CS365: Deep Learning

Mid Semester, Autumn 2023, 01.10.23

IIT Patna

1. **Attempt as many questions as you can. The maximum marks you can get is 60.**
2. **Do not write anything on the question paper.**
3. **Make necessary assumptions and state these clearly where required.**
4. **No clarification will be provided during the examination.**

Time: 2 Hrs Maximum marks: 60

1. Suppose you are given data $\{x_i, y_i\}_{i=1}^n$ and weights $\{c_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ and $c_i > 0$ for all i. Consider the following optimization problem: $\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n c_i (y_i - x_i^T \theta)^2$. Find a closed form expression for $\hat{\theta}$ assuming necessary matrices are invertible. For example, in linear regression we have the following expression $(X^T X)^{-1} X^T Y$ where X is a $n \times d$ matrix whose ith row is x_i^T and Y is a column vector where ith row is y_i. (8)

$$\hat{\theta} = (X^T C X)^{-1} X^T C Y$$

$$C = \begin{bmatrix} c_1 & & \\ & c_2 & \\ & & \ddots \\ & & & c_n \end{bmatrix}$$

2. (a) You forward propagate a batch of m examples in your network. The input $z = (z_1, \ldots, z_m)$ of the batch normalization layer has shape $(n_h = 3, m = 4)$, where n_h represents the number of neurons in the pre-batchnorm layer:

$$z = \begin{bmatrix} 12 & 14 & 14 & 12 \\ 0 & 10 & 10 & 0 \\ -5 & 5 & 5 & -5 \end{bmatrix}$$

What will be $\hat{z}$ (normalized value)? Express your answer as a matrix with shape 3×4.
(b) Suppose $\gamma = (1, 1, 1)^T$ and $\beta = (0, -10, 10)^T$. What is the final output y where $y = \gamma\hat{z} + \beta$? Express your answer as a matrix with shape 3×4. (2+2)

a)

$$\begin{bmatrix} -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \end{bmatrix}$$

b)

$$\begin{bmatrix} -1 & 1 & 1 & -1 \\ -11 & -9 & -9 & -11 \\ 9 & 11 & 11 & 9 \end{bmatrix}$$

3. Consider the following regression problem. The dataset for the problem is a set of n examples (x_i, y_i) where $i = 1, \ldots, n$ where x_i and y_i are real numbers for all i. The difficulty here is that we do not have access to the inputs or outputs directly. Also, we do not even know the number of examples in the dataset. We are, however, able to get a few numbers computed from the data. Let $\boldsymbol{w}^* = [w_0^*, w_1^*]$ be the least squares solution. Alternatively, we can say $\boldsymbol{w}^*$ minimizes the following

$$J(\boldsymbol{w}) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2$$

For this problem, you can assume that the solution is unique. If $\boldsymbol{w}^*$ is a least square solution then find simplify the following expression $\frac{1}{n} \sum_{i=1}^{n} (y_i - w_0^* - w_1^* x_i)(w_1^* x_i + w_0^*)$. (6)

$= 0$

(A) $\frac{\partial J}{\partial w_0} = \frac{1}{2n} \sum (y_i - w_0 - w_1 x_i) = 0$ for w_0^*, w_1^*

(B) $\frac{\partial J}{\partial w_1} = \frac{1}{2n} \sum (y_i - w_0 - w_1 x_i) x_i = 0$ for $w_0^* w_1^*$

$w_1^* (B) + w_0^* (A) = 0$

4. Let $X_1, \ldots, X_n$ be i.i.d. data from a uniform distribution over the disc of radius $\theta \in \mathbb{R}^2$. Thus, $X_i \in \mathbb{R}^2$ and

$$p(x, \theta) = \begin{cases} \frac{1}{\pi\theta^2} & \text{if } \|x\| \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\|x\| = \sqrt{x_1^2 + x_2^2}$. Please find the maximum likelihood estimate of θ. (4)

$$\hat{\theta} = \max_{1 \leq i \leq n} \|X_i\|$$

5. Consider the following function $f(x,y) = x^2 + 2xy + y^3$. Some one has computed Taylor series expansion of f at $(x_0, y_0) = (1, 2)$ as follows

$$\begin{aligned} f(x,y) &= 13 && (A)\\ &+ 3(x-1) + 7(y-2) && (B)\\ &+ (x-1)^2 + (x-1)(y-2) + 3(y-2)^2 && (C)\\ &+ 2(y-2)^3 && (D)\\ &+ \epsilon(x,y) && (E) \end{aligned}$$

where $|\epsilon(x,y)| > 0$ denotes the error due to higher order terms. Check the validity of each term (A)-(E). If any term is incorrect, please provide the correct expression along with justification. (2+2+2+2+2)

A - correct

B - $6(x-1) + 14(y-2)$

C - $(x-1)^2 + 2(x-1)(y-2) + 6(y-2)^2$

D - $(y-3)^2$

E → 0

6. Babli implements a neural network that has two inputs x_1 and x_2 and a single neuron that performs linear combination of the inputs with the parameters a and b ie., $\hat{y} = ax_1 + bx_2$ where $\hat{y}$ is the estimated value of actual y. The loss function is calculated as $L(y, \hat{y}) = (y - \hat{y})^2$. She applied gradient descent to learn the weight parameters. Afterwards, Babli learns that the true label is $y = x_1 + x_2$.
(a) Suppose a_0 and b_0 are the initial values of the weights, and a_k and b_k are the weights at iteration k. Give equations for the updated weights a_{k+1}, b_{k+1} in terms of current iteration's weights a_k, b_k, the step size parameter η, and the inputs x_1, x_2.
(b) Babli sees that when she fixed $x_1 = 1$, $x_2 = 1$ and ran a few iterations of gradient descent starting with $a_0 = 2$, $b_0 = 2$, she recorded that the two weights oscillated back and forth, as follows (a_k, b_k): $(2, 2), (0, 0), (2, 2), (0, 0), \ldots$. Was this observation by mistake or, if not, what value of η could have generated it?
(c) Babli sees that when she fixed $x_1 = 1, x_2 = 1$ and ran a few iterations of gradient descent starting with $a_0 = 2$, $b_0 = 0$, she recorded that the two weights remained unchanged. Was this observation by mistake or, if not, what value of η could have generated it?
(d) Babli sees that when she fixed x_1 and x_2 and ran a few iterations of gradient descent with $\eta = 0.01$ starting with $a_0 = b_0 = 2$, she recorded that b stayed unchanged, but a decayed to 1 gradually. Was this observation by mistake or, if not, what value of x_1 and x_2 could have generated it? (2+2+2+2)

a) $a_{k+1} = a_k - 2\eta[(a_k - 1)x_1^2 + (b_k - 1)x_1x_2]$

$b_{k+1} = b_k - 2\eta[(b_k - 1)x_2^2 + (a_k - 1)x_1x_2]$

b) $\eta = 1/2$

c) $\eta = 5$

d) $x_1 = 4$ (other non zero +ve values also work)

$x_2 = 0$

7. (a) What is early-stopping approach?
(b) Prove that early-stopping method can act as L_2 regularizer. [No need to derive the expression for L_2 regularization.]
(c) Prove that on an average there are 66% unique elements in a sample of size n when the sample is drawn with replacement policy from a set of n distinct elements. (4+8+8)

8. [**Bonus problem**] A simple perceptron or a neuron may be viewed as a linear separator. Thus, when we have a single neuron it divides the space into two regions. This was discussed in the class. Consider a situation where the input dimension is m and we have n neurons. What is the maximum number of regions that can be generated by n neurons in m dimension space? (10)

Done in the class.

Still a bonus problem!