CS7643: Deep Learning
Spring 2020
Problem Set 0

Instructor: Zsolt Kira
TAs: Rahul Duggal, Jiachen Yang, Sameer Dharur, Yinquan Lu
Patrick Grady, Anishi Mehta
Discussions: https://piazza.com/gatech/spring2020/cs4803dl7643a/home

Due: Tuesday, Jan 14, 11:55pm

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, it is mandatory to use the LaTeX template provided on the class webpage (https://www.cc.gatech.edu/classes/AY2020/cs7643_fall/assets/ps0.zip). For every question, there is only one correct answer. To mark the correct answer, change `\choice` to `\CorrectChoice`
   - For Section 2: Proofs, each problem/sub-problem is in its own page. This section has 5 total problems/sub-problems, so you should have 5 pages corresponding to this section. Your answer to each sub-problem should fit in its corresponding page.
   - For Section 2, LaTeX'd solutions are strongly encouraged (solution template available at https://www.cc.gatech.edu/classes/AY2020/cs7643_fall/assets/ps0.zip), but scanned handwritten copies are acceptable. If you scan handwritten copies, please make sure to append them to the pdf generated by LaTeX for Section 1.

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) 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:

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

   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:

$$p(x) = \begin{cases} 4x & 0 \leq x \leq 1/2 \\ -4x + 4 & 1/2 \leq x \leq 1 \end{cases} \tag{2}$$

   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 $C(x) = Pr(X \leq x)$.]

   ● $C(x) = 2x^2$ **for** $0 \leq x \leq 1/2$
   ○ $C(x) = -2x^2 + 4x - 3/2$ 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 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 $\hat{\mathbf{x}} = (5, -1, 6, 12, 7, -5)$.

○ $\begin{bmatrix} 0.157 \\ 0.0 \\ 0.131 \\ -0.065 \\ -0.846 \\ -0.846 \end{bmatrix}$ ● $\begin{bmatrix} 0.157 \\ 0 \\ 0.131 \\ -0.065 \\ -0.314 \\ -0.314 \end{bmatrix}$ ○ $\begin{bmatrix} 0.031 \\ 0 \\ 0.026 \\ -0.013 \\ -0.062 \\ -0.062 \end{bmatrix}$ ○ $\begin{bmatrix} -0.468 \\ 0 \\ -0.390 \\ 0.195 \\ 0.937 \\ 0.937 \end{bmatrix}$

5. (2 points) Which of the following functions are convex?

○ $||\mathbf{x}||_{\frac{1}{2}}$

○ $\min_i \mathbf{a}_i^T \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^n$

● $\log(1 + \exp(\mathbf{w}^T \mathbf{x}_i))$ **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 $x_1 ... x_n$ 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: $\hat{y} = \operatorname{argmin}_y \sum_{i=1}^n (y - x_i)^2$

○ B: $\hat{y} = \operatorname{argmin}_y \sum_{i=1}^n |y - x_i|$

○ 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.]

Let $f(x) = \log_e x - x + 1, (x > 0)$

$$f'(x) = \frac{1}{x} - 1$$
$$f''(x) = \frac{1}{x^2} < 0$$

As the second derivative is always negative, $f'(x)$ is monotonically decreasing. While $f'(x)$ is positive on $(0,1)$ and negative on $(1,\infty)$, we have:

$$f(x)_{max} = f(1) = 0$$
$$(\log_e x - x + 1)_{max} = 0$$

Therefore the equality holds if and only if $x = 1$.

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

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

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

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.

$$\begin{aligned} -KL(p, q) &= \sum_{i=1}^{k} p_i \log(\frac{q_i}{p_i}) \\ &\leq \sum_{i=1}^{k} p_i(\frac{q_i}{p_i} - 1) \quad = 0 \end{aligned}$$

Thus, we have:

$$KL(p, q) \geq 0$$

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

When $\frac{q_i}{p_i} = 1$ holds for all i, the equality above holds.

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

Let $p_1 = p_2 = p_3 = p_4 = 1/4$ and $q_1 = q_2 = 1/16, q_3 = q_4 = 7/16$.

$$\begin{aligned} KL(p, q) &= -0.4133 \neq \\ KL(q, p) &= -0.3164 \end{aligned}$$

9. (6 points) In this question, you will prove that cross-entropy loss for a softmax classifier is convex in the model parameters, thus gradient descent is guaranteed to find the optimal parameters. Formally, consider a single training example $(\mathbf{x}, y)$. Simplifying the notation slightly from the implementation writeup, let

$$\mathbf{z} = W\mathbf{x} + \mathbf{b}, \tag{10}$$

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{11}$$

$$L(W) = -\log(p_y) \tag{12}$$

Prove that $L(\cdot)$ is convex in W.

[*Hint:* One way of solving this problem is "brute force" with first principles and Hessians. There are more elegant solutions.]

Suppose the bias here is zero: $\boldsymbol{b} = 0$.

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$$

To compute the gradient of p_j , we suppose $x \in \mathbb{R}_N$, and w_n for the nth class.
When $n \neq j$:

$$\nabla_{w_n} p_j = -\frac{e^{w_j^T x} e^{w_n^T x}}{\left(\sum_{l=1}^k e^{w_l^T x}\right)^2} x = -p_j p_n x$$

When $n = j$:

$$\nabla_{w_j} p_j = \left(\frac{e^{w_j^T x}}{\sum_{l=1}^k e^{w_l^T x}} - \frac{e^{w_j^T x}}{\sum_{l=1}^k e^{w_l^T x}} \frac{e^{w_j^T x}}{\sum_{l=1}^k e^{w_l^T x}}\right) x = p_j(1 - p_j)x$$

So the first derivative of L is :

$$\nabla_{w_n} L(\cdot) = -(x(\mathbf{1}\{y = n\} - a_n))$$

And the second derivative:

$$\begin{aligned}\nabla^2_{w_n} L(\cdot) &= -\nabla_{w_n}(x(\mathbf{1}\{y = n\} - a_n)) \\ &= a_n(1 - a_n)xx^T\end{aligned}$$

As $x_i x_i^T$ is positive semi-definite and $a_n(1 - a_n) < 0$, the second derivative of L is positive semi-definite. Therefore, L is convex in W