CIS5200: Machine Learning Spring 2025

Homework 0

Release Date: January 16, 2025 *Due Date: January 30, 2025*

1 Written Questions

A1 Solution:

1. **TRUE.**

   Proof. Let A be an invertible matrix. Then:

$$
\begin{aligned}
AA^{-1} &= I && \text{(by definition of inverse)} \\
\det(A)\det(A^{-1}) &= \det(I) = 1 && \text{(determinant is multiplicative)} \\
\det(A^{-1}) &= \frac{1}{\det(A)} && \text{(}\det(A) \neq 0 \text{ since } A \text{ is invertible)}
\end{aligned}
$$

   □

2. **TRUE.**

   Proof. For any $n \times n$ matrix A:

   - Let P be the matrix that diagonalizes A, so $P^{-1}AP = J$ where J is block-diagonal in Jordan Canonical Form
   - The diagonal entries of J are the eigenvalues $\lambda_1, ..., \lambda_n$ of A
   - By trace cyclicity: $tr(A) = tr(PJP^{-1}) = tr(P^{-1}PJ) = tr(J) = \sum_{i=1}^{n} \lambda_i$

   Therefore, the trace equals the sum of eigenvalues. □

3. **FALSE.** Counterexample:

   - Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$
   - This matrix has rank 1 since its columns are linearly dependent but not all zero
   - The characteristic equation is $\lambda^2 = 0$, so both eigenvalues are 0
   - Therefore, a matrix can have rank k but fewer than k non-zero eigenvalues

A2 Solution:

1. **Find the nullspace of A:**

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

   Observe that row 2 is twice row 1, so we only need to consider:

$$2x_1 - x_2 = 0$$

   Therefore: $x_2 = 2x_1$

   The nullspace is spanned by:

$$\text{Nullspace}(A) = \text{span}\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$$

2. **Is $[1,1]^\top$ in the row space? No.** To verify, we attempt to solve $A^\top x = [1,1]^\top$:

$$\begin{aligned} 2x_1 + 4x_2 &= 1 \\ -x_1 - 2x_2 &= 1 \end{aligned}$$

   The second equation is not a multiple of the first equation, but the right-hand sides are equal, making the system inconsistent. Therefore, $[1,1]^\top$ is not in the row space of A.

A3 Solution:

1. **Eigenvalues and Eigenvectors:**

   - The characteristic equation $\det(A - \lambda I) = 0$ gives:

$$\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0$$

   - Expanding: $(2-\lambda)^2 - 1 = 0$
   - Solving: $\lambda = 1, 3$
   - The corresponding eigenvectors are:

$$\begin{aligned} v_1 &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \quad \text{for } \lambda_1 = 1 \\ v_2 &= \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{for } \lambda_2 = 3 \end{aligned}$$

2. **PSD Check:** Yes, A is positive definite since all eigenvalues (1 and 3) are positive.

3. **SVD:** Since A is symmetric, its SVD uses the eigenvectors as both left and right singular vectors:

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

   The singular values are the absolute values of the eigenvalues: $\sigma_1 = 3, \sigma_2 = 1$

A4 Solution:

1. **For $f(x) = \frac{1}{1+\exp(-w^\top x)}$ for column vector w:**

   - Let $z = w^\top x$. Note that $f(z)$ is the logistic function
   - Using the chain rule:
   $$\nabla_x f(x) = \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial x}$$
   - The derivative of the logistic function is $\frac{\partial f}{\partial z} = f(z)(1 - f(z))$
   - Since $z = w^\top x$, we have $\frac{\partial z}{\partial x} = w$
   - Combining these: $\nabla_x f(x) = f(x)(1 - f(x))w$

2. **For $f(x) = \|Ax - b\|_2^2$ for matrix $A \in \mathbb{R}^{n \times n}$ and vector b:**

   - First expand the squared norm:
   $$f(x) = (Ax - b)^\top (Ax - b) = x^\top A^\top Ax - 2b^\top Ax + b^\top b$$
   - Taking the gradient with respect to x:
     - $\frac{\partial}{\partial x}(x^\top A^\top Ax) = 2A^\top Ax$
     - $\frac{\partial}{\partial x}(-2b^\top Ax) = -2A^\top b$
     - $\frac{\partial}{\partial x}(b^\top b) = 0$
   - Therefore: $\nabla_x f(x) = 2A^\top (Ax - b)$

A5 Solution:

1. **Conditions for passing through origin:**

   Proof. The hyperplane passes through the origin (0) if and only if:
   $$\begin{aligned} w^\top (0) + b &= 0 \\ b &= 0 \end{aligned}$$
   Therefore, the hyperplane passes through the origin if and only if $b = 0$. □

2. **Distance from point x_0 to hyperplane:**

   Proof. Let p be any point on the hyperplane, so $w^\top p + b = 0$. The vector from x_0 to p is $(p - x_0)$.

   The distance d is found by projecting the vector $(p - x_0)$ onto the unit normal vector $\frac{w}{\|w\|_2}$. We use the unit normal vector because it points perpendicular to the hyperplane, and its length of 1 ensures we get the true distance. The projection gives us:
   $$d = \left| \frac{w^\top}{\|w\|_2}(p - x_0) \right|$$

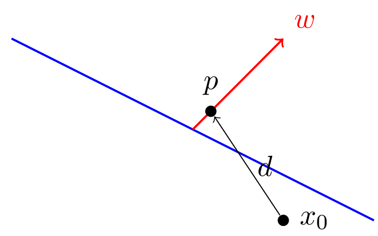

Figure 1: Distance from point x_0 to hyperplane

$$
\begin{aligned}
&= \frac{|w^\top p - w^\top x_0|}{\|w\|_2} \\
&= \frac{|(-b) - w^\top x_0|}{\|w\|_2} \quad \text{(since } w^\top p = -b\text{)} \\
&= \frac{|w^\top x_0 + b|}{\|w\|_2}
\end{aligned}
$$

This is the shortest distance because any other path from x_0 to the hyperplane would have a component parallel to the hyperplane, making it longer than the perpendicular path. □

A6 Solution:

1. **Maximum value of $\|x\|_2$ when $\|x\|_\infty = 1$:**

   Proof. When $\|x\|_\infty = 1$, each component satisfies $|x_i| \leq 1$. The maximum $\|x\|_2$ occurs when all entries are at their maximum magnitude of 1:

$$
\begin{aligned}
\|x\|_2 &= \sqrt{\sum_{i=1}^{n} x_i^2} \\
&\leq \sqrt{\sum_{i=1}^{n} 1^2} \quad \text{(since each } |x_i| \leq 1\text{)} \\
&= \sqrt{n}
\end{aligned}
$$

   This bound is achieved when $x = [\pm 1, \pm 1, ..., \pm 1]^\top$. □

2. **Minimum value of $\|x\|_1$ when $\|x\|_2 = 1$:**

   Proof. By the Cauchy-Schwarz inequality applied to vectors x and the all-ones vector 1:

$$|\langle x, 1 \rangle| \leq \|x\|_2 \|1\|_2$$

   Therefore:

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \geq \sqrt{\sum_{i=1}^{n} x_i^2} = \|x\|_2 = 1$$

where the first inequality follows from Cauchy-Schwarz and the last equality uses our assumption that $\|x\|_2 = 1$. This bound is achieved when exactly one component is ± 1 and all others are 0. Therefore, the minimum value of $\|x\|_1$ is 1. $\square$

A7 Solution:

1. **For $f(x) = x^3$:**

   Proof. A function is convex if and only if its second derivative is non-negative everywhere. For $f(x) = x^3$:

   $$f''(x) = 6x$$

   Since $f''(x)$ is negative for $x < 0$ and positive for $x > 0$, $f(x) = x^3$ is not convex on $\mathbb{R}$. $\square$

2. **For $f(x) = x^4 + \alpha x^2$:**

   Proof. Computing the second derivative:

   $$f''(x) = 12x^2 + 2\alpha$$

   For convexity, we need $f''(x) \geq 0$ for all $x \in \mathbb{R}$. Since $12x^2 \geq 0$ for all x, we only need:

   $$2\alpha \geq 0 \implies \alpha \geq 0$$

   Therefore, $f(x)$ is convex if and only if $\alpha \geq 0$. $\square$

A8 Solution:

- Let S denote "email is spam" and F denote "email is flagged as spam"
- We know:

$$\begin{aligned} P(S) &= 0.2 \text{ (prior probability of spam)} \\ P(\neg S) &= 0.8 \text{ (prior probability of legitimate email)} \\ P(F|S) &= 0.9 \text{ (true positive rate)} \\ P(F|\neg S) &= 0.05 \text{ (false positive rate)} \end{aligned}$$

- By Bayes' theorem:

$$\begin{aligned} P(S|F) &= \frac{P(F|S)P(S)}{P(F|S)P(S) + P(F|\neg S)P(\neg S)} \\ &= \frac{(0.9)(0.2)}{(0.9)(0.2) + (0.05)(0.8)} \\ &= \frac{0.18}{0.22} \approx 0.82 \end{aligned}$$

- Therefore, when the system flags an email as spam, it is correct approximately 82% of the time

A9 Solution:

1. **Distribution of $Y = \sum_{i=1}^{n} a_i X_i$ for fixed constants a_i:**

   Proof. Since each $X_i \sim N(\mu_i, \sigma_i^2)$ is independent:

   - For any constant a, $aX \sim N(a\mu, a^2\sigma^2)$
   - Sum of independent normal variables is normal
   - Therefore:

   $$\mathbb{E}[Y] = \sum_{i=1}^{n} a_i \, \mathbb{E}[X_i] = \sum_{i=1}^{n} a_i \mu_i$$
   $$Var(Y) = \sum_{i=1}^{n} Var(a_i X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

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

   □

2. **For $P(\max_{1 \le i \le n} X_i > 2)$ when $\mu_i = 0$ and $\sigma_i^2 = 1$:**

   Proof.

   $$\begin{aligned} P(\max X_i > 2) &= 1 - P(\text{all } X_i \le 2) \\ &= 1 - \prod_{i=1}^{n} P(X_i \le 2) \quad \text{(independence)} \\ &= 1 - (\Phi(2))^n \end{aligned}$$

   where Φ is the standard normal CDF. □

A10 Solution:

1. **Expected number of rolls to see a 6:**

   Proof. Let X be the number of rolls until first 6. Each roll is independent with:

   $$P(\text{success}) = p = \frac{1}{6}$$
   $$\mathbb{E}[X] = \frac{1}{p} = 6 \text{ rolls}$$

   This follows from the geometric distribution, which models number of trials until first success. □

2. **Expected number of rolls to see a 6 followed by a 6:**

   Proof. This is a two-stage problem that can be solved using conditional expectation:

- First, we need to roll a 6, which from part (a) takes an expected 6 rolls
- After getting a 6, we have two possibilities on the next roll:
  - Get a 6 (probability $\frac{1}{6}$) - adds 1 roll and we're done
  - Get any other number (probability $\frac{5}{6}$) - adds 1 roll and we start over

Let E be the expected number of rolls needed to see a 6 followed by a 6. By the law of total expectation:

$$\begin{aligned} E &= 6 + \frac{1}{6}(1) + \frac{5}{6}(1 + E) \\ &= 6 + \frac{1}{6} + \frac{5}{6} + \frac{5}{6}E \end{aligned}$$

Solving for E:

$$\begin{aligned} \frac{1}{6}E &= 6 + 1 \\ E &= 42 \text{ rolls} \end{aligned}$$

$\square$