CSCI5525 Assignment 1 Answer

Daihui DOU dou00005@umn.edu 5514178

October 7, 2018

1 Problem 1

(1) Let $\frac{\partial E[l(f(x),y)]}{\partial f(x)} = 0$, then

$$\int_y (f(x) - y)p(y|x)p(x)\mathrm{d}y = 0$$
$$\int_y f(x)p(y|x)p(x)\mathrm{d}y = \int_y yp(y|x)p(x)\mathrm{d}y$$

Since $f(x)$ and $p(x)$ is independent of y, and $\int_y p(y|x)\mathrm{d}y = 1$,

$$f(x)p(x)\int_y p(y|x)\mathrm{d}y = p(x)\int_y yp(y|x)\mathrm{d}y$$
$$\begin{aligned} f(x) &= \int_y yp(y|x)\mathrm{d}y \\ &= E[y|x] \end{aligned}$$

Thus the optimal $f(x)$ is $E[y|x]$, where $E[y|x] = \int_y yp(y|x)\mathrm{d}y$.

(2) Let $\frac{\partial E[l(f(x),y)]}{\partial f(x)} = 0$, then

$$\int \mathrm{sgn}(f(x) - y)p(y|x)p(x)\mathrm{d}y = 0$$
$$\int_{-\infty}^{f(x)} p(y|x)p(x) = \int_{f(x)}^{\infty} p(y|x)p(x)$$

which means the optimal $f(x)$ is the median of the distribution of y, i.e., $p(y \leq f(x)|x) = 0.5$.

2 Problem 2

The expectation for y is given by:

$$E[y] = \sum_{i=1:M} \int_{x \in \mathcal{R}_i} p(y_i, x)\mathrm{d}x$$

The error rate for class $y = C_j$ equals to the rate that x is in $\mathcal{R}_k$ where $k \neq j$. Thus

$$\begin{aligned} err[y = C_j] &= \sum_{i=1:M, i \neq j} \int_{x \in \mathcal{R}_i} p(y_i, x)\mathrm{d}x \\ &= \sum_{i=1:M, i \neq j} \int_{x \in \mathcal{R}_i} p(C_j|x)p(x)\mathrm{d}x \\ &= \sum_{i=1:M} \int_{x \in \mathcal{R}_i} p(C_j|x)p(x)\mathrm{d}x - \int_{x \in \mathcal{R}_j} p(C_j|x)p(x)\mathrm{d}x \end{aligned}$$

Proved.

3 Problem 3

3.1 (a)

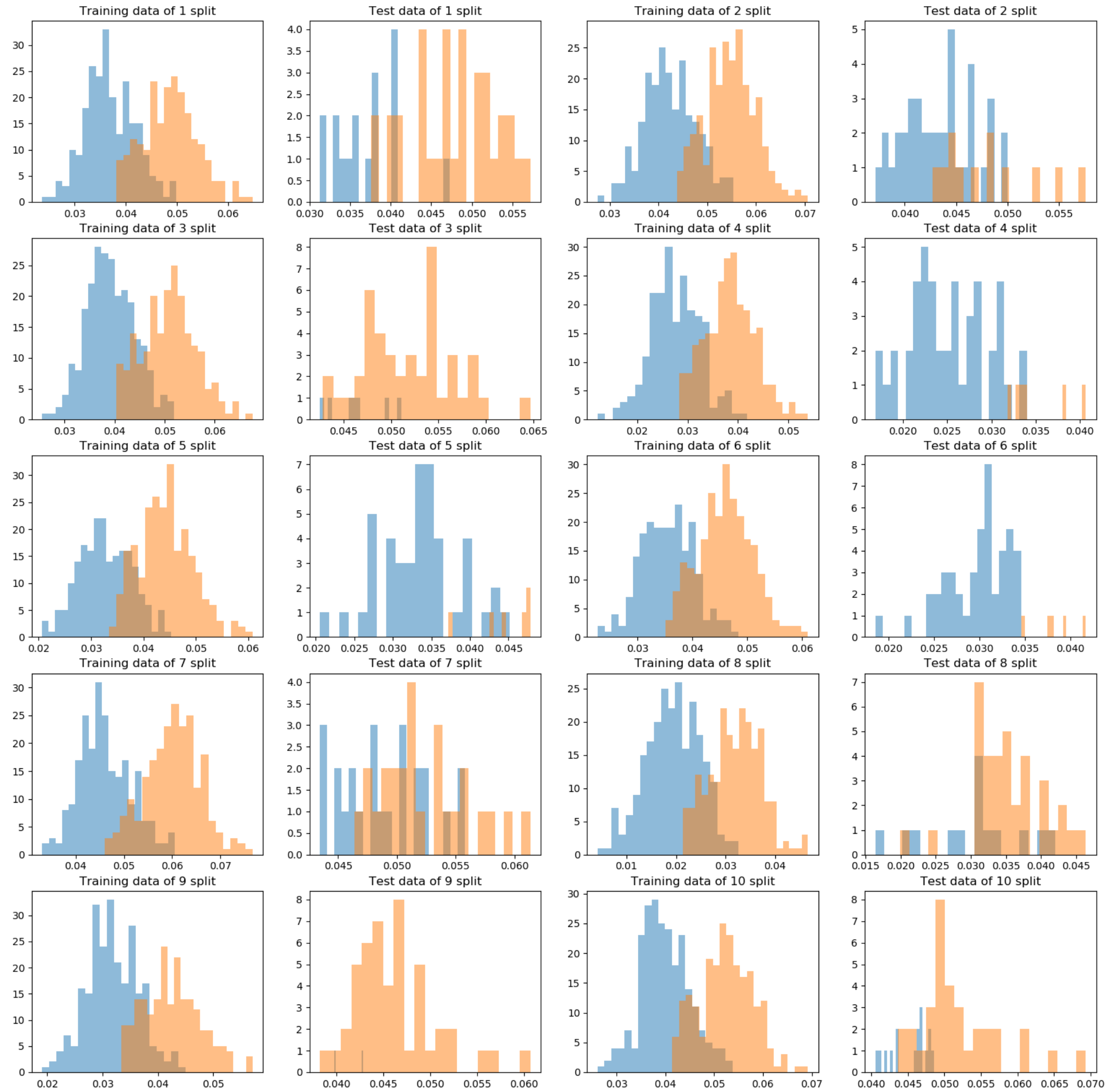
First modify the target T of boston dataset so that $p(y=0)=0.5$ and $p(y=1)=0.5$. And then Fisher's linear discriminant analysis (LDA) is applied to project boston dataset to 1 dimension.
Within-class covariance S_w is given by:

$$S_w = \sum_{\mathbf{x}_n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(x_n - \mathbf{m}_1)^T + \sum_{\mathbf{x}_n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

where m_i is mean of class i, and weight vector $\mathbf{w}$ is given by:

$$\mathbf{w} \propto S_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

and the projected data $f(x) = \mathbf{w}^T\mathbf{x}$.
The results of projected training data and test data of 10 cross-validation are below. Noted that LDA generally separated the two classes.



3.2 (b)

No, since boston dataset has two classes after modification and LDA can project datasets to a maximum of $K-1$ dimensions, where K is the number of classes.

This is because the maximum rank of between-class covariance S_b is K-1, and thus for $S_w^{-1}S_b$ there are at most $K-1$ eigenvalues and corresponding $K-1$ eigenvectors. The subspace dimension of the projected data is given by the eigenvectors, which is at most $K-1$.

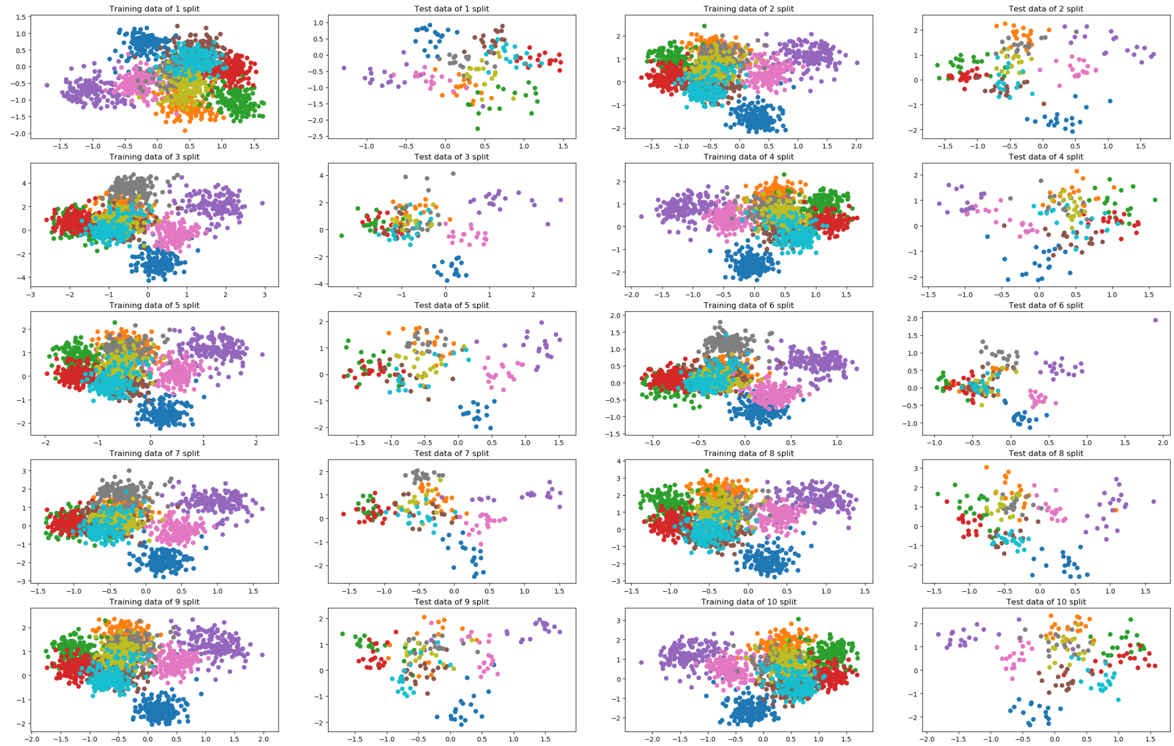
3.3 (c)

Between-class and within-class covariance S_w is given by:

$$S_b = \sum_{i=1}^{K} N_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$

$$S_w = \sum_{i=1}^{K} \sum_{x_n \in C_i} (\mathbf{x}_n - \mathbf{m}_i)(x_n - \mathbf{m}_i)^T$$

where K is the number of classes, N_i is the number of instances in class i, m_i is mean of class i.
The weight vector $\mathbf{w}$ is given by the top 2 eigenvectors of $S_w^{-1}S_b$ by the largest magnitude, and the projected data $f(x) = \mathbf{w}^T\mathbf{x}$.
The results of projected training data and test data of 10 cross-validation are below.



The testing error followed by Gaussian generative modeling is bellow

```
Test error of 1 split: [0.31666667]
Test error of 2 split: [0.27222222]
Test error of 3 split: [0.46666667]
Test error of 4 split: [0.40555556]
Test error of 5 split: [0.35]
Test error of 6 split: [0.45555556]
Test error of 7 split: [0.37777778]
Test error of 8 split: [0.36312849]
Test error of 9 split: [0.31284916]
Test error of 10 split: [0.31284916]
Test error mean: 0.36332712600869027
Test error std: 0.06076858471325458
```

4 Problem 4

4.1 (a) Logistic regression (LR)

Logistic regression is a discriminative model, and the class posteriors are given by:

$$p(C_k|\mathbf{x}) = \frac{\exp(a_k)}{\Sigma_j \exp(a_j)}$$

where $a_k = \mathbf{w}_k^T \mathbf{x}$. Then the likelihood is given by:

$$p(\mathbf{y}|\mathbf{w}_1, \ldots, \mathbf{w}_k) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\mathbf{x}_n)^{y_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_{nk}^{y_{nk}}$$

The optimization problem of $\mathbf{w}$ is solved by Iteratively Reweighted Least Squares (IRLS). Here the multi-class problem is treated as one-versus-the-rest problem. The update of $\mathbf{w}$ is then given by:

$$\begin{aligned}\mathbf{w}^{new} &= \mathbf{w}^{old} - H^{-1}(\mathbf{w}^{old})\nabla E(\mathbf{w}^{old}) \\ &= (X^T R X)^{-1} X^T R \mathbf{z}\end{aligned}$$

where $\mathbf{z} = X\mathbf{w}^{old} - R^{-1}(\pi - \mathbf{y})$, and R is a diagonal matrix initialized with $r^i = 1$, which is updated by

$$r_i = \frac{1}{\max(\delta, |y_i - X^i \mathbf{w}|)}$$

where δ is some small value.

4.2 (b) Naive-Bayes with marginal Gaussian distributions (GNB)

For K-class problem of Naive-Bayes, the posterior probability for class k:

$$p(C_k|\mathbf{x}) = \frac{p(\mathbf{x}|C_k)p(C_k)}{p(\mathbf{x})} = \frac{\exp(a_k)}{\Sigma_j^K \exp(a_j)}$$

were a_k is given by:

$$a_k = \log p(\mathbf{x}|C_k) + \log p(C_k)$$

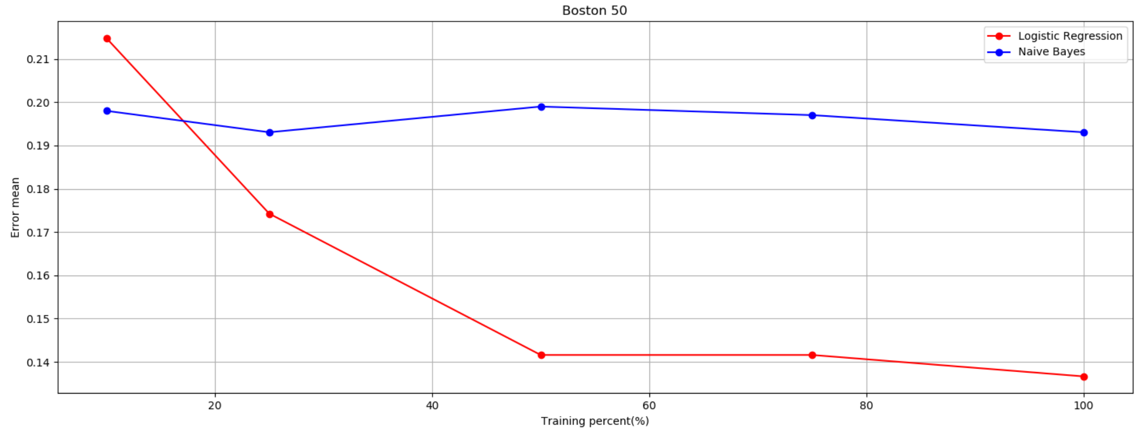
And for marginal Gaussian distribution, $p(x_i|C_k)$ is given by:

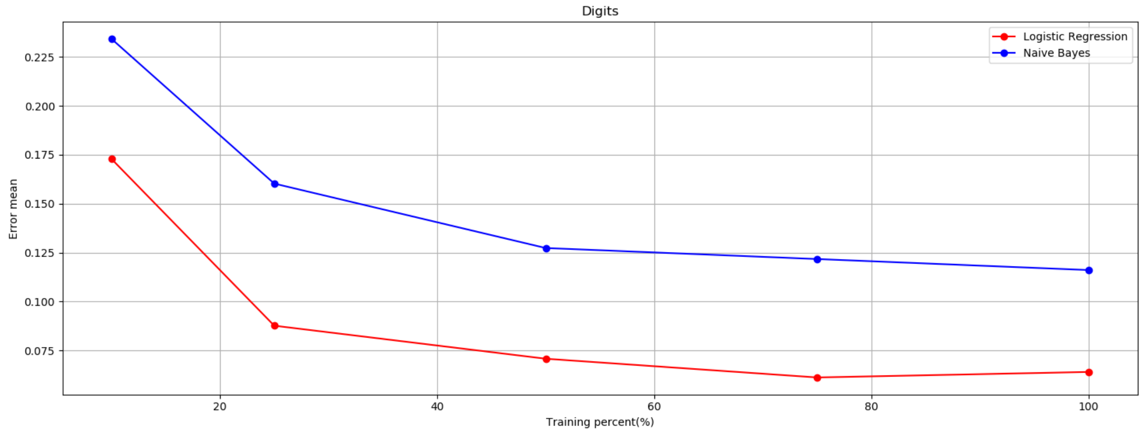
$$p(x_i|C_k) = \mathcal{N}(\mu_{ik}, \sigma_{ik}^2)$$

$$p(\mathbf{x}|C_k) = \prod_{n=1}^{D} p(x_i|C_k) = \frac{1}{(2\pi)^{D/2}(\Pi_{i=1}^{D}\sigma_{ik})} \exp\big(-\sum_{i=1}^{D} \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2}\big)$$

where μ_k and σ_k^2 are means and stds of class k of training data.

The results of LR and GNB are printed below:


Boston 50
Logistic Regression
Naive Bayes
Error mean
0.21
0.20
0.19
0.18
0.17
0.16
0.15
0.14
20
40
60
80
100
Training percent(%)


Digits
Logistic Regression
Naive Bayes
Error mean
0.225
0.200
0.175
0.150
0.125
0.100
0.075
20
40
60
80
100
Training percent(%)