ECS 132 Fall 2020: Assignment 8 - 20 pts

December 13, 2020

Instructions

1. You may in no circumstances upload your homework to private tutoring websites such as CourseHero or Chegg. Remember all material related to this course is a property of the University of California and posting them is a violation of the copyright laws.
2. If you refer to a source (either a book or the internet), you must cite it.
3. You are highly urged to work on these problems on your own. See the homework grading policy. Not getting the answer correct has very low penalty. However, trying it and then figuring out where you went wrong will really help you understand the material better and you will be much better prepared for the exams. If you do discuss with others, you must list their names.
4. Write your answers in R Markdown and submit the knitted pdf on Gradescope; for due date and other details see the Homework Policy and Schedule.

1 Problem

In a newspaper trivia column, L. M. Boyd (Boyd, L. M.: The Grab Bag (syndicated newspaper column), The San Francisco Chronicle (July 17, 1999)) ponders why lie detector results are not admissible in court. His answer is that "lie detector tests pass 10 percent of the liars and fail 20 percent of the truth-tellers." If you use these percentages and take $D = 1$ to mean being deceitful and $T = 1$ to mean failing the test, what are the numerical values of the sensitivity and specificity for such a lie detector test?

1.1 Answer - 2 pt

We define the following two random variables

$$D = \begin{cases} 1 & \text{person is a liar} \\ 0 & \text{person is a truth teller} \end{cases}$$

$$T = \begin{cases} 1 & \text{fails lie detector test} \\ 0 & \text{passes lie detector test} \end{cases}$$

10% of the liars pass test. This implies $P(T = 0)|(D = 1) = 0.1$. This implies that $P(T = 1)|(D = 1) = 0.9$. Thus the sensitivity $\eta = 0.9$. Similarly, 20% of truth tellers fail the test. This implies that $P(T = 1)|(D = 0) = 0.2$. This implies that the specificity $\theta = P(T = 0)|(D = 0) = 1 - P(T = 1)|(D = 0) = 0.8$.

2 Problem

In this problem we will plot the ROC curve for the example that we did class. We had defined the following two random variables

$$D = \begin{cases} 1 & \text{if sample is infected} \\ 0 & \text{if sample is not infected} \end{cases}$$

$$T = \begin{cases} 1 & \text{test is positive} \\ 0 & \text{test is negative} \end{cases}$$

The ROC curve plots the True Positive Rate (which is also the sensitivity η) against the False Positive Rate (which is also $1 - \theta$ where $theta$ is the specificity). Recall that

$$\begin{aligned} \eta &= \text{Sensitivity or also referred to as True Positive Rate} \\ &= P(T = 1|D = 1) \\ &= \frac{P(T = 1 \cap D = 1)}{P(D = 1)} \\ &= \frac{\text{True Positives}}{\text{Prevalance}} \end{aligned}$$

$$\begin{aligned} \theta &= \text{Specificity or also referred to as True Negative Rate} \\ &= P(T = 0|D = 0) \\ &= \frac{P(T = 0 \cap D = 0)}{P(D = 0)} \\ &= \frac{\text{True Negatives}}{1 - \text{Prevalance}} \end{aligned}$$

We assume that $X|(D = 0) \sim Norm(50, 10)$ and $X|(D = 1) \sim Norm(70, 15)$. In R there are functions to obtain both the probability density function (pdf) and the cumulative distribution function (cdf) of a random variable that has a Normal distribution. These functions are `dnorm` and `pnorm`. Review the help files to see the parameters of these functions. In this problem you are required to do the following

1. Plot the pdf and the cdf of $X|(D = 1) \sim Norm(70, 15)$ for values 20 to 120.
2. Let x^* denote the cutoff value such that if $X > x^*$ then the test is positive test otherwise it is considered be negative. Plot the ROC curve for values of x^* ranging from $52 < x \leq 65$.
3. If False Positive Rate and False Negative Rates are equally bad, determine the value of x^*.

2.1 Answer - 6 ptd (1.5 pt each for PDF, CDF, ROC, and sensitivity plot/intersection)

1. Use the following codes to draw the pdf

```
> x <- c(20:120)
> plot(x, dnorm(x, 70, 15)
```

   The pdf is shown in 1

   and for the cdf

Figure 1: PDF of Normal(70,15).

```
> x <- c(20:120)
> plot(x, pnorm(x, 70, 15)
```

The cdf is shown in Figure 2

Figure 2: CDF of Normal(70,15).

2. The ROC plots the True Positive Rate (TPR) (y-axis) against the False Positive Rate (FPR) (x-axis). This the same as η vs $1 - \theta$. For this part use the following code

```
# ROC Receiver-operator Characteristic plot

  x = seq(53,65,1)
  eta = 1 - pnorm(x, 70, 15)
  theta = pnorm(x, 50, 10)
  cbind(x, eta, theta)[x >= 52 & x <= 65, ]
  #cbind(x, eta, theta)
  plot(1-theta, eta, xlim=c(0,1), ylim=c(0,1), pch=19)
  lines(c(0,1),c(1,0))
```

The ROC plot is shown in

Figure 3: The ROC plot for x^* between 52 and 65.

3. If False Positive and False Negative are equally bad, determine the value of x^*.

   Plot η and θ on the same plot and find out were they intersect. From the plot and looking at the actual values you can find $x^* = 58$. This is shown in Figure 4.

Figure 4: Sensitivity η and Specificity θ plotted against x^* . They intersect at 58.

3 Problem

A certain person goes for a run each morning. When he leaves the house for his run, he is equally likely to go out either the front or the back door and similarly when he returns, he is equally likely to go to either the front or the back door. The runner owns 3 pairs of running shoes which he takes off after the run at whichever door he happens to be. If there are no shoes at the door from which he leaves to go running, he runs barefoot. We are interested in determining the proportion of time that he runs barefoot.

1. Set this problem up as a discrete time Markov chain. Give the states and the one-step transition probability matrix.
2. Determine the proportion of days that he runs barefoot.

3.1 Answer - 3 pts (1.5 pt for setting up probability space(s), 1.5 pt for working problem)

The state of the system can specified by a tuple (i, j) where i denotes the number of shoes in the front door and j denotes the number of shoes in the back door. Since there are 3 shoes, the possible states are $\{(1, 2), (2, 1), (0, 3), (3, 0)\}$ which will be denoted as $0, 1, 2, 3$, respectively. Now, we can write down the one step transition probability matrix $\mathbf{P}$

$$\mathbf{P} = \begin{bmatrix} p_{0,0} & p_{0,1} & p_{0,2} & p_{0,3} \\ p_{1,0} & p_{1,1} & p_{1,2} & p_{1,3} \\ p_{2,0} & p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,0} & p_{3,1} & p_{3,2} & p_{3,3} \end{bmatrix}$$

Since, the jogger is equally likely to go out from a door and equally likely to return to a particular door, we get the following

$$\mathbf{P} = \begin{bmatrix} 1/2 & 1/4 & 1/4 & 0 \\ 1/4 & 1/2 & 0 & 1/4 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 3/4 \end{bmatrix}$$

To obtain the steady state probabilities $\pi = [\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \boldsymbol{\pi_3}]$ we solve $\pi = \pi\mathbf{P}$ with the additional constrain that $\sum_{i=0}^{3} \pi_i = 1$. If we do that we will get following equations.

$$\begin{aligned} \pi_0 &= 0.5\pi_0 + 0.25\pi_1 + 0.25\pi_2 && (1) \\ \pi_1 &= 0.25\pi_0 + 0.5\pi_1 + 0.25\pi_3 && (2) \\ \pi_2 &= 0.25\pi_0 + 0.75\pi_2 && (3) \\ \pi_3 &= 0.25\pi_1 + 0.75\pi_3 && (4) \\ \pi_0 + \pi_1 + \pi_2 + \pi_3 &= 1 && (5) \end{aligned}$$

Staring at equations 3 and 4 we see that $\pi_1 = \pi_3$ and $\pi_2 = \pi_0$. Using these in Equation 1, we get $\pi_1 = \pi_0$ and finally using Equation 5, we get $\pi_0 = \pi_1 = \pi_2 = \pi_3 = 0.25$.

The probability that the jogger runs barefoot is $0.5 \times \pi_2 \ + \ 0.5 \times \pi_3 = 0.25$.

4 Problem

In a Binary Symmetric (Communication) Channel (BSC) data is sent data is sent using bits 0 and 1. When the source and the destination are far apart, there are repeaters that decode the bit and transmit generate a signal. However due to noise, there decoding error, i.e., there a probability α that a bit 0 will be decoded as (and hence transmitted) as 1. Similarly, β is the probability that a bit 1 will be decoded as (and hence transmitted as) 1. Let X_0 be the bit's initial parity and and let X_n be the bits parity after the nth repeater.

1. Construct the one-step transition matrix for this Markov Chain.
2. Suppose the input stream to this communication channel consists of 80% 0s and 20% 1s. Determine the proportion of 0s and 1s after the first repeater.
3. Under the same input values as in (b) determine the proportions of 0s and 1s existing the 5th relay.

4.1 Answer - 3 pts (1 pt each for part 1, 2, 3)

1. Construct the one-step transition matrix for this Markov Chain.

   This is a two state Markov Chain. As before, we denote $p_{ij} = P(X_n = j|X_{n-1} = i)$. We will adopt the notation that we had used in class (In stating the problem I had switched the notation, i.e., $p_{11} = \beta$). In particular,

$$\begin{aligned} \mathbf{P} &= \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}. \\ &= \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}. \end{aligned}$$

2. Suppose the input stream to this communication channel consists of 80% 0s and 20% 1s. Determine the proportion of 0s and 1s after the first repeater.

   The proportion of 0 at output of Repeater 1, denoted by $P_{R1}(0)$ will be

$$\begin{aligned} P_{R1}(0) &= P(0 \text{ transmitted and } 0 \text{ received}) + P(1 \text{ transmitted and } 0 \text{ received}) \\ &= p_{00} \times P(0 \text{ transmitted}) + p_{10} \times P(1 \text{ transmitted}) \\ &= 0.8 \times (1-\alpha) + 0.2 \times \beta \end{aligned}$$

$$\begin{aligned} P_{R1}(1) &= P(0 \text{ transmitted and } 1 \text{ received}) + P(1 \text{ transmitted and } 1 \text{ received}) \\ &= p_{01} \times P(0 \text{ transmitted}) + p_{11} \times P(1 \text{ transmitted}) \\ &= 0.8 \times \alpha + 0.2 \times (1-\beta) \\ &= \text{show that this equal to } 1 - P_{R1}(0) \end{aligned}$$

3. Under the same input values as in (b) determine the proportions of 0s and 1s existing the 5th relay.

   Find $\mathbf{P}^4$ which will give $p_{ij}(4)$, i.e., probability that at the 5th repeater, the decoded bit is j given that source transmits bit i. We know what is $\mathbf{P}^4$ from Problem 1.

$$\mathbf{P}^4 = \frac{1}{\alpha+\beta} \begin{bmatrix} \beta & \alpha \\ \beta & \alpha \end{bmatrix} + \frac{(1-\alpha-\beta)^4}{\alpha+\beta} \begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}$$

$$\begin{aligned} P_{R5}(0) &= P(0 \text{ transmitted at source and } 0 \text{ received at the R5}) + P(1 \text{ transmitted at source and } 0 \text{ received at R5}) \\ &= p_{00}(4) \times P(0 \text{ transmitted}) + p_{10}(4) \times P(1 \text{ transmitted}) \end{aligned}$$

   Similarly,

$$\begin{aligned} P_{R5}(0) &= P(0 \text{ transmitted at source and } 1 \text{ received at the R5}) + P(1 \text{ transmitted at source and } 1 \text{ received at R5}) \\ &= p_{01}(4) \times P(0 \text{ transmitted}) + p_{11}(4) \times P(1 \text{ transmitted}) \end{aligned}$$

5 Problem

The number of automobile sold weekly at a certain dealership is a random variable with expected value 16. Give an upper bound to the probability that the next week sales exceed 18.

5.1 Answer - 2 pts

Let the random variable X denote the number of automobile sold in a week. Using the Markov inequality, we have

$$\begin{aligned} P(X \geq 19) &\leq \frac{E(X)}{19} \\ &\leq \frac{16}{19} \\ &\leq .84 \end{aligned}$$

6 Problem

From past experience, a professor knows that the test scores of a student taking the final exam is a random variable with mean 75 and the variance is 16. How many students would have to take the exam to ensure with probability 0.9 that the class average would be within 5 of 75? You need to use the Central Limit Theorem.

6.1 Answer - 2 pts

In this problem we use the Central Limit Theorem (CLT). If $X_1, X_2, \ldots X_n$ are independent and identically distributed random variable with mean μ and variance σ^2, then $\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$ is standard Normal, i.e., $N(0,1)$ as $n \to \infty$. In the above problem we are given that

$$P(|\frac{\sum_{i=1}^n X_i}{n} - \mu| < 5) = 0.9$$

which can be written as

$$\begin{aligned} P(-\frac{5\sqrt{n}}{4} \leq \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}4} \leq \frac{5\sqrt{n}}{4}) &= 0.9 \\ P(-\frac{5\sqrt{n}}{4} \leq Z_n \leq \frac{5\sqrt{n}}{4}) &= 0.9 \\ \Phi(\frac{5\sqrt{n}}{4}) - \Phi(-\frac{5\sqrt{n}}{4}) &= 0.9 \\ 2\Phi(\frac{5\sqrt{n}}{4}) - 1 &= 0.9 \\ \Phi(\frac{5\sqrt{n}}{4}) &= 0.95 \end{aligned}$$

Now use the standard Normal table to find the value of z for which the $\Phi(z) = 0.95$ and then solve for n.

7 Problem

A supercomputer system has 50 jobs in the queue that will be executed in sequence. The times required to execute the jobs are independent and follow a common distribution that has mean 20 seconds and standard deviation 4 seconds. Approximate the probability that the supercomputer system will complete executing at least 25 of the jobs in the first 450 seconds. [Use the Central Limit Theorem]

7.1 Answer - 2 pts

If we let X_i be the RV that denotes the time it takes to supercomputer system to execute job i,

$$X = \sum_{i=1}^{25} X_i$$

is the time it takes to grade the first 25 jobs. Since the supercomputer will execute at leats 25 jobs in the first 450 minutes of work if the time to grade the first 25 jobs is less than or equal to 450 minutes, we need to find the $P(X \leq 450)$. To approximate the probability we use the Central Limit Theorem. We are given that $E(X_i) = \mu = 20$ seconds and standard deviation $\sigma = 4$ seconds.

$$\begin{aligned}
P(X \leq 450) &= P(\frac{X - n\mu}{\sqrt{n}\sigma} \leq \frac{450 - n\mu}{\sqrt{n}\sigma}) \\
&= P(\underbrace{\frac{X - 25 \times 20}{5 \times 4}}_{= N(0,1) \text{ by Central Limit Theorem}} \leq \frac{450 - 25 \times 20}{5 \times 4}) \\
&= P(Z \leq -2.5) \\
&= P(Z \geq 2.5) \\
&= 1 - \Phi(2.5) \\
&\approx 0.006
\end{aligned}$$