Section 8

Testing simple hypotheses. Bayes decision rules.

Let us consider an i.i.d. sample $X_1, \ldots, X_n \in \mathcal{X}$ with unknown distribution $\mathbb{P}$ on $\mathcal{X}$. Suppose that the distribution $\mathbb{P}$ belongs to a set of k specified distributions, $\mathbb{P} \in \{\mathbb{P}_1, \ldots, \mathbb{P}_k\}$. Then, given a sample $X_, \ldots, X_n$ we have to decide among k simple hypotheses:

$$\begin{cases} H_1: & \mathbb{P} = \mathbb{P}_1 \\ H_2: & \mathbb{P} = \mathbb{P}_2 \\ \quad\vdots & \\ H_k: & \mathbb{P} = \mathbb{P}_k. \end{cases}$$

In other words, we need to construct a decision rule

$$\delta : \mathcal{X}^n \to \{H_1, \ldots, H_k\}.$$

Let us note that sometimes this function δ will be random because when several hypotheses are 'equally likely' it will make sense to pick among them randomly. This idea of a *randomized* decision rule will be explained more precisely in the following lectures, but for now we will simply think of δ as a function of the sample.

Suppose that the ith hypothesis is true, i.e. $\mathbb{P} = \mathbb{P}_i$. Then the probability that a decision rule δ will make an error is

$$\mathbb{P}(\delta \neq H_i | H_i) = \mathbb{P}_i(\delta \neq H_i),$$

which is called the *error of type* i or type i error. When $k = 2$, i.e. we consider two hypotheses H_1 and H_2, the error of type 1

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1)$$

is also called the *size* or *level of significance* of the decision rule δ and

$$\beta = 1 - \alpha_2 = 1 - \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2(\delta = H_2)$$

is called the *power* of δ.

In order to construct a good decision rule, we need to decide how to compare decision rules. Ideally, we would like to make the errors of all types as small as possible. However, typically there is a trade-off among errors and it is impossible to minimize them simultaneously. A decision rule δ will create a partition of space $\mathcal{X}^n$ into k disjoint subsets $A_1, \ldots, A_k$ such that

$$\delta(X_1, \ldots, X_n) = H_j \quad \text{if and only if} \quad (X_1, \ldots, X_n) \in A_j.$$

Increasing a set A_j will decrease the error of type j since

$$\alpha_j = \mathbb{P}(A_j^c) = 1 - \mathbb{P}(A_j)$$

and, therefore, in this sense k simple hypotheses compete with each other. Of course, it is possible to give an example in which all errors are zero. For example, if all distributions $\mathbb{P}_1, \ldots, \mathbb{P}_k$ concentrate on disjoint subsets of $\mathcal{X}$ then one observation is enough to predict the correct hypothesis with no error.

One way to compare decision rules would be to assign weights $\xi(1), \ldots, \xi(k)$ to the hypotheses and consider a *weighted error*

$$\xi(1)\alpha_1 + \ldots \xi(k)\alpha_k = \xi(1)\mathbb{P}(\delta \neq H_1|H_1) + \ldots + \xi(k)\mathbb{P}(\delta \neq H_k|H_k).$$

In the next section we will construct decision rules that minimize this weighted error.

In the case of two simple hypotheses H_1 and H_2 it is more common to construct 'good' decision rules based on a different criterion. Before we describe this criterion, let us first see that in many practical problems different types of errors have very different meanings.

Example. Suppose that a medical test is done to determine if a patient is sick. Then based on the data from the test we have to decide between two hypotheses:

$$H_1 : \text{ positive}; H_2 : \text{ negative}.$$

Then the error of type one $\mathbb{P}(\delta = H_2|H_1)$ means that we determine that the patient is sick when he is not and the error of type two $\mathbb{P}(\delta = H_1|H_2)$ means that we determine that a patient is not sick when he is. Clearly, these errors are of a very different nature. In the first case a patient will not get a necessary treatment. In the second case a patient might get unnecessary and potentially harmful treatment. However, in the second case additional tests can be done whereas in the first case the sickness may be completely overlooked. This means that it may be more important to control the error of type 1 in this case.

Example. Radar missile detection/recognition. Suppose that based on a radar image we decide between a missile and a passenger plane:

$$H_1 : \text{ missile}, H_2 : \text{ not missile}.$$

Then the error of type one $\mathbb{P}(\delta = H_2|H_1)$, means that we will ignore a missile and error of type two $\mathbb{P}(\delta = H_2|H_1)$, means that we will possibly shoot down a passenger plane (which happened before). It depends on the situation to decide which error is more important to control.

Another example could be when 'guilty' or 'not guilty' verdict in court is decided based on some data. Presumption of innocence means that 'no guilty' hypothesis is a more important *null hypothesis* and the error of type $\mathbb{P}(\text{'guilty'}|\text{'not guilty'})$ should be controlled. When a drug company comes up with a new drug, it is their responsibility to prove that a drug works significantly better than a sugar pill, so a 'more important' null hypothesis in this case is that a drug does not work better.

These examples illustrate that in many situations a particular hypothesis is more important in a sense that the error corresponding to this hypothesis should be controlled. We will assume that H_1 is this hypothesis. Let $\alpha \in [0,1]$ be the largest possible error of type one that we are willing to accept, which means that we will only consider decision rules in the class

$$K_\alpha = \{\delta : \alpha_1 = \mathbb{P}_1(\delta \neq H_1) \le \alpha\}.$$

It now makes sense that among all decision rules in this class we should try to find a decision rule that makes the error of type two, $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$, as small as possible. We will show how to construct such decision rules in the following lectures but, first, we will construct decision rules that minimize the weighted error.

Bayes decision rules.

Given hypotheses $H_1, \ldots, H_k$ let us consider k nonnegative weights $\xi(1), \ldots, \xi(k)$ that add up to one $\sum_{i=1}^k \xi(i) = 1$. We can think of weights ξ as a priori probability on the set of k hypotheses that represent their relative importance. Then the *Bayes error* of a decision rule δ is defined as

$$\alpha(\xi) = \sum_{i=1}^k \xi(i)\alpha_i = \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i),$$

which is simply a weighted error. We would like to make the Bayes error as small as possible.

Definition: *Decision rule δ that minimizes $\alpha(\xi)$ is called a* Bayes decision rule.

Next theorem constructs Bayes decision rules in terms of p.d.f. or p.f. of $\mathbb{P}_i, 1 \le i \le k$.

Theorem. *Assume that each distribution $\mathbb{P}_i$ has p.d.f or p.f. $f_i(x)$. A decision rule δ that predicts H_j when*

$$\xi(j)f_j(X_1)\ldots f_j(X_n) = \max_{1\le i\le k} \xi(i)f_i(X_1)\ldots f_i(X_n)$$

is a Bayes decision rule.

In other words, we choose hypotheses H_j if it maximizes the weighted likelihood function

$$\xi(i)f_i(X_1)\ldots f_i(X_n)$$

among all hypotheses. If this maximum is achieved simultaneously on several hypotheses we can pick any one of them, or at random.

Proof. Let us rewrite the Bayes error as follows:

$$\begin{aligned}
\alpha(\xi) &= \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i) \\
&= \sum_{i=1}^k \xi(i) \int I(\delta \neq H_i) f_i(x_1)\ldots f_i(x_n) dx_1 \ldots dx_n
\end{aligned}$$

$$
\begin{aligned}
&= \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\ldots f_i(x_n)\Big(1 - I(\delta = H_i)\Big) dx_1 \ldots dx_n \\
&= \sum_{i=1}^{k} \xi(i) \underbrace{\int f_i(x_1)\ldots f_i(x_n) dx_1 \ldots dx_n}_{\text{this joint density integrates to 1 and } \sum \xi(i) = 1} \\
&\quad - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\ldots f_i(x_n) I(\delta = H_i) dx_1 \ldots dx_n \\
&= 1 - \int \sum_{i=1}^{k} \xi(i) f_i(x_1)\ldots f_i(x_n) I(\delta = H_i) dx_1 \ldots dx_n.
\end{aligned}
$$

To minimize this Bayes error we need to maximize this last integral, but we can actually maximize the sum inside the integral

$$\xi(1) f_1(x_1)\ldots f_1(x_n) I(\delta = H_1) + \ldots + \xi(k) f_k(x_1)\ldots f_k(x_n) I(\delta = H_k)$$

by choosing δ appropriately. For each $(x_1, \ldots, x_n)$ decision rule δ picks only one hypothesis which means that only one term in this sum will be non zero, because if δ picks H_j then only one indicator $I(\delta = H_j)$ will be non zero and the sum will be equal to

$$\xi(j) f_j(x_1)\ldots f_j(x_n).$$

Therefore, to maximize the integral δ should simply pick the hypothesis that maximizes this expression, exactly as in the statement of the Theorem. This finishes the proof.

□

Let us write down a Bayes decision rule in the case of two simple hypotheses H_1, H_2. For simplicity of notations, given a sample $X = (X_1, \ldots, X_n)$ we will denote the joint p.d.f. or p.f. by

$$f_i(X) = f_i(X_1)\ldots f_i(X_n).$$

Then the Bayes decision rule that minimizes the weighted error

$$\alpha = \xi(1)\mathbb{P}_1(\delta \neq H_1) + \xi(2)\mathbb{P}_2(\delta \neq H_2)$$

is given by

$$\delta = \begin{cases} H_1 : & \xi(1) f_1(X) > \xi(2) f_2(X) \\ H_2 : & \xi(2) f_2(X) > \xi(1) f_1(X) \\ H_1 \text{ or } H_2 : & \xi(1) f_1(X) = \xi(2) f_2(X). \end{cases}$$

Equivalently,

$$\delta = \begin{cases} H_1 : & \frac{f_1(X)}{f_2(X)} > \frac{\xi(2)}{\xi(1)} \\ H_2 : & \frac{f_1(X)}{f_2(X)} < \frac{\xi(2)}{\xi(1)} \\ H_1 \text{ or } H_2 : & \frac{f_1(X)}{f_2(X)} = \frac{\xi(2)}{\xi(1)}. \end{cases} \tag{8.0.1}$$

(Here $\frac{1}{0} = +\infty$, $\frac{0}{1} = 0$.) This type of test is often called a *likelihood ratio test* because it is expressed in terms of the ratio $f_1(X)/f_2(X)$ of likelihood functions.

Example. Suppose we have one observation X_1 and two simple hypotheses

$$H_1 : \mathbb{P} = N(0,1) \quad \text{and} \quad H_2 : \mathbb{P} = N(1,1).$$

Take equal weights

$$\xi(1) = \frac{1}{2} \text{ and } \xi(2) = \frac{1}{2}.$$

Then a Bayes decision rule δ that minimizes

$$\frac{1}{2}\mathbb{P}_1(\delta \neq H_1) + \frac{1}{2}\mathbb{P}_2(\delta \neq H_2)$$

is given by

$$\delta(X_1) = \begin{cases} H_1 : & \frac{f_1(X)}{f_2(X)} > 1 \\ H_2 : & \frac{f_1(X)}{f_2(X)} < 1 \\ H_1 \text{ or } H_2 : & \frac{f_1(X)}{f_2(X)} = 1. \end{cases}$$

This decision rule has a very intuitive interpretation. If we look at the graphs of these p.d.f.s (figure 8.1) the decision rule picks the first hypothesis when the first p.d.f. is larger, $x \leq 0.5$, and otherwise picks the second hypothesis, $x > 0.5$

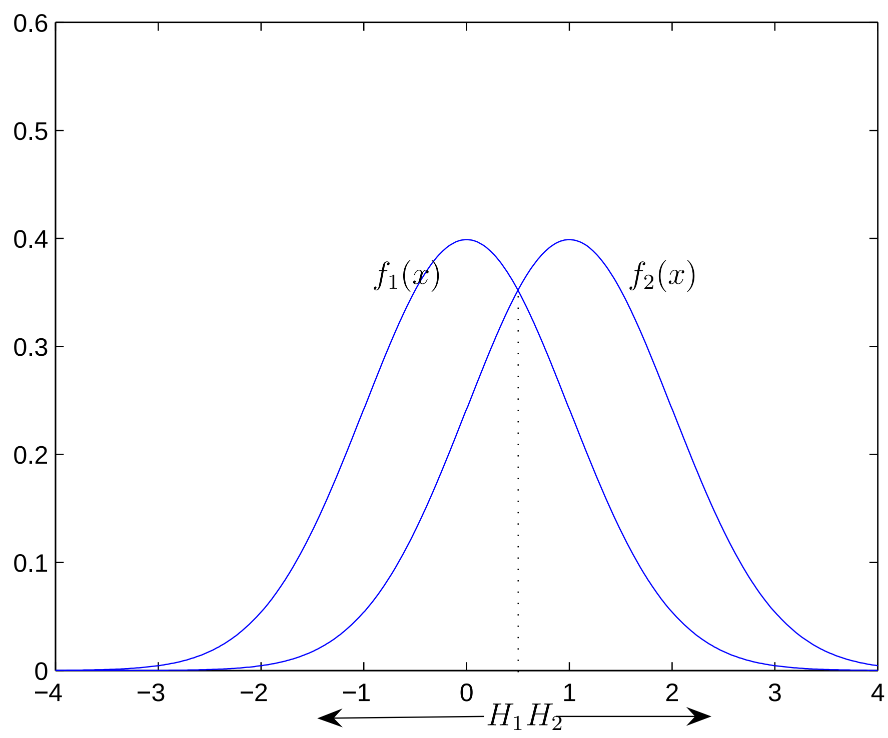

Figure 8.1: Bayes decision rule.

Example. Let us a general example case of n observations $X = (X_1, \ldots, X_n)$, two simple hypotheses $H_1 : \mathbb{P} = N(0,1)$ and $H_2 : \mathbb{P} = N(1,1)$, and arbitrary a priori weights $\xi(1), \xi(2)$. Then Bayes decision rule is given by (8.0.1). The likelihood ratio can be simplified:

$$\begin{aligned}\frac{f_1(X)}{f_2(X)} &= \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n X_i^2} \Big/ \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2}\sum_{i=1}^n (X_i-1)^2} \\ &= e^{\frac{1}{2}\sum_{i=1}^n ((X_i-1)^2 - X_i^2)} = e^{\frac{n}{2} - \sum_{i=1}^n X_i}.\end{aligned}$$

Therefore, the decision rule picks the first hypothesis H_1 when

$$e^{\frac{n}{2} - \sum X_i} > \frac{\xi(2)}{\xi(1)} \quad \text{or, equivalently,} \quad \sum X_i < \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

Similarly, we pick the second hypothesis H_2 when

$$\sum X_i > \frac{n}{2} - \log \frac{\xi(2)}{\xi(1)}.$$

In case of equality, we pick either H_1 or H_2.

$\square$