CS7643: Deep Learning
Fall 2021
Problem Set 0

Instructor: Dhruv Batra
TAs: Joanne Truong, Amol Agrawal, Andrew Szot,
Bhavika Devnani, Jordan Rodrigues, Swornim Baral
Discussions: https://piazza.com/gatech/fall2021/cs48037643

Due: Thursday, Aug 26, 11:59pm ET

Instructions

1. We will be using Gradescope to collect your assignments. Please read the following instructions for submitting to Gradescope carefully! Failure to follow these instructions may result in parts of your assignment not being graded. We will not entertain regrading requests for failure to follow instructions.
    - For Section 1: Multiple Choice Questions, please upload your answers directly to the Gradescope assessment. For every question, there is only one correct answer. If you haven't been added to the Gradescope, please fill out the form at https://qfreeaccountssjc1.az1.qualtrics.com/jfe/form/SV_3KUjzRkmrUwlWxo and/or post on Piazza
    - For Section 2: Proofs - This section has 7 total problems/sub-problems (Q9, Q9a - Q9c and Q10a - Q10c) and your answer to each sub-problem should start on a new page. Please be sure to mark the corresponding pages to the correct question numbers marked on the PS0 outline while submitting on Gradescope.
    - For Section 2, LaTeX'd solutions are strongly encouraged (solution template available at https://www.cc.gatech.edu/classes/AY2022/cs7643_fall/assets/ps0.zip), but scanned handwritten copies are acceptable. If you scan handwritten copies, please mark the pages correctly as mentioned above.
2. Hard copies are **not** accepted.
3. 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) Consider the tables below that display infection rates for a disease in two independent regions given vaccine status.

Region	Pop.	Vaccination Rates	% of Population Infected
Cityville	874,961	77.0%	0.36%
Townsland	578,759	37.7%	1%

Region	% of Infected people that are Vaccinated	% of Infected people that are Unvaccinated
Cityville	27.8%	72.2%
Townsland	5.0%	95.0%

It appears that infected individuals in Cityville are much more likely to be vaccinated than in Townsland. Given these tables, would a vaccinated individual be less likely to be infected in Cityville or Townsland?

○ Cityville ○ Townsland

2. (1 point)

Given a (possibly) biased coin with $P(Heads) = p$ and $P(Tails) = 1 - p$, first determine the method to generate a fair outcome (50:50) in the fewest amount of flips using this coin.

What is the expected number of coin flips required (in terms of p) to produce a fair outcome using this method?

○ $\frac{1}{p(1-p)}$ ○ $\frac{1}{1+p^2}$ ○ $\frac{2p}{1-p}$ ○ A fair outcome cannot be generated with a biased coin

3. (1 point) X is a continuous random variable with probability density function:

$$p(x) = \begin{cases} 2x^3/81 & 0 \le x \le 3 \\ 2(x-3)/8 & 3 \le x \le 5 \end{cases} \tag{1}$$

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

○ $C(x) = x^4/162$ for $0 \le x \le 3$
○ $C(x) = x^2/8 - 3x/4 + 13/8$ for $3 \le x \le 5$
○ All of the above
○ None of the above

4. (2 point) A random variable x in standard normal distribution has the following probability density:

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

Evaluate the following integral:

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx - c)dx \tag{3}$$

[*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

5. (2 points) Consider the following function of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(\mathbf{x}) = \sigma\left(\log\left(5\left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6)\right)\right) + \frac{1}{2}\right) \tag{4}$$

where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

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

○ $\begin{bmatrix} 0 \\ 0.031 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.157 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.357 \\ 0.268 \\ -0.214 \\ -0.894 \\ -0.894 \end{bmatrix}$ ○ $\begin{bmatrix} 0 \\ 0.357 \\ 0.268 \\ -0.214 \\ -0.447 \\ -0.447 \end{bmatrix}$

6. (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: $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$
   - ○ $\log(1 + \exp(\mathbf{w}^T \mathbf{x}))$ for $\mathbf{w} \in \mathbb{R}^d$
   - ○ All of the above

7. (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

8. (3 points) Prove that

$$\log_e x \leq x - 1, \qquad \forall x > 0 \tag{6}$$

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

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

9. (6 points) Consider two discrete probability distributions p and q over k outcomes:

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} q_i = 1 \tag{7a}$$

$$p_i > 0, q_i > 0, \quad \forall i \in \{1, \ldots, k\} \tag{7b}$$

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

$$KL(p, q) = \sum_{i=1}^{k} p_i \log \left(\frac{p_i}{q_i} \right) \tag{8}$$

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: $KL(p, q) \neq KL(q, p)$

10. (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:

$$f(\mathbf{x}) = \log\left(\sum_{i=1}^{n} e^{x_i}\right) \tag{9}$$

   (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:

$$\max\{x_1, \ldots, x_n\} \leq f(\mathbf{x}) \leq \max\{x_1, \ldots, x_n\} + \log(n) \tag{10}$$