

Exer 12 MI

Higor Gabriel de Freitas

October 2025

Feito e formatado com ajuda do ChatGPT

1 Definições Iniciais

Para uma única amostra com C classes:

- **Logits (Entradas da Camada de Saída):** z_k , onde $k \in \{1, \dots, C\}$.
- **Saída Prevista (Softmax):** $\hat{y}_k = \text{Softmax}(z_k)$.
- **Rótulo Verdadeiro (One-Hot):** $y_k \in \{0, 1\}$.

1.1 Função de Perda de Entropia Cruzada

A perda L para uma amostra é dada por:

$$L = - \sum_{k=1}^C y_k \log(\hat{y}_k)$$

1.2 Função de Ativação Softmax

A probabilidade prevista \hat{y}_k é:

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{j=1}^C e^{z_j}}$$

2 Cálculo do Gradiente na Camada de Saída

Usando a Regra da Cadeia:

$$\frac{\partial L}{\partial z_k} = \sum_{j=1}^C \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_k}$$

2.1 Passo 1: Derivada da Perda em relação à Saída

$$\frac{\partial L}{\partial \hat{y}_j} = \frac{\partial}{\partial \hat{y}_j} \left(- \sum_{i=1}^C y_i \log(\hat{y}_i) \right) = -y_j \frac{1}{\hat{y}_j}$$

2.2 Passo 2: Derivada da Softmax em relação ao Logit

Esta derivada tem dois casos:

Caso 1: $j = k$ (Derivada em relação a si mesma)

$$\frac{\partial \hat{y}_k}{\partial z_k} = \hat{y}_k (1 - \hat{y}_k)$$

Caso 2: $j \neq k$ (Derivada em relação a outro logit)

$$\frac{\partial \hat{y}_j}{\partial z_k} = -\hat{y}_j \hat{y}_k$$

2.3 Passo 3: Combinando e Simplificando

Substituindo os passos 1 e 2 na Regra da Cadeia:

$$\begin{aligned} \frac{\partial L}{\partial z_k} &= \frac{\partial L}{\partial \hat{y}_k} \frac{\partial \hat{y}_k}{\partial z_k} + \sum_{j \neq k} \frac{\partial L}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial z_k} \\ &= \left(-y_k \frac{1}{\hat{y}_k} \right) (\hat{y}_k (1 - \hat{y}_k)) + \sum_{j \neq k} \left(-y_j \frac{1}{\hat{y}_j} \right) (-\hat{y}_j \hat{y}_k) \\ &= -y_k (1 - \hat{y}_k) + \sum_{j \neq k} y_j \hat{y}_k \\ &= -y_k + y_k \hat{y}_k + \sum_{j \neq k} y_j \hat{y}_k \\ &= -y_k + \hat{y}_k \left(y_k + \sum_{j \neq k} y_j \right) \end{aligned}$$

Como o vetor y é One-Hot, $\left(y_k + \sum_{j \neq k} y_j \right) = \sum_{j=1}^C y_j = 1$.

3 Resultado Final (Gradiente na Camada de Saída)

A expressão do gradiente para o logit z_k é:

$$\frac{\partial L}{\partial z_k} = \hat{y}_k - y_k$$

Em notação vetorial, o erro de retropropagação δ na camada de saída é:

$$\delta_{saída} = \frac{\partial L}{\partial \mathbf{z}} = \hat{\mathbf{y}} - \mathbf{y}$$

4 Atualização dos Pesos da Camada de Saída

Usando este resultado, os gradientes para os pesos $W^{(L)}$ e viéses $b^{(L)}$ da camada de saída L são dados por:

Gradiente do Peso W_{kl} (conexão da ativação a_l para o logit z_k):

$$\frac{\partial L}{\partial W_{kl}} = \frac{\partial L}{\partial z_k} \cdot \frac{\partial z_k}{\partial W_{kl}} = (\hat{y}_k - y_k) \cdot a_l$$

Gradiente do Viés b_k :

$$\frac{\partial L}{\partial b_k} = \frac{\partial L}{\partial z_k} = \hat{y}_k - y_k$$