Homework 2: Applied Probability and Statistics

University of Arizona CSC 380: Principles of Data Science

Homework due at 11:59pm on September 16

This assignment will strengthen your understanding of the fundamental concepts in applied probability and statistics. The questions in this assignment will build on lecture material as well as the assigned readings from (Wasserman, L. 2004. "All of Statistics").

Deliverables Submit your responses as a PDF along with an archive of any code to D2L by the stated deadline. Standard archive formats such as ZIP, BZIP, TAR will be accepted. **Show all work along with answers.** This is for your benefit as incorrect answers may receive partial credit if the work demonstrates understanding.

Problem 1: Joint, Conditional, Marginal Probability (1 point)

Suppose we throw a fair six-sided die twice in a row. Let A be a random variable representing the number on the first throw, and B be the number on the second throw. Let S be the sum of both throws. What are the following probabilities?

a) $P(A = 1, B = 2 \mid S = 3)$

b) $P(S = 3 \mid A = 1, B = 2)$

c) $P(A = 1, B = 2 \mid S = 4)$

d) $P(A = 1 \mid S = 6)$

e) $P(A = 3 \mid S = 6)$

f) $P(A = 1 \mid S = 2)$

g) $P(S = 10)$

h) $P(S = 6)$

Problem 2: Random Monkey Redux (2 points)

The random monkey from HW1 is back! To refresh your memory, our monkey types on a 26-letter keyboard, using only lowercase letters. It is well-known that monkeys type uniformly at random from the alphabet. Given that, answer the following questions:

a) Let $X_i = 1$ be the event that the monkey types the word "proof" starting on the i^{th} letter, and $X_i = 0$ otherwise. What is the marginal probability $P(X_i = 1)$? Note: I am looking for an actual number or a fraction, not a formula.

b) If the monkey types 1,000,000 characters, then what is the ***expected number of times*** *the sequence "proof" appears?*

Unlike on HW1, we are looking for the **expectation** of the number of occurrences of "proof". While this may appear challenging, I assure you that it is surprisingly straightforward and well within your capability with a little guidance. To make this question a little easier I have provided some hints below:

- *The word "proof" cannot start on letter i if it has already started on any of the four letters preceding it. So X_i depends on variables $X_{i-1}, X_{i-2}, X_{i-3}, X_{i-4}$. By the same reasoning it is also dependent on variables $X_{i+1}, X_{i+2}, X_{i+3}, X_{i+4}$. X_i is independent of all other variables.*

- *Recall that the marginal probability can be calculated using the law of total probability,*

$$P(X_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_{10^6}} P(x_1, \ldots, x_{i-1}, X_i, x_{i+1}, \ldots, x_N)$$

  The formula above just sums over all other variables, except X_i. Since X_i is independent of most RVs, this simplifies to,

$$P(X_i) = \sum_{x_{i-4}^{i-1}, x_{i+1}^{i+4}} P(x_{i-4}^{i-1}, X_i, x_{i+1}^{x_{i+4}})$$

  where x_{i-4}^{i-1} is shorthand for the series $x_{i-4}, x_{i-3}, x_{i-2}, x_{i-1}$. You will find that many terms in this sum go to zero.

- *The expected number of occurrences is given by $\mathbf{E}\left[\sum_{i=1}^{10^6} X_i\right]$. Recall that expectation is a linear operator. If you don't know what that statement implies then I recommend reviewing lecture slides from 9/7.*

Problem 3: Discrete Approximation (3 points)

In continuous probability, we often need to solve messy integrals. For example, in this class we might need to use integrals to evaluate the probability of an event under a cumulative distribution function (CDF). Rather than solve this by hand, we can approximate it using discrete intervals. This problem will explore discrete approximation of integrals using a Gaussian model. Recall that the probability density function of a Gaussian is,

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

In the questions below, we will use Python to form a discrete approximation of this model, and evaluate associated probabilities.

a) *Form a discrete approximation of the Normal PDF with mean $\mu = 70$ and standard deviation $\sigma = 2$. To do this, create an array x of evenly spaced values in the range $[68, 80]$ at increments of 2. Create an array p containing values of the PDF at each location x. Plot the result as a bar chart together with a curve at more finely spaced intervals (e.g. 0.01). You may find the following functions useful:* `numpy.arange` *and* `matplotlib.pyplot.bar`.

b) *The bar chart above is a discrete approximation of the continuous PDF. We will use it to approximate $P(68 < X \leq 80)$. Recall that this is the CDF and so,*

$$P(68 < X \leq 80) = \int_{68}^{80} \mathcal{N}(x \mid \mu, \sigma^2)\, dx.$$

We will approximate this integral using a Riemann sum. Let N be the number of grid points in your array x. The spacing between grid points is Δx and let the probability values at the i^{th} point be p_i. The Reimann sum approximation is,

$$P(68 < X \leq 80) \approx \sum_{i=1}^{N} p_i\, \Delta x$$

What is the value of the Reimann sum approximation to $P(68 < X \leq 80)$?

c) *Now, reduce the spacing $\Delta x = 0.01$ and recompute the discrete approximation of $P(68 < X \leq 80)$. How do the two approximations compare? What is the practical downside of smaller spacing?*

d) *Repeat the steps above to show the distribution over the range $[20, 120]$ and compute $P(20 \leq X < 120)$. What is the value? This interval should contain almost all of the probability in this distribution, i.e. the event is almost certain.*