STAT 479: Homework 1

Due: 11:59PM January 31, 2025 by Canvas

Part 1: Probability Basics

1. **Conditional Probabilities** (10 points)

   Suppose we roll two dice. Let A be the event "the sum is greater than 7," and B be the event "the first die shows a 4." Select the correct answers for the following probabilities:

   (a) $P(A)$:

   A. $\frac{5}{12}$
   B. $\frac{7}{12}$
   C. $\frac{1}{6}$
   D. $\frac{1}{2}$

   (b) $P(A \cap B)$:

   A. $\frac{1}{6}$
   B. $\frac{1}{12}$
   C. $\frac{5}{36}$
   D. $\frac{1}{4}$

   (c) $P(A|B)$:

   A. $\frac{1}{3}$
   B. $\frac{5}{12}$
   C. $\frac{2}{3}$
   D. $\frac{3}{4}$

   (d) Are A and B independent?

   A. Yes, $P(A \cap B) = P(A) \cdot P(B)$.
   B. No, $P(A \cap B) \neq P(A) \cdot P(B)$.

Answer: Write your solution here. For multiple choice questions, only the letter answer is required.

(a)

(b)

(c)

(d)

2. **Bayes' Rule** (10 points)

A medical test for a rare disease has the following properties, where T is the test and D is the disease:

- $P(T^+|D) = 0.95$,
- $P(T^+|\neg D) = 0.02$,
- $P(D) = 0.001$.

(a) Using Bayes' rule, compute $P(D|T^+)$. Select the correct answer:

A. 0.32
B. 0.047
C. 0.019
D. 0.001

(b) At what $P(D)$ would the test give 95% confidence?

A. 0.10
B. 0.20
C. 0.30
D. 0.50

Answer:

(a)

(b)

3. **Continuous Random Variables** (10 points)

The probability density function (PDF) of a continuous random variable X is:

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Is $f(x)$ a valid PDF? Select the correct answer:

A. Yes, $\int_{-\infty}^{\infty} f(x)dx = 1$.

B. No, $f(x)$ does not integrate to 1.

(b) Compute $P(0.25 \le X \le 0.75)$. Select the correct answer:

A. 0.25
B. 0.50
C. 0.75
D. 1.00

(c) Determine the expected value $E[X]$. Select the correct answer:

A. 0.25
B. 0.33
C. 0.50
D. 0.67

(d) Determine the variance of X. Select the correct answer:

A. 0.056
B. 0.111
C. 0.167
D. 0.222

Answer:
(a)
(b)
(c)
(d)

4. **Joint and Marginal Probabilities** (10 points)

Two random variables X and Y have the following joint probability mass function (PMF):

$$P(X = x, Y = y) = \begin{cases} \frac{x+y}{10}, & x, y \in \{1, 2\}, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Which property must hold for $P(X, Y)$ to be a valid joint PMF?

A. $P(X = x, Y = y) \ge 0$ for all x, y.
B. $\sum_x \sum_y P(X = x, Y = y) = 1$.
C. Both (A) and (B).

D. None of the above.

(b) Compute the marginal probability $P(X = 1)$.

A. 0.30
B. 0.35
C. 0.40
D. 0.45

(c) Compute the conditional probability $P(Y = 2 \mid X = 2)$.

A. 0.40
B. 0.50
C. 0.60
D. 0.70

Answer:

(a)

(b)

(c)

Part 2: Estimation

5. **MLE and MAP for a Possibly Biased Coin** (10 points)

Suppose you are flipping a coin that might be biased (i.e., the probability of heads θ is unknown and not necessarily 0.5). You flip the coin n times and observe k heads.

(a) Using the Bernoulli likelihood function, what is the Maximum Likelihood Estimate (MLE) for θ?

A. $\frac{n}{k}$
B. $\frac{k}{n}$
C. $k \cdot n$
D. $1 - \frac{k}{n}$

(b) Suppose you have prior knowledge that the coin is likely close to fair, modeled using a Beta prior $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior distribution is:

$$\text{Posterior}(\theta|\text{data}) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}.$$

What is the MAP estimate for θ?

A. $\frac{k}{n}$
B. $\frac{k+\alpha-1}{n+\alpha+\beta-2}$
C. $\frac{k}{n+\alpha+\beta}$
D. $\frac{k+\alpha}{n+\beta}$

Answer:
(a)
(b)

Part 2: Estimation

6. **MLE and MAP for a Possibly Biased Coin** (10 points)

Suppose you are flipping a coin that might be biased (i.e., the probability of heads θ is unknown and not necessarily 0.5). You flip the coin n times and observe k heads.

(a) Using the Bernoulli likelihood function, what is the Maximum Likelihood Estimate (MLE) for θ?

A. $\frac{n}{k}$
B. $\frac{k}{n}$

C. $k \cdot n$

D. $1 - \frac{k}{n}$

(b) Suppose you have prior knowledge that the coin is likely close to fair, modeled using a Beta prior $\theta \sim \text{Beta}(\alpha, \beta)$. The posterior distribution is:

$$\text{Posterior}(\theta|\text{data}) \propto \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}.$$

What is the MAP estimate for θ?

A. $\frac{k}{n}$

B. $\frac{k+\alpha-1}{n+\alpha+\beta-2}$

C. $\frac{k}{n+\alpha+\beta}$

D. $\frac{k+\alpha}{n+\beta}$

Answer:

(a)

(b)

(c)

(d)

7. **Poisson Distribution** (10 points)

The Poisson distribution models the probability of observing a count x_i with the rate parameter λ:

$$P(x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

(a) Which of the following represents the likelihood function $L(\lambda)$ for a single observation x_1?

A. $\frac{\lambda^{x_1} e^{-\lambda}}{x_1!}$

B. $\lambda x_1 e^{-\lambda}$

C. $\lambda^{x_1-1} e^{-\lambda}$

D. λe^{-x_1}

(b) Which of the following represents the likelihood function $L(\lambda)$ for n independent observations $x_1, x_2, \ldots, x_n$?

A. $\prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$

B. $\prod_{i=1}^{n} \lambda x_i e^{-\lambda}$

C. $\sum_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$
D. $\prod_{i=1}^{n} \lambda^{x_i - 1} e^{-\lambda}$

(c) Which of the following represents the log-likelihood function $\log L(\lambda)$?
A. $\sum_{i=1}^{n} x_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$
B. $n\lambda - \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log(x_i!)$
C. $n \log \lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$
D. $\sum_{i=1}^{n} x_i \lambda - n\lambda$

(d) What is the MLE for λ, the rate parameter?
A. $\frac{\sum_{i=1}^{n} x_i}{n}$
B. $\sum_{i=1}^{n} x_i$
C. $n \cdot \sum_{i=1}^{n} x_i$
D. $\frac{n}{\sum_{i=1}^{n} x_i}$

Answer:
(a)
(b)
(c)
(d)

8. **Deriving Backpropagation from MLE** (10 points)

Consider a neural network with one hidden layer. The network's output is given by:

$$\hat{y} = \sigma(w_2 \cdot h),$$

where:

- $h = \sigma(w_1 \cdot x)$,
- $\sigma(z)$ is the sigmoid activation function defined as $\sigma(z) = \frac{1}{1+e^{-z}}$,
- w_1 and w_2 are weights,
- x is the input.

Assume that the training data (x, y) are drawn i.i.d. from a distribution, and the network is trained using Maximum Likelihood Estimation (MLE). For binary classification, the likelihood of the data is given by:

$$P(y|x, w_1, w_2) = \hat{y}^y (1 - \hat{y})^{1-y},$$

where $\hat{y}$ is the predicted probability for the positive class.

(a) Which of the following represents the negative log-likelihood $\mathcal{L}$?

A. $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$

B. $y \log(1 - \hat{y}) + (1 - y) \log \hat{y}$

C. $\hat{y} \cdot y + (1 - \hat{y}) \cdot (1 - y)$

D. $y \cdot \hat{y} + (1 - y) \cdot (1 - \hat{y})$

(b) What is the gradient of $\mathcal{L}$ with respect to w_2?

A. $(\hat{y} - y) \cdot h$

B. $(y - \hat{y}) \cdot h$

C. $\hat{y} \cdot (1 - h)$

D. $y \cdot (1 - \hat{y})$

(c) What is the gradient of $\mathcal{L}$ with respect to w_1?

A. $(\hat{y} - y) \cdot w_2 \cdot \sigma'(w_1 \cdot x) \cdot x$

B. $(\hat{y} - y) \cdot \sigma'(w_1 \cdot x) \cdot x$

C. $(\hat{y} - y) \cdot w_2 \cdot x$

D. $(\hat{y} - y) \cdot \sigma'(x) \cdot w_2$

Answer:

(a)

(b)

(c)