Recitation 10: Statistics

Brown University CS145: Probability & Computing

April 25, 2016

1 Bertsekas & Tsitsiklis 8.8

The probability of heads of a given coin is known to be either q_0 (hypothesis H_0) or q_1 (hypothesis H_1). We toss the coin repeatedly and independently, and record the number of heads before a tail is observed for the first time. We assume that $0 < q_0 < q_1 < 1$, and that we are given prior probabilities $P(H_0)$ and $P(H_1)$. For parts (a) and (b), we also assume that $P(H_0) = P(H_1) = \frac{1}{2}$.

a) Calculate the probability that hypothesis H_1 is true, given that there were exactly k heads before the first tail.

b) Consider the decision rule that decides in favor of hypothesis H_1 if $k \geq k^*$, where k^* is some nonnegative integer, and decides in favor of hypothesis H_0 otherwise. Give a formula for the probability of error in terms of k^*, q_0 and q_1. For what value of k^* is the probability of error minimized?

c) Assume that $q_0 = 0.3$, $q_1 = 0.7$, and $P(H_1) > 0.7$. How does the optimal choice of k^* (the one that minimizes the probability of error) change as $P(H_1)$ increases from 0.7 to 1.0.?

Solution

See attached matlab code for a plot of the probabilities under each hypothesis.

a) Let K be the number of heads observed before the first tail and let $p_{K|H_i}(k)$ be the PMF of K when hypothesis H_i is true. Therefore

$$p_{K|H_i}(k) = (1 - q_i)q_i^k.$$

Using Bayes' rule, we obtain

$$\begin{aligned} P(H_1|K = k) &= \frac{p_{K|H_1}(k)P(H_1)}{p_K(k)} \\ &= \frac{\frac{1}{2}(1 - q_1)q_1^k}{\frac{1}{2}(1 - q_1)q_1^k + \frac{1}{2}(1 - q_0)q_0^k} \\ &= \frac{(1 - q_1)q_1^k}{(1 - q_1)q_1^k + (1 - q_0)q_0^k}. \end{aligned} \tag{1}$$

b) An error occurs in two cases: if H_0 is true and $K \geq k^*$, or if H_1 is true and $K < k^*$. So the probability of error, denoted by p_e is

$$\begin{aligned} p_e &= P(K \geq k^* | H_0)P(H_0) + P(K < k^* | H_1)P(H_1) \\ &= \sum_{k=k^*}^{\infty} p_{K|H_0}(k)P(H_0) + \sum_{k=0}^{k^*-1} p_{K|H_1}(k)P(H_1) \\ &= P(H_0)\sum_{k=k^*}^{\infty}(1-q_0)q_0^k + P(H_1)\sum_{k=0}^{k^*-1}(1-q_1)q_1^k \\ &= P(H_0)(1-q_0)\frac{q_0^{k^*}}{1-q_0} + P(H_1)(1-q_1)\frac{1-q_1^{k^*}}{1-q_1} \\ &= P(H_1) + P(H_0)q_0^{k^*} - P(H_1)q_1^{k^*} \\ &= \frac{1}{2}(1 + q_0^{k^*} - q_1^{k^*}) \end{aligned} \tag{2}$$

To find value k^* that minimizes p_e, we temporarily treat k^* as a continuous variable and differentiate p_e with respect to k^*. See attached matlab code for a plot of p_e with respect to k^*. Setting the derivative to zero:

$$\frac{\mathrm{d}p_e}{\mathrm{d}k^*} = \frac{1}{2}((\log q_0)q_0^{k^*} - (\log q_1)q_1^{k^*}) = 0$$

and the solution is

$$\bar{k} = \frac{\log(|\log q_0|) - \log(|\log q_1|)}{|\log q_0| - |\log q_1|}$$

c) As in part (b), we have

$$p_e = P(H_1) + P(H_0)q_0^{k^*} - P(H_1)q_1^{k^*}$$

Consider the case where $P(H_1) = 0.7, q_0 = 0.3$ and $q_1 = 0.7$, using the calculations in part(b), we have $\bar{k} \approx 0.43$, thus the optimal value of k^* is either 0 or 1. We find that with either choice the probability of error p_e is the same and equal to 0.3, thus either choice minimizes the probability error.

Note that $\bar{k}$ decreases as $P(H_1)$ increases from 0.7 to 1.0, so the choice $k^* = 0$ remains optimal in this range. As a result, we always decide in favor of H_1 and the probability of error is $p_e = P(H_0) = 1 - P(H_1)$.

2 Bertsekas & Tsitsiklis 9.21

A normal random variable X is known to have a mean of 60 and a standard deviation equal to 5 (hypothesis H_0) or 8 (hypothesis H_1).

a) Consider a hypothesis test using a single sample x. Let the rejection region be of the form

$$R = \{x \mid |x - 60| > \gamma\}$$

for some scalar γ. Determine γ so that the probability of false rejection of H_0 is 0.1. What is the corresponding false acceptance probability?

b) Consider a hypothesis test using n independent samples $x_1, ..., x_n$. Let the rejection region be of the form

$$R = \left\{ (x_1, ..., x_n) \Big| |\frac{x_1 + ... + x_n}{n} - 60| > \gamma \right\}$$

where γ is chosen so that the probability of false rejection of H_0 is 0.1. How does the false acceptance probability change with n? What can you conclude about the appropriateness of this type of test?

c) Derive the structure of the LRT using n independent samples $x_1, \ldots, x_n$.

Solution

We have two hypotheses H_0 and H_1, under which the observation PDFs are

$$f_X(x; H_0) = \frac{1}{\sqrt{2\pi} \cdot 5} \exp\left\{ \frac{-(x-60)^2}{2 \cdot 25} \right\}$$

and

$$f_X(x; H_0) = \frac{1}{\sqrt{2\pi} \cdot 8} \exp\left\{ \frac{-(x-60)^2}{2 \cdot 64} \right\}$$

See attached matlab code for a plot of the probabilities under each hypothesis.

a) The probability of false rejection of H_0 is

$$P(x \in R; H_0) = 2\left(1 - \Phi(\frac{\gamma}{5})\right) = 0.1$$

which yields that $\gamma = 8.25$. The acceptance region of H_0 is $\{x | 51.75 < x < 68.25\}$ and the probability of false acceptance is

$$P(51.75 < x < 68.25; H_1) = 2\Phi\left(\frac{68.25 - 60}{8}\right) - 1 = 0.697 \tag{3}$$

b) Let $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. To determine γ, we set

$$P(\bar{X} \notin R; H_0) = 2\left(1 - \Phi(\frac{\gamma\sqrt{n}}{5})\right) = 0.1$$

which yields

$$\gamma = \frac{\Phi^{-1}(0.95)}{\sqrt{n}}$$

The acceptance region is

$$R = \left\{ x | 60 - \frac{\Phi^{-1}(0.95)}{\sqrt{n}} < x < 60 + \frac{\Phi^{-1}(0.95)}{\sqrt{n}} \right\}$$

and the probability of false acceptance of H_0 is

$$P\{\bar{X} \in R; H_1\} = 2\Phi(\frac{\Phi^{-1}(0.95)/\sqrt{n}}{8/\sqrt{n}}) - 1 = 0.697$$

We observe that, even if the probability of false rejection is held constant, the probability of false acceptance of H_0 does not decrease with n increasing. This suggests that the form of acceptance region we have chosen is inappropriate for discriminating between these two hypotheses.

c) Consider now the LRT, let $L(x)$ be the likelihood ratio and ξ be the critical value. We have

$$L(x) = \frac{f_X(x_1, ..., x_n; H_1)}{f_X(x_1, ..., x_n; H_0)} = \frac{8}{5} \exp\left\{\frac{39}{3200}\sum_{i=1}^{n}(x_i - 60)^2\right\}$$

and the rejection region is

$$\left\{x | \exp\left\{\frac{39}{3200}\sum_{i=1}^{n}(x_i - 60)^2 > \frac{5\xi}{8}\right\}\right\}$$