Introduction to Machine Learning, Spring 2025
Homework 1

(Due March 7, 2025 at 11:59pm (CST))

February 20, 2025

1. Please write your solutions in English.
2. Submit your solutions to the course Gradescope.
3. If you want to submit a handwritten version, scan it clearly.
4. Late homeworks submitted within 3 days of the due date will be marked down 25% each day cumulatively Homeworks submitted more than 3 days after the due date will not be accepted unless there is a valid reason, such as a medical or family emergency.
5. You are required to follow ShanghaiTech's academic honesty policies. You are allowed to discuss problems with other students, but you must write up your solutions by yourselves. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious penalties.

1. [15 points] [Math review(Linear Algebra)] Rank properties. Suppose a matrix $A \in \mathbb{R}^{m\times n}$, prove:

   (a) $\mathbf{rank}(A) \leq \min(m, n)$ [4 points] (Hint: You just need to consider the foundamental transformation of the matrix. And how **rank** is defined.)

   (b) $\mathbf{rank}(AB) \leq \min(\mathbf{rank}(A), \mathbf{rank}(B))$. [4 points]

   (c) $\mathbf{rank}(A^\top A) = \mathbf{rank}(A)$. [4 points] (Hint: consider the identity between **rank** and **nullity**)

   (d) What does the rank of a matrix essentially refer to? (Hint: consider the correspondence to the singularvalues.) [3 points]

Solution

(a) The rank of a matrix is the same with the number of leading ones after row echelon form. The number of leading ones must be less than or equal to the minimum of the number of rows and the number of columns.
Since the row echelon form of a matrix A is a matrix R with the same rank as A, we have $\mathbf{rank}(A) \leq \min(m, n)$.

(b) Let A be an $m \times n$ matrix, and denote its column vectors as: $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$, where $\mathbf{a}_i$ is the i-th column of A.

The j-th column of AB can be expressed as: $(AB)_j = A\mathbf{b}_j = \sum_{i=1}^{n} b_{ij}\mathbf{a}_i$, where $\mathbf{b}_j$ is the j-th column of B. This shows that the columns of AB are linear combinations of the columns of A. Thus:

$$\text{rank}(AB) \leq \text{rank}(A)$$

Let B be an $n \times p$ matrix, and denote its row vectors as: $B = \begin{bmatrix} \mathbf{b}_1^\top \\ \mathbf{b}_2^\top \\ \vdots \\ \mathbf{b}_n^\top \end{bmatrix}$, where $\mathbf{b}_i^\top$ is the i-th row of B.

The i-th row of AB can be expressed as: $(AB)_i^\top = \mathbf{a}_i^\top B = \sum_{j=1}^{n} a_{ij}\mathbf{b}_j^\top$, where $\mathbf{a}_i^\top$ is the i-th row of A. This shows that the rows of AB are linear combinations of the rows of B. Thus:

$$\text{rank}(AB) \leq \text{rank}(B)$$

Combining the above results, we have:

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$$

(c) Since $\forall A \in \mathbb{R}^{m\times n}$, we have $\mathbf{rank}(A) + \mathbf{nullity}(A) = n$.
Then we have $\mathbf{rank}(A^\top A) + \mathbf{nullity}(A^\top A) = n$.
Thus we only need to prove $\mathbf{nullity}(A^\top A) = \mathbf{nullity}(A)$.

$\Rightarrow$: $\forall \mathbf{x} \in \mathbb{R}^n$, if $\mathbf{x} \in \mathbf{Null}(A^\top A)$, i.e. $A^\top A\mathbf{x} = \mathbf{0}$, then we have $\mathbf{x}^\top A^\top A\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x}^\top(A^\top A\mathbf{x}) = \mathbf{0} \Rightarrow \|A\mathbf{x}\|_2^2 = 0 \Rightarrow A\mathbf{x} = \mathbf{0} \Rightarrow \mathbf{x} \in \mathbf{Null}(A)$.

$\Leftarrow$: $\forall \mathbf{x} \in \mathbb{R}^n$, if $\mathbf{x} \in \mathbf{Null}(A)$, i.e. $A\mathbf{x} = \mathbf{0}$, then we have $A^\top A\mathbf{x} = A^\top\mathbf{0} = \mathbf{0} \Rightarrow \mathbf{x} \in \mathbf{Null}(A^\top A)$.

So the null space of $A^\top A$ and A are the same, which means $\mathbf{nullity}(A^\top A) = \mathbf{nullity}(A)$.

(d) The number of non-zero singular values.

2. [15 points] [Math review(Linear Algebra)] Eigenvalue Decomposition(EVD) / Spectral Decomposition. Find the geometric and algebraic multiplicity of each eigenvalue of the following matrix A, and determine whether A is diagonalizable. If A is diagonalizable, then find a matrix P that diagonalizes A, and find the diagonal matrix Λ such that $\Lambda = P^{-1}AP$. Furthermore, if A could be orthogonally diagonalized, write Λ as $\Lambda = P^{\top}AP$.

(a) $A = \begin{bmatrix} -1 & 4 & -2 \\ -3 & 4 & 0 \\ -3 & 1 & 3 \end{bmatrix}$. [5 points]

(b) $A = \begin{bmatrix} 5 & 0 & 0 \\ 1 & 5 & 0 \\ 0 & 1 & 5 \end{bmatrix}$. [5 points]

(c) $A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$. [5 points]

Solution

(a) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 3 & -5 & 3 \\ -3 & 9 & -6 \\ 0 & -4 & 4 \end{bmatrix} \begin{bmatrix} -1 & 4 & -2 \\ -3 & 4 & 0 \\ -3 & 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & \frac{2}{3} & \frac{1}{4} \\ 1 & 1 & \frac{3}{4} \\ 1 & 1 & 1 \end{bmatrix}$

(b) $|A| = (\lambda - 5)^3 = 0 \Rightarrow \lambda_1 = \lambda_2 = \lambda_3 = 5$.

But $A - 5I = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $\mathbf{rank}(A - 5I) = 2, nullity(A - 5I) = 2$, so A is not diagonalizable.

(c) $A^{\top} = A$, so A must be orthogonally diagonalizable.

$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 8 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{bmatrix}$

3. [15 points] [Math review(Calculus)] Differential calculus of functions.

   (a) Suppose $f : \mathbb{R} \to \mathbb{R}$ is a second order differentiable function. Write down the second order Taylor expansion of $f(x)$ around x_0 with Peano's form of the remainder. [3 points]

   (b) Suppose $f : \mathbb{R} \to \mathbb{R}$ is a second order differentiable function. Write down the first order and second order's mean value theorem for $f(x)$ related to point x and $x + d$. [4 points]

   (c) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a second order differentiable function. Write down the second order Taylor expansion of $f(\mathbf{x})$ around $\mathbf{x}_0$ with Peano's form of the remainder. [4 points] (Hint: the first order derivative is the gradient ∇f, the second order derivative is the Hessian $\nabla^2 f$.)

   (d) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a second order differentiable function. Write down the first order and second order's mean value theorem for $f(\mathbf{x})$ related to point $\mathbf{x}$ and $\mathbf{x} + \mathbf{d}$. [4 points]

Solution

(a) $f(x) = f(x_0) + f'(x_0)(x - x_0) + \dfrac{f''(x_0)}{2!}(x - x_0)^2 + o((x - x_0)^2)$.

(b) First order: $\exists \xi \in (0, 1),\ f(x + d) = f(x) + f'(x + \xi d)d$.

Second order: $\exists \xi \in (0, 1),\ f(x + d) = f(x) + f'(x)d + \dfrac{1}{2} f''(x + \xi d)d^2$.

(c) $f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \dfrac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2)$.

(d) First order: $\exists \xi \in (0, 1),\ f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + \xi \mathbf{d})^\top \mathbf{d}$.

Second order: $\exists \xi \in (0, 1),\ f(\mathbf{x} + \mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{d} + \dfrac{1}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{x} + \xi \mathbf{d})\mathbf{d}$.

4. [15 points] [Math review(Probability and Statistics)] Suppose we have two random variables, A and B, where A takes on values a_1, a_2 and B takes on values b_1, b_2. Let $P(A = a_1) = 0.5$ and $P(B = b_1) = 0.5$.

   (a) Suppose a_1 and b_2 are disjoint (mutually exclusive). [3 points]
   1. What is $P(a_1, b_2)$?
   2. What is $P(a_1, b_1)$?
   3. What is $P(a_1 \mid b_2)$?

   (b) Suppose instead that A, B are independent. [3 points]
   1. What is $P(a_1, b_2)$?
   2. What is $P(a_1, b_1)$?
   3. What is $P(a_1 \mid b_2)$?

   (c) A student is looking at her activity tracker (Fitbit/Apple Watch) data and she notices that she seems to sleep better on days that she exercises. They observe the following:

Exercise	Good Sleep	Probability
yes	yes	0.3
yes	no	0.2
no	no	0.4
no	yes	0.1

   1. What is the P(GoodSleep = yes | Exercise = yes)?
   2. Why doesn't P(GoodSleep = yes, Exercise = yes) = P(GoodSleep = yes) $\cdot$ P(Exercise = yes)?
   3. The student merges her activity tracker data with her food logs and finds that the P(Eatwell = yes | Exercise = yes, GoodSleep = yes) is 0.25. What is the probability of all three happening on the same day? [3 points]

   (d) What is the expectation of X where X is a single roll of a fair 6-sided die ($S = \{1, 2, 3, 4, 5, 6\}$)? What is the variance of X? [3 points]

   (e) Imagine that we had a new die where the sides were $S = \{3, 4, 5, 6, 7, 8\}$. How do the expectation and the variance compare to our original dice? [3 points]

Solution

(a) 1. $P(A = a_1, B = b_2) = 0$
2. $P(A = a_1, B = b_1) = p(b_1 \mid a_1)\, p(a_1) = 0.5$ since $p(b_1 \mid a_1) = 1$
3. $P(A = a_1 \mid B = b_2) = 0$

(b) 1. $p(a_1, b_2) = 0.25$
2. $p(a_1, b_1) = 0.25$ since now $p(b_1 \mid a_1) = 0.5$
3. $p(a_1 \mid b_2) = 0.5$

(c) 1. P(GoodSleep = yes | Exercise = yes) $= \dfrac{0.3}{0.3 + 0.2} = 0.6$
2. Good Sleep and Exercise are not independent.
3. P(Eatwell = yes, Exercise = yes, GoodSleep = yes) = P(Eatwell = yes | Exercise = yes, GoodSleep = yes) $*$ P(Exercise = yes, GoodSleep = yes) = 0.075

(d) $\mathbb{E}$[X] = 3.5
Var[X] = 2.917
For variance, we can do $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$ or use the equivalent formulation $\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. In the first method, this gives $\dfrac{1}{6} \sum\limits_{x \in \{1,2,3,4,5,6\}} (x - 3.5)^2$

(e) $\mathbb{E}$[X] = 5.5
Var[X] = 2.917, note $\text{Var}[X + a] = \text{Var}[X]$ for scalar a

5. [20 points] [Math review(Information Theory)] Example of joint entropy. Let $p(x, y)$ be given by

X \ Y	0	1
0	$\frac{1}{3}$	$\frac{1}{3}$
1	0	$\frac{1}{3}$

Find:

(a) $H(X), H(Y)$. [4 points]
(b) $H(X \mid Y), H(Y \mid X)$. [4 points]
(c) $H(X, Y)$. [3 points]
(d) $H(Y) - H(Y \mid X)$. [1 points]
(e) $I(X; Y)$. [1 points]
(f) Draw a Venn diagram for the quantities in parts (a) through (e). [3 points]
(g) When is the mutual information $I(X; Y) = 0$? [4 points]

Solution

From the table, we can get that

$$P(X=0) = \frac{2}{3}, P(X=1) = \frac{1}{3}$$
$$P(Y=0) = \frac{2}{3}, P(Y=1) = \frac{1}{3}$$
$$P(X=0|Y=0) = 1, P(X=1|Y=0) = 0$$
$$P(X=0|Y=1) = \frac{1}{2}, P(X=1|Y=1) = \frac{1}{2}$$
$$P(Y=0|X=0) = \frac{1}{2}, P(Y=1|X=0) = \frac{1}{2}$$
$$P(Y=0|X=1) = 0, P(Y=1|X=1) = 1$$

(a)

$$H(X) = H\left(\frac{2}{3}, \frac{1}{3}\right) = \log 3 - \frac{2}{3} = 0.918 \text{ bits}$$
$$H(Y) = H\left(\frac{1}{3}, \frac{2}{3}\right) = \log 3 - \frac{2}{3} = 0.918 \text{ bits}$$

(b)

$$\begin{aligned} H(X|Y) &= P(Y=0)H(X|Y=0) + P(Y=1)H(X|Y=1) \\ &= \frac{1}{3}H(1,0) + \frac{2}{3}H\left(\frac{1}{2}, \frac{1}{2}\right) \\ &= \frac{1}{3}0 + \frac{2}{3}\log 2 \\ &= \frac{2}{3} \\ &= 0.667 \text{ bits} \end{aligned}$$

Similarly, we can get that

$$\begin{aligned} H(Y|X) &= P(X=0)H(Y|X=0) + P(X=1)H(Y|X=1) \\ &= \frac{2}{3}H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{3}H(0,1) \\ &= 0.667 \text{ bits} \end{aligned}$$

(c)

$$\begin{aligned} H(X,Y) &= H(X) + H(Y|X) \\ &= \left(\log 3 - \frac{2}{3}\right) + \frac{2}{3} \\ &= \log 3 \\ &= 1.585 \text{ bits} \end{aligned}$$

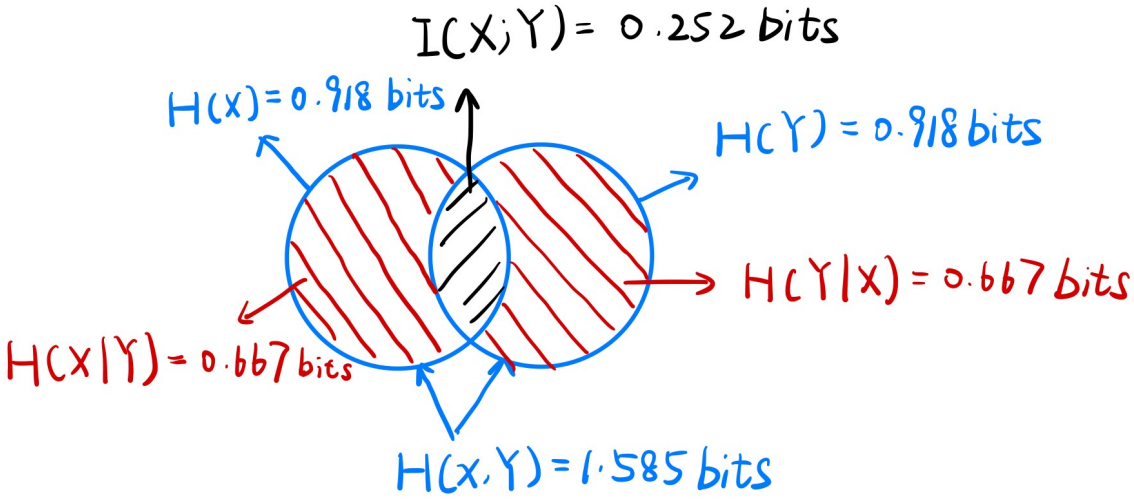
(d)

$$\begin{aligned} H(Y) - H(Y|X) &= H(Y) - (P(X=0)H(Y|X=0) + P(X=1)H(Y|X=1)) \\ &= \left(\log 3 - \frac{2}{3}\right) - \frac{2}{3} \\ &= 0.252 \text{ bits} \end{aligned}$$

(e)

$$I(X;Y) = H(Y) - H(Y|X) = 0.252 \text{ bits}$$

(f) The Venn diagram for the quantities are shown below.



(g) $I(X;Y) = KL\left(p(x,y)||p(x)p(y)\right) = 0$ if and only if $p(x,y) = p(x)p(y)$, which means X and Y are independent.