Practice Problems

CS215

September, 2024

Problem 1

Here is a nice application of basic probability. I hope you will like it! Suppose some n individuals have arrived at a lab for RTPCR testing. An RTPCR test involves extracting the nasal mucus of the individual and testing it within an RTPCR machine. Suppose that the probability that any individual will test positive is p, and let us assume that the test results across all n individuals are independent. Instead of individually testing each person, we follow the two-step procedure described below to save on the number of tests: (1) We divide the people into n/g groups, each of size g where we assume that g divides n. Small, equal-volume portions of the mucus samples of all individuals belonging to the same group are mixed together. This mixture is tested, thus leading to n/g independent tests, one per group. (2) If the mixture tests negative (non-infected), then all group members are declared negative. If the mixture tests positive (infected), then each member of the group is individually tested in a second round of tests.

It is known that the mixture of different mucus samples, or using small portions of the sample, has no influence on the probability p or on test accuracy. This procedure is called Dorfman pooling and it was widely used during the COVID-19 pandemic.

(a) What is the expected total number of tests? Note that an individual test counts as one test, and the test of a mixture also counts as one test.

(b) Now suppose that exactly $k \ll n$ individuals are infected. In this scenario, what is the number of tests required in Dorfman's method in the worst case? For what value of g, expressed in terms of n and k, will this worst case number of tests be minimized? What is the number of tests in that case?

Solution

(a) The total number of tests is n/g in the first round. The probability of any one individual being infected is p. Hence the probability of any one individual being non-infected is $1-p$. The probability of obtaining a group of g individuals who are all non-infected is $(1-p)^g$, and hence the probability of obtaining a group of g individuals containing at least one infected individual is $1-(1-p)^g$. As the total number of groups is n/g, the expected number of groups with at least one infected member is $n/g \times [1-(1-p)^g]$. Hence the expected number of total tests is $n/g + g \times n/g \times [1-(1-p)^g] = n/g + n\,[1-(1-p)^g]$.

(b) In the worst case, each of the k infected people will lie in a different group. So each of these k groups will have to be tested in the second round, and each member of these k groups will be tested individually. This will give rise to $k \times g$ more tests. So the total number is $n/g + kg$ for the worst case number of tests. If the worst case number has to be optimal, we set the derivative of this number (w.r.t. g) to zero, giving rise to $-n/g^2 + k = 0$, that is, $g = \sqrt{n/k}$. The number of tests in this case will be $2\sqrt{nk}$.

Problem 2

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then express the CDF of $Y = aX + b$ in terms of the CDF of X. Also write down the PDF of Y. Here a, b are non-zero constants. If the PDF of X is $f_X(\cdot)$ and $Y = aX + b$ as before, write down an expression for the PDF of Y, i.e. $f_Y(\cdot)$ in terms of $f_X(\cdot)$.

Solution

In the case where $a > 0$, we have

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \leq (y - b)/a) = F_X((y - b)/a).$$

In the case that $a < 0$, we have

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P(X \geq (y - b)/a) = 1 - F_X((y - b)/a).$$

This is the expression for the CDF of Y.
The PDF is obtained by taking the derivative w.r.t. x, giving $f_Y(y) = f_X((y - b)/a)/a$ when $a > 0$ and

$$f_Y(y) = -f_X((y - b)/a)/a \quad \text{when } a < 0.$$

This yields

$$f_Y(y) = f_X((y - b)/a)/|a|.$$

Now since

$$f_X(x) = \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}},$$

we have

$$f_Y(y) = \frac{e^{-(y-b-a\mu)^2/(2a^2\sigma^2)}}{|a|\sigma\sqrt{2\pi}} = \frac{e^{-(y-b-a\mu)^2/(2a^2\sigma^2)}}{|a|\sigma\sqrt{2\pi}}.$$

Problem 3

The entropy of a discrete random variable X is defined as

$$H(X) = -\sum_{i=1}^{K} p_i \log p_i$$

where $p_i = P(X = i)$ and $\sum_{i=1}^{K} p_i = 1$; $\forall i, 0 \leq p_i \leq 1$. In this definition, $0 \log 0$ is considered to be 0. For which PMF (i.e. for what values of $\{p_i\}_{i=1}^{K}$) will the entropy be maximum? What is this maximum value? Derive your answer by setting the first derivatives of the entropy to 0. Obtain the sign of the second derivatives, i.e. sign of $\frac{\partial^2 H}{\partial p_i^2}$. (Note: Given what you have learned so far, you will be able to find only a local maximum. But it turns out that the unexpectedness measure is a concave function, due to which a local maximum is also the global maximum. You are not expected to prove that it is a concave function.) For what PMF will the entropy be the least? Give an intuitive answer (it is not so easy to prove your answer for the minimum, in a quiz/exam). What is this minimum value?

Solution

We have

$$H(X) = -\sum_{i=1}^{K-1} p_i \log p_i - p_K \log p_K = -\sum_{i=1}^{K-1} p_i \log p_i + (1 - \sum_{i=1}^{K-1} p_i) \log(1 - \sum_{i=1}^{K-1} p_i).$$

Hence, we have for $j \in \{1, \ldots, K-1\}$ that

$$\frac{\partial H(X)}{\partial p_j} = -(1 + \log p_j) - \left(-1 - \log\left(1 - \sum_{i=1}^{K-1} p_i\right)\right) = -\log p_j + \log\left(1 - \sum_{i=1}^{K-1} p_j\right).$$

Setting this derivative to zero, we have $p_j = 1 - \sum_{i=1}^{K-1} p_j = p_K$. Thus, all the p_j values are equal to $1/K$, i.e. we have a discrete uniform PMD. The second derivative is given as

$$\frac{\partial^2 H(X)}{\partial p_j^2} = -1/p_j - 1/\left(1 - \sum_{j=1}^{K-1} p_j\right)$$

which is clearly less than 0, and thus the second derivative test is passed. Thus, a discrete uniform PMF maximizes the entropy. The least possible entropy occurs if $p_j = 1$ for some $j \in \{1, \ldots, K\}$ with the other values being all 0. This is called a Kronecker delta function. The entropy is always non-negative and because $0 \leq p_i \leq 1$ and hence $-p_i \log p_i$ is always positive.

Problem 4

Let Y be a Gaussian random variable with mean μ and variance σ^2. Derive the CDF and PDF of the random variable $X = |Y|$. Also derive $E\left(X^2\right)$.

Solution

We have $F_X(x) = P(X \leq x) = P(|Y| \leq x) = P(-x \leq Y \leq x) = \int_{-x}^{+x} f_Y(y)dy = F_Y(x) - F_Y(-x) = \Phi((x-\mu)/\sigma) - \Phi((-x-\mu)/\sigma)$ where Φ stands for the CDF of a zero-mean Gaussian random variable with unit variance. The PDF of X is given by $f_X(x) = \frac{1}{\sigma}[\phi((x-\mu)/\sigma) + \phi((-x-\mu)/\sigma)] = \frac{1}{\sigma}[\phi((x-\mu)/\sigma) + \phi((x+\mu)/\sigma)]$ where ϕ stands for the PDF of a zero-mean Gaussian random variable with unit variance, and where we use the symmetry of ϕ. This yields $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\left[\exp\left(-(x-\mu)^2/2\sigma^2\right) + \exp\left(-(x+\mu)^2/2\sigma^2\right)\right]$.

Problem 5

In a certain town, there exist 100 rickshaws out of which 1 is red and 99 are blue. A person XYZ observes a serious accident caused by a rickshaw at night and remembers that the rickshaw was red in color. Hence, the police arrest the driver of the red rickshaw. The driver pleads innocence. Now, a lawyer decides to defend the hapless rickshaw driver in court. The lawyer ropes in an opthalmologist to test XYZ's ability to differentiate between the colors red and blue, under illumination conditions similar to those that existed that fateful night. The opthalmologist suggests that XYZ sees red objects as red 99% of the time and blue objects as red 2% of the time. What will be the main argument of the defense lawyer? (In other words, what is the probability that the rickshaw was really a red one, when XYZ observed it to be red?)

Solution

Let R_R, R_B be the events that the rickshaw was red, blue respectively. Let X_R, X_B be the events that XYZ perceived a rickshaw to be red, blue respectively. We have $P(X_R \mid R_R) = 0.99, R(X_R \mid R_B) = 0.02, P(R_R) = 0.01, P(R_B) = 0.99$. We need to evaluate $P(R_R \mid X_R) = P(X_R \mid R_R) P(R_R) / P(X_R) . P(X_R) = P(X_R \mid R_R) P(R_R) + P(X_R \mid R_B) P(R_B) = 0.99 \times 0.01 + 0.02 \times 0.99 = 0.99 \times 0.03$. Hence $P(R_R \mid X_R) = \frac{0.99 \times 0.01}{0.99 \times 0.03} = 1/3$. In other words, the probability that the rickshaw was red when XYZ observed it to be red is only $1/3$. In other words, it is more probable that the rickshaw was a blue one, based on the available data!

Problem 6

If X and Y are two independent continuous random variables with PDFs f_X and f_Y respectively, then prove that the PDF of $Z = XY$ is given by $f_Z(z) = \int_{-\infty}^{+\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx$.

Solution

The CDF of Z is given by:

$$\begin{aligned} F_Z(z) = P(Z \leq z) &= P(XY \leq z, X \geq 0) + P(XY \leq z, X \leq 0) \\ &= P(Y \leq z/X, X \geq 0) + P(Y \geq z/X, X \leq 0) \\ &= (*) \int_0^{+\infty} f_X(x) \int_{-\infty}^{z/x} f_Y(y) dy dx + \int_{-\infty}^{0} f_X(x) \int_{z/x}^{\infty} f_Y(y) dy dx \\ &= \int_0^{+\infty} f_X(x) \left(F_Y(z/x) - 0\right) dx + \int_{-\infty}^{0} f_X(x) \left(1 - F_Y(z/x) dx\right. \end{aligned}$$

where the step marked $(^*)$ follows due to independence. Taking derivatives w.r.t. z, we obtain the PDF of Z as follows:

$$\begin{aligned} f_Y(z) &= \int_0^{+\infty} f_X(x) f_Y(z/x)/x dx - \int_{-\infty}^{0} f_X(x) f_Y(z/x)/x dx \\ &= \int_0^{+\infty} f_X(x) f_Y(z/x)/|x| dx + \int_{-\infty}^{0} f_X(x) f_Y(z/x)/|x| dx \\ &= \int_{-\infty}^{+\infty} f_X(x) f_Y(z/x)/|x| dx \end{aligned}$$