MATH4432 Notes

SmokingPuddle58

September 19, 2024

This is the lecture note typed by SmokingPuddle in September, 2024. It mainly contains what professor mentions starting from year 3. For the contents of the first two weeks, I will try my best to include as much as possible.

The main reference source comes from the professor himself, lecture notes, tutorial notes, and also from the Internet if necessary.

Please inform me if there is any errors, better within the semester or I will have a very high chance of forgetting the contents.

Theorems, Corollary, Lemma, Proposition

Definitions

Examples

Warnings / Remarks

Proofs, Answers

Some special symbols, notations and functions that will appear in this note:

$\mathbb{C}$	Set of complex numbers
$\mathbb{R}$	Set of real numbers
$\mathbb{Z}$	Set of integers
$\mathbb{Q}$	Set of rational numbers

Contents

1 Overview

Before we start, we shall clarify some of the notations that will be used.

Consider the following expression:

$$P(X = x)$$

If we say *r.v.* (Random variable) X, we actually means the name of the variable, while for x, we means the realization for such *r.v.*

Suppose we are now observing some quantitative response Y and also input variable X, consisting of p features, which can be expressed as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_p \end{bmatrix}$$

where $X_1, ..., X_p$ are random variables. Then the relation between Y and X can be expressed as:

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

where ε is the error term, and f is a deterministic function. We call such model the population level model, or ground truth model. (i.e. The number of samples is infinitely many)

Remark 1.1
Note that Y, ε are all random variables, while X is a collection of random variables.

If we want to consider a sample level (the realization of the random variables), then the equation becomes:

$$y_i = \quad f(x_i) + \varepsilon_i \quad i = 1, ..., n$$

where x_i can be a vector like the following:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{ip} \end{bmatrix}$$

and n is the sample size.

Remark 1.2
In machine learning, vectors usually means **column vectors, but not row vectors**.

For example, consider the equation $f(x) = a_1x_1 + a_2x_2 + a_3x_3$. If we know that $a = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$, $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$, then $f(x) = a^\intercal x$, where $a^\intercal$ is the transpose of the vector a.

Now let's go back to the ground truth model, which is $Y = f(X) + \varepsilon$. Suppose we want to construct f from the data, then we will have:

$$\hat{Y} = \hat{f}(X = x)$$

for any observed x.

Suppose we are interested in the difference between the data and the observed prediction, then we are interested in the value of $\mathbb{E}(Y - \hat{Y})^2$, the expected square error.

Remark 1.3
Both Y and $\hat{Y}$ are random variable, since $\hat{Y}$ comes from the randomness of the data chosen. Thus we are not interested in the value of $(Y - \hat{Y})^2$, since it is not a fixed value.

Theorem 1.1

$$\mathbb{E}(Y - \hat{Y})^2 = \underbrace{\mathbb{E}(f(X) - \hat{f}(X))^2}_{\text{Reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{Irreducible}}$$

Proof.