1 Multiple Choice 1 point

$l(y, \hat{y}) = |y - \hat{y}| = |y - w^T x|$

Suppose we are using squared loss l(y, y-hat) = (y - y-hat)^2 in a regression problem with a linear prediction model: y-hat = $\mathbf{w}^T\mathbf{x}$.
What is the derivative of the absolute loss w.r.t. **w** for a single example with feature vector **x** and true label y?

Note: Because the absolute loss is not differentiable at zero, you can assume that y is not equal to y-hat in this problem. Note that the sign function for a real valued input x (where x is not zero) is defined as sign(x) = +1 for x > 0 and sign(x) = -1 for x < 0

- (selected) $\mathbf{x}(\mathbf{w}^T\mathbf{x} - y)$
- $\mathbf{x}(y - \text{softmax}(\mathbf{w}^T\mathbf{x}))$
- $\mathbf{x}(\text{softmax}(\mathbf{w}^T\mathbf{x}) - y)$
- $\mathbf{x}(y - \mathbf{w}^T\mathbf{x})$

$\nabla_w |y - w^T x|$ ← 这个才是正确答案

$= \underbrace{-x}_{\boxed{\nabla_w (y - w^T x)}} \cdot \underbrace{(\text{sign}(w^T x - y))}_{\boxed{\nabla_u |u|}}$

2 Multiple Choice 1 point

Suppose in a classification problem, the true label **y** and the predicted label **y**-hat are both hard labels. Which of the following correctly describes the nature of the cross-entropy loss in this special case?

全是0，有一个1

- It is zero
- It is +infinity
- (selected) It is 0 if **y** is equal to **y**-hat and +infinity otherwise
- It is 0 if **y** is equal to **y**-hat and 1 otherwise

$L = \sum_j -y_j \ln \hat{y}_j$

$y_j = 1, \hat{y}_j \to 0 \Rightarrow -(-\infty) = \infty$

$y_j = 0, \hat{y}_j \to 0 \Rightarrow 0$

如果选择 ← $\underline{y_j = 0, \hat{y}_j = 1 \Rightarrow 0}$

那么 $y_j = 1$ 的地方 $\hat{y}_j = 0 \Rightarrow +\infty$

$y_j = 1, \hat{y}_j = 1 \Rightarrow 0$

因而只要1 label 不相同，一定有 $L = +\infty$；否则 $L = 0$

3 Multiple Choice 1 point

$l(y,\hat{y}) = \sum_j -y_j \ln \hat{y}_j \quad (= -\ln \hat{y}_\alpha,\ \alpha \text{ is true})$

Recall that cross-entropy loss l(**y**, **y**-hat) is defined as the sum of -y_j log(y-hat_j) over j. What is the partial derivative of cross-entropy loss w.r.t. y-hat_j?

- (●) -y_j / y-hat_j
- ( ) softmax(**y**-hat)$_j$ - y_j
- ( ) -y-hat_j / y_j
- ( ) softmax(**y**)$_j$ - **y**-hat_j

$$\frac{\partial l}{\partial \hat{y}_j} = \begin{cases} 0, & y_j = 0 \\ (-\ln \hat{y}_j)' = \frac{-1}{\hat{y}_j} = \boxed{\frac{-y_j}{\hat{y}_j}}, & y_j = 1 \end{cases}$$

$(y_j = 0$ 时也 $= \frac{-y_j}{\hat{y}_j} = 0)$

4 Multiple Choice 1 point

Suppose **y**-hat = softmax(**o**) and I send o_1, the first component of **o**, off to $\boxed{\textbf{+infinity}}$ (plus infinity) while keeping other components unchanged. What happens to **y**-hat_1, the first component of **y**-hat?

- ( ) It goes to -infinity
- ( ) It goes to +infinity
- (●) It goes to 1
- ( ) It goes to 0

$$\hat{y}_1 = \frac{e^{x_1}}{e^{x_1} + e^{x_2} + \dots + e^{x_d}} = 1$$

∞ ∞

5 Multiple Choice 1 point

Suppose **y**-hat = softmax(**o**) and I send o_1, the first component of **o**, off to **-infinity** (minus infinity) while keeping other components unchanged. What happens to **y**-hat$_1$, the first component of **y**-hat?

- It goes to -infinity
- It goes to +infinity
- (selected) It goes to 0
- It goes to 1

6 Multiple Choice 1 point

Suppose I have a uniform distribution on N outcomes. What is the entropy (in bits) of this distribution?

- (selected) $\log_2(N)$
- $\log_2(1/N)$
- $1/N \log_2(N)$
- $1/N \log_2(1/N)$

7 Multiple Choice 1 point

Which of the following is NOT true regarding the relationship between TensorFlow and Keras?

- (selected) TensorFlow offers a more a high level API (application programming interface) than Keras
- First release of Keras was before the first release of Tensorflow
- Keras was originally built on top of Theano, a precursor of TensorFlow
- Keras API occupies a front-and-center place in TensorFlow 2.0

8 Multiple Choice 1 point

What is a feature of TensorFlow that is not all supported by NumPy?

- (selected) Ability to numerically compute gradients of any differentiable expression
- Ability to create vectors and matrices
- Ability to manipulate vectors and matrices
- A vast library of predefined mathematical functions such a sqrt, log, exp, etc.

9 Multiple Choice 1 point

What does the "G" in "GPU" stand for?

- (selected) graphics
- grammar
- gradient
- google

10 Multiple Choice 1 point

The name "Jupyter" includes a references to 3 major programming languages for data science. Which of the following is one of them?

- (●) R
- ( ) Java
- ( ) JavaScript
- ( ) Ruby