Softmax classifier's gradient

szymon.brych@gmail.com

May 2017

We have a softmax-based loss function component given by:

$$L_i = -log\left(\frac{e^{f_{y_i}}}{\sum_{j=0}^{n} e^{f_j}}\right)$$

Where:
1. Indexed exponent f is a vector of scores obtained during classification 2. Index y_i is proper label's index where y is column vector of all proper labels for training examples and i is example's index

Objective is to find:

$$\frac{\partial L_i}{\partial f_k}$$

Let's break down L_i into 2 separate expressions of a loss function component:

$$L_i = -log(p_{y_i})$$

And vector of normalized probabilities:

$$p_k = \frac{e^{f_k}}{\sum_{j=0}^{n} e^{f_j}}$$

Let's substitute sum:

$$\sigma = \sum_{j=0}^{n} e^{f_j}$$

For $k = y_i$ using quotient rule:

$$\frac{\partial p_k}{\partial f_{y_i}} = \frac{e^{f_k}\sigma - e^{2f_k}}{\sigma^2}$$

For $k \neq y_i$ during derivation e^{f_k} is treated as constant:

$$\frac{\partial p_k}{\partial f_{y_i}} = \frac{-e^{f_k}e^{f_{y_i}}}{\sigma^2}$$

Going further:

$$\frac{\partial L_i}{\partial p_k} = -\left(\frac{1}{p_{y_i}}\right)$$

Using chain rule for derivation:

$$\frac{\partial L_i}{\partial f_k} = -\left(\frac{1}{\frac{e^{f_k}}{\sigma}}\right)\frac{\partial p_k}{\partial f_{y_i}} = -\left(\frac{\sigma}{e^{f_k}}\right)\frac{\partial p_k}{\partial f_{y_i}}$$

Considering k and y_i, for $k = y_j$ after simplifications:

$$\frac{\partial L_i}{\partial f_k} = \frac{e^{f_k} - \sigma}{\sigma} = \frac{e^{f_k}}{\sigma} - 1 = p_k - 1$$

And for $k \neq y_j$:

$$\frac{\partial L_i}{\partial f_k} = \frac{e^{f_k}}{\sigma} = p_k$$

These two equations can be combined using Kronecker delta:

$$\frac{\partial L_i}{\partial f_k} = p_k - \delta_{k y_i}$$