STSCI 3740/5740 Machine Learning and Data Mining

Dr. Nayel Bettache　　　　　　　　　　　　　　　　　　　Homework 1, **due Oct 10, 11:59pm**

Problem 1 (6 points)

1. Express $\text{Var}(X_1 - X_2)$ through the variances and covariances of X_1, X_2 (assuming all variances exist).

2. Assume that $X_1, ..., X_n$ are i.i.d. real-valued random variables with finite variances. Show that
$$\text{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\text{Var}(X_1).$$

3. Assume that X, Y are independent random variables with $\mathbb{E}[X] = 0, \mathbb{E}[Y] = 1, \text{Var}(X) = 1, \text{Var}(Y) = 2$. Compute $\mathbb{E}[(3X + Y)(5Y + 2X - 1)]$

Problem 2 (8 points)

Solve Problem 1 of Chapter 2.4, Problem 1 of Chapter 3.7 and Problem 1 of Chapter 4.8 in the textbook *Introduction to Statistical Learning*(second edition).

Problem 3 (12 points)

Solve Problem 9 of Chapter 2.4, Problems 8 and 9 of chapter 3.7 and Problem 14 of chapter 4.8 in the textbook *Introduction to Statistical Learning*. Choose either R or Python. The data set Auto can be found on the webpage of the course

You may follow the code in Chapter 2.3.4 (ISLR) or Chapter 2.3.7 (ISLP) to load data. (**For this problem you shall submit a R notebook or a Python notebook explaining every single step of your code.**)

Problem 4 (5 points)
This question is required for STSCI 5740. It is optional for 3740, that means you can have some bonus points, if you get the correct answer.

Classification is a very important research area and has been extensively studied from a theoretical aspect. In many research papers, the focus is on how to bound the so-called excess risk of a classifier. In this question, we will first define the excess risk, and then study some mathematical properties of the excess risk.

We will follow the notation in the lectures. Assume that Y takes values in $\{0, 1\}$. Based on the lectures, we know the Bayes classifier is $f^*(x) = 1$ if $p_1(x) = P(Y = 1|X = x) > 1/2$ and $f^*(x) = 0$ otherwise. (I use a slightly different notation f^* to denote the Bayes classifier rather than $\hat{f}$ in the slides).

Since $p_1(x)$ depends on the unknown data distribution, the Bayes classifier is not implementable in practice. One way to construct a practical classifier is the following. Let us first use some model or algorithm to estimate $p_1(x)$. We call this estimator as $\hat{p}_1(x) \in [0,1]$. Then we can plug-in the Bayes classifier. So, we have the following classifier $\hat{f}(x) = 1$ if $\hat{p}_1(x) > 1/2$ and $\hat{f}(x) = 0$ otherwise.

Now, we define the *excess risk* of the classifier $\hat{f}(x)$ as

$$R(\hat{f}) - R(f^*),$$

where $R(f) = P(Y \neq f(X))$ is the misclassification error of f (unconditioning on X). In words, the excess risk is the difference between the misclassification errors of $\hat{f}$ and the Bayes classifier. Since the Bayes classifier has the smallest misclassification error (shown in the class), we know the excess risk is always nonnegative. We can claim that $\hat{f}$ is a good classifier, if its excess risk is close to 0. So, for any given classifier $\hat{f}$, we would like to know its excess risk or its upper bound at least.

Given the above background, please prove the following inequality regarding the excess risk

$$R(\hat{f}) - R(f^*) \leq 2E|\hat{p}_1(X) - p_1(X)|.$$

(If you take some more advanced ML courses in the future, you will see this is an important inequality.)

Hint: You may first prove the following identity

$$P(Y \neq \hat{f}(X))|X = x) - P(Y \neq f^*(X))|X = x) = |2p_1(x) - 1| \times I(f^*(x) \neq \hat{f}(x)),$$

where $I()$ is the indicator function.

Problem 5 (4 Bonus points)
This exercise is optional for everyone.

Assume that we have the regression model

$$Y = f(X) + \varepsilon,$$

where ε is independent of X and $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) = \sigma^2$. Assume that the training data $(x_1, y_1), ..., (x_n, y_n)$ are used to construct an estimate $\hat{f}$ of f. Given a new random vector (X, Y) (i.e., test data independent of the training data),

1. Show that

$$\mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] = \mathbb{V}(\hat{f}(X)|X = x) + \mathbb{E}\left(\left[\mathbb{E}[\hat{f}(X)|X = x] - f(X)\right]^2 |X = x\right). \tag{1}$$

   Hint: You may start from

$$\mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] = \mathbb{E}\left[\left(f(X) - \mathbb{E}[\hat{f}(X)|X = x] + \mathbb{E}[\hat{f}(X)|X = x] - \hat{f}(X)\right)^2 |X = x\right].$$

   Then do the square expansion.

2. Show that

$$\mathbb{E}[(Y - \hat{f}(X))^2 | X = x] = \mathbb{V}(\hat{f}(X)|X = x) + \mathbb{E}\left(\left[\mathbb{E}[\hat{f}(X)|X = x] - f(X)\right]^2 |X = x\right) + \sigma^2.$$

   Hint: The proof follows from the similar derivations shown in the lecture together with the equation (1) above.

3. Explain the bias-variance trade-off based on the above equation.

4. Explain the difference between training MSE and test MSE. Can expected test MSE be smaller than σ^2?

Problem 6 (3 Bonus points)
This exercise is optional for everyone.

Consider a classification problem. Assume that the response variable Y can only take value in $C = \{1, 2, 3\}$. For a fixed x_0, assume that the conditional probability of Y given $X = x_0$ follows

$$P(Y = 1|X = x_0) = 0.6; \quad P(Y = 2|X = x_0) = 0.3; \quad P(Y = 3|X = x_0) = 0.1.$$

1. Derive the Bayes classifier at $X = x_0$.

2. Derive the corresponding Bayes error rate.

3. Consider a naive classifier $\hat{f}(x_0)$, called random guessing. That is we use the computer to randomly pick one number from $C = \{1, 2, 3\}$ with equal probability as the label for x_0. Compute the expected test error rate of this classifier. Show that the Bayes error rate is smaller than the expected test error rate for random guessing.