CS 584-A: Derivative of Cross-Entropy Softmax

Jeremiah McReynolds

February 27, 2021

1 Derivative of Objective Function

Objective Function: $J(w;X,y) = -\frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{k} y_{ik}\log\left(\frac{\exp(f_k)}{\sum_c \exp(f_c)}\right) + \lambda\|w\|_2^2$.

Gradient: $\frac{\delta J}{\delta w_k} = \frac{1}{N}\sum_{i=1}^{n}\left(\frac{\exp(f_k)}{\sum_c \exp(f_c)} - y_i\right)x_i + 2\lambda w_k$

1.1 Softmax Derivative

$$\sigma_k = \frac{\exp(f_k)}{\sum_c \exp(f_c)}$$

$\frac{\delta \sigma_k}{\delta w_k}, i == k$

$$\begin{aligned}
\frac{\delta \sigma_k}{\delta w_k} &= \frac{\exp(f_k)\sum_c \exp(f_c) - \exp(f_k)\exp(f_k)}{[\sum_c \exp(f_c)]^2} \\
&= \frac{\exp(f_k)(\sum_c \exp(f_c) - \exp(f_k))}{[\sum_c \exp(f_c)]^2} \\
&= \frac{\exp(f_k)}{\sum_c \exp(f_c)}\frac{\sum_c \exp(f_c) - \exp(f_k)}{\sum_c \exp(f_c)} \\
&= \sigma_k\left(\frac{\sum_c \exp(f_c)}{\sum_c \exp(f_c)} - \frac{\exp(f_k)}{\sum_c \exp(f_c)}\right) \\
&= \sigma_k(1 - \sigma_k)
\end{aligned}$$

$\frac{\delta \sigma_k}{\delta w_k}$, i != k

$$\begin{aligned}
\frac{\delta \sigma_k}{\delta w_k} &= \frac{0 - \exp(f_i)\exp(f_k)}{[\sum_c \exp(f_c)]^2} \\
&= \frac{-\exp(f_i)\exp(f_k)}{[\sum_c \exp(f_c)]^2} \\
&= \frac{\exp(f_i)}{\sum_c \exp(f_c)}\frac{\exp(f_k)}{\sum_c \exp(f_c)} \\
&= -\sigma_i\sigma_k
\end{aligned}$$

$\begin{cases} i == k, \sigma_k(1-\sigma_k) \\ i != k, -\sigma_i\sigma_k \end{cases}$

1.2 Cross-Entropy Loss Derivative

$L(w; X, y) = -\sum_{k=1}^{k} y_{ik} log(\sigma_k) + \lambda||w||_2^2$

$$
\begin{gathered}
= -\sum_{k=1}^{k} y_{ik} \frac{\delta log(\sigma_k)}{\delta w_k} + \frac{\delta \lambda||w||_2^2}{\delta w_k} \\
= -\sum_{k=1}^{k} y_{ik} \frac{1}{\sigma_k} \frac{\delta \sigma_k}{\delta w_k} \frac{\delta w_k x_i}{\delta w_k} + 2\lambda w_k \\
= -y_i(1 - \sigma_i)x_i - \sum_{i!=k} y_k \frac{1}{\sigma_k}(-\sigma_i \sigma_k)x_i \\
= -y_i x_i + y_i \sigma_i x_i + \sum_{i!=k} y_k \sigma_i x_i \\
= y_i \sigma_i x_i + \sum_{i!=k} y_k \sigma_i x_i - y_i x_i \\
= \sigma_i x_i (y_i + \sum_{i!=k} y_k) - y_i x_i \\
= \sigma_i x_i - y_i x_i \\
\frac{\delta L(w; X, y)}{\delta w_k} = (\sigma_i - y_i)x_i + 2\lambda w_k
\end{gathered}
$$

1.3 Gradient

$J(w; X, y) = -\frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{k} y_{ik} \log\left(\frac{\exp(f_k)}{\sum_c \exp(f_c)}\right) + \lambda\|w\|_2^2$

$$
\begin{gathered}
J(w; X, y) = \frac{1}{N} \sum_{i=1}^{n} L(w; X, y) \\
\frac{\delta J}{\delta w_k} = \frac{1}{N} \sum_{i=1}^{n} \frac{\delta L(w; X, y)}{\delta w_k} \\
\frac{\delta J}{\delta w_k} = \frac{1}{N} \sum_{i=1}^{n} (\sigma_i - y_i)x_i + 2\lambda w_k
\end{gathered}
$$