CIS3333: Mathematics of Machine Learning Fall 2025

Homework 1

Due: September 10, 2025, 11:59 PM ET

Submission Instructions: Submit a single PDF (clear scans or photos compiled) to Gradescope and assign pages for each problem. Show key steps and justify answers.

Collaboration & AI Policy: You may discuss approaches with classmates, but write up your own solutions and list collaborators. If you use computational tools (including LLMs) for checking, cite them and ensure the reasoning is your own.

Problem 1: Spam Filtering (9 points)

Suppose you are designing a spam filter. You have collected data on a large number of emails and computed the following joint probability mass function (PMF) $P(X, Y)$ for two discrete random variables. The variable Y represents whether an email is Spam ($Y = 1$) or Ham ($Y = 0$). The variable X represents a count of "high-risk" keywords in the email, categorized into three levels: x_1 (0 keywords), x_2 (1–5 keywords), and x_3 (6 or more keywords). The table below gives $P(X = x_i, Y = y)$.

	$Y = 0$ (Ham)	$Y = 1$ (Spam)
$X = x_1$ (0)	0.63	0.07
$X = x_2$ (1–5)	0.08	0.02
$X = x_3$ (6+)	0.03	0.17

Based on this table, answer the following:

a. (2 points) What is the overall probability that a random email from this dataset is Spam?

b. (3 points) What is the probability distribution for the number of "high-risk" keywords (i.e., what are $P(X = x_1)$, $P(X = x_2)$, and $P(X = x_3)$)?

c. (3 points) Now, let's look at emails based on their content. For an email that has 0 keywords ($X = x_1$), what is the probability that it is Spam? Repeat these calculations for emails with 1–5 keywords ($X = x_2$) and for emails with 6 or more keywords ($X = x_3$).

d. (1 point) Which keyword category maximizes $P(Y = 1|X = x_i)$?

Problem 2: The Recommendation Algorithm (6 points)

You are an engineer at a music streaming service, and you are analyzing your personalized discovery playlist algorithm. You find that the playlist pulls songs from three different sources: S_1 (songs from the user's favorite artists), S_2 (songs from new artists in genres the user likes), and S_3 (songs that are trending globally).

- The probability that a randomly selected song is from each source is: $P(S = S_1) = 0.6$, $P(S = S_2) = 0.3$, and $P(S = S_3) = 0.1$.
- Each source has a different "skip rate" (the probability the user dislikes and skips the song):
  - Songs from S_1 have a 1% skip rate.
  - Songs from S_2 have a 2% skip rate.
  - Songs from S_3 have an 8% skip rate.

Based on this information, answer the following:

a. (3 points) What is the overall probability that the user will skip any given song on the playlist?

b. (3 points) Suppose the user skips a song. What is the probability that it came from the "Trending Globally" source (S_3)? (Hint: Use Bayes' rule.)

Problem 3: Probability Density Function (PDF) (5 points)

Consider a continuous random variable X with a probability density function (PDF) given by:

$$f(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

a. (3 points) Verify that this is a valid PDF. (Hint: To show a function is a valid PDF, we need to show it satisfies two conditions.)

b. (2 points) Does this PDF take on values greater than 1? If yes, for what values of x?