Introduction to Machine Learning, Spring 2025
Homework 2

(Due March 14, 2025 at 11:59pm (CST))

February 25, 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. [10 points] [Math review(Linear Algebra)] Singularvalue decomposition(SVD).
   $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$. Find the SVD of $A = U\Sigma V^\top$.
   Solution

2. [25 points] [Convex Optimization Basics] Norm for a vector $\mathbf{x} \in \mathbb{R}^n$ or for a matrix $X \in \mathbb{R}^{m \times n}$ is a function $f : \mathbb{R}^n \to \mathbb{R}$ or $f : \mathbb{R}^{m \times n} \to \mathbb{R}$, which is widely used in optimization. Take vector's norm as example: they have many properties: 1. $f(\mathbf{x}) \geq 0$, iff $\mathbf{x} = \mathbf{0}$ the equality holds; 2. $f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x})$ for any $\alpha \in \mathbb{R}$; 3. $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. And matrices norms are similar, but you **should not** use them directly in this problem.

   (a) Proof: Any vector norm $f : \mathbb{R}^n \to \mathbb{R}$ is convex. [Hint: Consider the definition of convex function: $\forall \theta \in [0, 1], f(\theta \mathbf{x} + (1 - \theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y})$.] [5 points]

   (b) Let $f(X) = \|X\|_2$ be the spectral norm of a matrix $X \in \mathbb{R}^{m \times n}$, defined as the largest singular value of X. Prove that $f(X)$ is convex. [Hints:1.$\lambda_{\max}(A) = \sup_{\mathbf{y} \in \mathbb{R}^n} \frac{\mathbf{y}^\top A \mathbf{y}}{\|\mathbf{y}\|_2^2}$, 2.$\forall \mathbf{y}$, if $g(X, \mathbf{y})$is convex in X, then $f(X) = \sup_{\mathbf{y} \in \mathbb{R}^n} g(X, \mathbf{y})$ is convex.] [10 points]

   (c) Let $f(X) = \sum_{i=1}^{r} \sigma_i(X)$ be the the nuclear norm of a matrix $X \in \mathbb{R}^{m \times n}$,where $\sigma_i(X)$ are the singular values of X. Prove that $f(X)$ is convex. [Hint: $\|X\|_* = \sup_{\|Z\|_2 \leq 1} < Z, X >$] [10 points]

   Solution

3. [10 points] [Log-sum Inequality] Recall Jensen's inequality $f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ if f is convex for any random variable X. Proof the log-sum inequality:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are positive numbers.

Solution

4. [10 points] [Convexity of Mutual Information] From the definition of the mutual information $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$, we know that $I(X;Y)$ is a function of $p(x,y)$. This is because we can obtain $p(x)$ and $p(y)$ by computing the marginal distribution of $p(x,y)$. However, $I(X;Y)$ is a non-convex and non-concave function of $p(x,y)$. Which is not a good property for optimization.
In some specific cases, $p(x)$ as given. Then $I(X;Y)$ is a function of $p(y|x)$. Prove that $I(X;Y)$ is a convex function of $p(y|x)$.
[Hints:]

- Log-sum Inequality when $n = 2$:

$$(a_1 + a_2) \log \frac{a_1 + a_2}{b_1 + b_2} \leq a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}$$

- Consider 3 mutual information terms $I_1(X;Y), I_2(X;Y), I_\lambda(X;Y)$, which are separately computed from distributions $p_1(y|x), p_2(y|x), p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x), \lambda \in [0,1]$. Then only need to prove that $I_\lambda(X;Y) \leq \lambda I_1(X;Y) + (1-\lambda) I_2(X;Y)$.

Solution

5. [20 points] [Decision Tree] A dataset is given below. Now we want to discover the relationship between the features and the target variable by using a Decision Tree.

Outlook (X_1)	Temperature (X_2)	Humidity (X_3)	Play Tennis? (Y)
sunny	hot	high	no
overcast	hot	high	yes
rain	mild	high	yes
rain	cool	normal	yes
sunny	mild	high	no
sunny	mild	normal	yes
rain	mild	normal	yes
overcast	hot	normal	yes

(a) Using the dataset above, calculate the mutual information for each feature (X_1, X_2, X_3) to determine the root node for a Decision Tree trained on the above data.

- What is $I(Y; X_1)$? [3 points]
- What is $I(Y; X_2)$? [3 points]
- What is $I(Y; X_3)$? [3 points]
- What feature should be split on at the root node? [1 points]

(b) Calculate what the next split should be. [5 points]

(c) Draw the resulting tree. [5 points]

Solution