Due Date : January 12 (11pm), 2020

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 LaTeX and the template we provide when writing your answers. You may reuse most of the notation shorthands, equations and/or tables. See the assignment policy on the course website for more details.*
- *You will be using Gradescope, you should have received an invitation email (sent to the email address you use for StudiuM). Otherwise reply to the thread named "Gradescope Sign Up" on Piazza to let the TA know your "name", "email" and "student ID" (matricule) as shown on StudiuM to ask the TA to add you on Gradescope.*
- *Submit this test submission on Gradescope (you don't need to complete it, but are recommended to do it for self evaluation).*

Question 1. Let $\sigma(x) = \frac{1}{1+e^{-x}}$. Find $\frac{d\sigma}{dx}$ using the definition of the derivative (i.e. taking the limit of difference quotients).

Answer 1.

$$\begin{aligned}
\frac{d\sigma(x)}{dx} &= \lim_{h\to 0} \frac{\sigma(x+h) - \sigma(x)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}\left(\frac{1}{1+e^{-x-h}} - \frac{1}{1+e^{-x}}\right) \\
&= \lim_{h\to 0} \frac{1}{h}\left(\frac{1+e^{-x}-1-e^{-x-h}}{1+e^{-x}+e^{-x-h}+e^{-2x-h}}\right) \\
&= \frac{e^{-x}}{1+2e^{-x}+e^{-2x}} \cdot \lim_{h\to 0}\left(\frac{1-e^{-h}}{h}\right) && \text{(constant multiple, product rule)} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \cdot \lim_{h\to 0}\left(\frac{1-(1-h+o(|h|))}{h}\right) && \text{(Taylor's theorem)} \\
&= \frac{e^{-x}}{(1+e^{-x})^2} \cdot \lim_{h\to 0}\left(\frac{h+o(|h|))}{h}\right) \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x)(1-\sigma(x))
\end{aligned}$$

Question 2. Compute the gradients of $||\boldsymbol{x}||_2^2$, $||\boldsymbol{x}-\boldsymbol{a}||_2$, $||\boldsymbol{X}||_F^2$, and $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ w.r.t the input vector $\boldsymbol{x}$ and and matrix $\boldsymbol{X}$.

Answer 2.
$||\boldsymbol{x}||_2^2$: Rewriting $||\boldsymbol{x}||_2^2 = \sum_i x_i^2$, we can compute the partial derivative $\frac{\partial}{\partial x_j}\sum_i x_i^2 = 2x_j$. This means in vector form, $\nabla_{\boldsymbol{x}}||\boldsymbol{x}||_2^2 = 2\boldsymbol{x}$.

$||\boldsymbol{x}-\boldsymbol{a}||_2$: Similarly, taking the partial derivative yields

$$\begin{aligned}\frac{\partial}{\partial x_j}||\boldsymbol{x}-\boldsymbol{a}||_2 &= \frac{\partial}{\partial x_j}\sqrt{||\boldsymbol{x}-\boldsymbol{a}||_2^2} = \frac{1}{2\sqrt{||\boldsymbol{x}-\boldsymbol{a}||_2^2}}\cdot\frac{\partial}{\partial x_j}||\boldsymbol{x}-\boldsymbol{a}||_2^2 \\ &= \frac{1}{2||\boldsymbol{x}-\boldsymbol{a}||_2}\cdot\frac{\partial||\boldsymbol{x}-\boldsymbol{a}||_2^2}{\partial(x_j-a_j)}\cdot\frac{\partial(x_j-a_j)}{\partial x_j} = \frac{1}{2||\boldsymbol{x}-\boldsymbol{a}||_2}\cdot 2(x_j-a_j)\cdot 1\end{aligned}$$

which means $\nabla_{\boldsymbol{x}}||\boldsymbol{x}-\boldsymbol{a}||_2 = \frac{\boldsymbol{x}-\boldsymbol{a}}{||\boldsymbol{x}-\boldsymbol{a}||_2}$.

$||\boldsymbol{X}||_F^2$: We can also express the matrix Frobenius norm as a sum $\sum_{ij} x_{ij}^2$. The linearity of summation gives $\nabla_{\boldsymbol{X}}||\boldsymbol{X}||_F^2 = 2\boldsymbol{X}$.

$\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$: We can view the quadratic formula as $\boldsymbol{x}^\top\boldsymbol{a}$ where $\boldsymbol{a}$ depends on $\boldsymbol{x}$ through $\boldsymbol{a} := \boldsymbol{A}\boldsymbol{x}$. Then by product rule, we have

$$\nabla_{\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{a} = \boldsymbol{a} + \left(\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}}\right)^\top\boldsymbol{x}$$

where $\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}}$ is the Jacobian matrix $(\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}})_{ij} = \frac{\partial a_i}{\partial x_j} = \frac{\partial\sum_k A_{ik}x_k}{\partial x_j} = A_{ij}$, which means

$$\nabla_{\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^\top\boldsymbol{x} = (\boldsymbol{A}+\boldsymbol{A}^\top)\boldsymbol{x}$$

Question 3. 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 $\mathrm{Var}(X) = \mathbb{E}_Z[\mathrm{Var}(X|Z)] + \mathrm{Var}_Z(\mathbb{E}[X|Z])$. (Hint : $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}[X|Y]]$.)

Answer 3.

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

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

Question 4. Recall the density function of the uniform distribution on $[a,b]$ for $a<b$ is equal to $\frac{1}{b-a}$ for $x\in[a,b]$ and 0 elsewhere.

1. Use the density function to compute the mean and variance of a uniform distribution on $[a,b]$.
2. For integer $n>0$, derive a formula to compute the moment $\mathbb{E}[X^n]$ for X uniformly distributed between a and b.

Answer 4. The density function for uniform random variable can be written as $f(x) = \frac{1}{b-a}\delta_{[a,b]}(x)$, where $\delta_{[a,b]}(x) = 1$ if $x\in[a,b]$ and 0 otherwise. This means we can write the integration $\int_{\mathbb{R}} f(x)g(x)dx$ as $\frac{1}{b-a}\int_a^b g(x)dx$ for any integrable function g.

1. For the first moment, $g(x) = x$, so we have

$$\mathbb{E}[X] = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\cdot\frac{1}{2}(b^2-a^2) = \frac{a+b}{2}$$

We can decompose the variance as $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. For the second moment, we have $g(x) = x^2$ and

$$\mathbb{E}[X^2] = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\cdot\frac{1}{3}(b^3-a^3) = \frac{a^2+ab+b^2}{3}$$

Combining this with the variance decomposition, we have

$$\begin{aligned}\mathrm{Var}(X) &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4}\\ &= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12}\\ &= \frac{a^2-ab+b^2}{12} = \frac{(b-a)^2}{12}\end{aligned}$$

2. Generally, for the n'th moment, $g(x) = x^n$, we have

$$\mathbb{E}[X^n] = \frac{1}{b-a}\cdot\frac{1}{n+1}(b^{n+1}-a^{n+1}) = \frac{1}{n+1}\sum_{j=0}^{n} b^j a^{n-j}$$

Question 5. 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.

Answer 5.

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.

Question 6. 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 6.

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