Probability for Machine Learning

Please try to give clear and detailed answers.

Due date: 23/02/2022

Part I - Variance, covariance and correlation (3+1+1+1 points)
X and Y represent the size of 2 different bones in a extinct species. 5 fossil specimens of this species are available, we put the measurements of X and Y in the table below with some other values (including their means).

i	x_i	y_i	x_i^2	y_i^2	$x_i^2 y_i^2$
1	44	40	1936	1600	1760
2	65	60	4225	3600	3900
3	71	59	5041	3481	4189
4	75	65	5625	4225	4875
5	87	77	7569	5929	6699
μ	68.4	60.2	4879.2	3767	4284.6

a) What are Var(X), Var(Y) and Cov(X,Y)? [hint: in some cases, the properties are more useful than the definition]
b) What is the correlation corr(X,Y) between X and Y?
c) What can you say about the relationship between X and Y?
d) If the correlation was 0, what could you say about X and Y?

Part II – Random variable, Central Limit Theorem and probability distribution (2+2+2+2 points)
It's your first day as a data scientist in your new company. Your first job is to check a few things about a dataset which consists in the list of grades of a 1000 students (from all over the country) who took a national exam.

a) If we assume that you have no other data than this and that your boss would like you to "provide ASAP" the expected value of the grades of all the students who took this exam (not just for the 1000 students of this sample but the real expected value for all the population of students who took the exam), how exactly would you proceed since you only have access to that single dataset?

We assume now that the grade of a specific student can take any value between 0 and 20 (which is not the case in the real world, for example no one gives a grade of exactly Pi points or 10.000000000000000001), each one being as likely as any other.

b) If the grade of a student can be considered as a random variable X, what is the probability distribution of X?
c) What is the probability for a student to have a grade of 12 points or more?
d) What is the expected value E(X)

Part III - Conditional Probability and Bayes Theorem (2+2+2 points)

We know that in a small country with a population of 10 million people, 100 000 have an illness XYZ. There is a test for that disease with the following numbers :

- if a person is sick, the test is positive with a probability of 99%
- if a person is not sick, the test is positive with a probability of 5%

a) What is the probability to be healthy for a person whose test results are negative? What do you think of that probability?

b) What is the probability to be sick for a person whose test results are positive? What do you think of that probability?

c) Let's say the test results of a person are positive and she decides to take it another time : what is the probability of being healthy if this second test results are also positive? What is your interpretation of the value of that probability?