NEW STUFF

Keywords:

- Convexity in Optimization
  (Convex / Strict Convex)
- Hessian, posi-definite matrix
- Derivative of Softmax
- Integral of pmf

CS7643: Deep Learning
Fall 2017
Homework 0

Instructor: Dhruv Batra
TAs: Michael Cogswell, Abhishek Das, Zhaoyang Lv
Discussions: http://piazza.com/gatech/fall2017/cs7643

Due: Thursday, Aug 24, 11:55pm

Instructions

1. Please upload your answer sheet on Canvas with the following format:
   FirstName_LastName_HWx.pdf.
   LaTeX'd solutions are preferred (solution template available at cc.gatech.edu/classes/AY2018/cs7643_fall/assets/sol0.tex), but scanned handwritten copies are acceptable. Hard copies are not accepted.

2. We generally encourage you to collaborate with other students.

   Exception: HW0 is meant to serve as a background preparation test. You must NOT collaborate on HW0.

1 Probability and Statistics

1. (1 point) 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? Are we expected to make money?

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

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

   What is the equation for the corresponding cumulative density function (cdf) $C(x)$?

   [*Hint:* Recall that CDF is defined as $C(x) = Pr(X \leq x)$.]

HW

1) payout = $\begin{cases} \$1 & x = 1 \\ -\$1/4 & x \neq 1 \end{cases}$ $\quad x \in \{1, 1, 2, 3, 4, 5, 6\}$

We will have the expect of the payout to know if this is a good bet or not.

$$E(X) = \sum x_i \cdot p(x_i)$$

X: random variable

x_i: the value $X = x$

$p(x_i)$: the probability $X = x_i$

$$\Rightarrow E(X) = \frac{1}{6}(1) + 5 \cdot \left(\frac{1}{6}(-1/4)\right)$$

$$= -\frac{1}{24} \Rightarrow \text{We will need to give money}$$

(Concrete)

Note: cdf: culmulative density func (both discrete, continuous)

pmf: proba mass func (Discrete)

pdf: proba density func

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

$$\text{cdf} = \begin{cases} \int_0^{1/2} 4x\,dx & 0 \leq x \leq 1/2 \\ \int_{1/2}^{} -4x + 4\,dx & 1/2 \leq x \leq 1 \end{cases} = \begin{cases} 2x^2 & 0 \leq x \leq 1/2 \\ -2x^2 + 4x + c & 1/2 \leq x \leq 1 \end{cases}$$

3. (1 point) Recall that the variance of a random variable is defined as $Var[X] = E[(X-\mu)^2]$, where $\mu = E[X]$. Use the properties of expectation to show that we can rewrite the variance of a random variable X as

$$Var[X] = E[X^2] - (E[X])^2 \tag{3}$$

4. (1 point) A random variable x in standard normal distribution has following probability density

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{4}$$

Evaluate following integral

$$\int_{-\infty}^{\infty} p(x)(ax^2 + bx + c)dx \tag{5}$$

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

2 Proving Stuff

5. (2 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.]

6. (3 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.

(a) Using the results from Q5, show that $KL(p,q)$ is always positive.

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

[*Hint:* This question doesn't require you to know anything more than the definition of $KL(p,q)$ and the identity in Q5]

3) $Var[x] = E[(X-\mu)^2]$ $\quad \mu = E[X]$

Prove: $Var[X] = E[X^2] - (E[X])^2$

We have, $E[(X-\mu)^2]$

$= E[(X-\mu)(X-\mu)] = E[X^2 - 2\mu X + \mu^2]$

$= E[X^2] - 2\mu E[X] + E[\mu^2]$

↳ $E[E[X] \cdot E[X]] = E[X]^2$

$= E[X^2] - 2E[X]^2 + E[X]^2$

$= E[X^2] - E[X]^2$

4) $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ $\quad x \sim N(0,1)$

Evaluate: $\int_{-\infty}^{\infty} p(x)(ax^2+bx+c)\,dx$ (1)

(2) (1) $= E[aX^2 + bX + c] = aE[X^2] + bE[X] + E[c]$

$E[X^2] = Var[X] + (E[X])^2$

$= 1 + 0 = 1$

(2) $aE[X^2] + bE[X] + E[c]$

$= a + c$

Note: $f(x) = ax^2 + bx + c$

⇒ for continuous function:

$E[f(X)] = \int_{-\infty}^{\infty} p(x) \cdot f(x)\,dx$

2. Proving Stuff.

5. $\log_e x \leq x - 1 \quad \forall x \in 0$

$\Leftrightarrow \log_e x - x + 1 \leq 0$ (1)

We have (1): $\log_e x - x + 1$

$(1)' = \frac{1}{x} - 1$ if $\frac{1}{x} - 1 = 0$

$\Leftrightarrow \frac{1}{x} = 1 \Rightarrow x = 1$

We have:

x	$-\infty$		0		1		$+\infty$
$(1)'$		$-$	$+\infty$	$+$	0	$-$	
(1)			$\| -\infty$		0		$\| -\infty$

$\Rightarrow \log_e x - x + 1 \leq 0 \quad \forall x > 0$

$\Rightarrow \log_e x \leq x + 1 \quad \forall x > 0$

6. $KL(p,q) = \sum_{i=1}^{k} p_i \cdot \log\left(\frac{p_i}{q_i}\right) -$

$\Leftrightarrow p_i \cdot \log\left(\frac{p_i}{q_i}\right) = p_i \cdot \log(p_i) - p_i \cdot \log(q_i)$

$= p_i (\log(p_i)) + p_i(-\log(q_i) + q_i - 1 - q_i + 1)$

$= p_i [\log(p_i)] + p_i [-\log(q_i) + q_i - 1] + p_i(q_i + 1)$

Wrong

$x - 1 - \log_e x > 0$

a) $\sum_{i=1}^{k} p_i \cdot \log\left(\frac{p_i}{q_i}\right) = \sum_{i=1}^{k} p_i \cdot -\log\left(\frac{q_i}{p_i}\right)$

(1)

(1) $= p_i\left(-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1 - \frac{q_i}{p_i} + 1\right)$

$= p_i\left(-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1\right) + p_i\left(\frac{-q_i}{p_i} + 1\right)$

$= p_i\left(-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1\right) + (-q_i + p_i)$

$\Rightarrow KL = \sum_{i=1}^{k} p_i\left(-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1\right) + \sum_{i=1}^{k}(-q_i) + \sum_{i=1}^{k}(p_i)$

$\geq 0 \qquad = -1 \qquad = 1$

$\Rightarrow KL = \geq 0 + (-1) + (1)$

$\Rightarrow KL \geq 0$

b) Because $p_i\left(-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1\right)$ (1) ≥ 0

$\Leftrightarrow$ For $KL = 0$; (1) $= 0 \ \forall i$

$\Rightarrow$ 2 cases:

$$\begin{cases} p_i = 0 \ \forall i \\ -\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1 = 0 \quad \forall i \end{cases}$$

As we have proven above: $-\log\left(\frac{q_i}{p_i}\right) + \frac{q_i}{p_i} - 1 = 0$ iff $\frac{q_i}{p_i} = 1$

$\Leftrightarrow p_i = q_i \Rightarrow KL = 0$ when $p_i = q_i \ \forall i$ (1)

$p_i = 0 \ \forall i$ (2)

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

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

(2) $KL[q,p] = \sum_{i=1}^{k} q_i \cdot \log\left(\frac{q_i}{p_i}\right)$

An example: $p_1 = \frac{1}{3}$ $p_2 = \frac{2}{3}$ $q_1 = \frac{1}{4}$ $q_2 = \frac{3}{4}$

$$KL(p,q) = \frac{1}{3} \cdot \log\left(\frac{1}{3} \cdot \frac{4}{1}\right) + \frac{2}{3} \cdot \log\left(\frac{2}{3} \cdot \frac{4}{3}\right)$$

$$= \frac{1}{3} \log\left(\frac{4}{3}\right) + \frac{2}{3} \cdot \log\left(\frac{8}{9}\right) \approx 0,0174$$

$$KL(q,p) = \frac{1}{4} \cdot \log\left(\frac{1}{4} \cdot \frac{3}{1}\right) + \frac{3}{4} \cdot \log\left(\frac{3}{4} \cdot \frac{3}{2}\right)$$

$$= \frac{1}{4} \cdot \log\left(\frac{3}{4}\right) + \frac{3}{4} \log\left(\frac{9}{8}\right) \simeq 0,0164$$

3 Calculus

7. (3 points) Consider the following function of $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$:

$$f(\mathbf{x}) = \sigma\left(\log\left(5\left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6)\right)\right) + \frac{1}{2}\right) \tag{9}$$

where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{10}$$

Evaluate $f(\cdot)$ at $\hat{\mathbf{x}} = (5, -1, 6, 12, 7, -5)$. Then, compute the gradient $\nabla_{\mathbf{x}} f(\cdot)$ and evaluate it at the same point.

4 Softmax Classifier

8. (5 points) Implement a Softmax classifier (from scratch, no ML libraries allowed), and train it (via SGD) on CIFAR-10:
cc.gatech.edu/classes/AY2018/cs7643_fall/hw0-q8/.

9. (3 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

$$\mathbf{z} = W\mathbf{x} + \mathbf{b}, \tag{11}$$

$$p_j = \frac{e^{z_j}}{\sum_k e^{z_k}}, \tag{12}$$

$$L(W) = -\log(p_y) \tag{13}$$

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.]

3. CALCULUS

$$f(x) = 6 \left(\log \left(5 \left(\max\{x_1, x_2\} \cdot \frac{x_3}{x_4} - (x_5 + x_6) \right) + \frac{1}{2} \right) \right)$$

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

$\hat{x} = (5; -1; 6; 12; 7; -5)$ $\quad \nabla_x f(\cdot)$

$f(\hat{x}) \approx 0{,}80\,4756$

Gradient at point $\hat{x}$ $\quad x_1 = 5 \quad x_2 = -1 \Rightarrow \max\{x_1, x_2\} = 5$

$6'(x) = 6(x)(1 - 6(x))$

$\log'(x) = \frac{1}{x}$

4. Softmax Classifier

$$z = Wx + b, \qquad L(W) = -\log(P_y)$$

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

$W \rightarrow (z) \rightarrow P_j \rightarrow L(W)$

$$\frac{dL(W)}{dW} = \frac{dL(W)}{dP_j} \cdot \frac{dP_j}{dz} \cdot \frac{dz}{dW}$$

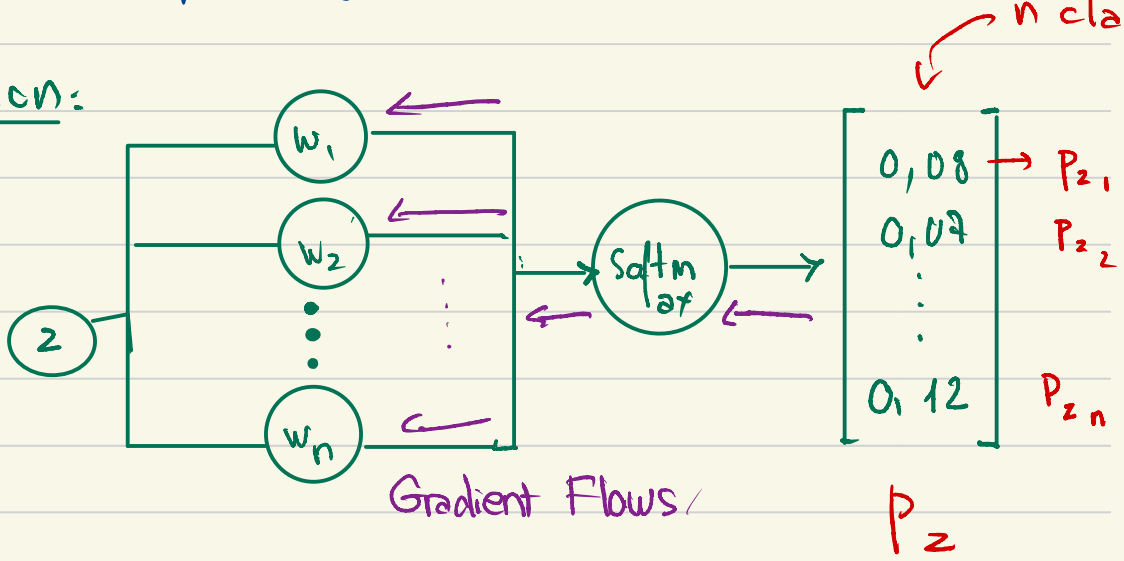
Assume there are n labels:

⇒ $x = [1, m]$ (1 samples m features)

$W = [m, n]$ (From m features to n classes)

z_j : sample z, class i

Visualization:



$$\frac{dL(W)}{dW} = -\frac{d\log(P_y)}{dW} = -\frac{1}{P_y} \cdot \frac{dP_y}{dW}$$

$$\frac{dP_y}{dW} = \frac{dP_{z_i=y}}{dW} =$$

We have: Softmax derivatives

$$\frac{\partial S(z_i)}{\partial z_j} = \begin{cases} -S(z_i)(1 - S(z_i)) & \text{if } j = i \\ -S(z_i) \cdot S(z_j) & \text{if } j \neq i \end{cases}$$