CSci5525 Homework 2

Rihan Chen

Oct 1, 2015

Problem 1

As

$$E_{(x,y)}[l(f(x),y)] = \int_x \int_y l(f(x),y)p(x,y)dydx = \int_x \{\int_y l(f(x),y)p(x|y)dy\}p(x)dx$$

. where $\int_y l(f(x),y)p(x|y)dy$ is actully $E(g(y)|x)$. For the quadratic loss function, it is actually $E((f(x)-y)^2|x)$.

Proof: Assume: $\mu(x) = E(y|x)$, then the conditional expectation can be writen as:

$$E((y-f(x))^2|x) = E(((y-\mu(x)) + (\mu(x)-f(x)))^2|x)$$

$$\begin{aligned}&= E((y-\mu(x))^2|x) + (\mu(x)-f(x))^2 + 2E((y-\mu(x))(\mu(x)-f(x))|x)\\ &= Var(y|x) + (\mu(x)-f(x))^2 + 2(E(y|x)-\mu(x))(\mu(x)-f(x))\\ &= Var(y|x) + (\mu(x)-f(x))^2\end{aligned}$$

Hence, we can choose $f(x) = \mu(x) = E(y|x)$ in order to minimize the conditional expection,furthermore,minimize the loss function.Therefore, we conclude that the optimal $f(x)$ is $E(y|x)$

Problem 2

Let $X = x$, the conditional error probablity error probability $P(f(X) \neq Y|X = x)$ for any f is expressed by:

$$\begin{aligned}&= 1 - P(f(X) = Y|X = x)\\ &= 1 - P(Y = 1, f(X) = 1|X = x) - P(Y = -1, f(X) = -1|X = x)\\ &= 1 - [I(f(x) = 1)P(Y = 1|X = x) + I(f(x) = -1)P(f(X) = -1|X = x)\\ &= 1 - [I(f(x) = 1)\eta(x) + I(f(x) = -1)(1-\eta(x))]\end{aligned}$$

where $\eta(x)$ is *P(1/X)* and I is the indicator function.

Thus $P(f(X) \neq Y|X = x) - P(f^*(X) \neq Y|X = x)$ is given by:

$$\begin{aligned}&= \eta(x)[I(f^*(x) = 1) - I(f(x) = 1)] + (1-\eta(x))[I(f^*(x) = -1) - I(f(x) = -1)]\\ &= \eta(x)[I(f^*(x) = 1) - I(f(x) = 1)] + (1-\eta(x))[I(f(x) = 1) - I(f^*(x) = 1)]\\ &= (2\eta(x)-1)[I(f^*(x) = 1) - I(f(x) = 1)]\\ &\geq 0\end{aligned}$$

If $\eta(x) > \frac{1}{2}$ the first and the second term are nonnegative.As $I(f^*(x) = 1)$ is 1, the second term is greater than 0 obviously. if $\eta(x) \leq \frac{1}{2}$, the first term and second term are nonpositive. As $I(f^*(x) = 1)$ is 0, the second term is less than 0 obviously. Hence, we for that for each $X = x$ the conditional probability $P(f(X) \neq Y|X = x)$ is the smallest. Hence, if we take expectation for this conditional probability, it is also the smallest. In conclusion, we prove that *bayes classifier* has the smallest error rate.

Problem 3

i

The Fisher's discriminant in this case is used for dimension reduction. The formula for *within-class* for multiple class problem is defined as follow:

$$S_w = \sum_{k=1}^{k} S_k$$

The S_k stands for the *within-class* for each class and m_k is the center for each class. In our example, the *k = 4*.

$$S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n$$

For the the Fisher's discrimant method, we also need the *between class* variance as well, which is defined as follow:

$$S_B = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T$$

$$m = \sum_{n=1}^{N} x_n$$

m is the center for the whole training set. As we can see that S_B matrix actually has *rank = k-1*. Therefore, for $S_W^{-1} S_B$, its *rank* is at most 3, which is decided by linaer algebra theorem.Then, we extract the eigenvectors corresponding to this non-zero eigenvalues of $S_W^{-1} S_B$, which is three in our case. Actually, we map the X, which has higher dimension to, Y, which consists of those eigenvectors and have lower dimension.

From now on, each our observation can be represented by *3* features in our case. And we can build our generative gaussian model according to these three features for each class. We assume that each class in our case is following multivariate gaussian model

$$Y \sim N(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \Sigma)$$

For sample version, the mean and variance can be defined as follow:

$$\hat{\mu}_i = \frac{1}{N_k} \sum_{n \in C_k} y_{n,i}$$

$$\hat{\Sigma} = \frac{1}{N_k} \sum_{n \in C_k} (y_n - \hat{\mu}_k)(y_n - \hat{\mu}_k)^T$$

Here, we actually build four gaussian models each of which is multivariate guassian with dimension three. For prediction, we just calculate the posterior probablility by these four models and choose the one that give us the maximum posterior probability.

ii

For the least squares linear discriminant, it's quite simple, we just to use *1-of-K* coding for the response, then anything else is the same as linear regression by least square method. The coefficients have closed form, which is: $\hat{W} = (XX)^{-1}X \cdot Y$ Different from simple linear regression, where W is actually a vector. In our case, W is actually a matrix which has 4 rows, one for each class. For prediction, we just choose the class that give us the largest y in the vector Y. The reason is that the y for each class is asymtotically the conditional expecion of y given x.

Table 1: Results For Fisher Linear Discriminant

Train/Test	1	2	3	4	5	
Train	0.046144	0.046335	0.046755	0.046870	0.0457250	
Test	0.059369	0.058682	0.05250515	0.0555937	0.0603981	
	6	7	8	9	10	standard error
Train	0.046488	0.04669	0.0466021	0.0454199	0.0464114	0.00425050
Test	0.0511325	0.0480439	0.053878	0.0607413	0.060055	0.00425050

Table 2: Results for Least Square Linear Discriminant

Train/Test	1	2	3	4	5	
Train	0.0532377	0.0540005	0.054801311	0.0530852	0.05358096	
Test	0.0607412	0.0559369	0.055593686	0.0607412	0.06245710	
	6	7	8	9	10	standard error
Train	0.0521699	0.053886	0.053352	0.0537716	0.053657	0.00064810
Test	0.06760467	0.0607413	0.0614276	0.05628002	0.06142759	0.00425050

From the result of the two methods, we can see that the *Fisher Discriminant Analysis* which combine with gaussian model seems better than *Least Square Discriminant Analysis* somehow. But the difference is not huge.

Problem 4

In order to optimize the logistic regression, gradient descent method is taken. The iteration for gradient descent is as follow:

$$\theta_{j+1} = \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

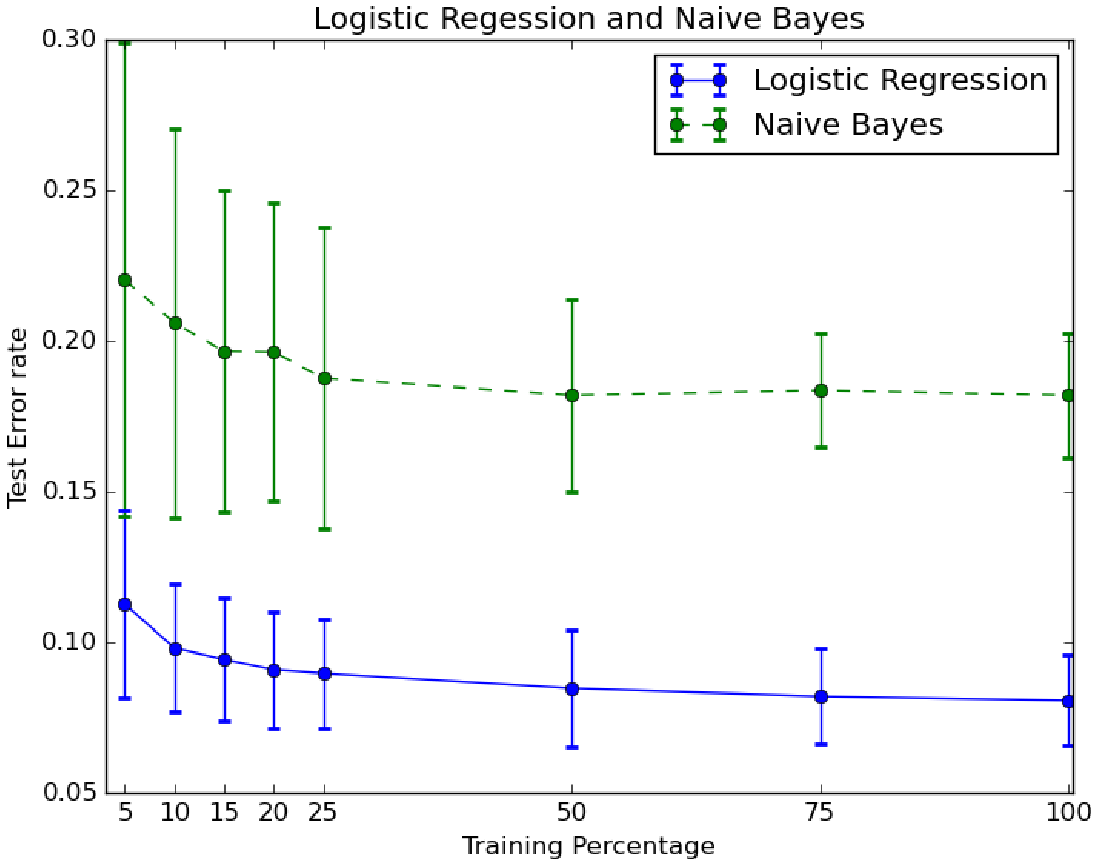
$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

where m stands for the number of observations and j stands for the j th feature. For the Naive Bayes, it is quite simple. For each class, only the mean and variance for each variance need to be calculated:

$$\bar{\mu}_{k,j} = \frac{1}{N_k} \sum_{i \in C_k} x_j^{(i)}$$

$$\sigma_{k,j}^2 = \frac{1}{N_k} \sum_{i \in C_k} (x_j^{(i)} - \bar{\mu}_{k,j})^2$$

where j stands for the j th feature, C_k stands for k th class. By the assumption of Naive Bayes, each class form a multivariate normal distribution with covariance matrix that is diagnal.Therefore, we only need to calculate the mean and variance for each feature respectively.



From the plot, we can see that with the size of training set increasing, the test error is decreasing gradually for both models. For the two methods, we see that logistic regression is much better than Naive Bayes for this dataset in terms of lower error rate and narrower error bars (which actually a 95% confidence interval under normal assumption). It may be due to that the Naive Bayes model has assumptions that cannot be achieved by the dataset, such as it assume that each feature is normally distributed and independent with each other.For logistic regression, which has less assumption compared with Naive Bayes model, it fits the data well in this case.