MATH4432 Notes

SmokingPuddle58

September 20, 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 **4**

1.1 Introduction 4

1.2 Estimation of f 5

1 Overview

1.1 Introduction

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 will be 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 $\hat{Y}$ is the prediction that is learnt from the data, and data comes from the random sample chosen from ground truth model. Thus we are not interested in the value of $(Y - \hat{Y})^2$, since it is not fixed.

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.
To prove the equation, we have to assume ε is independent of both $f(X)$ and $\hat{f}(X)$.

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) + \varepsilon - \hat{f}(X))^2 \\
&= \mathbb{E}(f(X) - \hat{f}(X) + \varepsilon)^2 \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \mathbb{E}(\varepsilon^2)
\end{aligned}$$

Since for any random variable X, and its expected value, $E(X) = \mu$, we have: (Covered in MATH2411)

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X - \mu)^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(\mu^2) - 2\mathbb{E}(X\mu) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu\mathbb{E}(X) \\
&= \mathbb{E}(X^2) + \mu^2 - 2\mu^2 \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}(f(X) - \hat{f}(X))^2 + \mathbb{E}(\varepsilon^2) \\
&= \mathbb{E}(f(X) - \hat{f}(X))^2 + \text{Var}(\varepsilon)
\end{aligned}$$

To conclude, you can only reduce the error for the reducible part, by making your model approximate the ground truth model as well as possible, while we can really do not much on the irreducible part.

1.2 Estimation of f

There are two methods for estimating f, namely parametric, and non-parametric methods.

For parametric method, we assume that f can be described by a set of parameters, such that once all of the parameters are known, then the model is known.

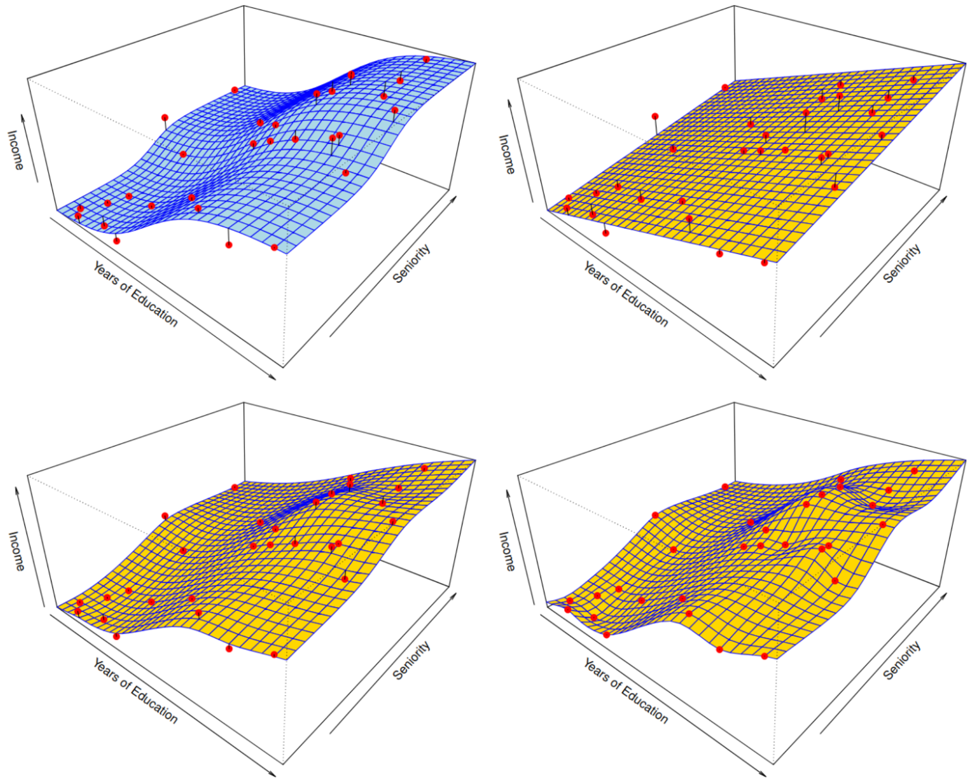
Example 1.1
One of the most simple assumption is that f is linear in X, which can be described as:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

For non-parametric method, we do not pre-specify the form of the model (Can be linear, non-linear. tree and neural network). To control the flexibility, we can always tune the parameters.

The advantages and disadvantages are listed in the following table.

	Parametric	Non-parametric
Advantage	Easy to solve and understand	More flexible and sometimes more powerful
Disadvantage	The model may be too simple to fit into the data	The model may be too flexible, there may be overfitting



Ground truth model	Underfitting
Good estimation	Overfitting (Fitting into the noise)

The above image shows the example of a ground truth model, a good estimation, overfitting and underfitting. It is also included in the lecture note.

In statistical machine learning context, to prevent overfitting, we shall introduce the concept of regularization (Covered in detail later), in which we want to find f, where f satisfies:

$$f = \arg\min_f \ \text{Loss}(Y, f) + \lambda R(f)$$

where λ is a tuning parameter (weight) for the regularizer $R(f)$ of the function f.

The introduction of the regularizer is trying to control the complexity / flexibility of the function f to prevent overfitting issue.