Lecture 18

Testing hypotheses.

(Textbook, Chapter 8)

18.1 Testing simple hypotheses.

Let us consider an i.i.d. sample $X_1, \dots, X_n$ with distribution $\mathbb{P}$ on some space $\mathcal{X}$, i.e. X's take values in $\mathcal{X}$. Suppose that we don't know $\mathbb{P}$ but we know that it can only be one of possible k distributions, $\mathbb{P} \in \{\mathbb{P}_1, \dots, \mathbb{P}_k\}$.

Based on the data $X_, \dots, X_n$ we have to decide among k simple hypotheses:

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

We call these hypotheses simple because each hypothesis asks a simple question about whether $\mathbb{P}$ is equal to some particular specified distribution.

To decide among these hypotheses means that given the data vector,

$$X = (X_1, \dots, X_n) \in \mathcal{X}^n$$

we have to decide which hypothesis to pick or, in other words, we need to find a decision rule which is a function

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

Let us note that sometimes this function δ can be random because sometimes several hypotheses may look equally likely and it will make sense to pick among them randomly. This idea of a randomized decision rule will be explained more clearly as we go on, but for now we can think of δ as a simple function of the data.

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

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

which we will call *error of type* i or type i error.

In the case when we have only two hypotheses H_1 and H_2 the error of type 1

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

is also called *size* or *level of significance* of decision rule δ and one minus type 2 error

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

is called the *power* of δ.

Ideally, we would like to make errors of all types as small as possible but it is clear that there is a trade-off among them because if we want to decrease the error of, say, type 1 we have to predict hypothesis 1 more often, for more possible variations of the data, in which case we will make a mistake more often if hypothesis 2 is actually the true one. In many practical problems different types of errors have very different meanings.

Example. Suppose that using some medical test we decide is the patient has certain type of decease. Then our hypotheses are:

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

Then the error of type one is

$$\mathbb{P}(\delta = H_2 | H_1),$$

i.e. we predict that the person does not have the decease when he actually does and error of type 2 is

$$\mathbb{P}(\delta = H_1 | H_2),$$

i.e. we predict that the person does have the decease when he actually does not. Clearly, these errors are of a very different nature. For example, in the first case the patient will not get a treatment that he needs, and in the second case he will get unnecessary possibly harmful treatment when he doesn not need it, given that no additional tests are conducted.

Example. Radar missile detection/recognition. Suppose that an image on the radar is tested to be a missile versus, say, 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 2

$$\mathbb{P}(\delta = H_2|H_1),$$

means that we will possibly shoot down a passenger plane (which happened before).

Another example could be when guilty or not guilty verdict in court is decided based on some tests and one can think of many examples like this. Therefore, in many situations it is natural to control certain type of error, give garantees that this error does not exceed some acceptable level, and try to minimize all other types of errors. For example, in the case of two simple hypotheses, given the largest acceptable error of type one $\alpha \in [0,1]$, we will look for a decision rule in the class

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

and try to find $\delta \in K_\alpha$ that makes the error of type 2, $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$, as small as possible, i.e. maximize the power.

18.2 Bayes decision rules.

We will start with another way to control the trade-off among different types of errors that consists in minimizing the weighted error.

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 an apriori probability on the set of our 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. Of course, we want to make this weigted error as small as possible.

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

Next theorem tells us how to find this Bayes decision rule in terms of p.d.f. or p.f. or the distributions $\mathbb{P}_i$.

Theorem. *Assume that each distribution $\mathbb{P}_i$ has p.d.f or p.f. $f_i(x)$. Then*

$$\delta = H_j \text{ if } \xi(j)f_j(X_1)\ldots f_j(X_n) = \max_{1\leq i\leq k} \xi(i)f_i(X_1)\ldots f_i(X_n)$$

is the 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:

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

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

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

by choosing δ appropriately. For each $(x_1, \dots, x_n)$ 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

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

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.

□