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.

CIS5200: Machine Learning	Fall 2025
Homework 0	
Release Date: August 27, 2025	*Due Date: September 08, 2025*

Name: Yi Fan Cai

PennKey: cai03

Honor Statement: By signing my name below, I affirm that I will not give or receive unauthorized assistance on this homework, and that all submitted work is my own.

Signature: Yi Fan Cai Date: 2025/9/4

Note: This document is a read-only file. To create an editable version click on Menu in the top left corner of your screen and choose the Copy Project option.

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

1. T

since invertible, we have $det(A)det(A^{-1}) = det(AA^{-1}) = det(I) = 1 \Rightarrow det(A^{-1}) = \frac{1}{det(A)}$

2. T

To calculate the eigenvalue, we use $p(\lambda) = det(\lambda I - A)$.

which also can be written as $p(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n)$

for the coefficient of λ^{n-1}, the first one is $-trace(A)$ & the second one is $-(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$

. which means $trace(A) = \lambda_1 + \cdots + \lambda_n$

3. F (give out an example) $\Rightarrow A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

for eigenvalues $\Rightarrow det(\lambda I - A) = 0 \Rightarrow \begin{vmatrix} \lambda & -1 \\ 0 & \lambda \end{vmatrix} = 0 \Rightarrow \lambda^2 = 0 \Rightarrow \lambda_1 = \lambda_2 = 0$

for rank $\Rightarrow$ it is 1 $0 \neq 1$ thus false

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.

1. $Ax = 0$

$\Rightarrow \begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 2x_1 - x_2 \\ 4x_1 - 2x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow 2x_1 = x_2$

the nullspace should be $null(A) = \{ t \begin{bmatrix} 1 \\ 2 \end{bmatrix} \mid t \in R \}$

2. the RREF is $\begin{bmatrix} 2 & -1 \\ 0 & 0 \end{bmatrix}$

so the row space is $\{ c[2 \ -1] \mid c \in R \}$

obviously, $[1, 1]^T$ is not in.

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?

1. $\det(\lambda I - A) = 0 \Rightarrow \begin{vmatrix} \lambda-2 & -1 \\ -1 & \lambda-2 \end{vmatrix} = 0$

$\Rightarrow \lambda^2 - 4\lambda + 4 - 1 = 0 \Rightarrow \lambda^2 - 4\lambda + 3 = 0 \Rightarrow (\lambda-3)(\lambda-1) = 0$

$\lambda_1 = 3 \quad \lambda_2 = 1$ are eigenvalues

for $\lambda_1 = 3$ the eigenvector is: $(\lambda I - A) \cdot v = 0$

$\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} v_1 - v_2 \\ -v_1 + v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow v_1 = v_2 \Rightarrow v = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$

for $\lambda_2 = 1$:

$\begin{bmatrix} -1 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow v = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \Rightarrow \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$

2. we have gotten the eigenvalues as $3 > 0$ & $1 > 0$, which are all positive, so is PSD

3. $A = U \Sigma V^T$

where U is $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$ & $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$

and since this is a symmetric matrix $V^T = U^T$

$A = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}^T$

for symmetric matrices, the singular values are equal to the absolute values of the eigenvalues

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 $\nabla_x f(x)$.

1. $\nabla_x f(x) = \nabla_x (1+e^{-w^T x})^{-1} = (-1)\cdot(1+e^{-w^T x})^{-2}\cdot(-e^{-w^T x}\cdot w)$

$= \frac{e^{-w^T x}}{(1+e^{-w^T x})^2}\cdot w$ or can write as $f(x)\cdot(1-f(x))\cdot w$

2. $f(x) = (Ax-b)^T(Ax-b) = (x^T A^T - b^T)(Ax-b)$

$= x^T A^T A x - x^T A^T b - b^T A x + b^T b.$

$= x^T A^T A x - 2b^T A x + b^T b.$

now calculate the $\nabla_x f(x)$

$$\Rightarrow \begin{cases} \nabla_x (x^T A^T A x) = (A^T A + (A^T A)^T)x = 2A^T A x \\ \nabla_x (2b^T A x) = 2A^T b. \\ \nabla_x (b^T b) = 0. \end{cases}$$

$\Rightarrow \nabla_x f(x) = 2A^T A x - 2A^T b = 2A^T(Ax-b)$

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?

1. which means passing through the $x=0$.
   $\Rightarrow w^T \cdot 0 + b = 0 \Rightarrow b = 0$

2. $d = \dfrac{|w^T x_0 + b|}{\|w\|}$

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 $\|x\|_1$?

1. the infinit norm means choosing the maximum absolute value and it is 1. which mean $\forall x_i$, we have $|x_i| \leq 1$
   for the max $\|x\|_2$, we let all $x_i = 1$, that is. $\sqrt{\sum_n (1)^2} = \sqrt{n}$

2. we have $\sqrt{\sum_{i=1}^{n} (x_i)^2} = 1$. for $\|x\|_1$ it means $\sum_{i=1}^{n} |x_i|$
   we know that $\sum_{i=1}^{n} (x_i)^2 \leq (\sum_{i=1}^{n} |x_i|)^2$
   $\Rightarrow \|x\|_1 \geq \|x\|_2$
   so the minimum of $\|x\|_1$ is 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 $\mathbb{R}$?

1. No.

for convex function, we need $\nabla^2 f(x) \geq 0$.

for $f(x) = x^3 \Rightarrow f'(x) = 3x^2 \Rightarrow f''(x) = 6x$.

when $\begin{cases} x \geq 0 & f''(x) \geq 0 \\ x < 0 & f''(x) < 0 \end{cases}$ $\Rightarrow$ thus not always non negative.

$\Rightarrow$ not convex on $\mathbb{R}$

2. $f(x) = x^4 + ax^2 \Rightarrow f(x)' = 4x^3 + 2ax \Rightarrow f(x)'' = 12x^2 + 2a$.

we need this to be non negative for $\forall x \in R$

since $12x^2 \geq 0$ we need $2a \geq 0 \Rightarrow a \geq 0$

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?

let the probability of legitimate email be $P(L)$

the probability of spam be $P(S)$

the probability of system flag spam is $P(F)$

we need to calculate : $P(S|F)$

$$= \frac{P(F|S) \cdot P(S)}{P(F)} = \frac{90\% \times 20\%}{P(F|S) \cdot P(S) + P(F|L) \cdot P(L)}$$

$$= \frac{18\%}{90\% \cdot 20\% + 5\% \cdot 80\%} = \frac{18}{18+4} = \frac{9}{11}$$

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 $P(\max_{1\le i\le n} X_i > 2)$? Express your answer in terms of the standard normal CDF Φ.

1. since X_i is independent random variable, the linear combine Y should also be gaussian distribution.

the. $E[Y] = \sum_{i=1}^n a_i \cdot \mu_i$

the $Var(Y) = \sum_{i=1}^n a_i^2 \cdot \sigma_i^2$

$\Rightarrow Y \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$

2. this means for $\forall X_i$, we have $X_i \sim N(0,1)$

$P(\max_{i\in[1,n]} X_i > 2) = 1 - P(\max_{i\in[1,n]} X_i \le 2)$

since independency of X_i, we calculate $P(X_i \le 2)$ which should be $\phi(2)$

$\Rightarrow P(\max_{i\in[1,n]} X_i > 2) = 1 - [\phi(2)]^n$.

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?

1. for a fair die the probability to see 6 in one turn is $\frac{1}{6}$. so, to let $P=1$, we need $P=\frac{1}{6}\times 6$

that means we need 6 rolls.

2. we can asseume two states,

$\begin{cases} A \text{ state: the last roll is not 6.} \\ B \text{ state: the last roll is 6, and we have one 6.} \end{cases}$

$\Rightarrow$. let a be the number of from state A to see two 6.

let b be the number of from state B to see two 6.

$\Rightarrow$ $\begin{cases} A = \underbrace{\tfrac{1}{6}}_{\text{we get a 6}}(1+\underbrace{B}_{\text{jump to state B}}) + \underbrace{\tfrac{5}{6}}_{\text{we do not get a 6.}}(1+A) \text{ stay in state A} \\ B = \underbrace{\tfrac{1}{6}}_{\text{we get the second 6.}}\cdot 1 + \tfrac{5}{6}(1+A) \text{ jump to state A} \end{cases}$

$\Rightarrow$ $A=\frac{1}{6}\left(1+\frac{1}{6}+\frac{5}{6}(1+A)\right)+\frac{5}{6}(1+A) \Rightarrow A=42$

$B=\frac{1}{6}+\frac{5}{6}+35=36$

so the answer is 42