Three layer nn

Yurii Mykhalchuk

June 2018

1 Forward pass

Scalar form for single input:

- $a_j^1 = \sum_{i=1}^{D} w_{ji}^1 \cdot x_i + b_j^1, j \in \{1..M\}$
  $z_j^1 = tanh(a_j^1), j \in \{1..M\}$
- $a_j^2 = \sum_{i=1}^{M} w_{ji}^2 \cdot z_i^1 + b_j^2, j \in \{1..L\}$
  $z_j^2 = \begin{cases} ReLU(a_j^2) + w_j^s x_j & \text{if } j =< D \\ ReLU(a_j^2) & \text{if } j > D \end{cases} j \in \{1..L\}$
- $a_j^3 = \sum_{i=1}^{L} w_{ji}^3 \cdot z_i^2 + b_j^3, j \in \{1..P\}$
  $y_j = softmax(a_j^3), j \in \{1..P\}$

Scalar form for mini-batch input of size B:

- $a_{jk}^1 = \sum_{i=1}^{D} w_{ji}^1 \cdot x_{ik} + b_j^1, j \in \{1..M\}, k \in \{1..B\}$
  $z_{jk}^1 = tanh(a_{jk}^1), j \in \{1..M\}, k \in \{1..B\}$
- $a_{jk}^2 = \sum_{i=1}^{M} w_{ji}^2 \cdot z_{ik}^1 + b_j^2, j \in \{1..L\}, k \in \{1..B\}$
  $z_{jk}^2 = \begin{cases} ReLU(a_{jk}^2) + w_j^s x_{jk} & \text{if } j =< D \\ ReLU(a_{jk}^2) & \text{if } j > D \end{cases} j \in \{1..L\}, k \in \{1..B\}$
- $a_{jk}^3 = \sum_{i=1}^{L} w_{ji}^3 \cdot z_{ik}^2 + b_j^3, j \in \{1..P\}, k \in \{1..B\}$
  $y_{jk} = softmax(a_{jk}^3), j \in \{1..P\}, k \in \{1..B\}$

Vector form for single input:

- $A_1 = W_1 X + b_1$
  $Z_1 = tanh(A_1)$

$$\text{Sizes: } \begin{cases} X = D \times 1 \\ W_1 = M \times D \\ b_1 = M \times 1 \\ A_1 = M \times 1 \\ Z_1 = M \times 1 \end{cases}$$

- $A_2 = W_2 Z_1 + b_2$
  $Z_2 = ReLU(A_2) + W_s X$
  $$\text{Sizes: } \begin{cases} W_2 = L \times M \\ b_2 = L \times 1 \\ A_2 = L \times 1 \\ W_s = L \times D \\ Z_2 = L \times 1 \end{cases}$$

- $A_3 = W_3 Z_2 + b_3$
  $Y = softmax(A_3)$
  $$\text{Sizes: } \begin{cases} W_3 = P \times L \\ b_3 = P \times 1 \\ A_3 = P \times 1 \\ Y = P \times 1 \end{cases}$$

Vector form for minibatch input of size B:

- $A_1 = W_1 X + b_1$
  $Z_1 = tanh(A_1)$
  $$\text{Sizes: } \begin{cases} X = D \times B \\ W_1 = M \times D \\ b_1 = M \times B \\ A_1 = M \times B \\ Z_1 = M \times B \end{cases}$$

- $A_2 = W_2 Z_1 + b_2$
  $Z_2 = ReLU(A_2) + W_s X$
  $$\text{Sizes: } \begin{cases} W_2 = L \times M \\ b_2 = L \times B \\ A_2 = L \times B \\ W_s = L \times D \\ Z_2 = L \times B \end{cases}$$

- $A_3 = W_3 Z_2 + b_3$
  $Y = softmax(A_3)$

$$\text{Sizes: } \begin{cases} Z_2 = L \times 1 \\ W_3 = P \times L \\ b_3 = P \times B \\ A_3 = P \times B \\ Y = P \times B \end{cases}$$

One-row equation:
$Y = softmax(W_3(ReLU(W_2 tanh(W_1 X + b1) + b2) + W_s X))$

2 Backpropagation

The last layer:
Let $\hat{Y}$ be a one-hot encoded vector with 1 in the place of the true class of an input and 0 otherwise. $Loss(Y, \hat{Y}) = -\sum_{j=1}^{P} \hat{y}_j \cdot log(y_j) = -log(y_j)$ because $\hat{y}_j$ is 0 for all classes except one which is a true class of the input

$$\delta_i^3 = \frac{\partial Loss(Y, \hat{Y})}{\partial a_i^3} = \frac{\partial Loss(Y, \hat{Y})}{\partial y_i} \cdot \frac{\partial softmax(A_3)}{\partial a_i^3} \quad i \in \{1..P\}$$

$$\frac{\partial Loss(Y, \hat{Y})}{\partial y_j} = \frac{\partial(-\sum_{j=1}^{P} \hat{y}_j \cdot log(y_j))}{\partial y_j} = -\frac{\hat{y}_j}{y_j} \quad j \in \{1..P\}$$

$$\frac{\partial softmax(A3)}{\partial a_j^3} = \frac{\partial \frac{e^{a_i^3}}{\sum_{l=1}^{L} e^{a_l^3}}}{\partial a_j^3} \quad i, j \in \{1..P\}$$

if i = j

$$\frac{\partial \frac{e^{a_i^3}}{\sum_{l=1}^{L} e^{a_l^3}}}{\partial a_j^3} = \frac{e^{a_i^3} \sum_{l=1}^{L} e^{a_l^3} - e^{a_j^3} e^{a_i^3}}{(\sum_{l=1}^{L} e^{a_l^3})^2} = \frac{e^{a_i^3}}{(\sum_{l=1}^{L} e^{a_l^3})} \cdot \frac{\sum_{l=1}^{L} e^{a_l^3} - e^{a_j^3}}{(\sum_{l=1}^{L} e^{a_l^3})} = softmax(a_i^3)(1 - softmax(a_j^3))$$

if i $\neq$ j

$$\frac{\partial \frac{e^{a_i^3}}{\sum_{l=1}^{L} e^{a_l^3}}}{\partial a_j^3} = \frac{0 \cdot \sum_{l=1}^{L} e^{a_l^3} - e^{a_i^3} e^{a_j^3}}{(\sum_{l=1}^{L} e^{a_l^3})^2} = -\frac{e^{a_i^3}}{(\sum_{l=1}^{L} e^{a_l^3})^2} \cdot \frac{e^{a_j^3}}{(\sum_{l=1}^{L} e^{a_l^3})^2} = -softmax(a_i^3) softmax(a_j^3)$$

We can write above expressions using Kroneker's delta:

$$\delta_{ij} = \begin{cases} 1 & if\, i = j \\ 0 & if\, i \neq j \end{cases}$$

$$\frac{\partial softmax(a_j^3)}{\partial a_j^3} = softmax(a_i^3)(\delta_{ij} - softmax(a_j^3))$$

Using all that derived equations:

$$\delta_i^3 = \frac{\hat{y}_i}{y_i} \cdot softmax(a_i^3)(\delta_{ij} - softmax(a_j^3)) \quad i, j \in \{1..P\}$$

$$\Delta w_{ij}^3 = \delta_i^3 \cdot \frac{\partial a_j^3}{\partial w_{ij}^3} = \delta_i^3 \cdot \frac{\partial(\sum_{j=1}^{L} w_{ij}^3 \cdot z_j^2 + b_i^3)}{\partial w_{ij}^3} = \delta_i^3 \cdot z_j^2 \quad i \in \{1..P\}, j \in \{1..L\}$$

Second layer:

$$\delta_k^2 = \delta_i^3 \cdot \frac{\partial(\sum_{j=1}^{L} w_{ij}^3 \cdot z_j^2 + b_i^3)}{\partial z_k^2} \cdot \frac{\partial(ReLU(a_k^2) + w_k^s x_k)}{\partial a_k^2} \quad i \in 1..P, k \in 1..L$$

$$\frac{\partial(\sum_{j=1}^{L} w_{ij}^3 \cdot z_j^2 + b_i^3)}{\partial z_k^2} = \sum_{m=1}^{P} w_{mk}^3$$

$$\frac{\partial(ReLU(a_k^2) + w_k^s x_k)}{\partial a_k^2} = \begin{cases} 1 & if a_k >= 0 \\ 0 & if a_k <= 0 \end{cases}$$

$$\Delta w_{ij}^2 = \delta_i^2 \cdot \frac{\partial a_k^2}{\partial w_{ij}^2} = \delta_i^2 \cdot \frac{\partial(\sum_{k=1}^{M} w_{ik}^2 \cdot z_j^1 + b_i^2)}{\partial w_{ij}^2} = \delta_i^2 \cdot z_j^1 \quad i \in \{1..L\}, j \in \{1..M\}, k \in \{1..L\}$$

First layer:

$$\delta_l^1 = \delta_i^2 \cdot \frac{\partial(\sum_{j=1}^{D} w_{ij}^1 \cdot x_i + b_j^1)}{\partial w_{ij}} \cdot \frac{\partial tanh(a_k^1)}{\partial a_k^1} \quad i \in \{1..L\}$$

I am too weak to finish this...