distribution of the features $p(\tilde{x} = 1|D)$ is less dependent on the training sample leading to larger error rate.
- Generally, the training error rates are lower than the testing error rates as the model is trained using the training data set.

c) Training and testing error rates for α = 1, 10 and 100

α	**1**	**10**	**100**
Training error rates	0.11549755301794454	0.1200652528548124	0.14192495921696574
Testing error rates	0.11588541666666667	0.119140625	0.13671875

Q2. Gaussian Naive Bayes

For the Gaussian Naive Bayes Classifier, we assume that the features are conditionally independent given a class label. We perform log transform on each feature using $log(x_{ij} + 0.1)$.

The naive bayes classifier can be defined by the class label prior and the feature likelihood and is estimated using the following equations:

$$p(\tilde{y} = 1|\tilde{x}, D) \propto \mathbf{log\,p}(y = 1|\lambda^{ML}) + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 1)$$

$$= log\lambda^{ML} + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 1)$$

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

$$= log(1 - \lambda^{ML}) + \sum_{j=1}^{D} log\ p(\tilde{x_j}\,|x_{i\in 1,j}, \tilde{y} = 0)$$

We do not assume any prior on the class label. Hence the class label prior λ can be estimated using maximum likelihood estimation:

$$\lambda^{ML} = \frac{N_1}{N} = \frac{Number\ of\ \ training\ samples\ with\ label\ 1}{Total\ number\ of\ training\ samples}$$

For the features, the posterior predictive distribution can be estimated by estimating the conditional means and variance of each feature and using the maximum likelihood estimation as a plug-in estimator for testing:

$$MLE\ of\ mean\ =\ \hat{\mu}\ =\frac{1}{N}\sum_{n\,=\,1}^{N} X_n$$

$$MLE\ of\ variance\ =\ \hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-[0.5(x-\mu)^2/\sigma^2]}$$

To predict the target label of the email data, we fit the classifier on the log transformed data and compare $p(\tilde{y} = 1|\tilde{x}, D)$ and $p(\tilde{y} = 0|\tilde{x}, D)$ and pick the label with the higher posterior predictive distribution.

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

Training Error Rates	0.16574225122349104
Testing Error Rates	0.16015625

Q3. Logistic Regression

For logistic regression, we estimate the parameters of the model $p(Y|X)$ from the training data and then compute $p(Y|X)$ to classify new samples

$$p(Y = 1|X) = \frac{1}{1 + e^{-\boldsymbol{W}^T X}}$$

$$p(Y = 0|X) = \frac{1}{1 + e^{\boldsymbol{W}^T X}}$$

To compute, $p(Y|X)$, first need to estimate the weight **W.** There is no close form solution to compute **W,** so we perform a numerical optimisation to find the best value $\widehat{\boldsymbol{W}}$ that minimises the negative log likelihood $NLL_{reg}(W)$.

$$NLL_{reg}(W) = NLL(W) + \frac{1}{2}\lambda W^T W$$

We use Newton's Methods with $l_2$regularization:

$$W_{k+1} = W_k - H_{reg}(W_k)^{-1} g_{reg}(W_k)$$

$$\boldsymbol{g_{reg}}(\boldsymbol{W}) = \boldsymbol{g}(\boldsymbol{W}) + \lambda \mathrm{W}$$

$$\boldsymbol{H_{reg}}(\boldsymbol{W}) = \boldsymbol{H}(\boldsymbol{W}) + \lambda \mathrm{I}$$

A bias term is introduced to prevent a decision boundary that passes through the origin. Hence $W = \begin{bmatrix} b \\ \omega \end{bmatrix}$, which are the bias and weight parameters.

Since we do not want to regularize the bias:

$$W_{k+1} = W_k - H_{reg}(W_k)^{-1} g_{reg}(W_k)$$

$$= W_k - \left(H(W_k) + \lambda \begin{pmatrix} 0 & 0 \\ 0 & I_D \end{pmatrix} \right)^{-1} \left(g(W_k) + \lambda \begin{pmatrix} 0 \\ \omega_k \end{pmatrix} \right)$$

To predict the target label of the email data we compare $p(Y = 1|X,)$ and $p(Y = 0|X)$ and pick the label with the higher posterior predictive distribution.

a) Plots of training and test error rates versus λ

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

- Generally, error rates increase as **λ** increases. However, at the beginning when **λ** is small, the error rates first decrease before subsequently increasing
- The regularization parameter **λ** reduces overfitting. Increasing lambda results in less overfitting however, it also add greater bias.

c) Training and testing error rates for λ = 1, 10 and 100.

λ	**1**	**10**	**100**
Training error rates	0.048939641109298535	0.04926590538336052	0.06264274061990212
Testing error rates	0.05859375	0.060546875	0.068359375

Q4: K-Nearest Neighbours (KNN)

For a KNN classifier, we find the K Nearest Neighbours for each sample data to compute the posterior probability:

$$p(Y = C|X) = \frac{k_c}{K}$$

where k_c is the number of data samples from the K neighbours that are from label c.

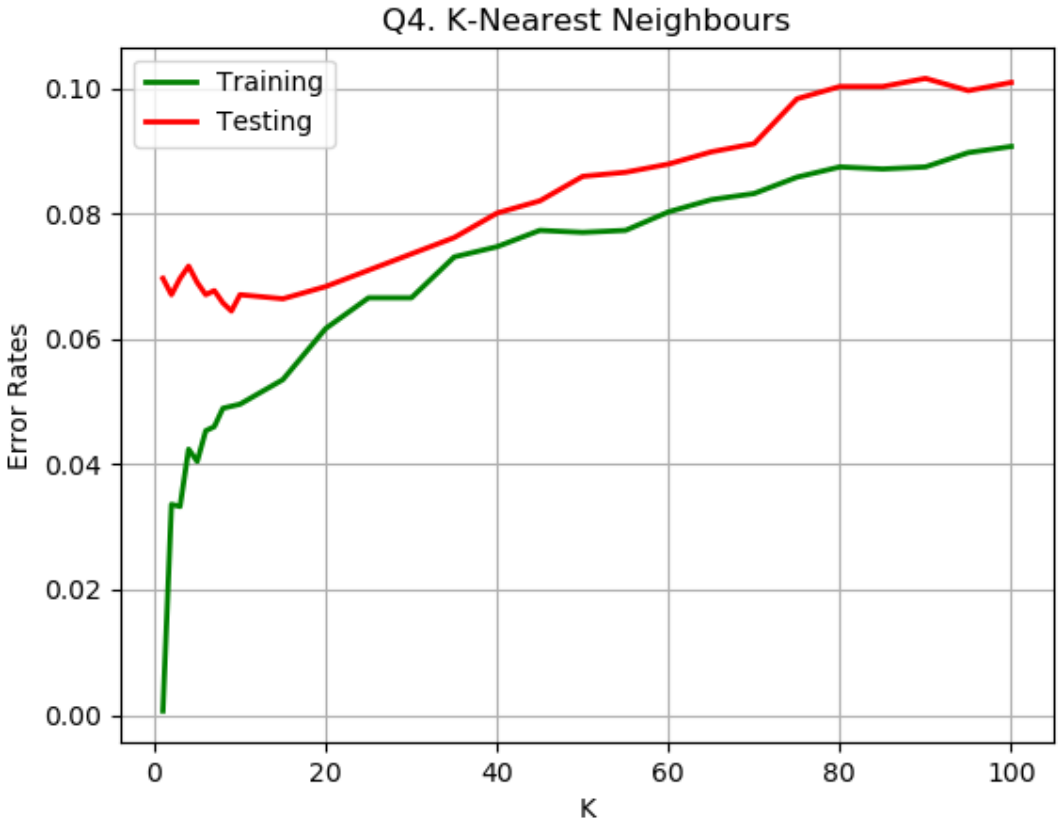
The distance between neighbours is measured using Euclidean distance:

$$dist(a, b) = \left(\sum_{j=1}^{D} |a_j - b_j|^2 \right)^{1/2}$$

To predict the target label of the email data we compare $p(Y = 1|X,)$ and $p(Y = 0|X)$ and pick the label with the highest probability. Repeat for different values of K.

There may exist cases where $p(Y = 1|X,) = p(Y = 0|X)$. When this occurs, we will assign a label of 1 (spam mail).

a) Plots of training and test error rates versus K



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

- At K = 1, the training error is very small and close to zero. At K=1, you chose the closest training sample in the dataset. When classifying the training set, the closest point to any training data is itself, hence training error is very small when K=1.
- However, at K=1, the test error, is very large, this implies that there is overfitting.
- As K increases, the training error increases drastically initially and increases gradually when K>10.
- The test errors are more stable and increases gradually as K increases.

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

K	1	10	100
Training error rates	0.0006525285481239804	0.04959216965742251	0.09070146818923328
Testing error rates	0.06966145833333333	0.06705729166666667	0.10091145833333333

Q5: Survey

I spent around **40 hours** to complete this assignment.