ComS_573 Machine Learning-Problem Set 3 Solution

1. (20 pts.) Consider the multiclass logistic regression model.

   (a) (5 pts.) Show that the derivatives of the softmax activation function are given by

$$\frac{\partial y_k}{\partial a_j} = y_k(\delta_{kj} - y_j).$$

   (b) (15 pts.) *Derive* the batch and single sample gradient descent weight update rules for minimizing the cross-entropy error function.

Solution: 1. In the multiclass logistic regression model, we have

$$y_k(\vec{x}) = \frac{e^{a_k}}{\sum_i e^{a_i}}.$$

where $a_k = \vec{w_k}^t \vec{x}$.

The derivative of the logistic sigmoid function computed as following:

$$\begin{aligned}
y_k(\vec{x}) &= \frac{e^{a_k}}{\sum_i e^{a_i}} \\
\frac{\partial y_k(\vec{x})}{\partial a_j} &= \frac{\frac{\partial e^{a_k}}{\partial a_j}(\sum_i e^{a_i}) - \frac{\partial \sum_i e^{a_i}}{\partial a_j}(e^{a_k})}{(\sum_i e^{a_i})^2} \\
&= \frac{\delta_{kj} e^{a_k}(\sum_i e^{a_i}) - e^{a_j} e^{a_k}}{(\sum_i e^{a_i})^2} \\
&= \frac{e^{a_k}}{\sum_i e^{a_i}} \cdot \frac{\delta_{kj}(\sum_i e^{a_i}) - e^{a_j}}{\sum_i e^{a_i}} \\
&= \frac{e^{a_k}}{\sum_i e^{a_i}} \cdot (\delta_{kj} - \frac{e^{a_j}}{\sum_i e^{a_i}}) \\
&= y_k(\delta_{kj} - y_j)
\end{aligned}$$

where

$$\delta_{kj} = \begin{cases} 1 & k = j \\ 0 & otherwise \end{cases}$$

2.Derive the batch and single sample gradient descent weight update rules for minimizing the cross-entropy error function. Let

$$y_{ik} = P(\omega_k|\vec{x_i}) = y_k(\vec{x_i}) = \frac{e^{a_{ik}}}{\sum_j e^{a_{ij}}};$$

where activations:

$$a_{ij} = \vec{w_j}^t \vec{x_i}$$

And the cross-entropy error function is :

$$E(W) = -\sum_i^n \sum_k^c t_{ik} ln y_{ik}$$

where t_{ik} is 1 of c encoding $(\vec{x}_i, \omega_k)$ is $(\vec{x}_i, \vec{t}_i)$, $t_{ik} = 1$ and $t_{ij} = 0$, if $j \neq k$.

Now by gradient descent algorithm and digression(lecture 05 page 13/28) chain rule, we have:

$$\frac{\partial E(W)}{\partial(\vec{w}_j)} = \sum_i \frac{\partial E(W)}{\partial(a_{ij})} \frac{\partial(a_{ij})}{\partial(\vec{w}_j)}$$

Where we can compute:

$$\frac{\partial(a_{ij})}{\partial(\vec{w}_j)} = \vec{x_i}$$