Statistical Machine Learning GU4241/GR5241
Homework 2

March 7, 2018

Problem 1

Let's denote $\hat{\beta}_{te}$ to be the least square estimator for testing data. In general it is not equal to $\hat{\beta}$, i.e., the least square estimator from testing data.

First, we consider the simple situation when M=N. Then we know

$$E[R_{tr}(\hat{\beta})] = \frac{N-p-1}{N}\sigma^2$$

$$E[R_{te}(\hat{\beta}_{te})] = \frac{N-p-1}{N}\sigma^2$$

By further noticing that $\hat{\beta}_{te}$ is the least square estimator for the testing data, which must minimize the square error, hence after taking expectation for all testing data, we have

$$E[R_{te}(\hat{\beta}_{te})] \leq E[R_{te}(\hat{\beta})]$$

This gives the inequality:

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})]$$

Then we prove that the expected testing error doesn't depend on the testing sample size M. This is because

$$E[R_{te}(\hat{\beta})] = E_{\hat{\beta}}E_{\tilde{x}_i,\tilde{y}_i}[\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2] = \frac{1}{M}\sum_{i=1}^{M}(E_{\hat{\beta}}E_{\tilde{x}_i,\tilde{y}_i}[(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2]).$$

The first expectation is taken on training data $\{x_i, y_i\}$ by considering $\hat{\beta}$ to be a random variable depending on the distribution of training.

Since $\{x_i, y_i\}$ and $\{\tilde{x}_i, \tilde{y}_i\}$ is iid for any i, $E_{\hat{\beta}}[E_{\tilde{x}_i,\tilde{y}_i}[(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2|\hat{\beta}]]$ does't change with i. Hence the expected testing error:

$$E[R_{te}(\hat{\beta})] = E_{\hat{\beta}}[E_{\tilde{x}_i,\tilde{y}_i}[(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2|\hat{\beta}]],$$

does not depend on sample size M. This finishes the proof.

Problem 2

(a)

The curve is a circle centered at $(-1, 2)$, with radius 2.

(b)

$\{(X_1, X_2) : (1 + X_1)^2 + (2 - X_2)^2 > 4\}$ is the set of points outside the circle. The other set correpondes to the region inside the circle.

(c)

$(0, 0)$: blue; $(-1, 1)$: red; $(2, 2)$: blue; $(3, 8)$: blue.

(d)

The decision boundary is characterized by $X_1^2 + 2X_1 + X_2^2 - 4X_2 + 1 = 0$, which is a linear in terms of X_1, X_1^2, X_2, and X_2^2.

Problem 3

The idea here is to illustrate how we get rid of potential over-fitting problems by dimension reduction in the high dimensional case. Each row in the dataset represents a handwritten zip-code with 16×16 pixels.

Table 2 summarizes the performance of all these 4 methods by comparing their training and testing accuracy. As a whole, we see that method 3 (LDA by averaging non-overlapping 2×2 pixels.) has the best testing accuracy. We can also say that the LDA with all 256 features has over-fitting problem by observing its high testing accuracy but relatively smaller testing accuracy. LDA with the first 49 principle components doesn't give a satisfactory result, even though these components have already accounted for 99.6% variance.

	Testing Accuracy	Training Accuracy
LDA with all 256 features	0.984	0.913
LDA with 49 principle components	0.954	0.909
LDA with features replaced by its 2×2 average	0.966	0.925
Multiple linear logistic regression	0.956	0.913

Table 1: Training and Testing Accuracy for 4 methods

Code:

You can check the code in the attached file "problem3.R".