Solutions + Correction key

201700080 Information Theory and Statistics
18 April 2023, 13:45 – 16:45

This test consists of 5 problems for a total of 28 points. All answers need to be justified. The use of a non-programmable calculator (not a "GR") is allowed. A handwritten single side A4 cheat sheet is allowed. No additional books or notes may be used.

1. Suppose you bought a large bag of Easter eggs containing white, dark and milk chocolate eggs. According to the label, the bag contains equally many of each type of eggs. However, you suspect that the proportions of white, dark and milk eggs are in fact 2/5, 1/5 and 2/5 respectively. In order to test the correctness of this presumption, you repeatedly draw a random egg from the bag, record its type and put it back. This results in a sample of observations $x_1, x_2, ..., x_n$, where x_i denotes the egg type of the ith draw.

   a. [1 pt] Give the Log-likelihood ratio function of your presumed distribution and the distribution claimed by the label corresponding to a single observation.
   Solution:
   Let your presumed distribution and the distribution claimed by the bag be denoted by P_1 and P_2 respectively. We have $P_1(x) = 2/5$ if $x \in \{\text{white}, \text{milk}\}$, $P_1(x) = 1/5$ if $x = \text{dark}$ and $P_2(x) = 1/3$ for each $x \in \{\text{white}, \text{dark}, \text{milk}\}$.

$$L(x) = \log \frac{P_1(x)}{P_2(x)} = \begin{cases} \log(6/5) \text{ if } x \in \{\text{white}, \text{milk}\}, \\ \log(3/5) \text{ if } x = \text{dark} \end{cases}$$

   b. [2 pt] Specify a binary hypothesis testing problem for choosing between the two distributions.
   Solution:

$$L(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} L(x_i)$$

   The optimal decision rule is of the form

$$\Psi_t^* = \begin{cases} 1, & \text{if } L(x_1, ..., x_n) \le t \\ 2, & \text{if } L(x_1, ..., x_n) > t \end{cases} \quad (1)$$

   for a threshold value $t \in \mathbb{R}$.

c. [3 pt] Verify Pinsker's inequality for the two distributions.

2. Consider the function

$$\rho(X,Y) = H(X|Y) + H(Y|X).$$

a. [2 pt] For random variables X, Y and Z, show that $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$.

Solution:

We have

$$\rho(X,Y) + \rho(Y,Z) = H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) \quad (2)$$

and

$$\rho(X,Z) = H(X|Z) + H(Z|X). \quad (3)$$

. For $H(X|Y) + H(Y|Z)$ we obtain

$$\begin{aligned} H(X|Y) + H(Y|Z) &\geq H(X|Y,Z) + H(Y|Z) \\ &= H(X,Y|Z) \\ &= H(X|Z) + H(Y|X,Z) \\ &\geq H(X|Z) \end{aligned}$$

and similarly $H(Z|Y) + H(Y|X) \geq H(Z|X)$. Hence, it follows that $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$.

b. [1 pt] Show that $\rho(X,Y) = 2H(X,Y) - H(X) - H(Y)$

Solution:

$\rho(X,Y) = H(X|Y)+H(Y|X) = H(X,Y)-H(Y)+H(X,Y)-H(X) = 2H(X,Y)-H(X)-H(Y)$

3. Consider compression of DNA data. The alphabet is $\mathcal{X} = \{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$ and $p(\mathtt{a}) = 0.3$, $p(\mathtt{c}) = 0.3$, $p(\mathtt{g}) = 0.25$, $p(\mathtt{t}) = 0.15$. Consider the code C for which $C(\mathtt{a}) = 00$, $C(\mathtt{c}) = 11$, $C(\mathtt{g}) = 010$ and $C(\mathtt{t}) = 100$.

a. [2 pt] Is C a prefix code? Is it a uniquely decodable code? Justify your answers.

Solution:

It is a prefix code and thus also uniquely decodable.

b. [3 pt] Construct a Huffman code for the source. Denote this code by C' and let L' denote its codewords lengths. Compute $\mathbb{E}[L']$.

Solution:

The Huffman code will create a prefix code with every character being encoded into a string of length 2. Example: $c(a) = 11, c(c) = 10, c(g) = 01, c(t) = 00$. No matter the choice of the code, we have $\mathbb{E}[L'] = 2$.

c. [2 pt] Is it possible to construct a prefix code for the source for which the expected codeword length is 1.9? Explain.

Solution:
No, Shannon's theorem.

P.T.O. (Please turn over)

4. Let X be a continuous random variable with probability density function $f_\theta(x) = \frac{x}{\theta}e^{-\frac{x^2}{2\theta}}$ for $x \geq 0$ and $f_\theta(x) = 0$ for $x < 0$. Here, θ is an unknown parameter. For this random variable it is known that $\mathbb{E}[X^2] = 2\theta$ and $\mathrm{Var}[X^2] = 4\theta^2$.
After observing n independent realizations of this random variable you want to estimate θ. One possible estimator is $\bar{\theta} = \frac{1}{2n}\sum_{i=1}^{n} x_i^2$.

a. [1 pt] Is $\bar{\theta}$ an unbiased estimator? Explain.
Solution:
Unbiased means that $\mathbb{E}(\bar{\theta}) - \theta = 0$ and therefore we must have that $\mathbb{E}(\bar{\theta}) = \theta$.

$$\mathbb{E}(\bar{\theta}) = \mathbb{E}(\frac{1}{2n}\sum_{i=1}^{n} x_i^2) = \frac{1}{2n}\sum_{i=1}^{n} \mathbb{E}(x_i^2) = \frac{1}{2}2\theta = \theta.$$

Hence we conclude that $\bar{\theta}$ is unbiased.

b. [1 pt] Compute the variance of $\bar{\theta}$. **Solution:**

$$\mathrm{Var}[\bar{\theta}] = \mathrm{Var}[\frac{1}{2n}\sum_{i=1}^{n} x_i^2] = \frac{1}{4n^2}\sum_{i=1}^{n}\mathrm{Var}[x_i^2] = \frac{4\theta^2}{4n} = \frac{\theta^2}{n}.$$

c. [3 pt] Assume that θ is unknown and compute the Fisher information $J(\theta)$ for a single observation. **Solution:**
$J(\theta) = -\mathbb{E}(\frac{\theta - x^2}{\theta^3}) = -(\frac{1}{\theta^2} - \frac{2}{\theta^2}) = \frac{1}{\theta^2}$

d. [3 pt] State the Cramer-Rao bound and discuss in your own words the implications of the results from a) and b). (*If you did not solve a) you may assume that $\bar{\theta}$ is unbiased and that Var$[\bar{\theta}] = 2\theta^2/n$. If you did not solve b) you may assume $J(\theta) = 1/\theta^2$*) **Solution:**
We must have that $\mathrm{var}(\bar{\theta}) \geq \frac{1}{J_n(\theta} = \frac{\theta^2}{n}$ for an unbiased estimator $\bar{\theta}$. This bound is achieved and therefore the estimator is optimal.

5. Consider a discrete channel defined as $Y = X + Z \bmod 11$, where $X \in \{0, 1, ..., 10\}$ and $Z \in \{1, 2, 3\}$. Suppose that the random variables X and Z are independent and $\mathbb{P}(Z = i) = 1/3$ for each $i = 1, 2, 3$.

a. [1 pt] Give the definition of channel capacity.
Solution:
$C = \max_{p_x} I(X;Y)$