Due Date : January 25 (10pm), 2019

Instructions

- *This assignment serves as a warm-up for the following assignments. You are not obliged to finish this assignment, but some of the results here might be useful for the upcoming assignments. Unless otherwise specified, you may use the results in this assignment directly in your answer in the future.*
- *Use a document preparation system such as LaTeX.*
- *You will be using Gradescope, you should have received an email to sign up, otherwise sign up for an account on gradescope.com and use course code 9EVRGV*
- *Submit this test submission on Gradescope (not necessary to complete and not marked)*

Question 1. Given any unit vector $\boldsymbol{n}$ (i.e. $||\boldsymbol{n}|| = 1$), we define the hyperplane $\mathcal{H}_{\boldsymbol{n}} := \{\boldsymbol{x} : \boldsymbol{n}^\top \boldsymbol{x} = 0\}$ for which $\boldsymbol{n}$ is known as the normal vector. For any vector $\boldsymbol{x}$, we define its projection into $\mathcal{H}_{\boldsymbol{n}}$ as $\pi_{\boldsymbol{n}}(\boldsymbol{x}) = \boldsymbol{x} - (\boldsymbol{x}^\top \boldsymbol{n})\boldsymbol{n}$.

1. Given two vectors $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$, take $\boldsymbol{n} = \frac{\boldsymbol{x}_2 - \boldsymbol{x}_1}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||}$. Show that $\pi_{\boldsymbol{n}}(\boldsymbol{x}_1) = \pi_{\boldsymbol{n}}(\boldsymbol{x}_2)$.
2. Let $\boldsymbol{w}$ be a vector and define $y_1 := \boldsymbol{x}_1^\top \boldsymbol{w}$ and $y_2 := \boldsymbol{x}_2^\top \boldsymbol{w}$. Show that $y_1 = y_2$ if and only if $\boldsymbol{w} \in \mathcal{H}_{\boldsymbol{n}}$.

*3. Let $\boldsymbol{X}$ be a n by p matrix whose rows $\boldsymbol{X}_{i,:}$ are all distinct. Show that there exists a vector $\boldsymbol{w}$ of length p such that the scalars $(\boldsymbol{X}\boldsymbol{w})_i$ are all distinct.

Answer 1.

1. By definition,

$$\begin{aligned}
\pi_{\boldsymbol{n}}(\boldsymbol{x}_1) - \pi_{\boldsymbol{n}}(\boldsymbol{x}_2) &= \boldsymbol{x}_1 - (\boldsymbol{x}_1^\top \boldsymbol{n})\boldsymbol{n} - \boldsymbol{x}_2 + (\boldsymbol{x}_2^\top \boldsymbol{n})\boldsymbol{n} \\
&= \boldsymbol{x}_1 - \boldsymbol{x}_2 - ((\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \boldsymbol{n})\boldsymbol{n} \\
&= \boldsymbol{x}_1 - \boldsymbol{x}_2 - \left((\boldsymbol{x}_1 - \boldsymbol{x}_2)^\top \left(\frac{\boldsymbol{x}_2 - \boldsymbol{x}_1}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||}\right)\right) \frac{\boldsymbol{x}_2 - \boldsymbol{x}_1}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||} \\
&= \boldsymbol{x}_1 - \boldsymbol{x}_2 - \frac{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||^2}(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{0}
\end{aligned}$$

2. $\boldsymbol{w} \in \mathcal{H}_{\boldsymbol{n}} \Leftrightarrow \boldsymbol{n}^\top \boldsymbol{w} = \frac{(\boldsymbol{x}_2 - \boldsymbol{x}_1)^\top \boldsymbol{w}}{||\boldsymbol{x}_2 - \boldsymbol{x}_1||} = 0 \Leftrightarrow \boldsymbol{x}_1^\top \boldsymbol{w} = \boldsymbol{x}_2^\top \boldsymbol{w} \Leftrightarrow y_1 = y_2$
3. Let $\boldsymbol{x}_i$ be the i'th row of $\boldsymbol{X}$. From the previous question, we know for any two rows $i \neq j$, $\{\boldsymbol{w} : \boldsymbol{x}_i^\top \boldsymbol{w} - \boldsymbol{x}_j^\top \boldsymbol{w} = 0\} = \mathcal{H}_{\boldsymbol{n}_{ij}}$ where $\boldsymbol{n}_{ij} = \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{||\boldsymbol{x}_i - \boldsymbol{x}_j||}$. Take $\boldsymbol{w} \in \mathbb{R}^p \setminus \bigcup_{i \neq j} \{\boldsymbol{w} : \boldsymbol{x}_i^\top \boldsymbol{w} - \boldsymbol{x}_j^\top \boldsymbol{w} = 0\} = \mathbb{R}^p \setminus \bigcup_{i \neq j} \mathcal{H}_{\boldsymbol{n}_{ij}}$, or equivalently,

$$\left\{\boldsymbol{w} \in \mathbb{R}^p : \boldsymbol{n}_{ij}^\top \boldsymbol{w} \neq 0 \quad \forall\, i \neq j\right\}$$

Since $\mathbb{R}^p$ minus finite union of $p-1$-dimensional hyperplanes is still infinite, this is a non-empty set.

Question 2. Recall the variance of X is $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

1. Let X be a random variable with finite mean. Show $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

2. Let X and Z be random variables on the same probability space. Show that $\text{Var}(X) = \mathbb{E}_Z[\text{Var}(X|Z)] + \text{Var}_Z(\mathbb{E}[X|Z])$. (Hint : $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$.)

Answer 2.

1. $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
2. From the previous question,

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \mathbb{E}_Z[\mathbb{E}[X^2|Z]] - \mathbb{E}_Z[\mathbb{E}[X|Z]]^2 \\ &= \mathbb{E}_Z[\text{Var}(X|Z) + \mathbb{E}[X|Z]^2] - \mathbb{E}_Z[\mathbb{E}[X|Z]]^2 \\ &= \mathbb{E}_Z[\text{Var}(X|Z)] + (\mathbb{E}_Z[\mathbb{E}[X|Z]^2] - \mathbb{E}_Z[\mathbb{E}[X|Z]]^2) \\ &= \mathbb{E}_Z[\text{Var}(X|Z)] + \text{Var}_Z(\mathbb{E}[X|Z])\end{aligned}$$

Question 3. Let $X \in \mathcal{X}$ be a random variable with density function f_X, and $g : \mathcal{X} \to \mathcal{Y}$ be continuously differentiable, where $\mathcal{X}$ and $\mathcal{Y}$ are subsets of $\mathbb{R}$. Let $Y := g(X)$, which is continuously distributed with density function f_Y.

1. Show that if g is monotonic, $f_Y(y) = f_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|$.
2. Let $f_X(x) = \mathbf{1}_{x\in[0,1]}(x)$ and $f_Y(y) = \mathbf{1}_{y\in[0,2]}(y) \cdot \frac{y}{2}$. Find a monotonic mapping g that translates f_X and f_Y.

*3. Let $N_Y = \{y \in \mathcal{Y} : g(x) = y,\ g(x)' = 0 \text{ for some } x \in \mathcal{X}\}$. Show that in general if $g'(x) = 0$ at most finitely many times, for $y \in \mathcal{Y} \setminus N_Y$,

$$f_Y(y) = \sum_{x\in\{x:g(x)=y\}} \frac{f_X(x)}{|g'(x)|}$$

4. Let $X \sim \mathcal{N}(0, 1)$, i.e. $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, and $g(x) = x^2$. Determine $f_Y(y)$.

Answer 3.

1. When g is monotonic and continuously differentiable, it is a bijection. Let g^{-1} be the inverse function. Let F_X be the cumulative distribution function (cdf) of X, and F_Y that of Y. We have, for every $y \in \mathcal{Y}$:

$$F_Y(y) = \begin{cases} \mathbb{P}(X < g^{-1}(y)) = F_X(g^{-1}(y)) & \text{if } g \text{ non-decreasing} \\ \mathbb{P}(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{if } g \text{ non-increasing} \end{cases}$$

By differentiating both sides of the equation, in both cases, we obtain :

$$f_Y(y) = \begin{cases} \frac{dg^{-1}(y)}{dy} f_X(g^{-1}(y)) & \text{if } g \text{ non-decreasing} \\ -\frac{dg^{-1}(y)}{dy} f_X(g^{-1}(y)) & \text{if } g \text{ non-increasing} \end{cases}$$

In both cases, this can be written as $f_Y(y) = f_X(g^{-1}(y)) \left|\frac{dg^{-1}(y)}{dy}\right|$

2. Using the formula shown in the previous question, we can look for a non-decreasing mapping g from $[0, 1]$ to $[0, 2]$ that satisfies $\frac{dg^{-1}(y)}{dy} = \frac{y}{2}$. It is straightforward then that $g : x \in [0, 1] \mapsto 2\sqrt{x}$ translates f_X to f_Y.

3. Intuitively, we want to split $\mathcal{X}$ into sub-intervals on which g (indexed by g_i) is monotonic, and select the g_i's that intersect with y horizontally.

   More rigorously, let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be disjoints intervals of $\mathcal{X}$ such that $\forall i\ \forall x \in \mathcal{X}_i\ g'(x) \neq 0$ and $\mathcal{X} \setminus \{x \in \mathcal{X} : g'(x) = 0\} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_n$. This amounts to splitting $\mathcal{X}$ into disjoints intervals on which g' does not take the value zero (hence monotonic). For each subset, we denote by g_i the restriction of g on $\mathcal{X}_i$, i.e. $g_i : x \in \mathcal{X}_i \mapsto g_i(x) = g(x)$. Note that all the functions g_i are monotonic by definition, and hence bijective.

   Let $y \in \mathcal{Y} \setminus N_Y$. The event $\{Y < y\}$ can be rewritten as the disjoint union of the events $\{X \in \mathcal{X}_1,\ g_1(X) < y\}, \ldots, \{X \in \mathcal{X}_n,\ g_n(X) < y\}$. Hence :

$$\begin{aligned}
F_Y(y) &= \sum_{i=1}^{n} \mathbb{P}(X \in \mathcal{X}_i,\ g_i(X) < y) \\
&= \sum_{i=1}^{n} \left(\mathbf{1}_{y \in g(\mathcal{X}_i)} \mathbb{P}(X \in \mathcal{X}_i,\ g_i(X) < y) + \mathbf{1}_{y > \sup g(\mathcal{X}_i)} \mathbb{P}(X \in \mathcal{X}_i) \right)
\end{aligned}$$

   Given that

$$\mathbb{P}(X \in \mathcal{X}_i,\ g_i(X) < y) = \begin{cases} F_X(g_i^{-1}(y)) - F_X(\inf \mathcal{X}_i) & \text{if } g_i \text{ non-decreasing} \\ F_X(\sup \mathcal{X}_i) - F_X(g_i^{-1}(y)) & \text{if } g_i \text{ non-increasing,} \end{cases}$$

   we have, similar to the first question :

$$\frac{d\mathbb{P}(X \in \mathcal{X}_i,\ g_i(X) < y)}{dy} = f_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right|$$

   Recall the inverse function theorem. Differentiating $F_Y(y)$ yields

$$\begin{aligned}
f_Y(y) &= \sum_{i=1}^{n} \mathbf{1}_{y \in g(\mathcal{X}_i)} f_X(g_i^{-1}(y)) \left| \frac{dg_i^{-1}(y)}{dy} \right| \\
&= \sum_{x \in \{x : g(x) = y\}} \frac{f_X(x)}{|g'(x)|}
\end{aligned}$$

4. We apply the formula of the previous question to obtain :

$$\begin{aligned}
f_Y(y) &= \left(\frac{f_X(\sqrt{y})}{|g'(\sqrt{y})|} + \frac{f_X(-\sqrt{y})}{|g'(-\sqrt{y})|} \right) \mathbf{1}_{(0,\infty)}(y) \\
&= \frac{\exp^{-y/2}}{\sqrt{2\pi y}} \mathbf{1}_{(0,\infty)}(y)
\end{aligned}$$

Question 4. Let Q and P be univariate normal distributions with mean and variance μ, σ^2 and m, s^2, respectively. Derive the entropy $H(Q)$, the cross-entropy $H(Q, P)$, and the KL divergence $D_{\mathrm{KL}}(Q||P)$.

Answer 4.

$$
\begin{aligned}
H(Q) &= -\mathbb{E}_{X\sim Q}[\log Q(X)] \\
&= -\mathbb{E}_{X\sim Q}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(X-\mu)^2}{2\sigma^2}\right] \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}_{X\sim Q}\left[(X - \mathbb{E}_{X\sim Q}[X])^2\right] \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathrm{Var}_{X\sim Q}\left[X\right] \\
&= \frac{1}{2}(1 + \log(2\pi\sigma^2)) \\
H(Q,P) &= -\mathbb{E}_{X\sim Q}[\log P(X)] \\
&= \frac{1}{2}\log(2\pi s^2) + \frac{1}{2s^2}\mathbb{E}_{X\sim Q}\left[(X-m)^2\right] \\
&= \frac{1}{2}\log(2\pi s^2) + \frac{1}{2s^2}\mathbb{E}_{X\sim Q}\left[(X-\mu)^2 + (\mu - m)(2X - m - \mu)\right] \\
&= \frac{1}{2}\log(2\pi s^2) + \frac{1}{2s^2}(\sigma^2 + (\mu - m)^2) \\
D_{\mathrm{KL}}(Q||P) &= H(Q,P) - H(Q) \\
&= \log(\frac{s}{\sigma}) + \frac{\sigma^2 - s^2 + (\mu - m)^2}{2s^2}
\end{aligned}
$$