ESPCI Paris

Report

Machine Learning

Submitted By :
Kaleb RONCATTI and
Gabriel CALLADO

Contents

1 Introduction

This report will include the resolution of the theoretical exercises proposed in the homework while the practical one will be provided in the Jupyter Notebook/HTML format. Concerning the theoretical questions that we tried to answer here, the exercise 2.4 was the only one that we were not able to finish completely.

2 Questions

2.1 Maximum Likelihood estimation

Firstly, for the generic experiment "tossing coins" we can start with the **Bernoulli Distribution** for each single experiment of tossing a coin only once, that is described as follows:

$$f(k_i; p) = \begin{cases} p & \text{if } k_i = 1 \text{ (Tail)}, \\ q = 1 - p & \text{if } k_i = 0 \text{ (Head)}. \end{cases}$$

Assuming that our experiments are independent $K_1, ..., K_n$ one from another and that they are all modelled by the same probability distribution described above (identically distributed random variables), rewriting the distribution can be rewritten as $f(k_i; p) = p^{k_i}(1-p)^{1-k_i}$ for $k_i \in \{0, 1\}$ we can agree that, the process of tossing n coins can be described by a product function and it is given by:

$$f(k_1, ..., k_n; p) = \prod_{i=1}^{n} p^{k_i}(1-p)^{1-k_i} = p^{k_1}(1-p)^{1-k_1}...p^{k_n}(1-p)^{1-k_n} =$$

$$= p^{k_1+k_2+...+k_n}(1-p)^{1-k_1+1-k_2+...+1-k_n} = p^{\sum k_i}(1-p)^{n-\sum k_i}$$

Subsequently, maximize this function will means to find the configuration (in our case, the probability p for a given n with a given number of heads and tails) for which the function is maximal, that is, the probability is maximal. Before normalizing we normally take the logarithm of $f(k_1, ..., k_n; p)$:

$$\mathcal{L} = \sum k_i ln(p) + \Big(n - \sum k_i\Big) ln(1-p)$$

Then, we can get the derivative with respect to p and equalize this derivative to 0 with the purpose of maximizing the Loglikelihood function:

$$\frac{\partial \mathcal{L}}{\partial p} = \frac{\sum k_i}{p} - \frac{n - \sum k_i}{1-p} = \frac{(1-p)\sum k_i - p\Big(n - \sum k_i\Big)}{p(1-p)} = \frac{\sum k_i - pn}{p(1-p)} = 0$$

which leads us to $\boxed{p_{LH} = \frac{\sum k_i}{n}}$ for the probability of tails. In our experiment, we have had 55 tails (consequently, 45 heads to equalize n=100) which means that the probability of our coin gives a tail is $p_{LH} = \frac{55.1+45.0}{100} = 0.55$.

For the confidence interval, we can obtain it by considering the information matrix and the Gaussian hypothesis that n is big enough (in our case, it is true since n=100). We can then consider that $p \to \mathcal{N}(p, I^{-1})$ with $I = -\mathbb{E}(\frac{\partial^2 \mathcal{L}}{\partial p^2})$ also called the Fisher information. This way we will have our confidence interval by:

$$\boxed{p_{LH} \pm g(1-\alpha/2)\frac{1}{\sqrt{I}}}$$

We must calculate the Fisher information and, for a 95% confidence interval we must take $\alpha = 1 - 0.95 = 0.05 = 5\%$ and because our confidence interval is double sided, we take $g(1-\alpha/2) = g(1-0.025) = g(0.975) = 1.96$. For the Fisher information:

$$\mathbb{E}(\frac{\partial^2 \mathcal{L}}{\partial p^2}) = \mathbb{E}(\frac{-\sum k_i}{p^2} + \frac{-n + \sum k_i}{(1-p)^2}) = \mathbb{E}(\frac{-np^2 - \sum k_i + 2p\sum k_i}{p^2(1-p)^2})$$

Thus, simplifying and after using that $p = p_{LH} = \frac{\sum k_i}{n}$:

$$I = -\mathbb{E}(\frac{\partial^2 \mathcal{L}}{\partial p^2}) = \mathbb{E}(\frac{np^2 + \sum k_i - 2p\sum k_i}{p^2(1-p)^2}) = \frac{n}{p_{LH}(1-p_{LH})}$$

which leads us to the following confidence interval:

$$p_{LH} \pm 1.96\sqrt{\frac{p_{LH}(1-p_{LH})}{n}} = 0.55 \pm 1.96\sqrt{\frac{0.55 \times 0.45}{100}} = 0.55 \pm 0.0975$$

As a conclusion, we can say that $\boxed{p_{LH} \in [0.4525, 0.6475]}$ and the coin should not be considered as unfair.

2.2 Bayesian estimation

The first thing that we must do in these kind of problems related with Bayesian estimation is to establish the hypothesis and understand what we are trying to answer. As Bob is tested for a disease, the test is said 90% reliable. In the Bayesian theory, the test being reliable means that P(Test result is +|Bob has the Disease)=0.9, thus P(Test result is +|Bob DOES NOT has the Disease)=0.1. Furthermore, around the age of Bob, we are told that 1% of the people has the disease, it means that P(Bob has the Disease)=0.01. We are interested in calculating the probability that Bob really has the disease, given that he was tested positive (and the test is not 100% reliable). In the Bayesian formulation, our interpretation of the problem leaded us to calculate the following probability: **P(Bob has the Disease|Test result was +)**. However, in the statement of the question it is not very clear what is meant to be calculated by "what is the probability that Bob has the disease?", because there is also the probability that Bob has the disease given that he was tested negative.

To solve this problem, we use the Bayes theorem:

$$P(Disease|+) = \frac{P(+|Disease).P(Disease)}{P(+)}$$

However, the pure information of P(+) is not given and must be calculated by:

$$P(+) = P(+|Disease)P(Disease) + P(+|notDisease)P(notDisease)$$

$$P(Disease|+) = \frac{P(+|Disease).P(Disease)}{P(+|Disease)P(Disease) + P(+|notDisease)P(notDisease)}$$

$$\boxed{P(Disease|+) = \frac{0.9 * 0.01}{0.9 * 0.01 + 0.1 * 0.99} = 0.0833 = 8.33\%}$$

In the first look, this probability may seems a low value. But, if we look at this carefully, we realise that by calculating the Bayesian probability we have considered that Bob was took randomly in the whole population, and because of the low prior 1% probability of being sick, the probability value a posteriori of 8.33% is coherent. The alternative option is that Bob has the disease given that he was tested negative and alternatively we can calculate a very low probability:

$$P(Disease|-) = \frac{P(-|Disease).P(Disease)}{P(-)} =$$

$$P(Disease|-) = \frac{P(-|Disease).P(Disease)}{P(-|Disease)P(Disease) + P(-|notDisease)P(notDisease)} =$$

$$\boxed{P(Disease|-) = \frac{0.1 * 0.01}{0.9 * 0.99 + 0.1 * 0.01} = 0.0011 = 0.11\%}$$

The following portion of the statement asks if it is preferable to reduce the number of false positives or negatives. Also, it says that a false positive rate is less negative results when the patient is positive. However, a false positive rate is a result of a test that indicates that a given condition exists (or is positive), when it does not.

When a patient is diagnosed as positive, the patient may have the disease or he may NOT have it (false positive) because the test is not 100% certain. In the case of a false positive, further examination will be realized and the symptoms of the patient will be observed until they realize that the patient does not have it. Nonetheless, when a patient is diagnosed as negative, the patient may really not have the disease or he may have it (false negative) because the test is not 100% sure. In the case of a false negative, the patient can be dismissed of treatment still having the disease. In the case of a contagious disease, he can be a vector of spreading and in the case of a deadly disease he can even die for lack of treatment.

We could think for example about a patient taking the HIV test. The consequences of a false positive (test is positive when he does not have HIV) would probably generate trauma but after further examination the doctors would discover that the patient is HIV-negative and not having HIV is ultimately a good thing. However, a false negative result (the patient has HIV but the test shows a negative result) would have difficult implications, meaning that the patient would be missing out on crucial treatments and runs a high risk of spreading the virus to others.

Hence, in the case of a disease, it is preferable to reduce the false negative rates, it means, having less negative results when the patients have indeed the disease.

2.3 Information Theory

We want to know the capacity of the channel and we know priory that this channel is memory less. From the lectures, its capacity of a discrete memory less channel is given by with I(X;Y) being the mutual information defined as $I(X;Y) = \sum_{x,y} P(Y|X)P(X)ln\frac{P(Y|X)}{P(Y)}$.

For the binary channel, we know that x can assume the values 0 and 1 and y can assume the values 0, 1 and $*$. Our inputs are equally likely and according to the statement of the question we'll have beforehand:

$$\begin{cases} P(X=1) = P(X=0) = \frac{1}{2} \\ P(Y=1) = P(Y=0) = (1-\epsilon)\frac{1}{2} \\ P(Y=*) = P(Y=*|X=0) = P(Y=*|X=1) = \epsilon \\ P(Y=0|X=0) = P(Y=1|X=1) = 1-\epsilon \\ P(Y=1|X=0) = P(Y=0|X=1) = 0 \end{cases}$$

Opening the sums for all terms we will have:

$$I(X;Y) = P(Y=0|X=0)P(X=0)ln\frac{P(Y=0|X=0)}{P(Y=0)} +$$

$$P(Y=0|X=1)P(X=1)ln\frac{P(Y=0|X=1)}{P(Y=0)} +$$

$$P(Y=1|X=0)P(X=0)ln\frac{P(Y=1|X=0)}{P(Y=1)} +$$

$$P(Y=1|X=1)P(X=1)ln\frac{P(Y=1|X=1)}{P(Y=1)} +$$

$$P(Y=*|X=0)P(X=0)ln\frac{P(Y=*|X=0)}{P(Y=*)} +$$

$$P(Y=*|X=1)P(X=1)ln\frac{P(Y=*|X=1)}{P(Y=*)} =$$

$$I(X;Y) = (1-\epsilon)(\frac{1}{2})ln(\frac{\cancel{1-\epsilon}}{\cancel{(1-\epsilon)}\frac{1}{2}}) + 0 + 0 + (1-\epsilon)(\frac{1}{2})ln(\frac{\cancel{1-\epsilon}}{\cancel{(1-\epsilon)}\frac{1}{2}}) + 0 + 0$$

This way, we conclude that $I(X;Y) = (1-\epsilon)ln(2)$ which implies that:

$$\boxed{C = max_{p(x)} I(X;Y) = (1-\epsilon)ln(2)}$$

Conversely, for q symbols, we need to realize the same process but now our x can assume the values $0, ..., i, ..., q$ with i being an integer between 0 and q:

$$\begin{cases} P(X=i) = \frac{1}{q} \\ P(Y=j) = (1-\epsilon)\frac{1}{q}; j \neq * \\ P(Y=*) = P(Y=*|X=i) = \epsilon \\ P(Y=i|X=i) = 1-\epsilon \\ P(Y=j|X=i) = 0; i \neq j, j \neq * \end{cases}$$

Such as before, all the crossed terms with $P(Y=j|X=i)$ will be zero and also, the terms with $ln\frac{P(Y=*|X=i)}{P(Y=*)}$ will be zero. This way:

$$I(X;Y) = (1-\epsilon)(\frac{1}{q})ln(\frac{\cancel{1-\epsilon}}{\cancel{(1-\epsilon)}\frac{1}{q}}) + (1-\epsilon)(\frac{1}{q})ln(\frac{\cancel{1-\epsilon}}{\cancel{(1-\epsilon)}\frac{1}{q}}) + ... + (1-\epsilon)(\frac{1}{q})ln(\frac{\cancel{1-\epsilon}}{\cancel{(1-\epsilon)}\frac{1}{q}})$$

$$I(X;Y) = (1-\epsilon)(\frac{1}{q})ln(q) + (1-\epsilon)(\frac{1}{q})ln(q) + ... + (1-\epsilon)(\frac{1}{q})ln(q)$$

Here, all the non-zero elements that we have in the sum are because of the sum with $X = Y = i$, we know that i goes from 0 until q, this way, counting the elements, we'll have $(q+1)$ elements.

$$I(X;Y) = (q+1)\left((1-\epsilon)(\frac{1}{q})ln(q)\right) = (1-\epsilon)(\frac{q+1}{q})ln(q)$$

$$\boxed{C = max_{p(x)} I(X;Y) = \frac{q+1}{q}(1-\epsilon)ln(q)}$$

Probably, the statement of the question would like the answer to be something in the style of $\boxed{C = max_{p(x)} I(X;Y) = (1-\epsilon)ln(q)}$ such as the first part of the exercise. But, for this, we would need to have q elements by the statement, and not (q+1) elements.

2.4 Maximum entropy method

We want to use the maximum entropy principle to find the form of the distribution for N binary sequences of length p.

One crucial information is directly given by the statement of the question when we are told that we constrain for each j saying that the average of σ_j is equal to the empirical mean. Also, we know that the sum of the probabilities must be equal to 1, this way we'll be able to construct two constraints and relate them to each other:

$$\begin{array}{l} \mu_j = \frac{1}{N}\sum_{i=1}^{N} \sigma_{ij} \\ 1 = \sum_{i=1}^{N} P(\sigma_{ij}) \end{array} \tag{1}$$

We would see by a lot of calculations that: if we write the entropy expression and we use the two constraints of the first equations (1) as Lagrange multipliers, followed by the differentiation of the whole expression, it would lead us to $P(\sigma_1, ..., \sigma_p) = \frac{e^{\sum h_j \sigma_j}}{Z}$ as stated by the question, so we can define that the form of P as $P(\sigma_j) = \frac{e^{h_j \sigma_j}}{z_j}$ because it leads to the right result:

$$P(\sigma_1, ..., \sigma_p) = \prod P(\sigma_j) = \prod_{j=1}^{p} \frac{e^{h_j \sigma_j}}{z_j} = \frac{e^{\sum h_j \sigma_j}}{z_1 ... z_p} = \frac{e^{\sum h_j \sigma_j}}{Z}$$

We can also write the mathematical expectation by using its definition ($E[X] = \sum_{i=1}^{k} x_i p_i$) and we can open the sum of probabilities equalized to 1. Note that both of these things will be done because we know that σ can assume the values of 1 and -1 and we must use the form of $P(\sigma_j)$ that we discovered.

$$\begin{cases} \mu_j = (+1)P(\sigma_j = 1) + (-1)P(\sigma_j = -1) \\ 1 = P(\sigma_j = 1) + P(\sigma_j = -1) \\ P(\sigma_j = +1) = \frac{e^{+h_j}}{z_j} \\ P(\sigma_j = -1) = \frac{e^{-h_j}}{z_j} \end{cases}$$

This way, we need to discover h_j and z_j on the previous expressions.

By simplifying these expression we can reduce ourselves into two equations as follows:

$$\begin{cases} \frac{1+\mu_j}{2} = \frac{e^{+h_j}}{z_j} \\ \frac{1-\mu_j}{2} = \frac{e^{-h_j}}{z_j} \end{cases}$$

By multiplying these two equations we'll get that $z_j = \frac{2}{\sqrt{1-\mu_j^2}}$ which leads us to $\boxed{Z = \prod_{j=1}^{p} z_j = \frac{2^p}{\prod_{j=1}^{p}\sqrt{1-\mu_j^2}}}$, by dividing the two equations we'll get $\boxed{h_j = \frac{1}{2} ln(\frac{1+\mu_j}{1-\mu_j})}$.

Now, if σ can assume the values of 0 and +1 we can apply the same process as we did before because the function $P(\sigma_1, ..., \sigma_p)$ keeps the same. In this case, the system of equations will be:

$$\begin{cases} \mu_j = (+1)P(\sigma_j = 1) + (0)P(\sigma_j = 0) \\ 1 = P(\sigma_j = 1) + P(\sigma_j = -0) \\ P(\sigma_j = +1) = \frac{e^{+h_j}}{z_j} \\ P(\sigma_j = 0) = \frac{1}{z_j} \end{cases}$$

By simplifying it again such as we did before:

$$\begin{cases} \mu_j = \frac{e^{+h_j}}{z_j} \\ 1 - \mu_j = \frac{1}{z_j} \end{cases}$$

Multiplying these equations we'll get that $z_j = \frac{1}{1-\mu_j}$ followed by $\boxed{Z = \prod_{j=1}^{p} z_j = \prod_{j=1}^{p} \frac{1}{1-\mu_j}}$.

Again, by dividing these equations in the same way, $\boxed{h_j = ln(\frac{\mu_j}{1-\mu_j})}$.