CrossEntropyLoss, KL Divergence, Entropy, and Logits

March 21, 2025

1 Introduction

This document explores CrossEntropyLoss, Kullback-Leibler (KL) Divergence, entropy $H(P)$, and the role of logits in machine learning, with detailed examples for two-class and three-class classification.

2 CrossEntropyLoss and KL Divergence

CrossEntropyLoss measures the difference between a true distribution P and a predicted distribution Q, while KL Divergence quantifies how Q diverges from P:

$$H(P, Q) = H(P) + D_{KL}(P||Q).$$

2.1 Two-Class Example

True label: class 1, $P = [0, 1]$, predicted $Q = [0.3, 0.7]$.

2.1.1 CrossEntropyLoss

$$H(P, Q) = -\sum_i P(i) \log(Q(i)) = -[0 \cdot \log(0.3) + 1 \cdot \log(0.7)] = -\log(0.7) \approx 0.3567$$

2.1.2 KL Divergence

$$D_{KL}(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right) = 1 \cdot \log\left(\frac{1}{0.7}\right) = 0.3567$$

Since $H(P) = 0$, $H(P, Q) = D_{KL}(P||Q)$.

3 Entropy $H(P)$

Entropy measures uncertainty:

$$H(P) = -\sum_i P(i)\log(P(i))$$

3.1 Two-Class Example

For $P = [0, 1]$:

$$H(P) = -[0 \cdot \log(0) + 1 \cdot \log(1)] = 0$$

(Convention: $0\log(0) = 0$).

For $P = [0.5, 0.5]$:

$$H(P) = -2 \cdot (0.5\log(0.5)) \approx 0.6931$$

3.2 Three-Class Example

For $P = [0, 1, 0]$:

$$H(P) = -[0 \cdot \log(0) + 1 \cdot \log(1) + 0 \cdot \log(0)] = 0$$

For $P = [0.33, 0.33, 0.33]$:

$$H(P) = -3 \cdot (0.33\log(0.33)) \approx 1.0974$$

4 Role of Logits

Logits are raw, unnormalized scores from a model, converted to probabilities via softmax:

$$Q(i) = \frac{e^{z_i}}{\sum_j e^{z_j}}.$$

They are fed directly into CrossEntropyLoss for stability and optimization.

4.1 Two-Class Logit Example

Logits: $z = [1.0, 2.0]$, true label: class 1 ($P = [0, 1]$).

4.1.1 Softmax

$$e^{1.0} \approx 2.718, \quad e^{2.0} \approx 7.389$$

$$\text{Sum} = 2.718 + 7.389 = 10.107$$

$$Q(0) = \frac{2.718}{10.107} \approx 0.269, \quad Q(1) = \frac{7.389}{10.107} \approx 0.731$$

$$Q = [0.269, 0.731]$$

4.1.2 CrossEntropyLoss

$$\log(Q(1)) = 2.0 - \log(10.107) \approx 2.0 - 2.313 = -0.313$$

$$\text{Loss} = -\log(Q(1)) \approx 0.313$$

If $z = [-1.0, 1.0]$:

$$e^{-1.0} \approx 0.368, \quad e^{1.0} \approx 2.718$$

$$\text{Sum} = 0.368 + 2.718 = 3.086$$

$$Q = [0.119, 0.881], \quad \text{Loss} = -\log(0.881) \approx 0.127$$

Lower loss reflects higher confidence in class 1.

4.2 Three-Class Logit Example

Logits: $z = [2.0, 1.0, -1.0]$, true label: class 1 ($P = [0, 1, 0]$).

4.2.1 Softmax

$$e^{2.0} \approx 7.389, \quad e^{1.0} \approx 2.718, \quad e^{-1.0} \approx 0.368$$

$$\text{Sum} = 10.475$$

$$Q = [0.705, 0.259, 0.035]$$

4.2.2 CrossEntropyLoss

$$\log(Q(1)) = 1.0 - \log(10.475) \approx -1.349$$

$$\text{Loss} = -\log(Q(1)) \approx 1.349$$

If $z = [0.5, 2.5, -0.5]$:

$$e^{0.5} \approx 1.649, \quad e^{2.5} \approx 12.182, \quad e^{-0.5} \approx 0.607$$

$$\text{Sum} = 14.438$$

$$Q = [0.114, 0.844, 0.042], \quad \text{Loss} = -\log(0.844) \approx 0.169$$

Higher logit for class 1 reduces the loss.

5 Conclusion

For one-hot P, $H(P) = 0$, making CrossEntropyLoss equal to KL Divergence. Logits enable efficient loss computation, with their values directly influencing prediction confidence and loss magnitude.