Probability For Machine Learning

Lecturer : Guillaume Euvrard
Email : guillaume.euvrard@epita.fr

Deadline : Sunday, 14 December 2019

Use of software tool, such as Python, is allowed and sometime necessary.
Please explain which software you use, and which functions from this software, to get your numerical results.
This may be done by sending me the file containing your program, or with a handwritten explanation in your worksheet.

Exercise 1

A factory builds computers, but it is not perfect : 1% out of the computers are defective.
A company buys 5 computers. Let X be the number of computers, out of the 5, which are defective.
Each defective computer results in a loss of 1000 € for the company. Let L be the total loss for the company.

Questions :

1. What are the possible values of X ?
2. Let k be a possible value of X. Determine the probability $P(X{=}k)$.
3. Determine the expectation and the standard deviation of X.
   Deduce the expectation and the standard deviation of L.

Exercise 2

A disease affects 1% of the population.
A medical test has been developed in order to evaluate whether a patient has the disease or not. The performances of the test are the following :

— 99% out of the people with the disease are positive at the test.
— 98% out of the people who don't have the disease are negative at the test.

Questions :

1. A patient is positive at the test. What is the probability that he/she has the disease ?
2. Let p be the answer of question 1. We suppose that the company who has developed the medical test wants to improve it in order to have a greater value of p.
   a. If 100% (instead of 99%) out of the people with the disease were positive at the test, what would be the value of p ?
   b. If 100% (instead of 98%) out of the people who don't have the disease were negative at the test, what would be the value of p ?
   c. The company tries to increase the second number : the percentage of people, out of those who don't have the disease, which are negative at the test. This percentage will increase from 98% to a value x%.
   What is the minimum value of x which would result in a value of p such that $p \geqslant 50\%$?

Exercise 3

In the city of Paris, the average rent for an apartment is 31 € per m^2 (for 1 month), with a standard deviation of 5 € per m^2. To be more accurate, if we pick an apartment randomly and define X as the rent per m^2 of this apartment, then

$$\mathrm{E}(X) = 31 \qquad \text{and} \qquad \mathrm{std}(X) = 5$$

A group of 30 students lives in Paris and each student rents an apartment. For $i \in \{1, \cdots, 30\}$, let X_i be the rent per m^2 that the i^{th} student pays each month. The variables X_i are supposed to be mutually independent and identically distributed.
Finally, let $\overline{X}$ be the average rent per m^2 in the group :

$$\overline{X} = \frac{X_1 + \cdots + X_{30}}{30}$$

Questions :

1. What are the expectation, the variance and the standard deviation of $\overline{X}$?
2. Provide a 95% prediction interval for $\overline{X}$. That is to say, determine a and b such that

$$P(a \leqslant \overline{X} \leqslant b) = 95\%$$

Exercise 4

An online shop has recorded the behavior of its customers. The statistics service has defined, for each customer, the variables X and Y about what the customer has bought during the 30 previous days :

$$\begin{array}{rcl} X & = & \text{number of books the customer has bought} \\ Y & = & \text{number of CD the customer has bought} \end{array}$$

Let $Z = \begin{pmatrix} X \\ Y \end{pmatrix}$. The analysis of the collected data has lead to a probability distribution of Z which is in the file "dsa_proba_data.xlsx"
Questions :

1. Calculate the expectation $\mathrm{E}(Z)$ of the random vector Z.
2. Calculate the covariance matrix $\mathrm{Cov}(Z)$.
3. Calculate the correlation $\mathrm{corr}(X, Y)$. Are the variables X and Y independent ?
4. We search for the best relation $Y = \alpha X + \beta$. The best relation is the relation which minimizes the expected error.
   We hence have to find the values of α and β which minimize

$$\mathrm{E}\left((Y - \alpha X - \beta)^2\right)$$

   What are the values of α and β ?