LUMS
A Not-for-Profit University

School of Science and Engineering

AI 501 Mathematics for Artificial Intelligence

ASSIGNMENT 4 – SOLUTIONS

Due Date: 11:55 pm, Tuesday, December 17, 2024.
Format: 6 problem, for a total of 100
Instructions:

- You are allowed to collaborate with your peers but copying your colleague's solution is strictly prohibited. This is not a group assignment. Each student must submit his/her own assignment.
- Solve the assignment on blank A4 sheets and staple them before submitting.
- Submit in-class or in the dropbox labeled AI-501 outside the instructor's office.
- **Write your name and roll no. on the first page.**
- Feel free to contact the instructor or the teaching assistants if you have any concerns.

- You represent the most competent individuals in the country, do not let plagiarism come in between your learning. In case any instance of plagiarism is detected, the disciplinary case will be dealt with according to the university's rules and regulations.
- We require you to acknowledge any use or contributions from generative AI tools. Include the following statement to acknowledge the use of AI where applicable.

  I have used [insert Tool Name] to [write, generate, plot or compute; explain specific use of generative AI] [number of times].

Problem 1 (20 marks)

Probability basics, Conditional Probability, Bayes theorem

1. Two factories supply light bulbs to the market. Bulbs from factory X work for over 5000 hours in 99% of cases, whereas bulbs from factory Y work for over 5000 hours in 95% of cases. It is known that factory X supplies 60% of the total bulbs available in the market.
   (a) What is the probability that a purchased bulb will work for longer than 5000 hours?
   (b) Given that a light bulb works for more than 5000 hours, what is the probability that it was supplied by factory Y?
   (c) Given that a light bulb does not work for more than 5000 hours, what is the probability that it was supplied by factory X?
2. A multiple choice exam has 4 choices for each question. The student has studied enough so that the probability they will know the answer to a question is 0.5, the probability that the student will be able to eliminate one choice is 0.25, otherwise all 4 choices seem equally plausible. If they know the answer they will get the question correct. If not they have to guess from the 3 or 4 choices. As the teacher you would like the test to measure what the student knows, and not how well they can guess. If the student answers a question correctly, what is the probability that they actually know the answer?
3. Suppose 30% of the women in a class received an A on the test and 25% of the men received an A. The class is 60% women. Given that a person chosen at random received an A, what is the probability this person is a woman?

Solution:

1. (a) Let H be the event a bulb works over 5000 hours, X be the event that a bulb comes from factory X, and Y be the event a bulb comes from factory Y. Then by the law of total probability:

$$P(H) = P(H \mid X)P(X) + P(H \mid Y)P(Y) = (0.99)(0.6) + (0.95)(0.4) = 0.974.$$

   (b)
$$P(Y \mid H) = \frac{P(H \mid Y)P(Y)}{P(H)} = \frac{(0.95)(0.4)}{0.974} \approx 0.39.$$

   (c)
$$P(X \mid H^c) = \frac{P(H^c \mid X)P(X)}{P(H^c)} = \frac{P(H^c \mid X)P(X)}{1 - P(H)} = \frac{(1 - 0.99)(0.6)}{1 - 0.974} = \frac{(0.01)(0.6)}{0.026} \approx 0.23.$$

2. Let C be the event a student gives the correct answer, K be the event a student knows the correct answer, E be the event a student can eliminate one incorrect answer, and G be the event a student has to guess an answer. Using Bayes theorem we have:

$$P(K \mid C) = \frac{P(C \mid K)P(K)}{P(C)} = \frac{P(C \mid K)P(K)}{P(C \mid K)P(K) + P(C \mid E)P(E) + P(C \mid G)P(G)}$$
$$= \frac{1 \cdot \frac{1}{2}}{1 \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4}} = \frac{24}{31} \approx 0.774,$$

   that is, approximately 77.4% of the students know the answer if they give the correct answer.

3. Let A be the event of receiving an A, W be the event of being a woman, and M the event of being a man. We are given $P(A \mid W) = 0.30$, $P(A \mid M) = 0.25$, $P(W) = 0.60$, and we want $P(W \mid A)$. From the definition:

$$P(W \mid A) = \frac{P(W \cap A)}{P(A)}.$$
$$P(W \cap A) = P(A \mid W)P(W) = (0.30)(0.60) = 0.18.$$
$$P(A) = P(W \cap A) + P(M \cap A).$$
$$P(M \cap A) = P(A \mid M)P(M) = (0.25)(0.40) = 0.10.$$

$$P(A) = P(W \cap A) + P(M \cap A) = 0.18 + 0.10 = 0.28.$$

$$P(W \mid A) = \frac{P(W \cap A)}{P(A)} = \frac{0.18}{0.28}.$$

Problem 2 (15 marks)

Discrete and Continuous random variables

In a machine learning application for spam email detection, a probabilistic model is used to decide whether an email is spam or not based on certain features. The model uses two types of random variables:

1. A continuous random variable X: The time it takes for the model to process an email. The Cumulative distribution function is given as:

$$F_X(x) = \begin{cases} 1 - \frac{a^3}{x^3} & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases}$$

   Find the density function, mean, and variance of the random variable X.

2. A discrete random variable Y: The number of spam-triggering keywords detected in an email. We know that $Y \sim \text{Poisson}(\lambda = 3)$. Determine the probability that at least 2 spam-triggering words are detected and the expected number of spam-triggering words in an email.

Solution:

1.
$$f_X(x) = \frac{dF_X(x)}{dx} = \begin{cases} 3a^3x^{-4} & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases}$$

   Also,
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^{\infty} x \cdot 3a^3x^{-4} dx = 3a^3 \int_a^{\infty} x^{-3} dx = 3a^3 \left[-\frac{1}{2}x^{-2} \right]_a^{\infty} = \frac{3a}{2}.$$

   Finally, we have
$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_a^{\infty} x^2 \cdot 3a^3x^{-4} dx = 3a^3 \int_a^{\infty} x^{-2} dx = 3a^3 \left[-x^{-1} \right]_a^{\infty} = 3a^2,$$

   so the variance is
$$\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 3a^2 - \left(\frac{3a}{2} \right)^2 = \frac{3a^2}{4}.$$

2. Probability is
$$P(X \geq 2) = 1 - P(X < 2) = 1 - (P(X = 0) + P(X = 1))$$
$$P(X = 0) = \frac{3^0 e^{-3}}{0!} = e^{-3}, \quad P(X = 1) = \frac{3^1 e^{-3}}{1!} = 3e^{-3}$$
$$P(X \geq 2) = 1 - (e^{-3} + 3e^{-3}) = 1 - 4e^{-3} \approx 0.801$$

   Expectation is
$$\mathbb{E}[X] = \lambda = 3$$

Problem 3 (15 marks)

Regularized Logistic Regression

The objective of logistic regression is to find a decision boundary between two or more distinct classes. We will focus on binary classification here. Given some data features $\mathbf{x} \in \mathbb{R}^{n+1}$ (with a 1 for the intercept term in the first position) for some $y \in \{0, 1\}$, we are essentially trying to learn a vector $\theta \in \mathbb{R}^{n+1}$ such that:

$$z = \mathbf{w}^T \mathbf{x}$$
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

And $\sigma(z) = 1$ when $y = 1$ and $\sigma(z) = 0$ when $y = 0$.
To find the θ, we are going to optimize the binary cross-entropy loss function with L2 regularization:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log\left(\sigma(\theta^T \mathbf{x}_i)\right) + (1 - y_i) \log\left(1 - \sigma(\theta^T \mathbf{x})\right) + \lambda ||\theta||_2^2$$

Since the minimizer of the cross-entropy loss above has has no analytical solution, it must be optimized via gradient descent. Derive the gradient of the regularized loss function with respect to θ.

Solution:

$$L = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\sigma(\vec{\theta}^T \vec{x}_i)) + (1 - y_i) \log(1 - \sigma(\vec{\theta}^T \vec{x}_i)) + \lambda ||\vec{\theta}||_2^2$$

Vectorize the loss function

$$L = -\frac{1}{N} \left[\vec{y} \log(\sigma(\vec{\theta}^T X)) + (1 - \vec{y}) \log(1 - \sigma(\vec{\theta}^T X)) \right] + \lambda ||\vec{\theta}||_2^2$$

$$\frac{\partial}{\partial \vec{\theta}} \log(\sigma(\vec{\theta}^T X)) = \frac{\partial}{\partial \vec{w}} \left(\frac{1}{1 + e^{-\vec{\theta}^T \vec{x}}} \right) = \frac{e^{-\vec{\theta}^T \vec{x}}}{(1 + e^{-\vec{w}^T \vec{x}})^2} \vec{x} = \sigma(-\vec{w}^T \vec{x}) \vec{x}$$

$$\frac{\partial \Theta}{\partial \vec{w}} = \frac{\partial}{\partial \vec{w}} \left(1 - \frac{1}{1 + e^{-\vec{w}^T \vec{x}}} \right) = -\sigma'(\vec{w}^T \vec{x}) \vec{x}$$

Problem 4 (20 marks)

Maximum Likelihood Estimate

1. Consider i.i.d drawing of random variables $X_1, X_2, \ldots, X_N$ from a Gaussian distribution with unknown mean and variance. Given the observation $X_1 = x_1, \ldots, X_N = x_N$, derive the MLE for the unknown mean and variance. Recall that the MLE for the unknown parameters can be obtained as
$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} f_{X_1,\ldots,X_N}\left(x_1, \ldots, x_N \middle| \mu, \sigma^2\right)$$
2. Now consider an i.i.d. drawing $X_1, X_2, \ldots, X_N$ from a Poisson distribution with unknown parameter λ (e.g., X_i could represent the number of customers arriving at a service desk per hour over a day). Given the observation $X_1 = x_1, \ldots, X_N = x_N$, show that the MLE for the unknown parameter is given as
$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
Recall that the MLE for λ will be obtained as
$$\hat{\lambda} = \arg\max_{\lambda} P_{X_1,\ldots,X_N}\left(x_1, \ldots, x_N \middle| \lambda\right)$$

Solution:

1.
$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} f_{X_1,\ldots,X_N}\left(x_1, \ldots, x_N \middle| \mu, \sigma^2\right)$$
$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(\frac{-(x_i-\mu)^2}{2\sigma^2}\Big)$$
take the natural log to convert the product of probabilities to a sum of its log. taking a log will still keep the MLE answer the same.
$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu,\sigma^2} -N\ln(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}$$
separate the equations to calculate $\hat{\mu}$ and $\hat{\sigma}^2$
$$\hat{\mu} = \arg\min_{\mu}\sum_{i=1}^{N}(x_i-\mu)^2, \hat{\sigma}^2 = \arg\max_{\mu} -N\ln(\sqrt{2\pi\sigma^2}) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}$$
first, we will calculate the $\hat{\mu}$ by taking the derivative wrt μ and equating it to 0.
$$\frac{d}{d\mu}\hat{\mu} = -2\sum_{i=1}^{N}(x_i-\hat{\mu}) = 0$$
$$\sum_{i=1}^{N} x_i - N\hat{\mu} = 0$$
$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$
now we will calculate the $\hat{\sigma}^2$ by taking the derivative wrt σ^2 and equating it to 0 as well.
$$\frac{d}{d\sigma^2}\hat{\sigma}^2 = -N\frac{1}{2\sigma^2} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^4} = 0$$
$$\frac{N}{2\hat{\sigma}^2} = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\hat{\sigma}^4}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

we have found the MLE of μ and σ^2.

2.

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

take the natural log to convert the product of probabilities to a sum of its log. taking a log will still keep the mle answer the same.

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{N} ln\Big(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\Big)$$

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{N} [ln(\lambda^{x_i}) + ln(e^{-\lambda}) - ln(x_i!)]$$

the term $ln(x_i!)$ goes away since it is not dependant on λ.

$$\hat{\lambda} = \arg\max_{\lambda} \sum_{i=1}^{N} [x_i ln(\lambda) - \lambda]$$

taking the derivative wrt λ and equating it to 0.

$$\frac{d}{d\lambda}\hat{\lambda} = \sum_{i=1}^{N} \Big(\frac{x_i}{\hat{\lambda}} - 1\Big) = 0$$

$$\frac{\sum_{i=1}^{N} x_i}{\hat{\lambda}} - N = 0$$

$$\hat{\lambda} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

the MLE for the parameter is shown/proven.

Problem 5 (15 marks)

You are tasked with building a binary classification model that predicts whether an email is **spam** (class 1) or **not spam** (class 0). Assume the following:

- The dataset consists of binary features $x \in \{0, 1\}$, where $x_i = 1$ indicates that a certain word appears in the email.
- The classification model uses the Bernoulli distribution for the features.
- Given a training dataset D of n emails, you are required to estimate the probability that a word appears in spam and non-spam emails using both MLE and MAP approaches. The training dataset D has the following word occurrences for a particular word w:

$$\text{Class 1 (Spam)} : n_1 = 20 \text{ emails, with } k_1 = 15 \text{ emails containing the word } w.$$

$$\text{Class 0 (Not Spam)} : n_0 = 30 \text{ emails, with } k_0 = 5 \text{ emails containing the word } w.$$

Solution:

The MLE estimate for the Bernoulli parameter θ (probability of a word appearing) is given by:

$$\theta_{\text{MLE}} = \frac{\text{Number of successes (word appearing)}}{\text{Total number of trials (emails)}}.$$

For Class 1 (Spam): We are given $n_1 = 20$ (total emails) and $k_1 = 15$ (emails containing the word w):

$$\theta_{\text{MLE, spam}} = \frac{k_1}{n_1} = \frac{15}{20} = 0.75.$$

For Class 0 (Not Spam): We are given $n_0 = 30$ (total emails) and $k_0 = 5$ (emails containing the word w):

$$\theta_{\text{MLE, not spam}} = \frac{k_0}{n_0} = \frac{5}{30} \approx 0.1667.$$

The MAP estimate incorporates a prior belief using the Beta distribution $\text{Beta}(\alpha, \beta)$, which is the conjugate prior for the Bernoulli likelihood. The MAP estimate is given by:

$$\theta_{\text{MAP}} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2},$$

where:

- k = number of successes (emails containing the word),
- n = total number of trials (emails),
- α, β = hyperparameters of the Beta prior.

Assuming a Uniform Prior: For a uniform prior $\text{Beta}(1, 1)$ ($\alpha = 1, \beta = 1$), the MAP estimate simplifies to:

$$\theta_{\text{MAP}} = \frac{k + 1 - 1}{n + 1 + 1 - 2} = \frac{k}{n}.$$

Thus, under the uniform prior, the MAP estimate is identical to the MLE estimate.

For Class 1 (Spam):

$$\theta_{\text{MAP, spam}} = \frac{k_1}{n_1} = \frac{15}{20} = 0.75.$$

For Class 0 (Not Spam):

$$\theta_{\text{MAP, not spam}} = \frac{k_0}{n_0} = \frac{5}{30} \approx 0.1667.$$

Problem 6 (15 marks)

Naive Bayes

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Given the dataset, calculate the probability that the answer is "Yes" or "No" for the following conditions: Outlook = Rainy, Temperature = Mild, Humidity = Normal, Windy = True. Compute the posterior probabilities to determine which class (Yes or No) has the higher probability. To solve this using the Naive Bayes algorithm, we calculate the posterior probabilities:

$$P(\text{Yes} \mid \text{conditions}) = \frac{P(\text{conditions} \mid \text{Yes}) \cdot P(\text{Yes})}{P(\text{conditions})},$$

and

$$P(\text{No} \mid \text{conditions}) = \frac{P(\text{conditions} \mid \text{No}) \cdot P(\text{No})}{P(\text{conditions})},$$

where:

$$P(\text{conditions} \mid \text{Yes}) = P(\text{Outlook} = \text{Rainy} \mid \text{Yes}) \cdot P(\text{Temperature} = \text{Mild} \mid \text{Yes})$$

$$\cdot P(\text{Humidity} = \text{Normal} \mid \text{Yes}) \cdot P(\text{Windy} = \text{True} \mid \text{Yes}),$$

and similarly:

$$P(\text{conditions} \mid \text{No}) = P(\text{Outlook} = \text{Rainy} \mid \text{No}) \cdot P(\text{Temperature} = \text{Mild} \mid \text{No})$$

$$\cdot P(\text{Humidity} = \text{Normal} \mid \text{No}) \cdot P(\text{Windy} = \text{True} \mid \text{No}).$$

The classification decision is based on the higher posterior probability.

Solution:

$$\text{Likelihood of Yes} = P(\text{Outlook} = \text{Rainy} \mid \text{Yes}) \cdot P(\text{Temp} = \text{Mild} \mid \text{Yes}) \cdot P(\text{Humidity} = \text{Normal} \mid \text{Yes})$$

$$\cdot P(\text{Windy} = \text{True} \mid \text{Yes}) \cdot P(\text{Yes}) = \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \mathbf{0.0141}.$$

$$\text{Likelihood of No} = P(\text{Outlook} = \text{Rainy} \mid \text{No}) \cdot P(\text{Temp} = \text{Mild} \mid \text{No}) \cdot P(\text{Humidity} = \text{Normal} \mid \text{No})$$

$$\cdot P(\text{Windy} = \text{True} \mid \text{No}) \cdot P(\text{No}) = \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{5} \cdot \frac{5}{14} = \mathbf{0.0103}.$$

After normalization, the probabilities are calculated as follows:

$$\text{Yes} = \frac{0.0141}{0.0141 + 0.0103} = \mathbf{0.58}$$

$$\text{No} = \frac{0.0103}{0.0141 + 0.0103} = \mathbf{0.42}$$

— End of Assignment —