Introduction to Deep Learning for Computer Vision Assignment 3: Simple Linear Classifier II

Paul Molina-Plant

January 28, 2019

Abstract

TODO

1 Derivation of the gradient for cross entropy

1.1 The cross entropy function

$$L_i = -log\frac{\sum_j t_{ij}e^{s_{ij}}}{\sum_k e^{s_{ik}}} \tag{1}$$

I assume log base e for (1).

$$t_{ij} = \delta_{ij} = \begin{cases} 1 & j = y_i \\ 0 & j \neq y_i \end{cases} \tag{2}$$

Softmax P_{ij}

$$P_{ij} = \frac{e^{s_{ij}}}{\sum_k e^{s_{ik}}} \tag{3}$$

Among the k classes, $j = y_i$ for exactly one of them. Therefore, we can ignore the other $k-1$ terms in the sum of (1).

$$L_i = -ln\frac{e^{s_{ij}}}{\sum_k e^{s_{ik}}} = -lnP_{ij} \tag{4}$$

1.2 Gradient of W

Computing the gradient $\frac{\partial L_i}{\partial W}$ via chain rule.

$$\frac{\partial L_i}{\partial W} = \frac{\partial L_i}{\partial P_{ij}}\frac{\partial P_{ij}}{\partial s_{ij}}\frac{\partial s_{ij}}{\partial W}$$

Linear classifier $f(x_i, W) = s_{ij}$.

$$s_{ij} = Wx_i + b_i \tag{5}$$

The partial derivative of s_{ij} wrt W is nonzero when $j = y_i$ and zero everywhere else. This condition is expressed with δ_{ij} (2).

$$\frac{\partial s_{ij}}{\partial W} = \delta_{ij}x_i$$

The partial derivative of P_{ij} (3) wrt W by applying the chain rule.

$$\begin{aligned}\frac{\partial P_{ij}}{\partial W} &= \frac{\partial}{\partial W}\frac{e^{s_{ij}}}{\sum_k e^{s_{ik}}} = \frac{\delta_{ij}x_i e^{s_{ij}} - e^{s_{ij}}x_i e^{s_{ij}}}{(\sum_k e^{s_{ik}})^2} \\ &= \left(\frac{e^{s_{ij}}}{\sum_k e^{s_{ik}}}\right)\left(\frac{\delta_{ij}x_i\sum_k e^{s_{ik}} - x_i e^{s_{ij}}}{\sum_k e^{s_{ik}}}\right) \\ &= P_{ij}(\delta_{ij}x_i - P_{ij}x_i) \\ &= P_{ij}x_i(\delta_{ij} - P_{ij})\end{aligned}$$

Finally the partial derivative of L_i (4) wrt W.

$$\begin{aligned}\frac{\partial L_i}{\partial W} &= \frac{\partial}{\partial W}(-lnP_{ij}) \\ &= -\frac{1}{P_{ij}}P_{ij}x_i(\delta_{ij} - P_{ij}) \\ &= (P_{ij} - \delta_{ij})x_i\end{aligned} \tag{6}$$

1.3 Gradient of b_i

Computing the gradient $\frac{\partial L_i}{\partial b_i}$ via chain rule.

$$\frac{\partial L_i}{\partial b_i} = \frac{\partial L_i}{\partial P_{ij}} \frac{\partial P_{ij}}{\partial s_{ij}} \frac{\partial s_{ij}}{\partial b_i}$$

The partial derivative of s_{ij} (5) wrt b_i is nonzero when $j = y_i$ and zero everywhere else. This condition is expressed with δ_{ij} (2).

$$\frac{\partial s_{ij}}{\partial W} = \delta_{ij}$$

The partial derivative of P_{ij} (3) wrt b_i by applying the chain rule.

$$\frac{\partial P_{ij}}{\partial b_i} = \frac{\partial}{\partial W} \frac{e^{s_{ij}}}{\sum_k e^{s_{ik}}} = \frac{-e^{s_{ij}} e^{s_{ij}}}{\left(\sum_k e^{s_{ik}}\right)^2} = -P_{ij}^2$$

Finally the partial derivative of L_i (4) wrt b_i.

$$\begin{aligned} \frac{\partial L_i}{\partial b_i} &= \frac{\partial}{\partial b_i} (-lnP_{ij}) \\ &= -\frac{1}{P_{ij}} (-P_{ij}^2) \\ &= P_{ij} \end{aligned} \tag{7}$$

Figure 1: Example caption.

Example Section.

1.4 Example Subsection

Example Sub Section.

Example paragraph.

1.5 Example Equation

This is how a equation looks like.

$$y = ax^2 + bx + c \;\; , \tag{8}$$

where an inline equation looks like $a = b$.

1.6 Example Figure

To put a figure, you can do as shown in Fig. 1.