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

$$A^\top A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

Then:

$$p(\lambda) = \det(\lambda I_2 - A) = \lambda^2 - 4\lambda + 3 = (\lambda - 3)(\lambda - 1)$$
$$p(\lambda) = 0 \Rightarrow \lambda_1 = 3, \lambda_2 = 1$$

1. Notice that here we require $\lambda_1 > \lambda_2$. So we get the two singular values of A are $\sigma_1 = \sqrt{3}, \sigma_2 = \sqrt{1} = 1$. So the diagonal matrix $\Sigma \in \mathbb{R}^{3\times 2}$ is:

$$\Sigma = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

2. Then we can get the orthogonal matrix $V \in \mathbb{R}^{2\times 2}$ by the eigenvectors of $A^\top A$:

$$\left(\lambda_1 I_2 - A^\top A\right) x = 0$$

Normalize the basis of the solution space, we get $v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$. Solve the homogeneous equation:

$$\left(\lambda_2 I_2 - A^\top A\right) \boldsymbol{x} = 0$$

Normalize the basis of the solution space, we get $v_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$. So we get:

$$V = \begin{bmatrix} v_1 & v_2 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}.$$

3. Now we get the orthogonal matrix $U \in \mathbb{R}^{3\times 3}$, we know that the last two columns of U are:

$$\boldsymbol{u}_1 = \frac{1}{\sigma_1} A\boldsymbol{v}_1 = \frac{\sqrt{3}}{3} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{6}}{3} \\ \frac{\sqrt{6}}{6} \\ \frac{\sqrt{6}}{6} \end{bmatrix},$$
$$\boldsymbol{u}_2 = \frac{1}{\sigma_2} A\boldsymbol{v}_7 = 1 \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} 0 \\ -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{21}}{2} \end{bmatrix}.$$

Then we need to find the third column vector $\boldsymbol{u}_3 \in \mathbb{R}^3$ such that $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$ forms an orthonormal basis of $\mathbb{R}^3$:

$$\boldsymbol{x} \cdot \boldsymbol{u}_1 = \frac{\sqrt{6}}{3} x_1 + \frac{\sqrt{6}}{6} x_2 + \frac{\sqrt{6}}{6} x_3 = 0$$
$$x \cdot \boldsymbol{u}_2 = 0x_1 - \frac{\sqrt{2}}{2} x_2 + \frac{\sqrt{2}}{2} x_3 = 0$$

Solve the linear equations:

$$\begin{bmatrix} \frac{\sqrt{3}}{3} & \frac{\sqrt{6}}{6} & \frac{\sqrt{G}}{6} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

We can get that a basis of $\operatorname{span}\{u_1, u_2\}^+$ is $x = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$, normalize it we can get that $u_3 = \dfrac{x}{\|x\|} = \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}$. Thus:

$$U = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{4}}{3} & 0 & -\frac{1}{\sqrt{3}} \\ \frac{\sqrt{6}}{6} & -\frac{\sqrt{2}}{2} & \frac{1}{\sqrt{3}} \\ \frac{\sqrt{6}}{3} & \frac{\sqrt{3}}{2} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

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

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

(a) Since f is a norm function, we have the properties of norm functions that,
1. $\forall\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$.
2. $\forall\mathbf{x} \in \mathbb{R}^n, \forall a \in \mathbb{R}, f(a\mathbf{x}) = |a|f(\mathbf{x})$.
So we have, $\forall\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall\theta \in [0,1]$,
From property 1., we can get that

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq f(\theta\mathbf{x}) + f((1-\theta)\mathbf{y})$$

From property 2., we can get that

$$f(\theta\mathbf{x}) = |\theta|f(\mathbf{x}) \text{ and } f((1-\theta)\mathbf{y}) = |1-\theta|f(\mathbf{y})$$

Since $\theta \in [0,1]$, so we have $|\theta| = \theta$ and $|1-\theta| = 1-\theta$,
So we can get that $\forall\mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall\theta \in [0,1]$,

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

So above all, from the defination, we can get that f is a convex function.

(b) $f(X) = \sqrt{\lambda_{\max}(X^\top X)} = \sqrt{\sup_{\mathbf{y}\in\mathbb{R}^n} \frac{\mathbf{y}^\top(X^\top X)\mathbf{y}}{\|\mathbf{y}\|_2^2}} = \sup_{\mathbf{y}\in\mathbb{R}^n} \sqrt{\frac{(X\mathbf{y})^\top(X\mathbf{y})}{\|\mathbf{y}\|_2^2}} = \sup_{\mathbf{y}\in\mathbb{R}^n} \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$

Let $g(X,\mathbf{y}) = \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$, then for a fixed $\mathbf{y}$, and $\forall X_1, X_2 \in \mathbb{R}^{m\times n}$, $\forall\theta \in [0,1]$, we have

$$\begin{aligned} g(\theta X_1 + (1-\theta)X_2, \mathbf{y}) &= \frac{\|\left[\theta X_1 + (1-\theta)X_2\right]\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \\ &\leq \frac{\|\theta X_1\mathbf{y}\|_2 + \|(1-\theta)X_2\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \\ &= \theta\frac{\|X_1\mathbf{y}\|_2}{\|\mathbf{y}\|_2} + (1-\theta)\frac{\|X_2\mathbf{y}\|_2}{\|\mathbf{y}\|_2} \\ &= \theta g(X_1,\mathbf{y}) + (1-\theta)g(X_2,\mathbf{y}) \end{aligned}$$

So when $\mathbf{y}$ is fixed, $g(X,\mathbf{y})$ is convex in X.

From the property of pointwise supremum, we have for each $\mathbf{y}$, $g(X,\mathbf{y}) = \frac{\|X\mathbf{y}\|_2}{\|\mathbf{y}\|_2}$ is convex in X.

Then $f(X) = \sup_{\mathbf{y}\in\mathbb{R}^n} g(X,\mathbf{y})$ is convex.

(c) Similarly with the relation between L_1 norm and L_∞ norm, we could guess that the nuclear norm is the dual norm of the spectral norm. And we could prove our guess.

Define $\|X\|_* = \sum_{i=1}^{r} \sigma_i(X)$ to be the nuclear norm, and $\|X\|_2 = \sigma_{\max}(X)$ to be the spectral norm.

i.e. we want to prove that

$$\|X\|_* = \sup_{\|Z\|_2\leq 1} < Z, X >$$

We firstly apply SVD to X, and we have $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with singular values on the diagonal. And suppose that X has r singular values.
Define $P = UV^\top = UI_nV^\top$, then we could find that all the singular values of P are 1, i.e. $\|P\|_2 = 1$.
Then we can get that

$$\begin{aligned}
\sup_{\|Z\|_2\le 1} < Z, X > &= \sup_{\|Z\|_2\le 1} \mathbf{Tr}(Z^\top X) \\
&\ge \mathbf{Tr}(P^\top X) && (\|P\|_2 = 1 \le 1) \text{ satisfies the constraint} \\
&= \mathbf{Tr}((UV^\top)^\top U\Sigma V^\top) \\
&= \mathbf{Tr}(V^\top VU^\top U\Sigma) && (\mathbf{Tr}(AB) = \mathbf{Tr}(BA)) \\
&= \mathbf{Tr}(\Sigma) \\
&= \sum_{i=1}^{r} \sigma_i(X) \\
&= \|X\|_*
\end{aligned}$$

i.e. we have proved that

$$\|X\|_* \le \sup_{\|Z\|_2\le 1} < Z, X > \tag{1}$$

Then we can prove the other direction of the inequality.

$$\begin{aligned}
\sup_{\|Z\|_2\le 1} < Z, X > &= \sup_{\|Z\|_2\le 1} \mathbf{Tr}\left(Z^\top \left(U\Sigma V^\top\right)\right) \\
&= \sup_{\|Z\|_2\le 1} \mathbf{Tr}\left(\left(V^\top Z^\top U\right)\Sigma\right) && (\mathbf{Tr}(AB) = \mathbf{Tr}(BA))
\end{aligned}$$

$$\begin{aligned}
\sup_{\|Z\|_2\le 1} < Z, X > &= \sup_{\|Z\|_2\le 1} \sum_{i=1}^{r} \sigma_i\left(X\right) \cdot \left(V^\top Z^\top U\right)_{ii} \\
&= \sup_{\|Z\|_2\le 1} \sum_{i=1}^{r} \sigma_i\left(X\right) \left(u_i^\top Z v_i\right) && (u_i, v_i \text{ are the } i\text{-th column of } U, V) \\
&= \sup_{\|Z\|_2\le 1} \sum_{i=1}^{r} \sigma_i\left(X\right) \left\|u_i^\top Z v_i\right\|_2 \\
&\le \sup_{\|Z\|_2\le 1} \sum_{i=1}^{r} \sigma_i\left(X\right) \|u_i\|_2 \|Z\|_2 \|v_i\|_2 && \text{(Cauchy-Schwarz Inequality)} \\
&\le \sum_{i=1}^{r} \sigma_i\left(X\right) && (\|Z\|_2 \le 1, \|u_i\|_2 = \|v_i\|_2 = 1) \\
&= \|X\|_*
\end{aligned}$$

i.e. we have proved that

$$\|X\|_* \ge \sup_{\|Z\|_2\le 1} < Z, X > \tag{2}$$

So combine the inequalities (1) and (2) we can get that the nuclear norm is the dual norm of the spectral norm. i.e.

$$\|X\|_* = \sup_{\|Z\|_2\le 1} < Z, X > \tag{3}$$

And then we prove the triangle inequality of the nuclear norm: $\forall X, Y \in \mathbb{R}^{m\times n}$,

$$\begin{aligned}
\|X + Y\|_* &= \sup_{\|Z\|_2\le 1} < Z, X + Y > && \text{(By the equation (3) we have proved)} \\
&= \sup_{\|Z\|_2\le 1} (< Z, X > + < Z, Y >) \\
&\le \sup_{\|Z\|_2\le 1} < Z, X > + \sup_{\|Z\|_2\le 1} < Z, Y > && \text{(By the property of supremum)} \\
&= \|X\|_* + \|Y\|_*
\end{aligned}$$

So we have proved that $\forall X, Y \in \mathbb{R}^{m\times n}$,

$$\|X+Y\|_* \leq \|X\|_* + \|Y\|_* \tag{4}$$

Suppose that the i-th eigenvalue of $X^\top X$ is $\lambda_i(X^\top X)$, and the i-th singular value of X is $\sigma_i(X)$. Which means that $\sigma_i(X) = \sqrt{\lambda_i(X^\top X)}$.
Then we have the i-th singular value of θX, where $\theta \in [0,1]$ is

$$\sigma_i(\theta X) = \sqrt{\lambda_i((\theta X)^\top(\theta X))} = \sqrt{\lambda_i(\theta^2 X^\top X)} = \theta\sqrt{\lambda_i(X^\top X)} = \theta\sigma_i(X)$$

So we have

$$\|\theta X\|_* = \sum_{i=1}^{r}\sigma_i(\theta X) = \sum_{i=1}^{r}\theta\sigma_i(X) = \theta\sum_{i=1}^{r}\sigma_i(X) = \theta\|X\|_*$$

So above all, we have proved that $\forall\theta \in [0,1]$,

$$\|\theta X\|_* = \theta\|X\|_* \tag{5}$$

With the above conclusions, we could prove that $\forall X_1, X_2 \in \mathbb{R}^{m\times n}, \theta \in [0,1]$

$$\begin{aligned}
f(\theta X_1 + (1-\theta)X_2) &= \|\theta X_1 + (1-\theta)X_2\|_* \\
&\leq \|\theta X_1\|_* + \|(1-\theta)X_2\|_* && \text{(By the inequality (4) we have proved)} \\
&= \theta\|X_1\|_* + (1-\theta)\|X_2\|_* && \text{(By the equality (5) we have proved)} \\
&= \theta f(X_1) + (1-\theta)f(X_2)
\end{aligned}$$

So above all, we have proved that the nuclear norm is convex.

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

We can construct a distribution X s.t.
the domain of X is $\frac{a_i}{b_i}, i \in \{1, 2, \cdots, n\}$, and the PMF of X is $P(X = \frac{a_i}{b_i}) = \frac{b_i}{\sum_{i=1}^{n} b_i}$.

Since $\forall i \in \{1, 2, \cdots, n\}, a_i > 0, b_i > 0$,
So $P(X = \frac{a_i}{b_i}) > 0$, and $\sum_{i=1}^{n} P(X = \frac{a_i}{b_i}) = 1$.
So it's a valid distribution.
So

$$\mathbb{E}(X) = \sum_{i=1}^{n} (\frac{a_i}{b_i}) \cdot P(X = \frac{a_i}{b_i}) = \sum_{i=1}^{n} \frac{a_i}{b_i} \cdot \frac{b_i}{\sum_{k=1}^{n} b_k} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

And constrtuct $f(x) = x \log x$, then $f'(x) = 1 + \log x$, $f''(x) = \frac{1}{x}$.
Since $x > 0$, so $f''(x) = \frac{1}{x} > 0$.
So $f(x)$ is strictly convex, so from the Jensen's inequality, we can get that

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$$

i.e.

$$\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \log \left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \leq \sum_{i=1}^{n} P(X = \frac{a_i}{b_i}) \cdot f(\frac{a_i}{b_i})$$

$$\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \log \left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right) \leq \sum_{i=1}^{n} \frac{b_i}{\sum_{k=1}^{n} b_k} \cdot (\frac{a_i}{b_i}) \log(\frac{a_i}{b_i})$$

Since $b_i > 0$, so $\sum_{i=1}^{n} b_i > 0$, so appointment $\sum_{i=1}^{n} b_i$ on both sides simultaneously, we can get that

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

i.e.

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

So above all, with such construction, we have proved the inequality

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

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

If $p(x)$ is fixed(given), $I(X;Y)$ is convex in $p(y|x)$.
define: 3 distributions: $p_1(y|x), p_2(y|x), p_\lambda(y|x)$, where
$p_\lambda(y|x) = \lambda p_1(y|x) + (1-\lambda)p_2(y|x)$, $p_1(x) = p_2(x) = p_\lambda(x) = p(x)$. So

$$I_i(X;Y) = \sum_{x,y} p_i(x,y) \log \frac{p_i(x,y)}{p_i(x)p_i(y)}, \quad i = 1,2$$

$$I_\lambda(X;Y) = \sum_{x,y} p_\lambda(x,y) \log \frac{p_\lambda(x,y)}{p_\lambda(x)p_\lambda(y)}$$

From log-sum Inequality with $n = 2$:

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

$$\begin{aligned}
I_\lambda(X;Y) &= \sum_{x,y} p_\lambda(x,y) \log \frac{p_\lambda(x,y)}{p_\lambda(x)p_\lambda(y)} \\
&= \sum_{x,y} p_\lambda(y|x)p_\lambda(x) \log \frac{p_\lambda(y|x)}{p_\lambda(y)} \\
&= \sum_x p(x) \sum_y \left(\underbrace{\lambda p_1(y|x)}_{a_1} + \underbrace{(1-\lambda)p_2(y|x)}_{a_2} \right) \log \frac{\lambda p_1(y|x) + (1-\lambda)p_2(y|x)}{\underbrace{\lambda p_1(y)}_{b_1} + \underbrace{(1-\lambda)p_2(y)}_{b_2}} \\
&\leq \sum_x p(x) \sum_y \left[\lambda p_1(y|x) \log \frac{\lambda p_1(y|x)}{\lambda p_1(y)} + (1-\lambda)p_2(y|x) \log \frac{(1-\lambda)p_2(y|x)}{(1-\lambda)p_2(y)} \right] \\
&= \lambda \sum_x p(x) \sum_y p_1(y|x) \log \frac{p_1(y|x)}{p_1(y)} + (1-\lambda) \sum_x p(x) \sum_y p_2(y|x) \log \frac{p_2(y|x)}{p_2(y)} \\
&= \lambda I_1(X;Y) + (1-\lambda) I_2(X;Y)
\end{aligned}$$

So we have proved that $I(X;Y)$ is convex in $p(y|x)$.

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

(a)

$$\mathrm{H(Y)} = -\frac{6}{8} * \log_2\left(\frac{6}{8}\right) - \frac{2}{8} * \log_2\left(\frac{2}{8}\right) \approx 0.811$$

- $I(Y; X_1) = 0.467$
  For attribute X_1,

$$\begin{aligned}
&- H(Y \mid X_1 = \text{ sunny }) = -\left[\frac{1}{3} * \log_2\left(\frac{1}{3}\right) + \frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right] \approx 0.918 \\
&- H(Y \mid X_1 = \text{ rain }) = 0 \\
&- H(Y \mid X_1 = \text{ overcast }) = 0 \\
&\Rightarrow H(Y \mid X_1) = \left[\frac{3}{8} * 0.918 + \frac{3}{8} * 0 + \frac{2}{8} * 0\right] \approx 0.344 \\
&\Rightarrow I(Y; X_1) \approx 0.811 - 0.344 = 0.467
\end{aligned}$$

- $I(Y; X_2) = 0.061$
  For attribute X_2,

$$\begin{aligned}
&- H(Y \mid X_2 = \text{ hot }) = -\left[\frac{1}{3} * \log_2\left(\frac{1}{3}\right) + \frac{2}{3} * \log_2\left(\frac{2}{3}\right)\right] \approx 0.918 \\
&- H(Y \mid X_2 = \text{ cool }) = 0 \\
&- H(Y \mid X_2 = \text{ mild }) = -\left[\frac{3}{4} * \log_2\left(\frac{3}{4}\right) + \frac{1}{4} * \log_2\left(\frac{1}{4}\right)\right] \approx 0.811 \\
&\Rightarrow H(Y \mid X_2) = \left[\frac{3}{8} * 0.918 + \frac{1}{8} * 0 + \frac{4}{8} * 0.811\right] \approx 0.75 \\
&\Rightarrow I(Y; X_2) \approx 0.811 - 0.75 = 0.061
\end{aligned}$$

- $I(Y; X_3) = 0.311$
  For attribute X_3,

$$
\begin{aligned}
&- H\left(Y \mid X_3 = \text{high}\right) = -\left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right) + \frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = 1 \\
&- H\left(Y \mid X_2 = \text{normal}\right) = 0 \\
&\Rightarrow H\left(Y \mid X_3\right) = \left[\frac{4}{8} * 1.0 + \frac{4}{8} * 0\right] = 0.5 \\
&\Rightarrow I\left(Y; X_3\right) \approx 0.811 - 0.5 = 0.311
\end{aligned}
$$

- Split on X_1 at the root node
  Since splitting on attribute X_1 gives the highest mutual information, the root node is X_1.

(b) From the above part, as we can see that the sub-datasets $\mathcal{D}_{(X_1=\text{ rain })}$ and $\mathcal{D}_{(X_1=\text{ overcast })}$ are pure, there will be no further splitting on those and we will place a leaf node with label assignment decided by majority vote classifier. So, we need to split only on the sub-dataset $\mathcal{D}_{(X_1=\text{ sunny })}$. Now, we will use only $\mathcal{D}_{(X_1=\text{ sunny })}$ to estimate the probabilities for the next split.
$H(Y) = -\frac{1}{3} * \log_2\left(\frac{1}{3}\right) - \frac{2}{3} * \log_2\left(\frac{2}{3}\right) \approx 0.918$
For attribute X_2,

- $H\left(Y \mid X_2 = hot\right) = 0$
- $H\left(Y \mid X_2 = \text{cool}\right) = 0$
- $H\left(Y \mid X_2 = \text{ mild }\right) = -\left[\frac{1}{2} * \log_2\left(\frac{1}{2}\right) + \frac{1}{2} * \log_2\left(\frac{1}{2}\right)\right] = 1$
  $\Rightarrow H\left(Y \mid X_2\right) = \left[\frac{2}{3} * 1.0 + \frac{1}{3} * 0\right] \approx 0.67$
  $\Rightarrow I\left(Y; X_2\right) \approx 0.918 - 0.67 \approx 0.25$

For attribute X_3,

- $H\left(Y \mid X_3 = high\right) = 0$
- $H\left(Y \mid X_3 = \text{ normal }\right) = 0$
  $\Rightarrow H\left(Y \mid X_3\right) = \left[\frac{2}{3} * 0 + \frac{1}{3} * 0\right] = 0$
  $\Rightarrow I\left(Y; X_3\right) \approx 0.918$

We split using attribute X_3 as it gives the highest mutual information.

(c)