Data Science (CSE558)

Mid-Sem

Time: 2 Hour **Full Marks: 70**

Instructions

1. **You are given 2 hours to answer all the questions.**
2. **Do not possess any electronic device except for a calculator during the exam.**
3. **Please read all questions carefully before writing your answers. Keep your answers concise.**
4. **Please clearly state any extra information (such as assumptions) used to reach your answers from the given questions. Write all the necessary steps untidy writing will fetch negative marks.**
5. **If you take any extra sheet then**
6. **If you have any questions; raise your hand, and we will come to you.**

All The Best!

1. **Consider the following null hypothesis (H_0) and alternate hypothesis (H_1). In terms of the judgement, state both Type-I and Type-II errors. (2 Marks)**
   H_0: Defendant did commit the crime.
   H_1: Defendant did not commit the crime.

2. **A drunk man is standing at the exit of a pub and wants to leave. Due to his condition, he takes a step either forward or backward. The probability of taking a forward step is 0.3. What is the probability that after 11 steps, he is 1 step behind the door? (3 Marks)**

3. **Consider a 3-faced biased die such that Pr(1) = Pr(3) = 0.25 and Pr(2) = 0.5. Use this biased die to mimic an unbiased 6-faced die? (3 Marks)**

4. **If an unbiased die is rolled 3 times then what is the probability of getting 3 consecutive heads. (5 Marks)**

5. **Consider 5 unbiased dice are thrown simultaneously. What is the probability that their sum will be divisible by 6? (5 Marks)**

6. **For a non-negative random variable X and a positive real number `a' from Markov's inequality, we know,**

a. **Using the above inequality, prove the following Chebyshev's inequality. (3 Marks)**

b. **Let are iid (independent, identically distributed) random variables where each of them share common mean and variance, i.e., μ and σ^2 respectively. Let be a random variable. Prove the following (weak law of large number) for any $\varepsilon > 0$. (5 Marks)**

7. **Suppose you toss a coin 100 times and you get 55 heads. Do you have enough evidence to reject the hypothesis that the coin is unbiased at 5% level of significance? (4 Marks)**

8. **Consider two sets of randomly sampled students from different years. The percentage of marks obtained by them in AI are in the following table. At a 5% level of significance test if there is enough evidence to suggest that mean of class 2020 and class 2021 are different.**
(5 Marks)

2020	**30**	**105**	**40**	**110**	**35**	**60**	**-15**	**70**	**20**	**65**	**50**	**55**	**25**
2021	**80**	**60**	**45**	**90**	**50**	**65**	**35**	**70**	**40**	**65**			

9. **Consider a game of luck where you can either win or lose. For every win you get Rs 1,000 and for every loss you lose Rs 100. In a single day you are allowed to play this game for any number of times until you win. In the last 7 days if you had won it in the 14^{th}, 9^{th}, 16^{th}, 3^{rd}, 11^{th}, 18^{th} and 13^{th} attempt.**
   a. **What is θ_{MLE}, the probability of winning the game? (4 Marks)**
   b. **If prior distribution is $P(\theta) = \theta^2 \cdot (1-\theta)^{14}$, then compute θ_{MAP}? (5 Marks)**

10. **Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed (i.i.d.) random variables and $X = X_1 + X_2 + \ldots + X_n$. Let, $E[X] = \mu$ and $Var(X) = \sigma^2$. Answer the following for a constant $c < \sigma$.**
    a. **Give an upper bound on $Pr(|X - E[X]| \geq c \cdot \sigma)$ using Chebyshev inequality. (2 Marks)**
    b. **For every $1 \leq i \leq n$ let, $|X_i - E[X_i]| \leq 1$. Given an upper bound on $Pr(|X - E[X]| \geq c \cdot \sigma)$ using Bernstein's inequality. (4 Marks)**
    c. **Compare both the bounds. (3 Marks)**

11. **A patient undergoes a test, and the result indicates a very rare disease that affects only 1% of the population. The test has a 99% probability of correctly returning a positive result for people who have the disease and a 10% probability of incorrectly returning a positive result for people who do not have the disease.**
    a. **What is the probability that the patient has the disease given that he tested positive? (4 Marks)**
    b. **He again retakes the tests and this time he tested negative. Now what is the probability that he does not have the disease given that this time he tested negative? (6 Marks)**

12. **Consider a music platform maintains bloom filter for each user to represent their 100 favourite songs. Explain, how to group users with similar music tastes in music for customized suggestions. The algorithm should be efficient and should not compare the song lists directly. (7 Marks)**

Rubrics:

Q1. Check the definition of Type1 and Type2 errors.

Type-I - False Positive - Reject $\mathbf{H_0}$

Type-II - False Negative - Accept $\mathbf{H_0}$

1 mark for the above definition

Statements:

The defendant did commit the crime but we lacked evidence to be proven guilty.

The defendant did not commit the crime but was proven guilty.

1 mark for the statements

Q2. (i) To identify no of forward steps = 5 and no of backward steps = 6 **1 mark**

(ii) (11 C 6)* (0.7 ^ 6) *(0.3 ^ 5) **or** (11 C 5)* (0.3 ^ 5) *(0.7 ^ 6) **1 mark**

(iii) correct answer **(0.132)** **1 mark**

Q3. (i) Identify the 6 events and these events should be unique.{ (1,2,3) , (1,3,2) , (2,1,3) , (2,3,1) , (3,1,2) , (3,2,1) } **1 mark**

(ii) check whether their probabilities are equal or not. **1 mark**

(iii) correct answer - **1/6**

(correct answer + discard unnecessary event **)** **1 mark**

(only correct answer **)** **0.5 mark**

Q6.

(a) $Y = |X - E[X]|$

$Pr(Y \geq a) = \frac{E[Y]}{a}$

$Pr(Y \geq a) = Pr(Y^2 \geq a^2) \leq \frac{E[Y^2]}{a^2}$ → 1 Mark

$E[Y^2] = E[(X - E[X])^2]$

$= E[X^2 + E[X]^2 - 2 \times E[X]]$

$= E[X^2] + \underbrace{E[E[X]^2]}_{\text{constant}} - E[2 \times E[X]]$

$= E[X^2] + E[X]^2 - 2E[X] \cdot E[X]$

$= E[X^2] - (E[X])^2$

$= Var(X)$

Proving Var → 1 Mark

stating $E[|X - E[X]|^2] = Var(X)$

$E[Y^2] = Var(X)$

$E[X^2] - (E[X])^2 = Var(X)$

→ 1 Mark

(b) $\lim_{n\to\infty} \frac{Var(\bar{X})}{\varepsilon^2} \to 0$

$Var(X_i) = \sigma^2 \quad \forall i \in [x]$

$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$ → 1 Mark

$= \sum_{i=1}^{n} Var\left(\frac{1}{n} X_i\right)$

$= \sum_{i=1}^{n} \frac{1}{n^2}\sigma^2$ → 1 Mark

OR

$= \sum_{i=1}^{n} E\left[\left(\frac{1}{n}X_i\right)^2\right] - \left(E\left[\frac{1}{n}X_i\right]\right)^2$

$= \sum_{i=1}^{n} \frac{1}{n^2}\left[\underbrace{E[X_i]^2 - (E[X_i])^2}_{\sigma^2}\right]$

$= \frac{\sigma^2}{n^2}$

$\lim_{n\to\infty} Pr(|X - E[X]| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$ → 1 Mark

$\lim_{n\to\infty} Pr(|X - E[X]| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} \to 0$

→ 2 Marks

Q7. **(i)** Correct Null Hypothesis i.e. Coin is unbiased **1 mark**

(ii) Correct test identified i.e. Z-Test for proportion **1 mark**

(iii) Correct identification of p and p_0 (-0.5 if calc is wrong despite correct p & p_0) **1 mark**

(iv) Correct conclusion i.e. Can't Reject H_0 **1 mark**

Q9).

Q9.

(9).

$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ P(D \mid \theta)$ → 1 mark

Probability of winning at ith attempt = $(1-\theta)^{i-1}\theta$ (geometric dist)

→ 0.5 mark

$P(D \mid \theta) = \prod_{i=1}^{7} p(d_i, \theta)$ (d_i is the no. of attempts)

$= \prod_{i=1}^{7} (1-\theta)^{d_i - 1}\theta$

→ 1 mark

Differentiating both sides & equating to 0

$\frac{d\,P}{d\theta} = \sum_{i=1}^{7} \frac{d\,(1-\theta)^{d_i-1}\theta^{*}}{d\theta} = 0$

1 mark

$\Rightarrow \sum_{i=1}^{7} (d_i - 1)(-1)(1-\theta)^{d_i-2}\theta^{*} + \sum_{i=1}^{7} (1-\theta^{*})^{d_i - 1}\,d\theta = 0$

$$\Rightarrow \sum_{i=1}^{7} (d_i - 1)\theta^* = \sum_{i=1}^{7} (1-\theta^*)$$

$$\Rightarrow \theta^* = \frac{7}{\sum_{i=1}^{7} d_i} = \frac{7}{84} = \frac{1}{12}$$

$$\Rightarrow \theta_{MLE} = \frac{1}{12}$$ - 0.5 mark

(Use double differentiation to confirm maxima)

(b) $\theta_{MAP} = \arg\max_{\theta} P(D|\theta) \cdot P(\theta)$

- 1 Mark

$$P(D|\theta) = \prod_{i=1}^{7} p(d_i, \theta)$$

$$= \prod_{i=1}^{7} (1-\theta)^{d_i - 1}\theta$$ - 1 mark

$$\Rightarrow \theta_{map} = \arg\max_{\theta} \left(\prod_{i=1}^{7} (1-\theta)^{d_i - 1}\theta \right) \cdot P(\theta)$$

$$= \arg\max_{\theta} \left(\prod_{i=1}^{7} (1-\theta)^{d_i - 1}\theta \right) \theta^2 (1-\theta)^{14}$$

- 1 Mark

1 mark

Differentiating both sides and equating to zero,

$$\frac{d\left((1-\theta)^{91}\theta^9\right)}{d\theta} = 0$$

$\Rightarrow 91\,(1-\theta)^{90}(-1)\theta^{9} + 9\theta^{8}(1-\theta)^{91} = 0$

$\Rightarrow 91\theta^{9} = 9 - 9\theta^{9}$

$\Rightarrow \theta^{9} = \frac{9}{100}$

$\Rightarrow \theta_{MAP} = \frac{9}{100}$ — 0.5 Mark

10.

(10)

(a) $P(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$

$k = c\sigma$

$P(|X-\mu| \geq c\sigma) \leq \frac{\sigma^2}{c^2\sigma^2}$ (1 marks)

$\leq \frac{1}{c^2}$ (2 marks)

(b) $Pr(|(X - E(X))| \geq t) \leq 2e\left(\frac{-t^2}{\sigma^2 + bt}\right)$

here $b = 1$

$t = c\sigma$

$\leq 2e\left(\frac{-c^2\sigma^2}{c^2\sigma^2 + c\sigma}\right)$ (2 marks)

eqn — (A) ——— (3)

if you did more step for simplification of the eqn (A) and get

$\leq 2e\left(\frac{-3c^2}{2(3c+1)}\right)$ or (3) $\frac{1}{e^{c^2}}$ then get 4 marks

(c) if you compare both of them without any logic. then got 1- marks.

if you give the reason like when c is increasing then How $\frac{1}{e^{c^2}}$ decrease exponentially. and compare $\frac{1}{c^2}$ and $\frac{1}{e^{c^2}}$ then got ___ (3) marks

11.)

11.) D: Having the disease

(a) P: Being Tested Positive

$P(D)$ = Probability of having disease = 0.01

$P(\sim D)$ = Probability of not having disease = 0.99

$P(P/D)$ = Prob that he tested +ve if he had the disease = 0.99

$P(P/\sim D)$ = Prob that he tested +ve even if he doesn't have the disease = 0.1

(1) Mark

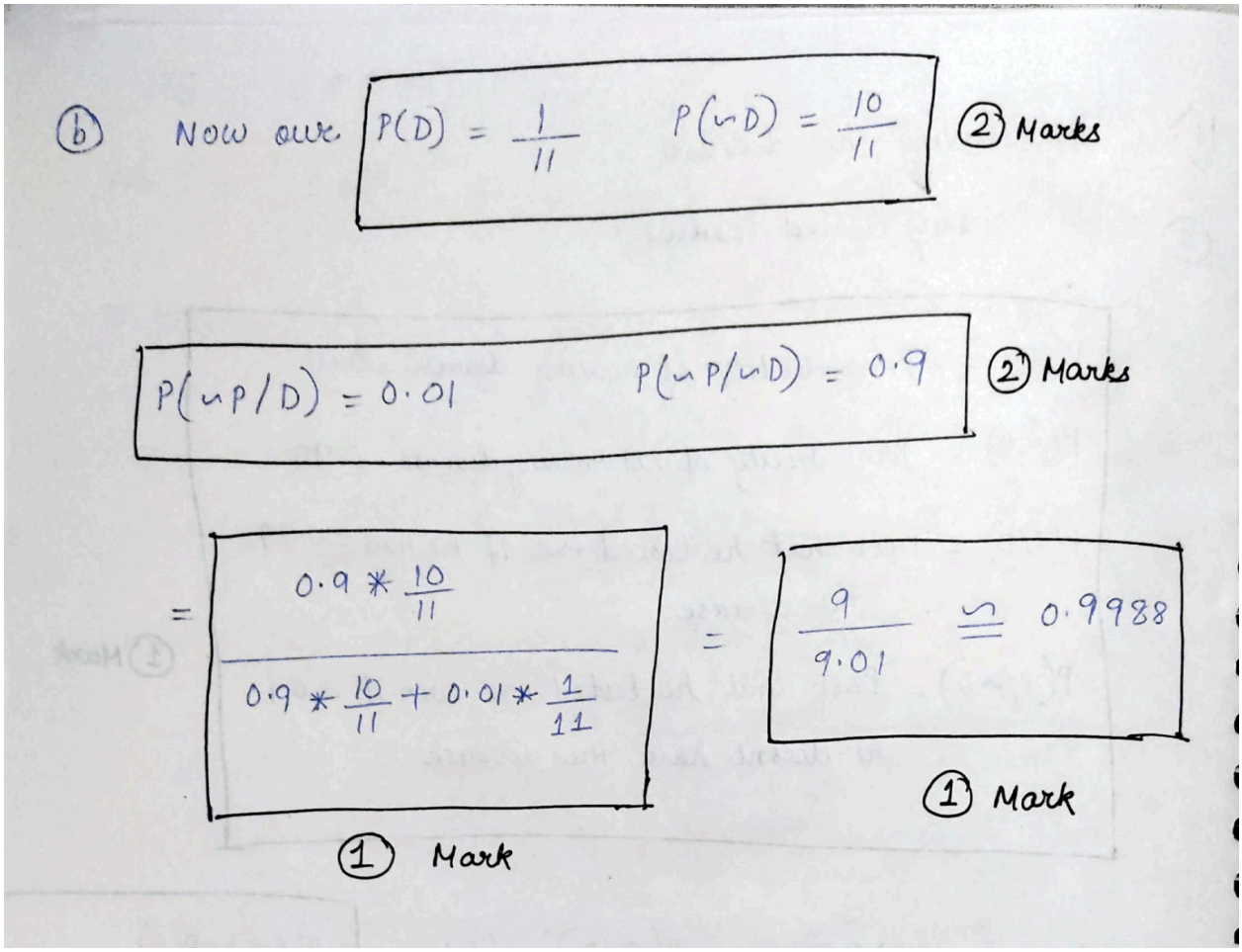
$$P(D/P) = \frac{P(D \cap P)}{P(P)} = \frac{P(P/D) \cdot P(D)}{P(P \cap D) \cdot P(P \cap \sim D)} = \frac{P(P/D) \cdot P(D)}{P(P/D) \cdot P(D) + P(P/\sim D) \cdot P(\sim D)}$$

(1) Mark

$$= \frac{0.99 * 0.01}{0.99 * 0.01 + 0.99 * 0.1} = \frac{0.01}{0.11} = \frac{1}{11} \simeq 0.0909$$

(1) Mark (1) Mark

Que.5



1.If the answer is correct but there is no logical explanation present: 2 Marks out of 5.
2.If the answer is correct with some logical explanation but not exactly correct: 4 Marks out of 5.
3.If the answer is correct with correct explanation: 5 Marks out of 5.

Que8.
1.Handling of outliers (steps must be mentioned in the answer): 1 Mark
2.Degree of freedom: 1 Mark
3.Formulas: 2 Marks
4. If hypothesis is accepted-1 Mark, if not then 0 Mark

Q12. Has correctly applied a bloom filter (on the definite finite quantity i.e. songs) : 4/
Has correctly applied a bloom filter (on the much higher quantity i.e. users) : 2/
Has not applied bloom filter / applied wrong : 0
Algorithm is less than $O(n^2)$ and is adheres to the above : 3/
Algorithm is $O(n^2)$ and is adheres to the above : 1/
Else : 0