

## Tasks

For a training example  $(x_i, y_i)$  with  $K$  classes:

$s = Wx_i = \text{score vector } s_j$  ( $s_j = \text{predicted score for class } j$ )

### Task 1: SVM Loss (35')

- 1) Derive SVM loss  $L_i^{SVM}$  for  $x_i$ ; set  $\Delta$  as the margin hyperparameter. (1')
- 2) Derive the gradient  $\frac{\partial L_i^{SVM}}{\partial s_k}$  for
  - $k = y_i$  (correct class) (5')
  - $k \neq y_i$  (incorrect class) (5')
 And explain:
  - why is  $\Delta$  only applied to incorrect classes? (3')
  - How does  $\Delta$  enforce a "safety margin"? (5')
- 3) Given scores  $s = [3, -1, 4]$  for classes ["cat", "dog", "bird"]; True class  $y_i$  is "cat" (index=0);  $\Delta = 1$ , compute and explain:
  - $L_i^{SVM}$  for each score (3')
  - How does  $L_i^{SVM}$  change if  $\Delta = 2$  (5')
- 4) Using the same score  $s = [3, -1, 4]$  and  $y_i = 0$ :
  - Compute  $\frac{\partial L_i^{SVM}}{\partial s}$  (3')
  - Interpret the gradient: Why are some values positive, negative, or zero? (5')

Please show the step-by-step process in the report for all the subtasks.

$$\begin{aligned}
 Q1/1: L_i^{SVM} &= \sum_{j \neq y_i} \max\{0, s_j - s_{y_i} + \Delta\} \\
 Q1/2/1: L_i^{SVM} &= \sum_{j \neq y_i} \max\{0, s_j - s_{y_i} + \Delta\} \text{ as } k = y_i \\
 \text{therefore } \frac{\partial L_i^{SVM}}{\partial s_k} &= \frac{\partial L_i^{SVM}}{\partial s_{y_i}} = \sum_{j \neq y_i} \frac{\partial \max\{0, s_j - s_{y_i} + \Delta\}}{\partial s_{y_i}} \\
 &= \sum_{j \neq y_i} -\mathbb{I}(s_j - s_{y_i} + \Delta > 0) \text{ as } \begin{cases} \frac{\partial s_j - s_{y_i} + \Delta}{\partial s_{y_i}} = -1 \\ \frac{\partial 0}{\partial s_{y_i}} = 0 \end{cases} \\
 &= - \sum_{j \neq y_i} \mathbb{I}(s_j - s_{y_i} + \Delta > 0), \\
 \text{where } \mathbb{I}(s_j - s_{y_i} + \Delta > 0) &= 1 \text{ if } s_j - s_{y_i} + \Delta > 0
 \end{aligned}$$

Q1/2/2:  $k \neq y_i$  then  $i + \Delta$

$$L_i^{SVM} = \sum_{j \neq y_i, k} \max\{0, s_j - s_{y_i} + \Delta\} + \max\{0, s_k - s_{y_i}\}$$

$$\frac{\partial L_i^{SVM}}{\partial s_k} = \begin{cases} 0, & \text{if } s_k - s_{y_i} + \Delta \leq 0 \\ 1, & \text{if } s_k - s_{y_i} + \Delta > 0 \end{cases}$$

as the front term is constant for  $s_k$

$$\text{Therefore } \frac{\partial L_i^{SVM}}{\partial s_k} = \mathbb{I}(s_k - s_{y_i} + \Delta > 0)$$

Q1/2/3: the  $\Delta$  is used for determine the decision boundary, when the predict class is true, which means  $s_i = s_{y_i}$ , the loss is 0, so it will not apply  $\Delta$  to make loss  $> 0$

Q1/2/4:  $\Delta$  sets a minimum separation in score space, so the Loss = 0 if  $s_j - s_{y_i} + \Delta \leq 0$ ; which means the correct class score have at least  $\Delta$  margin with predict score, which avoid noisy point.

$$\begin{aligned} Q1/3/1: L_i^{SVM} &= \max\{0, s_{dog} - s_{cat} + \Delta\} + \max\{0, s_{bird} - s_{cat} + \Delta\} \\ &= \max\{0, -1 - 3 + 1\} + \max\{0, 4 - 3 + 1\} \\ &= 2 \end{aligned}$$

$$Q_1/3/2 : \mathcal{L}_i^{SVM} = \max\{0, -1-3+2\} + \max\{0, 4-3+2\} \\ = 3$$

$$Q_1/4/1 : \bar{j}=1 (\text{dog}) : -1-3+1 \leq 0$$

from  $Q_1/2/2$  we know its gradient is 0

$$\bar{j}=2 (\text{bird}) : 4-3+1=2 > 0 \text{ so gradient is } 1$$

as  $\bar{j}=0$  we use  $Q_1/2/1$  know gradient is  $-1$ .

$$\text{so } \frac{\partial \mathcal{L}_i^{SVM}}{\partial S} = [-1, 0, 1]$$

$$Q_1/4/2 : \frac{\partial \mathcal{L}_i^{SVM}}{\partial S_p} > 0 \text{ if } S_p - S_{y_i} + \Delta > 0.$$

$$< 0 \text{ if } S_p = S_{y_i}$$

$$= 0 \text{ if } S_p - S_{y_i} + \Delta \leq 0.$$

**Task 2: Softmax Loss (35')**

1) Derive

- softmax probability  $p_j$  for class  $j$  (1')
- softmax loss  $L_i^{softmax}$  (1')

2) Derive the gradient  $\frac{\partial L_i^{softmax}}{\partial s_k}$  for

- $k = y_i$  (correct class) (5')
- $k \neq y_i$  (incorrect class) (5')

And explain:

- Why does minimizing  $L_i^{softmax}$  force  $p_{y_i} \rightarrow 1$ ? (3')

3) Given scores  $s = [3, -1, 4]$  for classes ["cat", "dog", "bird"]; True class  $y_i$  is "cat" (index=0), compute and explain:

- $p_j$  (3')
- $L_i^{softmax}$  for the given scores (3')
- What happens to  $L_i^{softmax}$  if all scores are scaled by 2 (i.e.,  $s_{new} = 2s$ ) (4')

4) Using the same score  $s = [3, -1, 4]$  and  $y_i = 0$ :

- Compute  $\frac{\partial L_i^{softmax}}{\partial s}$  (3')
- How does the gradient for the correct class differ from SVM? (7')

Please show the step-by-step process in the report for all the subtasks.

$$Q2/1/1: p_j = \frac{e^{s_j}}{\sum_{m=1}^K e^{s_m}}$$

$$Q2/1/2: L_i^{softmax} = -\log p_{y_i} = -\log\left(\frac{e^{s_{y_i}}}{\sum_{m=1}^K e^{s_m}}\right)$$

$$Q2/2/1: k = y_i \text{ then}$$

$$\frac{\partial L_i^{softmax}}{\partial s_k} = \frac{\partial (-\log p_{y_i})}{\partial s_k} = \frac{\partial (-\log \frac{e^{s_k}}{\sum_{m=1}^K e^{s_m}})}{\partial s_k} = -1 + \frac{\frac{\partial \sum_{m=1}^K e^{s_m}}{\partial s_k}}{\sum_{m=1}^K e^{s_m}}$$

$$Q2/2/2: k \neq y_i \text{ then}$$

$$\frac{\partial L_i^{softmax}}{\partial s_k} = \frac{\partial (-\log \frac{e^{s_k}}{\sum_{m=1}^K e^{s_m}})}{\partial s_k} = \frac{e^{s_k}}{\sum_{m=1}^K e^{s_m}}$$

Q2/2/3: Loss =  $-\log P_{y_i}$  to argumin  $-\log P_{y_i}$  we need.  
 calculate  $\operatorname{argmax} \log P_{y_i} \Rightarrow \operatorname{argmax} P_{y_i}$  as  $0 \leq P_{y_i} \leq 1$   
 so force  $P_{y_i} \rightarrow 1$

$$Q2/3/1: P_{cat} = \frac{e^3}{e^3 + e^{-1} + e^4} = 0.268$$

$$P_{dog} = \frac{e^{-1}}{e^3 + e^{-1} + e^4} = 0.004$$

$$P_{bird} = \frac{e^4}{e^3 + e^{-1} + e^4} = 0.728$$

$$P_{\hat{y}} = [0.268, 0.004, 0.728]$$

$$Q2/3/2: L = -\log P_{cat} = -\log 0.268 = 1.317.$$

$$Q2/3/3: P'_{cat} = \frac{e^6}{e^6 + e^{-2} + e^8} = 0.1188$$

$$P'_{dog} = \frac{e^{-2}}{e^6 + e^{-2} + e^8} = 0.0004$$

$$P'_{bird} = \frac{e^8}{e^6 + e^{-2} + e^8} = 0.8808$$

$$L_{new} = -\log 0.1188 = 2.249.$$

The loss increase, as the best score is in birds, which make probability get down

Q<sub>2</sub>/4/1: from Q<sub>2</sub>/1/1 and Q<sub>2</sub>/1/2 we know

$$\frac{\partial L}{\partial S} = [P_{cat} - 1, P_{dog}, P_{bird}] = [-0.732, 0.004, 0.728]$$

Q<sub>2</sub>/4/2: The SVM is  $[-1, 0, 1]$

- ① The softmax have range in  $[-1, 1]$ . in gradient, well SVM is not, condition to  $\mathbb{R}$ .
- ② SVM gradient is accounted discretely, with solid size of gradient, well softmax presents the probability

**Task 3: Comparative Analysis (30')**

- 1) Using the same score  $s = [3, -1, 4]$  and  $y_i = 0$ :
  - Compare the previous derived losses  $L_i^{SVM}$  and  $L_i^{softmax}$ , which is larger? (2')
  - Why? (3')
- 2) Using the gradients from task 1 and task 2:
  - How does SVM penalize "near-miss" errors (e.g.,  $s_j \approx s_{y_i}$ ) vs. "large-margin" errors? (2')
  - How does Softmax adjust probabilities for low-confidence predictions? (3')
- 3) Why is SVM loss called "max-margin" and Softmax "cross-entropy"? Please use your own words to define them and show your understanding of the definition and meaning of them. (8')
- 4) Compare between SVM loss and Softmax loss, which is better and why? (in this case, please not just focus on previous tasks and show your understanding of the two losses) (12')

$$Q3/1/1: L_i^{SVM} = 2 > L_i^{softmax} = 1.317$$

Q3/1/2: As bird score higher than cat, the  $\Delta = 1$  in this case  
 so  $4 - 3 + 1 = 2$  and softmax use  $-\log P_i$  to calculate if  $\Delta < 0.317$ .  
 the conclusion will change.

Q3/2/1:

- ~ for near-miss: the gradient for  $\xi$  is 1, the error is similar to 0
- for large-margin: also the gradient is 1, but loss is  $\xi - s_{y_i} + \Delta \gg 0$

Q3/2/2:

for true prediction its  $P_{y_i} - 1$ ; for false is  $P_{y_i}$   
 when low confidence, the high prob ~~are~~ have big gradient and the  
 small prob have small gradient.

Q3/3: SVM is maximize the true and false margin to optimize  
 $s_{y_i} \geq s_j + \Delta$

Softmax use Cross Entropy to predict the probability to true distribution

Q3/4: if data have noise  $\rightarrow$  SVM for stability  
 if need probability and soft prediction  $\rightarrow$  softmax