Quiz for Neural Computation

Due: Optional

Problem 1 (Gradient)

The gradient of a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is denoted by ∇f. Which of the following statements is correct?

(A) The gradient ∇f is a vector with positive elements

(B) The gradient ∇f is a function which maps vectors to vectors

(C) The gradient ∇f is a function which maps vectors to scalars

(D) The gradient ∇f is a vector with negative elements

Solution 1

The true answer is B. According to our definition, the gradient maps a vector to another vector, that is

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \mapsto \nabla f(\mathbf{x}) = \begin{pmatrix} \partial f(\mathbf{x})/\partial x_1 \\ \partial f(\mathbf{x})/\partial x_2 \\ \vdots \\ \partial f(\mathbf{x})/\partial x_d \end{pmatrix}$$

Problem 2 (Minibatch SGD)

Which of the following statement is not true (n is the sample size)?

(A) If the batch size is 1, then minibatch SGD becomes SGD

(B) If the batch size is n, then minibatch SGD (sampling without replacement) becomes gradient descent.

(C) If the batch size is n, then minibatch SGD (sampling with replacement) becomes gradient descent.

(D) As compared to SGD, minibatch SGD builds a better gradient estimator by using more examples.

Solution 2

The true answer is C. Since minibatch SGD can sample a point several times, then minibatch SGD (sampling with replacement) is not gradient descent. For example, if the sample size is $n = 3$. Then it is possible that the batch is z_1, z_1 and z_2. In this case, the stochastic gradient is $\frac{1}{3}(\nabla C_1(\mathbf{w}) + \nabla C_1(\mathbf{w}) + \nabla C_2(\mathbf{w}))$, which is different from the following gradient used in gradient descent

$$\frac{1}{3}(\nabla C_1(\mathbf{w}) + \nabla C_2(\mathbf{w}) + \nabla C_2(\mathbf{w}))$$

Problem 3 (Propagation)

Which of the following statement is not true?

(A) Forward propagation goes from the input layer to the output layer

(B) Backward propagation aims to to compute a gradient of a function

(C) Backward propagation is based on a chain rule

(D) Forward and backward propagation are two independent processes

Solution 3

The true answer is D. Forward propagation can compute the function value of some nodes. These values are used in backward propagation. Therefore, these two processes are not independent.

Problem 4 (Multi-Layer Perceptron)

Consider a fully-connected MLP with 4 layers: 1 input layer, 1 output layer and 2 hidden layers. Assume the input layer has 6 nodes, the two hidden layers have 5 and 10 nodes respectively, and the output layer has 3 nodes. How many trainable parameters are there in this MLP?

$$(A) : 100 \qquad (B) : 128 \qquad (C) : 160 \qquad (D) : 180$$

Solution 4

The true answer is B. According to the definition of MLPs, the trainable parameters include the weights and bias.

- Weight parameters include 3 matrices: $\mathbf{W}^2, \mathbf{W}^3, \mathbf{W}^4$. The size of these matrices are as follows

$$\mathbf{W}^2 \in \mathbb{R}^{6\times 5}, \quad \mathbf{W}^3 \in \mathbb{R}^{5\times 10}, \quad \mathbf{W}^4 \in \mathbb{R}^{10\times 3}$$

- Bias parameters include 3 vectors: $\mathbf{b}^2, \mathbf{b}^3, \mathbf{b}^4$. The size of these matrices are as follows

$$\mathbf{b}^2 \in \mathbb{R}^{5}, \quad \mathbf{b}^3 \in \mathbb{R}^{10}, \quad \mathbf{b}^4 \in \mathbb{R}^{3}$$

Therefore, the total number of parameters are

$$\underbrace{6 * 5 + 5 * 10 + 10 * 3}_{\text{weight parameters}} + \underbrace{5 + 10 + 3}_{\text{bias parameters}} = 110 + 18 = 128$$

Problem 5 (Minimization)

Let a, b, c, d be four numbers. Consider two points $\mathbf{x}_1 = (a, b)^\top$ and $\mathbf{x}_2 = (c, d)^\top$. Consider the following minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^2} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + 2\|\mathbf{x} - \mathbf{x}_2\|_2^2.$$

Which of the following is the minimizer?

$$(A) : \begin{pmatrix} \frac{a}{3} + \frac{c}{3} \\ \frac{b}{3} + \frac{d}{3} \end{pmatrix} \qquad (B) : \begin{pmatrix} \frac{a}{2} + \frac{c}{2} \\ \frac{b}{2} + \frac{d}{2} \end{pmatrix} \qquad (C) : \begin{pmatrix} \frac{a}{3} + \frac{2c}{3} \\ \frac{b}{3} + \frac{2d}{3} \end{pmatrix} \qquad (D) : \begin{pmatrix} \frac{2a}{3} + \frac{c}{3} \\ \frac{2b}{3} + \frac{d}{3} \end{pmatrix}$$

Solution 5

The true answer is C. The gradient of the objective function is

$$\nabla C(\mathbf{x}) = 2(\mathbf{x} - \mathbf{x}_1) + 4(\mathbf{x} - \mathbf{x}_2).$$

By the first-order optimality condition, we have

$$\nabla C(\mathbf{x}^*) = 0 \Longrightarrow 6(\mathbf{x}^*) = 2\mathbf{x}_1 + 4\mathbf{x}_2$$

and therefore

$$\mathbf{x}^* = \frac{1}{3}\mathbf{x}_1 + \frac{2}{3}\mathbf{x}_2 = \begin{pmatrix} \frac{a}{3} + \frac{2c}{3} \\ \frac{b}{3} + \frac{2d}{3} \end{pmatrix}$$

Problem 6 (Gradient Descent)

Consider a binary classification problem with the following training examples

$$\mathbf{x}^{(1)} = \begin{pmatrix} -0.5 \\ 0.25 \\ -0.8 \\ -1 \end{pmatrix} \qquad \mathbf{x}^{(2)} = \begin{pmatrix} -1 \\ -0.1 \\ -0.1 \\ -1 \end{pmatrix} \qquad \mathbf{x}^{(3)} = \begin{pmatrix} 0.5 \\ 0 \\ 0.25 \\ 0.1 \end{pmatrix} \qquad \mathbf{x}^{(4)} = \begin{pmatrix} -0.2 \\ -0.3 \\ 0.2 \\ 0 \end{pmatrix}$$

$$\mathbf{x}^{(5)} = \begin{pmatrix} -0.8 \\ 0 \\ -0.8 \\ -1 \end{pmatrix} \qquad \mathbf{x}^{(6)} = \begin{pmatrix} -0.15 \\ -0.5 \\ 0.05 \\ -0.25 \end{pmatrix} \qquad \mathbf{x}^{(7)} = \begin{pmatrix} -1 \\ 0 \\ -1 \\ -1 \end{pmatrix} \qquad \mathbf{x}^{(8)} = \begin{pmatrix} 0 \\ -0.25 \\ 0.25 \\ 0.1 \end{pmatrix}$$

$$y^{(1)} = 1 \qquad y^{(2)} = 1 \qquad y^{(3)} = -1 \qquad y^{(4)} = -1$$
$$y^{(5)} = 1 \qquad y^{(6)} = -1 \qquad y^{(7)} = 1 \qquad y^{(8)} = -1$$

Suppose we consider a linear model for classification $\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$, where $\mathbf{w} = (w_1, w_2, w_3, w_4)^\top \in \mathbb{R}^4$. We minimize the objective function (for simplicity we do not consider the bias in the linear model)

$$C(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} C_i(\mathbf{w}), \tag{1}$$

where

$$C_i(\mathbf{w}) = \begin{cases} 0, & \text{if } y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)} \geq 1 \\ \frac{1}{2}(1 - y^{(i)} \mathbf{w}^\top \mathbf{x}^{(i)})^2, & \text{otherwise.} \end{cases}$$

Suppose we run gradient descent with $\mathbf{w}^{(0)} = (0,0,0,0)^\top$ and step size $\eta_t = \eta = 0.5$. What is $\mathbf{w}^{(25)}$? (We only preserve three digits after the decimal point)

$$\begin{pmatrix} -0.721 \\ 1.262 \\ -1.204 \\ -0.484 \end{pmatrix} \qquad \begin{pmatrix} -0.527 \\ 1.220 \\ -1.047 \\ -0.445 \end{pmatrix} \qquad \begin{pmatrix} -0.536 \\ 1.260 \\ -1.257 \\ -0.565 \end{pmatrix} \qquad \begin{pmatrix} -0.637 \\ 1.120 \\ -1.056 \\ -0.395 \end{pmatrix}$$

(A) (B) (C) (D)

Solution 6

The true answer is B.

Problem 7 (Stochastic Gradient Descent)

Let us consider the Problem 6, e.g., the same objective function. Suppose we run SGD to minimize Eq. (1). Let $\mathbf{w}^{(0)} = (0, 0, 0, 0)^\top$ and $\eta_t = \eta = 0.1$. Let $i_t = (t \mod 8) + 1$, i.e., $i_0 = 1, i_1 = 2, \ldots, i_7 = 8, i_8 = 1, \ldots$. What is $\mathbf{w}^{(80)}$? (We only preserve three digits after the decimal point)

$$\begin{pmatrix}-0.469\\0.890\\-0.799\\-0.449\end{pmatrix} \quad \begin{pmatrix}-0.480\\0.706\\-0.879\\-0.549\end{pmatrix} \quad \begin{pmatrix}-0.569\\0.990\\-0.579\\-0.479\end{pmatrix} \quad \begin{pmatrix}-0.669\\0.706\\-0.764\\-0.462\end{pmatrix}$$

(A) (B) (C) (D)

Solution 7

The true answer is A.

Problem 8 (Momentum)

Let us consider the Problem 6, e.g., the same objective function. Suppose we run Momentum to minimize Eq. (1). Let $\mathbf{w}^{(0)} = (0, 0, 0, 0)^\top$ and $\eta_t = \eta = 0.5$. Let the parameter α in the Momentum be 0.5. What is $\mathbf{w}^{(25)}$? (We only preserve three digits after the decimal point)

$$\begin{pmatrix}-0.798\\1.939\\-1.697\\-0.408\end{pmatrix} \quad \begin{pmatrix}-0.798\\1.849\\-1.667\\-0.422\end{pmatrix} \quad \begin{pmatrix}-0.698\\1.839\\-1.657\\-0.402\end{pmatrix} \quad \begin{pmatrix}-0.598\\1.829\\-1.557\\-0.502\end{pmatrix}$$

(A) (B) (C) (D)

Solution 8

The true answer is D.

Problem 9 (Adaptive Gradient Descent)

Let us consider the Problem 6, e.g., the same objective function. Suppose we run AdaGrad to minimize Eq. (1). Let $\mathbf{w}^{(0)} = (0, 0, 0, 0)^\top$ and $\eta_t = \eta = 0.1$. Let the parameter δ in the AdaGrad be 10^{-6}. Let $i_t = (t \mod 8) + 1$, i.e., $i_0 = 1, i_1 = 2, \ldots, i_7 = 8, i_8 = 1, \ldots$. What is $\mathbf{w}^{(80)}$? (We only preserve three digits after the decimal point)

$$\begin{pmatrix}-0.478\\0.849\\-0.722\\-0.380\end{pmatrix} \quad \begin{pmatrix}-0.498\\0.858\\-0.612\\-0.360\end{pmatrix} \quad \begin{pmatrix}-0.398\\0.868\\-0.623\\-0.354\end{pmatrix} \quad \begin{pmatrix}-0.698\\0.862\\-0.632\\-0.352\end{pmatrix}$$

(A) (B) (C) (D)

Solution 9

The true answer is A.

Problem 10 (Perceptron)

Let us consider the dataset in Problem 6. Suppose we apply the Perceptron algorithm to find a linear model. Suppose we initialize $\mathbf{w}^{(0)} = (0, 0, 0, 0)^\top$. Suppose we go through the dataset once in this order: $(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots, (\mathbf{x}^{(8)}, y^{(8)})$. What is $\mathbf{w}^{(8)}$?

$$\begin{pmatrix} -0.45 \\ 0.70 \\ -0.82 \\ -0.70 \end{pmatrix} \qquad \begin{pmatrix} -0.40 \\ 0.75 \\ -0.84 \\ -0.70 \end{pmatrix} \qquad \begin{pmatrix} -0.35 \\ 0.75 \\ -0.85 \\ -0.75 \end{pmatrix} \qquad \begin{pmatrix} -0.48 \\ 0.70 \\ -0.75 \\ -0.85 \end{pmatrix}$$

(A) (B) (C) (D)

Solution 10

The true answer is C.