CIS5200: Machine Learning Fall 2025

Homework 0

Release Date: August 27, 2025 *Due Date: September 08, 2025*

There are two parts to Homework 0: written and programming. The goal of the written component is to give you an idea of the level of mathematical knowledge and maturity expected in this course. You should have seen all this material before; the goal of this homework is to encourage you to revisit some of the material and refresh your memory. The goal of the programming component is to help you get familiar with PyTorch and Google Colab, which we will use for the rest of the semester.

- In total, homework assignments will account for **15%** of the grade of the course. HW0 will count for **1%** of the grade. You will receive full credit for HW0 if you attempt all questions. For those requiring proofs, it is sufficient to explain why you believe it is true/false intuitively.
- Although we encourage collaboration on homework in general, **HW0 must be completed entirely on your own**. It is a test of your level of readiness to take this course, so please work on it independently. Use Ed Discussion only for clarifications on HW0.
- All written homework solutions should be **handwritten** and include your name, pennkey, and an honor statement ("I promise to follow the honor code").
- Submit your written HW0 as a single PDF file and submit your coding HW0 as an ipynb file through the same Gradescope assignment. Further instructions for submitting the programming component to Gradescope are included in the Colab notebook.
- The deadline is **11:59 PM ET**. Late HWs will be penalized **33%** per day (with 4 unpenalized late days). We will drop the lowest HW score (which could be a zero). In general, we will not offer exceptions for being sick, having job interviews, etc. Of course, if you have extreme extenuating circumstances such as an extended illness, please reach out to the instructors on Ed.

Here is a list of resources to help brush up on the mathematical background:

- General Review - Mathematics for Machine Learning by Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong, Cambridge University Press, 2020
- Linear Algebra Review
- Probability Review 1 and 2
- Additional Recommended Resources:
  - Matrix Calculus: The Matrix Calculus You Need For Deep Learning
  - Convex Optimization: Boyd and Vandenberghe, Chapter 2-3
  - Vector Calculus: Paul's Online Math Notes - Vector Calculus

Prerequisites: This course assumes familiarity with:

- Linear Algebra: matrix operations, eigenvalues/eigenvectors, vector spaces, matrix rank
- Multivariable Calculus: gradients, partial derivatives, chain rule
- Probability: basic probability rules, conditional probability, expectation, common distributions
- Optimization: convexity, critical points, gradient descent

Disclaimer: If you find HW0 to be very time consuming and extremely difficult, this course may not be right for you.

1 Written Questions

Q1 [*Linear Algebra*] Let A be a real-valued $n \times n$ matrix. Which of the following statements are true? Give a proof or counterexample.

1. If A is invertible, then $det(A^{-1}) = \frac{1}{det(A)}$
2. The sum of eigenvalues equals the trace of the matrix
3. If A has rank k, then exactly k eigenvalues are non-zero

Q2 [*Linear Algebra*] Let $A = \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix}$.

1. Find the nullspace of A.
2. Is the vector $[1, 1]^\top$ in the row space of A? Justify your answer.

Q3 [*Linear Algebra*] Consider the matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

1. What are the eigenvalues and corresponding eigenvectors of A?
2. Is A a PSD (positive semidefinite) matrix?
3. Since A is symmetric, its SVD can be written using eigenvalues and eigenvectors of A. Express the SVD of A using your answers from part (a). What do you notice about the relationship between singular values and eigenvalues in this case?

Q4 [*Calculus*] For column vector x ($n \times 1$ vector), answer the following questions:

1. Let $f(x) = \frac{1}{1+\exp(-w^\top x)}$ for column vector w, compute $\nabla_x f(x)$.
2. Let $f(x) = \|Ax - b\|_2^2$ for matrix $A \in \mathbb{R}^{n \times n}$ and n-dimensional vector b, compute $\nabla_x f(x)$.

Q5 [*Geometry*] Consider the hyperplane $w^\top x + b = 0$ for fixed vector $w \in \mathbb{R}^n$ and scalar $b \in \mathbb{R}$.

1. Under what conditions does the hyperplane pass through the origin?
2. What is the distance of any point x_0 from the hyperplane?

Q6 [*Vector Norms*] Let x be an n-dimensional vector.

1. If $\|x\|_\infty = 1$, what is the maximum possible value of $\|x\|_2$ in terms of n?
2. If $\|x\|_2 = 1$, what is the minimum possible value of $\|x\|_1$?

Q7 [*Convexity*] Consider the following functions:

1. Is $f(x) = x^3$ convex on $\mathbb{R}$? Justify your answer.
2. For what values of α is $f(x) = x^4 + \alpha x^2$ convex on $\mathbb{R}$?

Q8 [*Probability*] A spam detection system has the following properties:

- 80% of emails are legitimate (non-spam)
- For legitimate emails, the system has a 95% accuracy
- For spam emails, the system has a 90% accuracy

If the system flags an email as spam, what is the probability that it is actually spam?

Q9 [*Probability*] Let $X_1, X_2, ..., X_n$ be independent random variables where $X_i \sim N(\mu_i, \sigma_i^2)$.

1. Let $Y = \sum_{i=1}^{n} a_i X_i$ for fixed constants a_i. What is the distribution of Y? Express your answer in terms of μ_i, σ_i^2, and a_i.
2. If all $\mu_i = 0$ and $\sigma_i^2 = 1$, what is $P(\max_{1 \leq i \leq n} X_i > 2)$? Express your answer in terms of the standard normal CDF Φ.

Q10 [*Probability*] Consider rolling a fair six-sided die repeatedly.

1. What is the expected number of rolls needed to see a 6?
2. What is the expected number of rolls needed to see a 6 followed by a 6?

2 Programming Questions

Use the link here to access the Google Colaboratory (Colab) for the programming. Be sure to make a copy by going to "File", and "Save a copy in Drive". This assignment includes some cells with test cases for students to receive immediate feedback.

You do not need to write out or submit solutions to the problems below, but thinking through them will guide you through the coding homework and help you pass the test cases reliably.

Let $X \in \mathbb{R}^{m \times n}$ (rows $x_i^\top$), $w \in \mathbb{R}^n$, $y \in \{\pm 1\}^m$, $A \in \mathbb{R}^{n \times n}$, and let $\mathbf{1} \in \mathbb{R}^m$ denote the all-ones vector.

Matrix Operations / Batched Sample Gradients

Q[Grad1] Linear form. For $f(x_i; w) = w^\top x_i$,

1. derive the per-sample gradient $\nabla_{x_i} f \in \mathbb{R}^n$;
2. stack these to give the batched gradient $\nabla_X f(X) \in \mathbb{R}^{m \times n}$.

Example Solution (Grad1). Since $f(x_i; w) = \sum_{j=1}^n w_j x_{ij}$, we have $\nabla_{x_i} f = w \in \mathbb{R}^n$. Stacking across $i = 1, \ldots, m$ gives

$$\nabla_X f(X) = \begin{bmatrix} w^\top \\ \vdots \\ w^\top \end{bmatrix} = \mathbf{1}\, w^\top \in \mathbb{R}^{m \times n}.$$

Q[Grad2] Squared norm. For $f(x_i) = x_i^\top x_i = \|x_i\|_2^2$,

1. derive $\nabla_{x_i} f$;
2. write the batched gradient $\nabla_X f(X)$.

Q[Grad3] Squared error (least squares). For $f(x_i, y_i; w) = (y_i - w^\top x_i)^2$,

1. derive $\nabla_{x_i} f$;
2. write the batched gradient $\nabla_X f(X)$.

Q[Grad4] Logistic loss. For labels $y_i \in \{\pm 1\}$ and $f(x_i, y_i; w) = \log\big(1 + \exp(-y_i\, w^\top x_i)\big)$,

1. derive $\nabla_{x_i} f$;
2. write the batched gradient $\nabla_X f(X)$.

(Hint: you may express your answer using $\sigma(t) = 1/(1 + e^{-t})$.)

Q[Grad5] Quadratic form. For $f(x_i; A) = x_i^\top A x_i$,

1. derive $\nabla_{x_i} f$;
2. write the batched gradient $\nabla_X f(X)$.

Bonus: What is the simplified expression when A is symmetric?

Dataset Statistics (Means, Variances, Standardization)

Q[Stat1] Per-feature means. Define the per-feature *sample mean* vector $\mu \in \mathbb{R}^n$ of X.

1. Give the coordinate formula for μ_j.
2. Give a compact matrix expression using $\mathbf{1}$.
3. State the shape of μ.

Q[Stat2] Per-feature variances (unbiased). Using μ from Stat1, define the per-feature *unbiased* sample variance vector $v \in \mathbb{R}^n$ (i.e., ddof = 1).

1. Give the coordinate formula for v_j.
2. Give a vectorized matrix expression using $X_c = X - \mathbf{1}\mu^\top$.
3. Briefly explain how this differs from the *population* variance.

Q[Stat3] Standardization (population std). Let $\mu \in \mathbb{R}^n$ be the per-feature mean of $X \in \mathbb{R}^{m \times n}$ and let $s_{\text{pop}} \in \mathbb{R}^n$ be the per-feature *population* standard deviation,

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} X_{ij}, \qquad (s_{\text{pop}})_j = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (X_{ij} - \mu_j)^2}\,.$$

Define the standardized matrix

$$Z = (X - \mathbf{1}\mu^\top) \oslash s_{\text{pop}},$$

where $\oslash$ denotes columnwise division.

1. Write Z_{ij} explicitly.
2. Prove that each column of Z has mean 0.
3. Prove that each column of Z has *population variance* 1. (Optional: compute the *sample variance* of each column and show it equals $m/(m-1)$).
4. What happens if a column of X is constant (i.e., $(s_{\text{pop}})_j = 0$)? Briefly describe a practical code guard to avoid division by zero.