midterm

Graded

Student

Jeong Yong Yang

Total Points

73 / 100 pts

Question 1

(no title) **46** / 60 pts

1.1 **(no title)** **4** / 4 pts

- ✔ **– 0 pts** Correct
- **– 4 pts** Incorrect/Missing

1.2 **(no title)** **0** / 4 pts

- **– 0 pts** Correct
- ✔ **– 4 pts** Incorrect/Missing

1.3 **(no title)** **6** / 6 pts

- ✔ **– 0 pts** Correct
- **– 6 pts** Incorrect/Missing
- **– 2 pts** A option incorrectly marked
- **– 2 pts** B option incorrectly marked
- **– 2 pts** C option incorrectly marked

1.4 **(no title)** **8** / 10 pts

- **– 0 pts** Correct
- **– 5 pts** Incorrect/Missing distribution of X
- **– 3 pts** Incorrect/Missing Expectation
- **– 2 pts** Incorrect/Missing Variance
- **– 10 pts** Incorrect/Missing
- ✔ **– 2 pts** minor error
- **– 5 pts** Partial

1.5 **(no title)** **10** / 12 pts

– **0 pts** Correct

– **12 pts** Incorrect

✔ – **2 pts** A option marked incorrect

– **2 pts** B option marked incorrect

– **2 pts** C option marked incorrect

– **2 pts** D option marked incorrect

– **2 pts** E option marked incorrect

– **2 pts** F option marked incorrect

1.6 **(no title)** **12** / 12 pts

✔ – **0 pts** Correct

– **12 pts** Incorrect/Missing

– **4 pts** A part is wrong/missing

– **4 pts** B part is wrong/missing

– **4 pts** C part is wrong/missing

1.7 **(no title)** **Resolved** **6** / 12 pts

– **0 pts** Correct

✔ – **6 pts** Wrong - Recognized Monty Hall problem

– **6 pts** Wrong - Defined events

– **12 pts** Wrong

Regrade Request **Submitted on: Mar 08**

I mentioned that this problem is a Monty Hall problem. Can I please get partial credit on this problem? Can I please get response to this....?

Partial credits given now.

Reviewed on: Mar 13

Question 2

(no title) **20** / 20 pts

✔ **– 0 pts** Correct

prior

– 0 pts Click here to replace this description.

– 2 pts wrong/miss prior for health

– 2 pts wrong/miss prior for sick

conditional probability

– 8 pts wrong/miss conditional probability

Bayes and classification

– 3 pts wrong Naive Bayes classifier formula

– 2 pts wrong posterior for health

– 2 pts wrong posterior for sick

– 1 pt wrong/miss final conclusion

Question 3

(no title) **Resolved** **7** / 20 pts

+ 0 pts Incorrect

✔ **+ 4 pts** define events

✔ **+ 3 pts** apply Bayes

+ 2 pts partial credits for binomial

+ 3 pts Correct expressions for prior, likelihood etc

+ 4 pts correct posterior

+ 6 pts correct result

Regrade Request **Submitted on:** Mar 08

I think I mentioned the events within the Bayes and wrote correct expression for prior and likelihood. Can you please double check this?

I see, I will give you credits for define the events. But the likelihood is not correct so there will be no points for that.

Reviewed on: Mar 08

CS 365 - Foundations of Data Science
Midterm
February 23rd, 2023

Name: Jeong Yong Yang

BU ID: U95912941

Instructions

- This exam is CLOSED book, notes, and devices.
- The exam consists of 3 questions on 6 pages.
- **Only the first 7 pages will be graded.** Write only your final answer and justification in the space provided.
- Please answer all questions on this exam sheet.
- Only correct, and mathematically rigorous answers will receive full credit.
- Please read through all questions carefully and be sure that you understand the instructions before working on a problem.
- The exam will be scored out of 100 possible points.
- When you deal with numbers, just write a clean formula; don't attempt to evaluate it. No calculator is needed!
- Please ask if you have any questions.

GOOD LUCK!

1.) _____ (60 points)
2.) _____ (20 points)
3.) _____ (20 points)
Σ: _____ (100 points)

Reminders

- The pdf of the uniform random variable $X \sim Unif(a, b)$ is defined as

$$f_X(x) = \begin{cases} \frac{1}{b-a} \text{ if } a < x < b, \\ 0 \text{ otherwise} \end{cases}$$

- A discrete random variable X that follows Poisson distribution has the pmf $\mathbf{Pr}[X = x] = \lambda^x e^{-\lambda}/x!, x = 0, 1, \ldots$.
- A discrete random variable K that follows a geometric distribution has the pmf $\mathbf{Pr}[K = k] = p(1-p)^{k-1}, k = 1, 2, \ldots$.
- The pdf of a Gaussian variable $X \sim N(\mu, \sigma^2)$ is $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Question 1 *Short answers [60 pts]*

1) [4 pts] Let X be a real random variable. If $\text{Var}[X] = 0$, then there exists a value α such that $\Pr(X = \alpha) = 1$.

A. Always false.

B. It depends on whether X is discrete or continuous.

(C.) Always true.

2) [4pts] A real random variable must be either discrete or continuous.
A. False (B.) True

3) [6 pts] Encircle the correct answer(s) and cross out the wrong ones concerning the Weak Law of Large Numbers (WLLN):

~~A.~~ The mean μ can be unbounded.

(B.) WLLN can be used for estimating π.

(C.) WLLN can be used for constructing confidence intervals.

4) [10 pts] Suppose 10 people walk into a party. Due to covid-19, each pair $\{i, j\}$ shakes hands with probability p. Let X be the number of handshakes. What is the distribution of X? Provide expressions for the expectation and the variance of X.

The distribution is a binomial distribution, which is $n C_k p^k (1-p)^{n-k}$.

Expected value is $n \cdot E[X_i]$ while X_i is a bernoulli distribution for only one pair, giving expected value as $n \cdot p$.

Variance is $n \cdot \text{Var}[X_i] = n \cdot p \cdot (1-p)$

5) [12 pts] Encircle the correct answer(s) and cross out the wrong ones.

(A.) For any random variable X and non-negative value t the following inequality is true:

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

(B.) The expected value of the maximum of two uniform random variables is $\frac{2}{3}$.

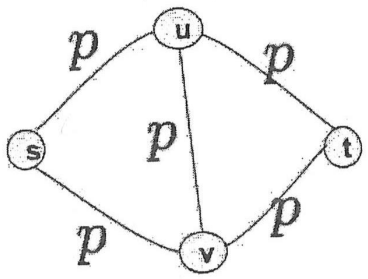
~~C.~~ The distribution of the sum of two uniform random variables in $(0, 1)$ is uniform in $(0, 2)$.

(D.) If U is a uniform random variable in $(0, 1)$ then $\lfloor nU \rfloor + 1$ is a discrete uniform random variable in $\{1, \ldots, n\}$.

~~E.~~ Chebyshev's inequality states that any non-negative random variable X satisfies

$$\mathbf{Pr}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}(X)}{t}.$$

(F.) Let X, Y, Z be uniform random variables in $(0, 1)$. The probability $\mathbf{Pr}[X + Y + Z \leq 1]$ is $\frac{1}{6}$.



p p p

p^3

Figure 1: See problem 5.

6) [12 pts] Consider the following 5 pipes connecting nodes s, u, v, t as shown in Figure 1. Each pipe fails with probability $1-p$ (i.e., succeeds with probability p), independently from the rest of the pipes. Define X to be an indicator random variable that is equal to 1 if and only if triangle (s, u, v) exists. Similarly, define Y to be an indicator random variable equal to 1 if and only if triangle (u, v, t) exist respectively. Here, we say that a triangle exists if all pipes that it consists of *do not* fail, or equivalently if they all succeed. Fill in the correct expressions:

(a) [4 pts] $\Pr(X = 0) =$ $1 - p^3$

(b) [4 pts] $\mathrm{E}(X) =$ p^3

(c) [4 pts] $\mathrm{Cov}(X, Y) =$ $p^5 - p^6$

7) [12 pts] Three prisoners, A, B, and C, are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days. The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

Discuss the following two ways of thinking. Specifically, analyze *using probabilities* whether the reasoning of the Warden and A's are correct or wrong.

• **Warden's thinking:** Clearly, either B or C will die, so I gave no new information to A. The probability of his death is still $\frac{1}{3}$.

• **A's thinking:** Given Warden's answer, either myself or C will get pardoned. Given that the pardoning was random, my chances went up from $\frac{1}{3}$ to $\frac{1}{2}$.

A's thinking is correct and Warden's thinking is wrong.
This is the Monty hall problem (similar version of it).
Since it is given that B is executed, there is only two choices left: A or C
that will be pardoned, which gives the probability of 1/2.
b) P(A lives) was $\frac{1}{3}$ in start. However, after B dies, we know that
P(A lives) when P(C dies), which happens at 1/2.

Question 2 (Naive bayes) *[20 points][1]*

You are working at the AI department of a big hospital, and you want to develop an automated system that can predict whether a person is sick or not given their symptoms. The doctors record four possible symptoms: running nose (N), coughing (C), reddened skin (R), and fever (F). So far, six patients have been tested, and the following table presents the findings, where + indicates the presence of a symptom, whereas - the lack of the symptom. For example, patient P1 has a running nose, coughing, reddened skin but no fever.

Patient	N	C	R	F	Label
P1	+	+	+	-	Sick
P2	+	+	-	-	Sick
P3	-	-	+	+	Sick
P4	+	-	-	-	Healthy
P5	-	-	-	-	Healthy
P6	-	+	+	-	Healthy

What would the Naive Bayes classifier predict for the new patient with $\langle N = +, C = +, R = -, F = - \rangle$. Show all the steps of your derivation.

Two labels/classes: sick or healthy

$P(\text{sick} \mid N=+, C=+, R=-, F=-) = \dfrac{P(N=+, C=+, R=-, F=- \mid \text{sick})\, P(\text{sick})}{P(N=+, C=+, R=-, F=-)}$

- since there is not enough sample size and $P(N=+, C=+, R=-, F=- \mid \text{sick})$ is 0, we apply naive bayes theorem and assume that each condition is conditionally independent given the class (sick or healthy)

$P(N=+ \mid \text{sick}) \cdot P(\text{sick}) \cdot P(C=+ \mid \text{sick}) P(\text{sick}) \cdot P(R=- \mid \text{sick}) P(\text{sick}) \cdot P(F=- \mid \text{sick}) P(\text{sick})$

$= \left(\frac{2}{3} \cdot \frac{1}{2}\right) \cdot \left(\frac{2}{3} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{3} \cdot \frac{1}{2}\right) \left(\frac{2}{3} \cdot \frac{1}{2}\right)$

Similarly, for healthy

$P(N=+ \mid \text{healthy}) P(\text{healthy}) \cdots P(F=- \mid \text{healthy}) P(\text{healthy})$

$= \left(\frac{1}{3} \cdot \frac{1}{2}\right) \cdot \left(\frac{1}{3} \cdot \frac{1}{2}\right) \cdot \left(\frac{2}{3} \cdot \frac{1}{2}\right) \cdot \left(1 \cdot \frac{1}{2}\right)$

Now, if we compare, we get

$\frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{2}{3}$ for sick

$\frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{3}$ for healthy ($P(\text{sick}) = P(\text{healthy})$ so they cancel)

$\frac{8}{34}$ for sick

$\frac{6}{34}$ for healthy

Since the Naive Bayes for sick gives higher probability, it predicts that the patient is sick given those conditions

[1] Keep all probabilities in the form of fractions, e.g., if you the fraction $\frac{1}{3}$ occurs keep it as is and do not write 0.333333 etc.

Question 3 *Bayes' rule [20 pts]*

A drawer contains 10 T-shirts in total, 5 red and 5 blue. A toddler removes a T-shirt from the drawer uniformly at random (uar). The toddler's parents, Alice and Bob are not allowed to see the contents of the drawer; namely they do not know what color was the removed T-shirt. What they are allowed to do is to draw repeatedly with replacement[2] from the drawer, as many times as they want.

- Alice draws 20 times in total, finding 12 red and 8 blue T-shirts.
- Bob draws only 4 times, and all the draws result in red.

Clearly, both have evidence to support that their toddler removed a blue T-shirt, so they agree on that. However, Alice is arguing that her empirical evidence is stronger than Bob's as she performed more experiments, while Bob is arguing that his empirical evidence is stronger as he only observed red.

Model the problem mathematically, and decide whose empirical evidence is stronger.

$$P(\text{red} \mid 5\text{red}, 4\text{blue}) = \frac{P(5\text{red}, 4\text{blue} \mid \text{red}) \cdot P(\text{red})}{P(5\text{red}, 4\text{blue} \mid \text{red}) \cdot P(\text{red}) + P(5\text{red}, 4\text{blue} \mid \text{blue}) \cdot P(\text{blue})}$$

$$= \frac{(\frac{1}{2}) \cdot 5/9}{\frac{1}{2} \cdot 5/9 + \frac{1}{2} \cdot 4/9} = \frac{5/9}{1} = 5/9$$

$$P(\text{blue} \mid 5\text{red}, 4\text{blue})$$

$$= \frac{P(5\text{red}, 4\text{blue} \mid \text{blue}) \cdot P(\text{blue})}{P(5\text{red}, 4\text{blue} \mid \text{blue}) \cdot P(\text{blue}) + P(5\text{red}, 4\text{blue} \mid \text{red}) P(\text{red})}$$

$$= \frac{4/9 \cdot \frac{1}{2}}{\frac{4}{9} \cdot \frac{1}{2} + 5/9 \cdot \frac{1}{2}} = 4/9$$

$$P(\text{red} \mid 5\text{blue}, 4\text{red}) = \frac{\frac{1}{2} \cdot 4/9}{4/9 \cdot \frac{1}{2} + 5/9 \cdot \frac{1}{2}} = 4/9$$

$$P(\text{blue} \mid 5\text{blue}, 4\text{red}) = \frac{5/9 \cdot \frac{1}{2}}{5/9 \cdot \frac{1}{2} + 4/9 \cdot \frac{1}{2}} = 5/9$$

- Alice ~~Bob~~

- Bob is correct since Alice has $(5/9)^{12} \cdot (4/9)^8$ while Bob has $(5/9)^4 \cdot (4/9)^0$. Bob has a higher probability.

[2] I.e., they draw a T-shirt uar, and then they put it back inside the drawer