ECON 5202
Decision Theory in Econometrics

Professor Yu-Chang Chen
LaTeX by Hung-Chun Li
National Taiwan University

Semester 112-2

Contents

Chapter 1

Classical Secretary Problem

1.1 Problem Statement

There are hundreds of people interviewing for your job. You interview them one by one and give each a score. Your goal is to find the best person, the one with the highest score. The restriction in this problem is that after you interview a person, you must decide whether to give this person an offer immediately after the interview. For example, if you want to marry the best person, you keep dating until you make a decision. You can decide whether to marry your current girlfriend or boyfriend, but you cannot go back and marry an ex-girlfriend or ex-boyfriend.

- n applicants, 1 secretary
- X_i: score of the i-th applicant (the higher the better)
- Interviews are in random order
- Restriction: have to make decision right after the interview
- Objective: find the best applicant

Given the data, should I offer the job to the person I just interviewed, or should I interview the next person?

$$\psi = (\psi_1(X_1), \psi_2(X_1, X_2), \ldots, \psi_n(X_1, \ldots, X_n)),\ \psi_i = \{0, 1\}$$

In conclusion, the goal is to find the optimal "stopping rule" that maximize the probability of finding the best person.

1.2 Illustration

Stopping Rule #1: always give given first person an offer.

$$\begin{aligned}&\psi = (\psi_1(X_1), \psi_2(X_1, X_2), \ldots, \psi_n(X_1, \ldots, X_n))\\ &\psi_1 = 1, \quad \psi_2 = \cdots = \psi_n = 0,\end{aligned}$$

Then the probability of winning is the probability of the first one is the one with highest score. Because the applicants come in random order, $Pr(\text{win}) = \frac{1}{n}$. Obviouly, the strategy is unlikily to be the success.

Stopping Rule #2: always skip the first one. For the rest of the interviews, make the offer if the person is best so far.

$$\begin{aligned}
\psi_1(X_1) &= 0, \\
\psi_2(X_1, X_2) &= \begin{cases} 1, & X_2 > X_1 \\ 0, & \text{o.w.} \end{cases} \\
&\vdots \\
\psi_n(X_1, \ldots, X_n) &= \begin{cases} 1, & X_n > X_1, \ldots, X_n > X_{n-1} \\ 0, & \text{o.w.} \end{cases}
\end{aligned}$$

Example.

When $n = 3$, $(X_1, X_2, X_3) = (5, 8, 10)$, there's 6 possible orders

1	②	3	→	fail, pick the smaller one
1	③	2	→	win
2	③	1	→	win
2	1	③	→	win
3	2	1	→	fail, skip the best one
3	1	2	→	fail, skip the best one

$\longrightarrow Pr(\text{win}) = \frac{1}{2} > \frac{1}{3}$

1.3 Solving CSP

The **Stopping Rule #2** suggests skipping the first applicant, but what if we also skip the second applicant, the third applicant, and so on? The more applicants we skip, the more confident we are that the next candidate who shows up is the best one among all the people interviewed so far. However, there's a trade-off: if we skip too many people, we might miss out on the best person. So, what is the optimal number of people to pass?

Definition 1.3.1: Threshold Rules

N_r: Starting from r-th applicant, we make an offer to candidate.

Claim

Threshold Rules N_r for some r is the optimal rule.

Proof for Claim.

Observation 1

w_j is the probability of winning using an optimal rule that pass first j applicants.

$$w_j \geq w_{j+1}$$

as

$$\{\text{rules that skip first } j\} \supset \{\text{rules that skip first } j+1\}$$
$$\implies w_j \text{ is decreasing in } j$$

Observation 2

When should we make an offer to a candidate, someone who is the best we seen so far? If the j-th applicant is the best so far

$$\begin{aligned} P(\text{Win if pass }) &= W_j \\ P(\text{Win if make an offer }) &= P(\text{the best applicant appears in the first } j \text{ interviews}) \\ &= \frac{j}{n} \end{aligned}$$

So we make an offer if

$$P(\text{Winning if make an offer}) \geq P(\text{Winning if don't make an offer})$$
$$\frac{j}{n} \geq w_j$$

So, base on Observation 1 and Observation 2, if for some j,

$$\begin{aligned} & \frac{j}{n} \geq w_j \\ \implies & \frac{j+1}{n} \geq \frac{j}{n} \geq w_j \geq w_{j+1} \\ \implies & \frac{j+1}{n} \geq w_{j+1}. \end{aligned}$$

This indicates that if it is optimal to make an offer to a candidate at some point j, it will also be optimal to make an offer to the candidate at $j+1$. Once we pass the threshold, it is always optimal to make an offer. Therefore, an optimal rule must be a threshold rule.

So far, we already know "an optimal rule must be a threshold rule". How about the optimal threshold r^*?

skip skip skip

$$\max \quad \frac{r-1}{n} \sum_{k=r}^{n} \frac{1}{k-1}.$$

When $n = 3$,

$$r^* = 2 \Longrightarrow \text{Pass } r^* - 1 = 1 \text{ applicants}$$
$$P(\text{Win}) = \frac{1}{2} \geq \frac{1}{3}.$$

When $n = 7$,

$$r^* = 3 \Longrightarrow \text{Pass } r^* - 1 = 2 \text{ applicants}$$
$$P(\text{Win}) \simeq 0.414 > \frac{1}{7}.$$

When $n = 10000$,

$$r^* = 3680 \Longrightarrow \text{Pass } r^* - 1 = 3679 \text{ applicants}$$
$$P(\text{Win}) \simeq 0.36.$$

When $n \to \infty$

$$\frac{r}{n} \to 0.368 = e^{-1}.$$

1.4 Concluding CSP

There's a few extension for CSP:

1 Secretary problem with full information

2 CSP with rejection

3 CSP with recalls

4 CSP with time cost

Chapter 2

Bayes Theorem

2.1 Review of Bayes Theorem

Theorem 2.1.1: Bayes Theorem

Let A, B be two events,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Proof.

$$\begin{aligned} P(A|B) &= \frac{P(B \cap A)P(A)}{P(B)} \\ &= \frac{P(A)P(B|A)}{P(B)} \end{aligned}$$

□

Bayes theorem can be think as a way to rationally take information into account. In many decision problem, our goal is to use the information we have to form a belief and make decision base on it. We call A the "prior", and threat B as a new "information" to update our belief. Given that B has occured, what is the probability of A?

Example.

For example, we are rolling a die,

$$A = \{\text{odd number}\} = \{1, 3, 5\}$$
$$B = \{\text{prime number}\} = \{2, 3, 5\}.$$

The probability of rolling an odd number and the probability of rolling an prime number are

$$P(A) = \frac{1}{2},\ P(B) = \frac{1}{2}.$$

However, if we have a new information that the outcome is an prime number, the probability of it being and odd number is

$$P(A|B) = \frac{2}{3}.$$

Example.

For example, we want to know whether the email is a spam,

$$A = \{\text{spam}\},\ P(A) = \frac{1}{4}$$
$$B = \{\text{send from NTU}\},\ P(B) = \frac{1}{3},\ P(B|A) = \frac{1}{2}$$

$$\begin{aligned} P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \\ &= \frac{(\frac{1}{2})(\frac{1}{4})}{\frac{1}{3}} = \frac{3}{8} \end{aligned}$$

Let X, Y be discrete random variables with pmf $f_X(x)$ and $f_Y(y)$.

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}$$

Proof.

$$\begin{aligned} f(x|y) &= P(X = x|Y = y) \\ &= \frac{P(Y = y|X = x)f(X = x)}{P(Y = y)} \\ &= \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \end{aligned}$$

□

2.2 Application of Bayes Theorem: Laplace's Rule

How do we trade-off between numbers of reviews and average rating? Suppose we toss coin 1 for 10 times and coin 2 for 5 times

$$X_1, X_2, \ldots, X_{10} \sim Ber(p_1) \longrightarrow \bar{x} = 0.5$$
$$Y_1, Y_2, \ldots, Y_5 \sim Ber(p_2) \longrightarrow \bar{y} = 0.6.$$

If our goal is to find which coin has higher probability of head. Obviously, $\bar{y} > \bar{x}$. But it doesn't take into account the number of the samples, which measure how confidence we are. Therefore, comparison based on the average itself is not very useful.

Fact 2.2.1

If a random variable $P \sim Beta(\alpha, \beta)$, it would have pdf and expected value given by

$$f_P(p) = \begin{cases} \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}, & 0 \le p \le 1, \\ 0, & \text{otherwise.} \end{cases}$$
$$E[P] = \frac{\alpha}{\alpha+\beta}.$$

When $\alpha = \beta = 1$, the pdf and expected value becomes

$$f_P(p) = \begin{cases} 1, & 0 \le p \le 1, \\ 0, & \text{otherwise.} \end{cases}$$
$$E[P] = \frac{1}{2}.$$

This is exactly the pdf of $U(0,1)$, so uniform distribution is a special case of beta distribution.

Before we have any information, we assume the data follow a uniform distribution, i.e., $P \sim U(0,1)$. We call this a "flat prior" because the pdf is flat, meaning that there is no particular value of P is more likely than any other; every P is equally likely.

Moreover, given P, the observations follow a Bernoulli distribution, $X_1, X_2, \ldots, X_n | P \overset{\text{iid}}{\sim} Ber(P)$. We call the pdf of this distribution, $f(x_1, x_2, \ldots, x_n|p) = \Pi_{i=1}^n p^{x_i}(1-p)^{1-x_i}$, called the likelihood.

"Prior" is the distribution of parameter of interest before observing the data, and "likelihood" is the probability of the observed data given a specific parameter value. Now, let's see how do I update the prior.

Once we see the data and update our belief, $f(p|\mathbf{x})$, we call this "posterior". According

to the Bayes Theorem, it is given by

$$\begin{aligned}
f(p|\mathbf{x}) &= \frac{f(\mathbf{x}|p)f(p)}{f(\mathbf{x})}. \\
f(\mathbf{x}|p)f(p) &= (\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i})(1) \\
&= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\
f(\mathbf{x}) &= \int_0^1 f(\mathbf{x}|p)f(p)\mathrm{d}p \\
f(p|\mathbf{x}) &= \frac{1}{f(\mathbf{x})} p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\
&\propto p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\
&= \underbrace{p^{(\sum_{i=1}^n x_i+1)-1}(1-p)^{(\sum_{i=1}^n (1-x_i)+1)-1}}_{\sim Beta(\alpha,\beta) \text{ where } \alpha=\sum_{i=1}^n x_i+1=n\bar{x}+1 \; \beta=\sum_{i=1}^n (1-x_i)+1=n(1-\bar{x})+1}
\end{aligned}$$

Thus, posterior follows

$$P|\mathbf{X} \sim Beta(n\bar{X}+1, n(1-\bar{X})+1).$$

My guess on P given data is

$$\begin{aligned}
E[P|\mathbf{X}] &= \frac{\alpha}{\alpha+\beta} \\
&= \frac{n\bar{X}+1}{n\bar{X}+1+n(1-\bar{X})+1} \\
&= \frac{n\bar{X}+1}{n+2}.
\end{aligned}$$

This is the Laplace's Rule. We will show that using $\frac{n\bar{X}+1}{n+2}$ is better than using the average $\bar{X}$ later.

Back to the question at the begining, which coin has higher probability of head?

$$\begin{aligned}
E[P|\mathbf{X}] &= \frac{10(0.5)+1}{10+2} = 0.5 \\
E[P|\mathbf{Y}] &= \frac{5(0.6)+1}{5+2} = 0.57.
\end{aligned}$$

2.3 Copernican's Rule

How can we predict a person's lifespan based on his/her current age?

We expected the prior of lifespan to be normally distributed, says around 80 and with standard deviation of 100

$$t_{max} \sim N(80, 100).$$

Our observations here is the current age t, the question is, what is the likelihood, $f(t|t_{max})$? The Copernican's Principle tells us that the exact timming that we met this person is

random, it can be any time in his/her life. So the likelihood should be uniform distributed

$$f(t|t_{max}) = \frac{1}{t_{max}}.$$

The posterior is given by

$$\begin{aligned} f(t_{max}|t) &= \frac{f(t|t_{max})f(t_{max})}{f(t)} \\ &= \frac{\frac{1}{t}f(t_{max})}{f(t)}. \end{aligned}$$

And the prediction according to the Laplace's Rule is

$$E[t_{max}|t] = \int t_{max} f(t_{max}|t) \mathrm{d}t_{max}.$$

But there is one significant difference here compared to Laplace's Rule example. In the Laplace's Rule example:

$$P \sim U(0,1),\ \mathbf{X}|P \sim Ber(P),\ P|\mathbf{X} \sim Beta(\cdot).$$

We call this a "conjugate model", where the prior and posterior are in the same class of distribution families. For example, if the prior and likelihood is normally distributed, then the posterior is also nomally distributed. The conjugate model is very easy to derive and convenient to calculate, but not all cases are conjugate. In this case, the posterior $f(t_{max}|t)$ is not normal distributed, so the posterior above does not have a closed form. We have to calculate it using a computer.

Chapter 3

Statistical Decision Problem

3.1 Decision problem and solution concepts

There is 3 componenets in a decision problem:

1 States s: The situation that occurs; for example, a rainy day or sunny day.

2 Action a: What decision you make; for example, whether to bring an umbrella or not.

3 Loss function $L(s, a)$: Function for loss given by states and actions, where lower values are better; for example,

$L(s,a)$	**Sunny**	**Rainy**
To bring	5	10
Not to bring	0	100

If we know the state, the decision would be easy. However, the question is that we don't know which state is going to occur.

One way is to address this is to implement the "mini-max" method, which choosees the action with the best worst outcome; in other words, the outcome with lowest maximum loss. In the previous example, the worst outcome if you bring the umbrella is 10, and the worst outcome if you don't bring the umbrella is 100. The decision maker would choose the best option, so he/she would bring the umbrella. This is a relativly conservative approach to make decision because, even if the worst case happens, it won't be too bad. The name of "mini-max" comes from the fact that we first find the maximum loss for each option and then choose the option with the minimum loss.

$$\min_a \max_s L(s, a).$$

Besides "mini-max", one can also use the "mini-min" method, but it's rarely used in practice.

The second method to define optimality is known as "mini-max regret". Regret $R(s, a)$ represents the loss associated with an action a given a particular state s , subtracted by the loss of the optimal action that could have been taken for that state. The decision maker would first calculate the maximum regret for each states, and find the action that

can minimize the maximum regret.

$$\begin{aligned}&\min_a \max_s R(s,a)\\ =&\min_a \max_s [L(s,a) - \max_a L(s,a)].\end{aligned}$$

For instance, on a sunny day, the optimal action is not bringing an umbrella, resulting in a regret of $5-0=5$ for bringing it, and $0-0=0$ for not bringing it. Conversely, on a rainy day, the optimal action is to bring an umbrella, yielding a regret of $10-10=0$ for bringing it, while not bringing it incurs a regret of $100-10=90$.

$R(s,a)$	**Sunny**	**Rainy**
To bring	5	0
Not to bring	0	90

So far, the two methods makes decisions solely based on outcomes, but incorporating probabilities can significantly improve decision making. The third method is call "minimize expected loss", where the expected loss represents the risk.

$$\min_a E_s[L(s,a)] = \min_a Risk(a)$$

For example, if the probability of raining is 0.8, the risk for each action is

$$\begin{aligned}&Risk(\text{to bring}) = 0.2(5) + 0.8(10) = 9\\ &Risk(\text{not to bring}) = 0.2(0) + 0.8(100) = 80.\end{aligned}$$

Since the risk for bringing the umbrella is smaller, the decision maker would decide to bring the umbrella.

3.2 Statistical decision problem

Besides states, actions, and loss function, there's two more componenets in a statistical decision problem. They are data and decision rule. For example, the data could be

$$X = \begin{cases}1, & \text{if sunny,}\\ 0, & \text{if cloudy.}\end{cases}$$

and

$$\begin{aligned}&P(\text{raining}|X=1) = 0.8\\ &P(\text{raining}|X=0) = 0.\end{aligned}$$

Our goal is to find a decision rule that telling us what is our decision and when $X=1$ and $X=0$, respectivly,

$$d : X \mapsto \{\text{to bring, not to bring}\}.$$

When $X=1$,

$$\implies P(\text{raining}) = 0.8 \implies \text{to bring the umbrella.}$$

When $X = 0$,

$$\implies P(\text{raining}) = 0 \implies \text{not to bring the umbrella.}$$

This is an optimal decision rule. We want to make decision based on data, data-driven decision making.

Example.

> **(estimation problem)**
> $$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, 1)$$
> The state μ is the unknown parameter we want to estimate. And, $\hat{\mu}$ is the estimator.
> $$\hat{\mu}|X_1, X_2, \ldots, X_n$$
> Given the data $X_1, X_2, \ldots, X_n$, we want to make guess, or an estimate, what is $\hat{\mu}$? This is exactly the decision rule.
> $$L(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$$
> Given the data, how can I find the best to minimize the lost. Our goal is to find the optimal decision rule, $\hat{\mu}$, to minimize the expected loss in some sense(mini-max, mini-max regret, min risk, ...).
> $$\min_{\hat{\mu}(\cdot)} E[(\hat{\mu}(X_1, X_2, \ldots, X_n) - \mu)^2]$$
> We can also use other definition of loss function, like $L(\mu, \hat{\mu}) = |\mu - \hat{\mu}|$. Depends on the loss funtion we choose, the optimal decision rule could be different.

3.3 Revisiting the Laplace Rule

Now, let's use the statistical decision problem frame work to revisit the Laplace Rule. Suppose the data and states is given by

$$\overbrace{X_1, X_2, \ldots, X_n}^{\text{data}} \overset{iid}{\sim} Ber(\overbrace{p_1}^{states})$$
$$\underbrace{Y_1, Y_2, \ldots, Y_m}_{\text{data}} \overset{iid}{\sim} Ber(\underbrace{p_2}_{states}).$$

The decision rules is

$$d(\mathbf{X}, \mathbf{Y}) \mapsto \{\text{coin 1, coin 2}\}.$$

If the decision maker made the right guess, he/she got 1; If the decision maker made the wrong guess, he/she got 0. So the loss function is

$$L((p_1, p_2), a) = \begin{cases} 0, & \text{if } p_1 > p_2, \ \ a = \text{coin 1} \\ 0, & \text{if } p_2 > p_1, \ \ a = \text{coin 2} \\ 1, & \text{if } p_1 > p_2, \ \ a = \text{coin 2} \\ 1, & \text{if } p_2 > p_1, \ \ a = \text{coin 1.} \end{cases}$$

When $p_1 > p_2$, given $d(\cdot)$,

$$\begin{aligned} Risk((p_1, p_2), d(\cdot)) &= E[Loss((p_1, p_2), d(\cdot))] \\ &= 0 \cdot P(Loss = 0) + 1 \cdot P(Loss = 1) \\ &= P(Loss = 1) \\ &= P(d(\cdot) = \text{coin 2}) \quad \text{(probability of choosing the wrong coin)} \\ &= P(d(\mathbf{X}, \mathbf{Y}) = \text{coin 2}). \end{aligned}$$

The 0-1 lost function is common in hypothetic testing.

Chapter 4

Estimation Problem

Now, let's dive into more details about statistical decision problem. In this chapter, we will try to understand estimation problem as a decision problem.

4.1 Introduction

Let's start from a simple example, suppose we have a coin and filp it for many times

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} Ber(p).$$

Our goal is to estimate p by $\hat{p}$, which is a function of $X_1, X_2, \ldots, X_n$, the data

$$\hat{p}(X_1, X_2, \ldots, X_n).$$

We want $\hat{p}$ to be as close as p as possible. So we define the loss function, for example, using the most common one – square loss

$$L(\hat{p}(\cdot), p) = (\hat{p} - p)^2.$$

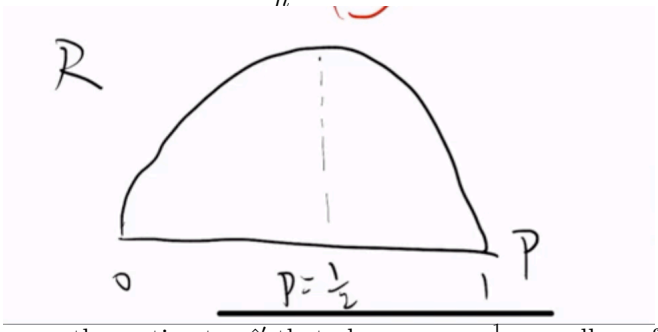
Consequetly, we define risk function as expected loss

$$R(\hat{p}(\cdot), p) = E_{X_1,\ldots,X_n}[L(\hat{p}(\cdot), p)]$$

Example.

Consider another estimator $\hat{p}$

$$\begin{aligned}
\hat{p}(X_1, X_2, \ldots, X_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
R(\hat{p}(\cdot), p) &= E[(\hat{p} - p)^2] \\
&= E[(\bar{X} - p)^2] \\
&= Var(\bar{X}) \quad (\because E\bar{X} = p) \\
&= \frac{1}{n} Var(X_i) \\
&= \frac{1}{n}(p)(1-p)
\end{aligned}$$

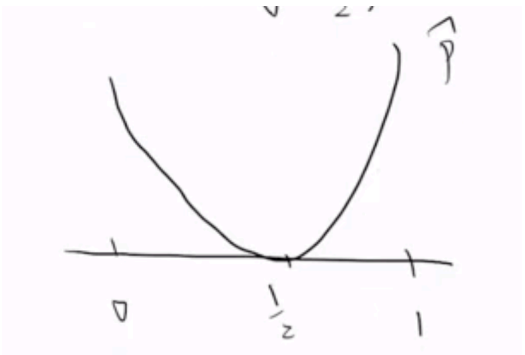


Now, consider another estimator $\hat{p}'$ that always guess $\frac{1}{2}$ regardless of the data.

$$\hat{p}'(X_1, X_2, \ldots, X_n) = \frac{1}{2}.$$

Intuitivly, this is not a good estimator, but it indeed is an estimator.

$$\begin{aligned}
R(\hat{p}'(\cdot), p) &= E[(\frac{1}{2} - p)^2] \\
&= E[(\frac{1}{2} - p)^2]
\end{aligned}$$



We got two estimators $\hat{p}$ and$\hat{p}'$. $\hat{p}$ is better than $\hat{p}'$ in some places, but in other places $\hat{p}'$ is better than $\hat{p}$. The question is – which is the better?

4.2 Choice of loss function

$$L(\hat{P}, P) = (\hat{P}, P)^2$$

There's a few reasons that MSE is a frequently used loss function:

1 Differentiable

2 Convex

3 MSE = Variance + Bias^2

$$\begin{aligned} MSE &= E[(\hat{P} - P)^2] \\ &= E[(\hat{P} - E[\hat{P}] + E[\hat{P}] - P)^2] \\ &= E[(\hat{P} - E[\hat{P}])^2 + 2(\hat{P} - E[\hat{P}])(\hat{P} - E[\hat{P}]) + (E[\hat{P}] - P)^2] \\ &= E[(\hat{P} - E[\hat{P}])^2] + 2E[(\hat{P} - E[\hat{P}])(\hat{P} - E[\hat{P}])] + (E[\hat{P}] - P)^2 \\ &= \underbrace{E[(\hat{P} - E[\hat{P}])^2]}_{Var(\hat{P})} + \underbrace{(E[\hat{P}] - P)^2}_{Bias^2(\hat{P})} \end{aligned}$$

$$L(\hat{P}, P) = |\hat{P}, P|$$

Fact 4.2.1

$$(1, 3, 5, 7, 10^6)$$

Consider the above case. When we use mean square error, $\bar{X_n}$ is sensitive to outlier. But when we use absolute square error, median is not that sensitive to outlier.

The choice of loss function does matter.

Definition 4.2.2: Huber Loss

$$L(\hat{P}, P) = \begin{cases} \frac{1}{2}(\hat{P} - P)^2, & |\hat{P} - P| \leq 1 \\ |\hat{P}| - \frac{1}{2}, & |\hat{P} - P| > 1 \end{cases}$$

Huber loss is an important loss function that conbining square error and absolute error, enjoying the advantages of both. For small deviation, it use square error. For large deviation, it use absolute error. It is robust to outliers, which means it is not sensitive to outliers. Also, Huber Loss is differentiable at zero, but absolute error is not. Besides, there's many others numerical properties that makes Huber Loss a good one.

4.3 Optimality

In the first section, we porpose two estimators, $\hat{p}$ and $\hat{p}'$:

$$\hat{p} = \bar{X_n}$$
$$\hat{p}' = \frac{1}{2}$$

And the corresponding risks are:

$$R(\hat{p}, p) = \frac{1}{n}p(1-p)$$
$$R(\hat{p}', p) = (p - \frac{1}{2})^2$$

Sometimes the $\hat{p}$ is better, but other times $\hat{p}'$ is better, it depends on p. However, the problem is that we don't know the actual p. Otherwise we could calculate the risk and choose the better estimator.

So, how do we define optimal estimator given that risk depends on state (p) in general and the state is unknown? We first try the following definition.

Definition 4.3.1: Possible definition of optimality

$\hat{p}^*$ is optimal if

$$R(\hat{p}^*, p) \leq R(\hat{p}, p), \quad \forall p \in [0,1],\ \hat{p}$$

In other words, the optimal estimator is the one that has a lower risk than any other estimator and in every possible state. Although this is a convincing definition, such estimator may not exists in general because the condition is too strong. Take the below case as an example,

$$\hat{p}' = \frac{1}{2} \implies R(\hat{p}', p = \frac{1}{2}) = 0$$
$$\hat{p}'' = \frac{1}{3} \implies R(\hat{p}'', p = \frac{1}{3}) = 0$$

These two unreasonable estimators have extremly good performace in some states. But, it doesn't perform well in most of the states. The current definition means that if we want to find an estimator with the lowest risk in every state, we need to find an estimator that has zero risk in every possible state:

$$R(\hat{p}^*, p) = 0, \quad \forall p \in [0,1]$$

However, it is obviously impossible. Therefore, we can restrict our target from among all estimators to only "reasonable" estimators, for example, the unbiased estimator, $E[\hat{p}] = \hat{p}$ First,

Definition 4.3.2

$\hat{p}^*$ is optimal if

$$R(\hat{p}^*, p) \leq R(\hat{p}, p), \quad \forall p \in [0,1],\ \hat{p} \text{ s.t. } E[\hat{p}] = \hat{p}$$

In some scenario, such best unbiased estimator may exist.

4.4 Bayes and Mini-max

Consider the following two estimators:

$$R(\hat{p_1}^*, p) < R(\hat{p_2}, p), \quad \text{for some } p$$
$$R(\hat{p_1}^*, p) > R(\hat{p_2}, p), \quad \text{for some } p$$

Given the fact that we don't know p, which estimator should we choose?

We have two weaker alternatives to define optimality. First, if we have some prior $f(p)$, we can choose the estimator by comparing the average risk, also called Bayesian Risk.

Definition 4.4.1: Bayes optimality

An estimator is a Bayes optimal if it minimize the Bayes risk.

$$\int R(\hat{p}, p) f(p) dp$$

where the $f(\cdot)$ is the prior.

The idea is to choose the estimator with smaller risk on average.

The second approach is called the "min-max approach". Here, we can choose the estimator with smaller "maximum risk".

Definition 4.4.2: Mini-max optimality

An estimator is mini-max optimal is it minimize

$$\max_{p \in P} R(\hat{p}, p)$$

The idea is to choose the estimator with smaller risk at the worst-case scenario.

Chapter 5

Emperical Bayes Rule

5.1

In the last chapter, we learned that an estimation problem consists of the following elements:

1. **Parameter**: The unknown quantity we are interested in estimating:

$$\theta \in \mathbb{R}$$

2. **Data**: Observations from a random variable:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_{X|\theta}(X \mid \theta)$$

3. **Prior**: The prior distribution of the parameter:

$$f_\theta(\theta)$$

4. **Estimator**: A decision rule that depends on the data:

$$\hat{\theta}(X_1, X_2, \ldots, X_n)$$

5. **Loss**: A measure of how far the estimator deviate from the true parameter, typically a random quantity depending on the data. An example is the mean square error:

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$$

6. **Risk**: The expected loss over the data, averaging over all possible outcomes. It measures how far the estimator is from the true parameter on average, typically a random quantity depending on the true parameter:

$$\begin{aligned} R(\hat{\theta}, \theta) &= E_{X_1, X_2, \ldots, X_n}[(\hat{\theta} - \theta)^2] \\ &= \int_x (\hat{\theta} - \theta)^2 f(x \mid \theta) dx \end{aligned}$$

If we want to compare the estimators, we can simply compare the risks. However, another problem arised: the expected loss generally depends on parameter (θ) because the distribution of data depends on θ.

To address this, we can compute the "Bayes Risk", which averages the risk over the prior distribution of θ:

$$\int_\theta R_{\hat{\theta}}(\theta) f(\theta) d\theta$$

Now, the Bayes Risk does not depends on the parameter θ but only on the estimator $\hat{\theta}$. We can compare the estimators by comparing their Bayes Risk.

5.2

Definition 5.2.1: Bayes Rule

Bayes Rule is the estimator that minimize the Bayes Risk

$$\hat{\theta}^b = \arg\min_{\hat{\theta}} \int_\theta R_{\hat{\theta}}(\theta) f(\theta) d\theta$$

$$\begin{aligned}
&\int_\theta R_{\hat{\theta}}(\theta) f(\theta) d\theta \\
&= \int_\theta \int_x (\hat{\theta} - \theta)^2 f(x \mid \theta) dx f(\theta) d\theta \\
&= \int_\theta \int_x (\hat{\theta} - \theta)^2 f(x \mid \theta) f(\theta) dx d\theta \\
&= \int_x \int_\theta (\hat{\theta} - \theta)^2 f(\theta \mid x) f(x) d\theta dx \\
&= \int_x \int_\theta (\hat{\theta} - \theta)^2 f(\theta \mid x) d\theta f(x) dx
\end{aligned}$$

The integral inside is the square loss integrated over the posterior. If we can minimize it for all x (find $\hat{\theta}(X_1, \ldots, X_n)$ that minimizes $\int_\theta (\hat{\theta} - \theta)^2 f(\theta \mid x) dx$), then we minimize the double integral and find the Bayes Rules.

5.3

$$\int_\theta (\hat{\theta} - \theta)^2 f(\theta \mid x) d\theta$$

is mathmatically same as

$$\min_{c \in \Re} \int (c - Z)^2 f(Z \mid x) dZ$$

where Z is a random variable with pdf $f(z)$.

$$\min_{c \in \Re} E[(c - Z)^2]$$

where Z is a random variable.

Claim

$$c^* = E[Z]$$

Proof for Claim.

let $c \in R$

$$\begin{aligned}
&E[(Z-c)^2] \\
=&E[(Z-E[Z]+E[Z]-c)^2] \\
=&E[(Z-E[Z])^2] + 2E[(Z-E[Z])(E[Z]-c)] + E[(E[Z]-c)^2] \\
=&Var(Z) + 2(E[Z]-c)\cdot 0 + (E[Z]-c)^2 \\
=&Var(Z) + (E[Z]-c)^2 \\
\geq&Var(Z)
\end{aligned}$$

$$\begin{aligned}
\implies &\min_{c\in\Re} E[(c-\theta)^2 \mid X=x] \\
&c = E[\theta \mid X=x] \\
\implies &\hat{\theta^b} = E[\theta \mid X=x] \\
&= \text{posterior mean}
\end{aligned}$$

In conclusion, how do we find the Bayes Rule, the estimator that minimizes Bayes Risk given prior $f(\theta)$ and likelihood $f(x \mid \theta)$

1. Calculate posterior
$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)}$$
2. Take expected value of θ with respect to $f(\theta \mid x)$
$$\mathbb{E}_{\theta|x}[\theta \mid x] = \int \theta f(\theta \mid x) \mathrm{d}\theta$$

Posterior mean is the Bayes Rule under square loss. If we are using other loss function, we may use difference estimator.

5.4 Example: Laplace Rule

Let's use Laplace Rule as an example to explain Bayes Rule. Recall that the problem here is that we want to estimate the probability of heads when flipping a coin.

$$X_1, X_2, \ldots, X_m \overset{iid}{\sim} Ber(p)$$
$$p \sim Uni(0,1).$$

Our goal is to find an estimator $\hat{p}(X_1, X_2, \ldots, X_m)$ that minimizes the Bayes Risk, which is the Bayes Rule.

The Bayes rule, which equals the posterior mean, is given by

$$\hat{p}^{\mathrm{b}} = \mathbb{E}[p \mid X_1, X_2, \ldots, X_n].$$

In the previous chapter, we show that $p \mid X_1, X_2, \ldots, X_n \sim Beta(\Sigma X_i + 1, n + 1 - \Sigma X_i)$.

$$\begin{aligned}\hat{p}^{\mathrm{b}} &= \frac{\alpha}{\alpha + \beta} \\ &= \frac{\Sigma X_i + 1}{\Sigma X_i + 1 + n + 1 - \Sigma X_i} \\ &= \frac{\Sigma X_i + 1}{n + 2}.\end{aligned}$$

Note that this estimator is similiar, but different from the MLE, which is

$$\hat{p}^{\mathrm{MLE}} = \frac{\Sigma X_i}{n}.$$

In the future, we will show that Bayes Rule is an estimator that better than MLE.

5.5 Estimating the prior

Everything we dicuss above is based on the assumption about the prior distribution. For example, in the case of Laplace Rule, we assumed that p is uniformly distributed within $[0, 1]$. But what if the prior distribution does not follow $\mathrm{Uni}(0, 1)$?

Fact 5.5.1

Bifferent prior distribution would lead to different Bayes Rule. When

$$\begin{aligned} p &\sim \mathrm{Beta}(\alpha, \beta) \\ X_1, X_2, \ldots, X_m &\overset{\mathrm{iid}}{\sim} \mathrm{Ber}(p) \\ p \mid X_1, X_2, \ldots, X_n &\sim \mathrm{Beta}(\Sigma X_i + \alpha, n - \Sigma X_i + \beta). \end{aligned}$$

The Bayes Rule is

$$\begin{aligned}\hat{p}^{\mathrm{b}} &= \frac{\alpha}{\alpha + \beta} \\ &= \frac{\Sigma X_i + \alpha}{\Sigma X_i + \alpha + n - \Sigma X_i + \beta} \\ &= \frac{\Sigma X_i + \beta}{n + \alpha + \beta}.\end{aligned}$$

And the Laplace Rule is the special case when $\alpha = \beta = 1$. ($\mathrm{Beta}(1, 1) = \mathrm{Uni}(0, 1)$.)

How do we choose our prior distribution? Is our prior distribution correct? We can

use "emperical Bayes". The idea is that we do not subjectivly choose the prior distribution but instead estimate it from data and use the estimated prior distribution to proceed the Bayesian analysis.

Soppose we have m coins, each with parameter p_i, where

$$p_1, p_2, \ldots, p_m \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta),$$

and α, β are unknown. For each coin, we toss it n times, then our data $X_1, X_2, \ldots, X_m$ would be given by

$$X_i = \#(\text{heads of } i\text{-th coin out of } n \text{ tosses}),$$

where

$$X_i \mid p_i \sim \text{Bin}(n, p_i).$$

The most straightforward way is to use MLE. We begin from

$$\begin{aligned}
&\underbrace{f(x_i \mid \alpha, \beta)}_{\text{probability of } X_i = x_i} \\
=&\int_0^1 \underbrace{f(x_i \mid p_i, \alpha, \beta)}_{\sim \text{Bin}(n, p_i)} \underbrace{f(p_i \mid \alpha, \beta)}_{\sim \text{Beta}(\alpha, \beta)} \mathrm{d}p_i \\
=&\int_0^1 \binom{n}{x_i} p_i^{x_i}(1-p_i)^{n-x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1}\mathrm{d}p_i \\
=&\binom{n}{x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \int_0^1 p_i^{x_i+\alpha-1}(1-p_i)^{n-x_i+\beta-1}\mathrm{d}p_i
\end{aligned}$$

Then we can reparameterize it into

$$\begin{aligned}
=&\binom{n}{x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \frac{\Gamma(x_i+\alpha)\cdot\Gamma(n-x_i+\beta)}{\Gamma(x_i+\alpha+n-x_i+\beta)} \\
&\underbrace{\int_0^1 \frac{\Gamma(x_i+\alpha+n-x_i+\beta)}{\Gamma(x_i+\alpha)\cdot\Gamma(n-x_i+\beta)} p_i^{x_i+\alpha-1}(1-p_i)^{n-x_i+\beta-1}\mathrm{d}p_i}_{\sim \text{Beta}(x_i+\alpha, n-x_i+\beta)} \\
=&\binom{n}{x_i} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \frac{\Gamma(x_i+\alpha)\cdot\Gamma(n-x_i+\beta)}{\Gamma(\alpha+n+\beta)}.
\end{aligned}$$

Finally, estimate α, β by

$$\max_{\alpha,\beta} \sum_{i=1}^{n} \ln f(x_i \mid \alpha, \beta).$$

In summary, we do not make too much assumptions about the prior distribution, but simply set it as a beta distribution with unknown parameters *alpha* and β. We use the likelihood to derive marginal distribution of X_i. Using the marginal PDF, we can find the $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}}$ through MLE.

Recall that the Bayes Rule of $\hat{p_i}$ is

$$\hat{p}_i^{\text{b}} = \frac{x_i + \alpha}{n + \alpha + \beta}.$$

The emperical Bayes Rule is simply replace α and β with $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}}$

$$\hat{p}_i^{\text{EB}} = \frac{x_i + \hat{\alpha}^{\text{MLE}}}{n + \hat{\alpha}^{\text{MLE}} + \hat{\beta}^{\text{MLE}}}.$$

5.6 Application: User ranking

The emperical Bayes method is very commonly use in real-world scenario, particularly for ranking purposes. For example, suppose you run a e-commerce platform and want to calculate the complaint rate for each seller on your platform to rank them accordingly. This method allows you to empirically estimate the prior distribution. By combining prior information with the observed data, you can achieve more accurate and reliable rankings.

$$n_i = \#(\text{transaction for seller } i)$$
$$x_i = \#(\text{complaint for seller } i)$$
$$\hat{p}_i^{\text{EB}} = \frac{x_i + \hat{\alpha}^{\text{MLE}}}{n_i + \hat{\alpha}^{\text{MLE}} + \hat{\beta}^{\text{MLE}}}.$$

And this method is better than using the sample mean

$$\hat{p}_i^{\text{EB}} = \frac{x_i}{n_i}.$$

We will demonstrate it in a future chapter.

Chapter 6

Finishing The Estimation Problem

6.1 A taste of mini-max

In the last chapter, we saw that **the posterior mean is the Bayes rule under square loss**. As long as we can calculate the posterior, finding Bayes rule is usually not hard. However, when it comes to mini-max, there is no universal procedure to find the mini-max estimator

$$\min_{\hat{\theta}} \max_{\theta} R(\hat{\theta}, \theta).$$

One way to prove some estimator is mini-max is through the following proposition:

Proposition 6.1.1

If $\hat{\theta}$ is a Bayes rule with constant risk ($R(\hat{\theta}, \theta) \equiv c \ \forall \theta$), then $\hat{\theta}$ is mini-max.

Proof for Proposition.

(Proof by contradiction)
Suppose $\hat{\theta}$ is a Bayes rule with constant risk but not mini-max

$$\begin{aligned}
&\implies \exists \hat{\theta}' \quad \text{s.t.} \max_{\theta} R(\hat{\theta}', \theta) \leq \max_{\theta} R(\hat{\theta}, \theta) \\
&\implies \text{Bayes risk of } \hat{\theta}' < \max_{\theta} R(\hat{\theta}', \theta) \quad (\because \text{average is smaller than max}) \\
&\qquad\qquad \leq \max_{\theta} R(\hat{\theta}, \theta) \\
&\qquad\qquad = \text{Bayes risk of } \hat{\theta} \quad (\because \hat{\theta} \text{ has constant risk}) \\
&\implies (\rightarrow\!\!\!\!\!\times\!\!\!\!\!\leftarrow)
\end{aligned}$$

6.2 Admissiblility

Definition 6.2.1: Admissible

$\hat{\theta}$ is admissible if there is no other estimator $\hat{\theta}'$ such that

$$R(\hat{\theta}', \theta) \leq R(\hat{\theta}, \theta) \quad \forall \theta$$

and

$$R(\hat{\theta}', \theta) < R(\hat{\theta}, \theta) \quad \text{for some } \theta.$$

If and estimator $\hat{\theta}$ is admissible, it means that it's always not worse than another estimator $\hat{\theta}'$. And if an estimator is not admissible, we call it inadmissable.

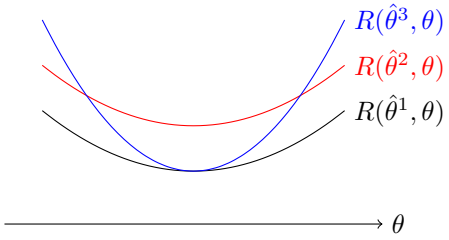


Figure 6.1: Visualizing admissiblility: $\hat{\theta}^1$ is admissable; $\hat{\theta}^2$ and $\hat{\theta}^3$ is inadmissable.

Remark.

> If an estimator is not admissible, it is bad. But even if some estimator is admissible, it does not necessarily means it is good.

Example.

> Consider an unreasonable estimator $\hat{\theta} = 3$. Since no other estimator can achieve zero risk for $\theta = 3$, $\hat{\theta}$ is admissible. But $\hat{\theta} = 3$ is still a very bad estimator.

6.3 Stein's Paradox

Theorem 6.3.1: Stein's Paradox

MLE for the mean of normal distribution is not admissible.

Consider estimating the mean of many normal distribution:

$$\begin{aligned} x_1 &\sim N(\mu_1, 1) \\ x_2 &\sim N(\mu_2, 1) \\ &\vdots \\ x_n &\sim N(\mu_n, 1) \end{aligned}$$

Our goal is the esitmate $\mu_1, \ldots, \mu_n$, there are a few way to do it:

1. MLE
$$\hat{\mu_i}^{\text{MLE}} = X_i$$

2. Bayesian
$$\begin{aligned}
\mu_1, \ldots, \mu_n &\overset{\text{iid}}{\sim} \underbrace{N(0, \gamma^2)}_{\substack{\text{Prior distribution} \\ \text{with } \gamma \text{ known}}} \\
\hat{\mu_i}^{\text{B}} &= \mathbb{E}[\mu_i \mid X_1, X_2, \ldots X_n] \\
&= \mathbb{E}[\mu_i \mid X_i] \\
&= (\frac{\gamma^2}{\gamma^2+1})X_i + (\frac{1}{\gamma^2+1})0 \\
&= (\frac{\gamma^2}{\gamma^2+1})X_i \\
&= (1 - \frac{1}{\gamma^2+1})X_i
\end{aligned}$$

3. Emperical Bayes
Simply replace the parameter with estimated parameter
$$\begin{aligned}
\mu_i \overset{\text{iid}}{\sim} N(0, \gamma^2) \quad &\Longleftrightarrow \quad X_i = \mu_i + \epsilon_i \\
X_i \mid \mu_i \sim N(\mu_i, 1) \quad &\qquad\quad \epsilon_i \sim N(0,1), \mu_i \perp\!\!\!\perp \epsilon_i \\
&\Longrightarrow \mathbb{V}(X_i) = \gamma^2 + 1 \\
s^2 &= \frac{1}{n-1}\sum (X_i - \bar{X})^2 \\
&\mathbb{E}[s^2] = \gamma^2 + 1 \\
&\hat{\gamma}^2 = s^2 - 1 \\
\Longrightarrow \hat{\mu_i}^{\text{EB}} &= (\frac{\hat{\gamma}^2}{\hat{\gamma}^2+1})X_i \\
&= (\frac{s^2-1}{s^2-1+1})X_i \\
&= (\frac{s^2-1}{s^2})X_i
\end{aligned}$$

4. James-Stein Estimator
Estimate $\frac{1}{\gamma^2+1}$
$$\begin{aligned}
\sum X_i &\sim (\gamma^2+1)\chi^2_n \\
\mathbb{E}[\frac{n-2}{\sum X_i^2}] &= \frac{1}{\gamma^2+1} \\
\hat{\mu_i}^{\text{JS}} &= (1 - \frac{n-2}{\sum X_i^2})X_i
\end{aligned}$$

For all $\mu_1, \cdots \mu_n$,

$$\hat{\mu_i}^{\text{JS}} = \begin{pmatrix} \hat{\mu_i}_1^{\text{JS}} \\ \vdots \\ \hat{\mu_i}_n^{\text{JS}} \end{pmatrix}$$

has smaller MSE than $\hat{\mu_i}^{\text{MLE}}$ for $n \geq 3$.

Corollary 6.3.2

MLE is admissible.

Chapter 7

Binary Decision Problem

7.1 Introduction

1. Hypothesis testing

$$X_1, X_2, \ldots, X_n \sim \mathrm{N}(\mu, 1)$$

$$\mathbf{H}_0 : \mu = 0$$
$$\mathbf{H}_1 : \mu \neq 0$$

$$d : X_1, \ldots, X_n \mapsto \{\text{Reject } \mathbf{H}_0, \text{Do not reject } \mathbf{H}_0\}$$

2. Mixture problem: the distribution of data is a mix of several distributions

$$X = gW + (1-g)Z, \quad \text{where } g \in \{0, 1\}$$
$$= \begin{cases} W, & \text{when } g = 1 \\ Z, & \text{when } g = 0 \end{cases}$$

where

$$W \sim \mathrm{N}(\mu_1, 1)$$
$$Z \sim \mathrm{N}(\mu_2, 1)$$

For instance, consider a dataset of prople's height. Heights of both males and females typically follow normal distributions, but with different mean values (μ_1 for males, μ_2 for females). The overall dataset is thus a mixture of these two distributions. In this scenario, we observe the height (X) but not the gender (g)

Given a set of observations $X_1, X_2, ..., X_n$, the goal is to classify each observation into one of the two groups (male or female in the height example):

$$d : X_1, \ldots, X_n \mapsto \{0, 1\}$$

3. Classification problem
$$(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$$
Both Y and X are observed. Our goal is to predict Y_{n+1} given X_{n+1}.

   The classification problem is very similar to the mixture problem, but the key difference lies in the availablility of labels in the data. Classification problems have associated labels, while mixture problems do not. Consequently, classification is considered supervised learning, and mixture problems are categorized as unsupervised learning.

7.2 A review of hypothesis testing

Hypothesis is statement about unknown parameter, θ. For example, it could be $\theta = 0$ or $\theta > 1$. Hypothesis testing involves using sample data to determine whether the hypothesis is correct. There are two types of hypotheses:

1. **Null hypothesis** ($\mathbf{H}_0$): This is the default assumption or the status quo
2. **Alternative hypothesis** ($\mathbf{H}_1$ or $\mathbf{H}_a$): This is the claim we want to test. It is the opposite of the null hypothesis.

For example:

1. $\mathbf{H}_0$: $\theta = 0$
2. $\mathbf{H}_1$: $\theta \neq 0$

Given the dataset $X_1, \ldots, X_n$, the goal is to construct a decision rule to determine whether to reject or fail to reject the null hypothesis based on observed data.

$$d(X_1, \ldots, X_n) \in \{\text{reject } \mathbf{H}_0, \text{ do not reject } \mathbf{H}_0\}.$$

One of the very common hypothesis testing is "t-test". Given a dataset of independent and identically distributed normal random variables:

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{N}$$

We aim to deterimine whether the population mean, μ, significantly differs from zero. If the sample mean, $\bar{X}_n$, is sufficiently distant from zero—that is, if $\mid \bar{X_n} - 0 \mid > c$ for some constant $c \in \mathbb{R}$—we reject the null hypothesis. Otherwise, we fail to reject it. Formally, the decision rule is:

$$d = \begin{cases} \text{reject } \mathbf{H}_0 & \text{if } \mid \bar{X_n} \mid > c \\ \text{do not reject} \mathbf{H}_0 & \text{o.w.} \end{cases}$$

Now, the problem is how do we determine the constant c?

The choice of c actually depends on properties of test. If threshold c becomes larger, what happens to the probability of type-I error and type-II error? When c increase, the sample mean need to be further alway from zero, so that we can reject. In other words, we are more conservative and tend to not reject unless the sample mean are too away from zero.

In this case, the probability of type-I error would decrease and the probability of type-II error would increase. Consequently, there's a trade-off between probability of type-I error and type-II error.

	$\mathbf{H}_0$ is true	$\mathbf{H}_1$ is true
reject $\mathbf{H}_0$	Type-I Error (False Positive)	✓
do not reject $\mathbf{H}_0$	✓	Type-II Error (False Negative)

Now, let's link hypothesis testing with what we learned in estimation. The counterparts of loss in estimation are Type-I Error and Type-II Error in testing. In both scenarios, our decision may not match the true state. Therefore, just as losses are random, errors are also random. The counterparts of risk in estimation is the probability of Type-I Error and the probability of Type-II Error in testing. In testing, risk generally depends on the true parameter, and the probabilities of errors also depend on the parameter (threshold). If the parameter or data generation process differs, the same testing problem could result in different probabilities of errors. Thus, like risk, these probabilities of errors are non-random, or in other words, deterministic.

Estimation	Testing
Loss= $(\hat{\theta} - \theta)^2$	Type-I Error and Type-II Error
Risk= $\mathbb{E}[(\hat{\theta} - \theta)^2]$	P(Type-I Error) and P(Type-II Error)

7.3 Le Cam Lower Bound

Definition 7.3.1: Simple Hypothesis

The simple hypothesis refers to hypothesis that only contains one point in both null hypothesis and alternative hypothesis.

Example.

> This is a simple hypothesis:
>
> $$\mathbf{H}_0 : \mu = 0$$
> $$\mathbf{H}_1 : \mu = 1.$$
>
> This is not simple hypothesis as $\mathbf{H}_1$ contains more than one point:
>
> $$\mathbf{H}_0 : \mu = 0$$
> $$\mathbf{H}_1 : \mu \neq 0.$$

Consider the following simple hypothesis testing scenario:

- $\mathbf{H}_0$: The data is drawn from the probability distribution $f(x \mid \theta_0)$.
- $\mathbf{H}_1$: The data is drawn from the probability distribution $f(x \mid \theta_1)$.

Given a sample of n independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$, we aim to determine whether the underlying population distribution is $f(x \mid \theta_0)$ or $f(x \mid \theta_1)$.

For instance, consider the task of determining whether a dataset is drawn from $\mathrm{N}(0, 1)$ or $\mathrm{N}(10, 1)$. Given the sample data is $\{0.5, -0.2, -0.6, 0.5, 0.3\}$, you probably would guess it is orginated from $\mathrm{N}(0, 1)$. Now consider another task of determining whether a dataset is drawn from $\mathrm{N}(0, 1)$ or $\mathrm{N}(0.01, 1)$. Obviously, this one is much harder than the last one because the distribution are much similar. And the harder means it larger probabilities of type-I error and type-II error.

But how do we measure the similarity between two distributions? Can we define a distance?

Definition 7.3.2: Total Variation Distance

Let $f(x)$ and $g(x)$ be two pdf. Then the total variation distance between two distributions is defined as:

$$\|f - g\|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathbb{R}} |f(x) - g(x)| \, \mathrm{d}x$$

Lemma 7.3.3

$$\begin{aligned} \|f - g\|_{\mathrm{TV}} &= \frac{1}{2} \int_{\mathbb{R}} |f(x) - g(x)| \, \mathrm{d}x \\ &= \sup_{A \subset \mathbb{R}} \int_A f(x) - g(x) \, \mathrm{d}x \end{aligned}$$

Proof for Lemma

$$\begin{aligned} & \sup_{A \subset \mathbb{R}} \int_A f(x) - g(x) \, \mathrm{d}x \\ = & \sup_{A \subset \mathbb{R}} [\int_A f(x) \, \mathrm{d}x - \int_A g(x) \, \mathrm{d}x] \\ = & \sup_{A \subset \mathbb{R}} [P_f(A) - P_g(A)] \\ = & \end{aligned}$$

In Figure 7.1, since the distributions are quite different, the total variation distance is closed to 1. In contrast, in Figure 7.2, the distributions are very similar, leading to a total variation distance is closed to 0.

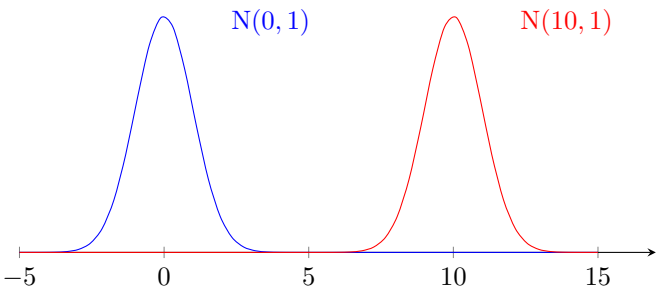


Figure 7.1: Large total variation distance

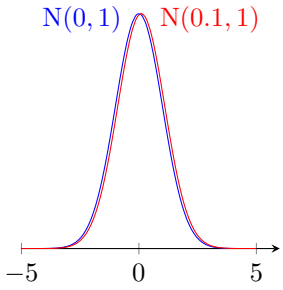


Figure 7.2: Small total variation distance

Theorem 7.3.4: Le Cam's Lower Bound

For simple hypothesis, the sum of probabilities of two error is must higher than some lower bound.

$$P(\text{Type-I Error}) + P(\text{Type-II Error}) \geq 1 - \|f - g\|_{\text{TV}}$$

Proof.

$$\begin{aligned} A &= \text{rejection region} \\ &= \{x \in X \mid \text{you reject when } x = X\} \end{aligned}$$

$$\begin{aligned} P(\text{Type-I Error}) &= P_f(A) \\ P(\text{Type-II Error}) &= P_g(A^c) \end{aligned}$$

$$\begin{aligned} &P(\text{Type-I Error}) + P(\text{Type-II Error}) \\ =&P_f(A) + P_g(A^c) \\ =&P_f(A) + 1 - P_g(A) \\ =&1 - (P_g(A) - P_f(A)) \\ \geq&1 - \|f - g\|_{\text{TV}} \quad (\because P_g(A) - P_f(A) \leq \|f - g\|_{\text{TV}}) \end{aligned}$$

□

The Le Cam's Lower Bound is quite intuitive. When distributions are similar, the total variation distance is small. This make it difficult to disinguish between the distributions, leading to inevitable errors and a relativly high lower bound. Conversely, if the distributions are different, the total variation distance is large. This make it easy to regonize the distributions, resulting a relatively low lower bound.

Chapter 8

The Classification Problem

8.1 Revisiting the judge's decision problem

Prediction ╲ Truth	$y = 1$	$y = 0$
$\hat{y} = 1$	0	c_2 (False Positive)
$\hat{y} = 0$	c_1 (False Negative)	0

Let's fix c_1, $c_2 \in \mathbb{R}$, and take a look on how probability of guilty $(p) = P(y = 1)$ affect decision.

$$L(\hat{y}, y) = \begin{cases} 0, & \hat{y} = y \\ c_1, & \hat{y} \neq y, y = 0 \ (\implies \hat{y} = 1, y = 0) \\ c_2, & \hat{y} = 0, y = 1 \end{cases}$$

$$R(\hat{y} = 1, p) = P(y = 0) \cdot c_1 = (1 - p)c_1$$
$$R(\hat{y} = 0, p) = P(y = 1) \cdot c_2 = pc_2$$

Choose $\hat{y} = 1$ if

$$\begin{aligned} &(1 - p)c_1 < pc_2 \\ \implies &c_1 < p(c_1 + c_2) \\ \implies &p > \frac{c_1}{c_1 + c_2}. \end{aligned}$$

We only perdict $\hat{y} = 1$ when the percentage of guilty (p) is high. And the cutoff $\frac{c_1}{c_1+c_2}$ depends on the judges' preference. When $c_1 = c_2$,

$$p > \frac{1}{2} \implies \hat{y} = 1.$$

The decision rule becomes predicting someone is guilty whenever the percentage of guilty is more than $\frac{1}{2}$.

8.2 Conditioning on X

In applying the dicussion in previous section to real-world questions, a common problem arises: p is unknown and needs to be estimated. Considering the following data

y	Gendr	Age
1	M	20
0	M	25
0	M	31
1	M	40
0	F	42

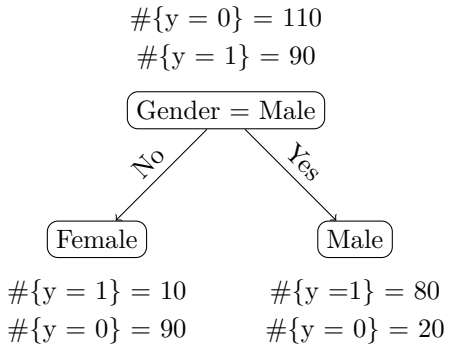
Here, p can be estimated as:

$$p = P(y = 1) = \frac{1}{n}\sum_{i=1}^{N} y_i.$$

But how can we include the covariates such as gender and age to make better decisions? Consider the "covariate-specific" decision rule:

$$P(y = 1 \mid X = x) \geq \frac{c_1}{c_1 + c_2}.$$

We can use these covariates to improve our decision-making. Consider the following scenario:



If $c_1 = c_2$, our rule would be

$$\begin{cases} \hat{y} = 1 & \text{for male} \\ \hat{y} = 0 & \text{for female} \end{cases} \text{ because } \begin{cases} p = 0.8 & \text{for male} \\ p = 0.1 & \text{for female} \end{cases}$$

However, if we do not condition on covariates, the probability of someone is guilty is $p = \frac{90}{90+110} = 0.45$. This means that if we send someone to jail, there's a 45% change of false positive. But if we know the person's gender is male, the probability of a false positive reduce to only 20%. This demonstrates how conditioning on covariates can significantly improve decision-making.

Definition 8.2.1: Entropy for Bernoulli distribution

$X \sim \text{Ber}(p)$

$$H(X) = -[p \ln p + (1-p)\ln(1-p)]$$

Conditioning on covariate is useful because it reduces entropy, a measure of how random or unpredictable a distribution is. If $p = \frac{1}{2}$, the distribution is highly unpredictable because there's an equal chance of X being 0 or 1. However, when p is closed to 0 (or 1), the distribution is highly predictable because there's a high probability of X being 0 (or 1). The notion of entropy captures how easy it is to predcit a binary outcome.

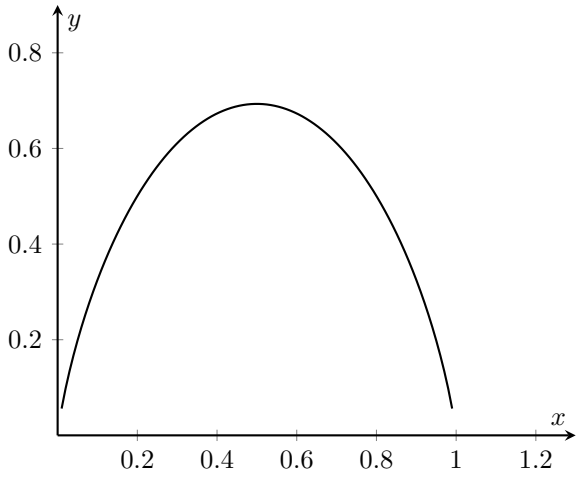


Figure 8.1: Entropy for Bernoulli distribution

8.3 Receiver operator characteristics curves

What's the effect of c?

	$y=1$	$y=0$
$\hat{y}=1$	TP	FP
$\hat{y}=0$	FN	TN

Definition 8.3.1: True Positive Rate

The true positive rate represents how many of observations you successfully labeled as positive amoung all positive observations.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

We call this "recall". The higher the better.

Definition 8.3.2: False Positive Rate

The false positive rate represents how many of observations you falsly labeled as positive amoung all negative observations.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

The lower the better.

There is a trade-off between high TRP or high FPR. If we increase the cutoff c,

$$\implies \hat{y} = 1 \text{ less often}$$
$$\implies \text{TPR decrease; FPR decrease}$$

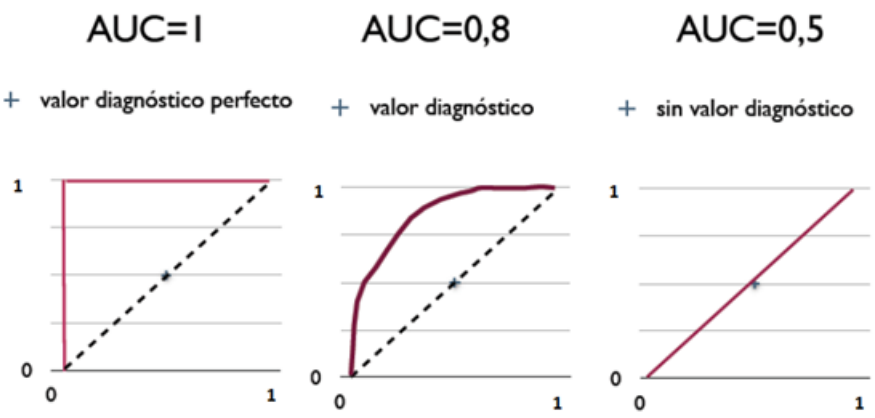
If we decrease the cutoff c, the opposite occurs,

$$\implies \hat{y} = 1 \text{ more often}$$
$$\implies \text{TPR increase; FPR increase}$$

Both FPR and TPR are increasing function of c. This means that as c increase, both TPR and FPR will also increase, and vice versa. It is impossible that one of them increase while the other decrease. This trade-off minor the Type-I and Type-II errors in hypothesis testing: if we increase the cutoff, the type-I error may decrease, but the type-II error will inevitablly increae.

We can use the Receiver Operating Characteristic (ROC) Curve, TPR on the y-axis and FPR on the x-axis to visualize the performance of a classification model across various threshold settings. This curve illustrates the trade-offs between detecting true positives and incorrectly identifying false positives as the classification threshold changes. By examining the ROC curve, we can determine how the model's performance varies with different cutoff values. The curve allows us to visually assess the balance between sensitivity and specificity at different levels, helping us choose an appropriate threshold based on our specific needs.

Then, we can calculate the Area Under the Curve (AUC). The AUC is a single scalar value representing the overall ability of the model to discriminate between positive and negative classes. An AUC of 0.5 indicates no discriminative ability (random guessing), while an AUC of 1.0 indicates perfect classification. The higher the AUC, the better the model is at distinguishing between the positive and negative classes.


AUC=1
+ valor diagnóstico perfecto
1
0
0
1
AUC=0,8
+ valor diagnóstico
1
0
0
1
AUC=0,5
+ sin valor diagnóstico
1
0
0
1

Chapter 9

Examples

9.1 Theorem System

Definition 9.1.1: Definition Name

A defintion.

Theorem 9.1.2: Theorem Name

A theorem.

Lemma 9.1.3: Lemma Name

A lemma.

Fact 9.1.4

A fact.

$$E = mc^2 \tag{9.1}$$

$$\vec{F} = m\vec{a} \tag{9.2}$$

$$\int_0^1 x^2\,dx = \frac{1}{3} \tag{9.3}$$

Corollary 9.1.5

A corollary.

Proposition 9.1.6

A proposition.

Claim

A claim.

Proof for Claim.

A reference to Theorem 9.1.2

Proof. Veniam velit incididunt deserunt est proident consectetur non velit ipsum voluptate nulla quis. Ea ullamco consequat non ad amet cupidatat cupidatat aliquip tempor sint ea nisi elit dolore dolore.

Laboris labore magna dolore eiusmod ea ex et eiusmod laboris. Et aliquip cupidatat reprehenderit id officia pariatur. □

Example.

> Nostrud esse occaecat Lorem dolore laborum exercitation adipisicing eu sint sunt et. Excepteur voluptate consectetur qui ex amet esse sunt ut nostrud qui proident non. Ipsum nostrud ut elit dolor. Incididunt voluptate esse et est labore cillum proident duis.

Some remark.

Remark.

> Some more remark.

9.2 Pictures

Figure 9.1: Waterloo, ON