CS7643: Deep Learning
Fall 2020
Problem Set 0

Instructor: Dhruv Batra
TAs: Prabhav Chawla, Yihao Chen, Sameer Dharur, Hrishikesh Kale,
Michael Piseno, Joanne Truong, Tianyu Zhan
Discussions: https://piazza.com/gatech/fall2020/cs48037643

Due: Thursday, August 20, 11:59pm ET

Instructions

1. We will be using a Gradescope Online Assignment for PS0. If you are not registered for this course, please fill the following form in order to be added to Gradescope and be able to submit PS0: https://forms.gle/jDoqyKa3zTeMbwqT7.

2. While we work through adding students to Gradescope, feel free to refer to this PDF to look at questions that are a part of PS0. However, PS0 submissions will be through the Gradescope Online Assignment only.

3. PS0 will not count towards your final grade; however, you are still required to make a submission.

4. We generally encourage you to collaborate with other students. You may talk to a friend, discuss the questions and potential directions for solving them. However, you need to write your own solutions and code separately, and *not* as a group activity. Please list the students you collaborated with.
   Exception: PS0 is meant to serve as a background preparation test. You must NOT collaborate on PS0.

1 Multiple Choice Questions

1. (1 point) true/false We are machine learners with a slight gambling problem (very different from gamblers with a machine learning problem!). Our friend, Bob, is proposing the following payout on the roll of a dice:

$$\text{payout} = \begin{cases} \$2 & x = 1 \\ -\$1/4 & x \neq 1 \end{cases} \tag{1}$$

where $x \in \{1, 2, 3, 4, 5, 6\}$ is the outcome of the roll, $(+)$ means payout to us and $(-)$ means payout to Bob. Is this a good bet i.e. are we expected to make money?

◯ True ◯ False

2. (1 point) X is a continuous random variable with the probability density function:

$$p(x) = \begin{cases} 8x & 0 \leq x \leq 1/2 \\ -2x + 1 & 1/2 \leq x \leq 1 \end{cases} \tag{2}$$

Which of the following statements are true about equation for the corresponding cumulative density function (CDF) $C(x)$?
[*Hint:* Recall that CDF is defined as $C(x) = Pr(X \leq x)$.]

◯ $C(x) = 4x^2$ for $0 \leq x \leq 1/2$
◯ $C(x) = -x^2 + x - 1/4$ for $1/2 \leq x \leq 1$
◯ All of the above
◯ None of the above

3. (2 point) A random variable x in standard normal distribution has the following probability density

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{3}$$

Evaluate following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \tag{4}$$

[*Hint:* We are not sadistic (okay, we're a little sadistic, but not for this question). This is not a calculus question.]

◯ a + b + c ◯ c ◯ a + c ◯ b + c

4. (2 points) Consider the following function of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(\mathbf{x}) = \sigma\left(\log\left(5\left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6)\right)\right) + \frac{1}{2}\right) \tag{5}$$

where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

Compute the gradient $\nabla_{\mathbf{x}} f(\cdot)$ and evaluate it at at $\hat{\mathbf{x}} = (-1, 3, 4, 5, -5, 7)$.

◯ $\begin{bmatrix} 0 \\ 0.031 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ◯ $\begin{bmatrix} 0 \\ 0.157 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ◯ $\begin{bmatrix} 0 \\ 0.358 \\ 0.269 \\ -0.215 \\ -0.846 \\ -0.846 \end{bmatrix}$ ◯ $\begin{bmatrix} 0 \\ 0.358 \\ 0.269 \\ -0.215 \\ -0.448 \\ -0.448 \end{bmatrix}$

5. (2 points) Which of the following functions are convex?
   - ◯ $\|\mathbf{x}\|_{\frac{1}{2}}$
   - ◯ $\min_{i=1}^{k} \mathbf{a}_i^T \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$, and a finite set of arbitrary vectors: $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$
   - ◯ $\log\left(1 + \exp(\mathbf{w}^T \mathbf{x})\right)$ for $\mathbf{w} \in \mathbb{R}^d$
   - ◯ All of the above

6. (2 points) Suppose you want to predict an unknown value $Y \in \mathbb{R}$, but you are only given a sequence of noisy observations $x_1, \ldots, x_n$ of Y with i.i.d. noise ($x_i = Y + \epsilon_i$). If we assume the noise is I.I.D. Gaussian ($\epsilon_i \sim N(0, \sigma^2)$), the maximum likelihood estimate ($\hat{y}$) for Y can be given by:
   - ◯ A: $\hat{y} = \text{argmin}_y \sum_{i=1}^{n} (y - x_i)^2$
   - ◯ B: $\hat{y} = \text{argmin}_y \sum_{i=1}^{n} |y - x_i|$
   - ◯ C: $\hat{y} = \frac{1}{n} \sum_{i=1}^{n} x_i$
   - ◯ Both A & C
   - ◯ Both B & C

2 Proofs

7. (3 points) Prove that

$$\log_e x \leq x - 1, \qquad \forall x > 0 \tag{7}$$

with equality if and only if $x = 1$.

[*Hint:* Consider differentiation of $\log(x) - (x - 1)$ and think about concavity/convexity and second derivatives.]

8. (6 points) Consider two discrete probability distributions p and q over k outcomes:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{8a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{8b}$$

The Kullback-Leibler (KL) divergence (also known as the *relative entropy*) between these distributions is given by:

$$KL(p, q) = \sum_{i=1}^{k} p_i \log \left(\frac{p_i}{q_i} \right) \tag{9}$$

It is common to refer to $KL(p, q)$ as a measure of distance (even though it is not a proper metric). Many algorithms in machine learning are based on minimizing KL divergence between two probability distributions. In this question, we will show why this might be a sensible thing to do.

[*Hint:* This question doesn't require you to know anything more than the definition of $KL(p, q)$ and the identity in Q7]

(a) Using the results from Q7, show that $KL(p, q)$ is always non-negative.

(b) When is $KL(p, q) = 0$?

(c) Provide a counterexample to show that the KL divergence is not a symmetric function of its arguments: $KL(p, q) \neq KL(q, p)$

9. (6 points) In this question, we will get familiar with a fairly popular and useful function, called the log-sum-exp function. For $\mathbf{x} \in \mathbb{R}^n$, the log-sum-exp function is defined (quite literally) as:

$$f(\mathbf{x}) = \log\left(\sum_{i=1}^{n} e^{x_i}\right) \tag{10}$$

(a) Prove that $f(\mathbf{x})$ is differentiable everywhere in $\mathbb{R}^n$.
[*Hint:* Multivariable functions are differentiable if the partial derivatives exist and are continuous.]

(b) Prove that $f(\mathbf{x})$ is convex on $\mathbb{R}^n$.
[*Hint:* One approach is to use the second-order condition for convexity.]

(c) Show that $f(\mathbf{x})$ can be viewed as an approximation of the max function, bounded as follows:

$$\max\{x_1, \ldots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \ldots, x_n\} + \log(n) \tag{11}$$