Statistical Machine Learning GU4241/GR5241

Spring 2018
https://courseworks.columbia.edu/

Homework 2

Due: Thursday, Feb. 22st, 2017

Homework submission: Please submit your homework electronically through Canvas by 11:59pm on the due date. You need to submit both the pdf file and your code (either in R or Python).

Problem 1 (Training Error vs. Test Error, ESL 2.9, 10 points)

In this problem, we want to use the least squares estimator to illustrate the point that the trainning error is generally an underestimate of the prediction error (or test error).

Consider a linear regression model with p parameters,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \text{ where } \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

We fit the model by least squares to a set of trainning data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn independently from a population. Let $\hat{\beta}$ be the least squares estimate obtained from the training data. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \cdots, (\tilde{x}_M, \tilde{y}_M)$ $(N \geq M > p)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^T \tilde{x}_i)^2$, prove that

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

where the expectations are over all that is random in each expression.

Hints:

- Consider the least squares estimate $\tilde{\beta}$ based on the test data.
- The expection of residual sum-of-squres $\sum_{i=1}^{N}\mathbb{E}(y_i - \hat{\beta}^T x_i)^2$ is $(N - p - 1)\sigma^2$.

Problem 2 (Non-linear Decision Boudary, ISL 9.2, 10 points)

We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$. We now investigate a non-linear decision boundary.

(a) Sketh the curve

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

(b) On your sketch, indicate the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 \leq 4.$$

(c) Suppose that a classifier assigns an observation to the blue class if

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

and to the red class otherwise. To what class is the observation $(0, 0)$ classified? What about $(-1, 1)$, $(2, 2)$ or $(3, 8)$?

(d) Argue that while the decision boundary in (c) is not linear in terms of X_1 and X_2, it is linear in terms of X_1, X_1^2, X_2 and X_2^2.

Problem 3 (LDA and Logistic Regression, 20 points)

The zipcode data are high dimensional, and hence linear discriminant analysis suffers from high variance. Using the training and test data for the 3s, 5s, and 8s, compare the following procedures:

1. LDA on the original 256 dimensional space.
2. LDA on the leading 49 principle components of the features.
3. LDA when you *filter* the data as follows. Each non-overlapping 2×2 pixel block is replaced by its average.
4. Multiple linear logistic regression using the same filtered data as in the previous question. [Use the `multinomial` family in the R package `glmnet`; use the solution at the end of the path].

Homework Problems. Compare the procedures with respect to training and test misclassification error. You need to report both training and test misclassification error in your submission.