# CSCI 145 Problem Set 4

September 16, 2025

## Submission Instructions

Please upload *your* work by **11:59pm Monday September 22, 2025.**

- You are encouraged to discuss ideas and work with your classmates. However, you **must acknowledge** your collaborators at the top of each solution on which you collaborated with others and you **must write** your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown. If you would like to use LaTeX, you can import the source of this document (available from the course webpage) to Overleaf.
- I recommend that you write your solutions to coding questions in a Jupyter notebook using Google Colab.
- You should submit your solutions as a **single PDF** via the assignment on Gradescope.

**Grading:** The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by **not** grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.

# Problem 1: Spam or Ham

In this problem, we will use Naive Bayes to identify emails as spam or ham (not spam). Consider the *bag-of-words* matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. Each row corresponds to one of $n$ emails, and each column corresponds to one of $d$ words. For email index $i \in \{1, \ldots, n\}$ and word index $j \in \{1, \ldots, d\}$, the corresponding entry in $\mathbf{X}$ is given by

$$[\mathbf{X}]_{i,j} = \begin{cases} 1 & \text{word } j \text{ appears in email } i \\ 0 & \text{else.} \end{cases} \tag{1}$$

You can find code to load a bag-of-words matrix here.

## Part A: Computing Likelihood

Split the bag-of-words matrix $\mathbf{X}$ and the labels $\mathbf{y} \in \{0, 1\}^n$ into training and testing sets, using an $80 - 20$ random split.

Compute the likelihoods for the *training data* $\mathbf{p}^{(1)} \in [0, 1]^n$ and $\mathbf{p}^{(0)} \in [0, 1]^n$ as described in class, *without* using a for loop.

## Part B: Computing Log Posteriors

Compute the log of the posteriors for the *testing data* using matrix multiplication between $\mathbf{X}$, $\log(\mathbf{p}^{(1)})$, and $\log(\mathbf{p}^{(0)})$.

Why are we using the *log* likelihoods? What happens to the evidence?

## Part C: Accuracy

Use the log posteriors to make predictions for the test set. What is the accuracy?