Com S 573 HW 2 Karak Choudhury

Problem 1:

a) For the multiclass logistic regression, we know,

$$P(\omega_k | \underset{\sim}{x}) = y_k(\underset{\sim}{x}) = \frac{\exp(a_k)}{\sum_i \exp(a_i)}$$

where, $a_k = \underset{\sim}{\omega}_k' \underset{\sim}{x}$

So, $$\frac{\partial y_k(\underset{\sim}{x})}{\partial a_j} = \frac{1}{\left(\sum_i e^{a_i}\right)^2}\left[\sum_i e^{a_i} \frac{\partial e^{a_k}}{\partial a_j} - e^{a_k} \frac{\partial \sum_i e^{a_i}}{\partial a_j}\right]$$

$$= \frac{1}{\left(\sum_i e^{a_i}\right)^2}\left[e^{a_k} \sum_i e^{a_i} \frac{\partial a_k}{\partial a_j} - e^{a_k} e^{a_j}\right]$$

let $\delta_{k,j} = \begin{cases} 1 & \text{if } k=j \\ 0 & \text{o.w.} \end{cases}$

$$= \frac{1}{\left(\sum_i e^{a_i}\right)^2}\left[\delta_{k,j} e^{a_k} \sum_i e^{a_i} - e^{a_k} e^{a_j}\right]$$

$$= \frac{e^{a_k}}{\sum_i e^{a_i}}\left[\delta_{k,j} - \frac{e^{a_j}}{\sum_i e^{a_i}}\right]$$

$$= y_k(\delta_{k,j} - y_j) \quad \text{(Proved)}$$

Problem 1:

b) We have,

$$P(\omega_k | \underset{\sim}{x}_i) = y_k(\underset{\sim}{x}_i) = \frac{e^{a_{ik}}}{\sum_j e^{a_{ij}}}$$

Where, $a_{ij} = \underset{\sim}{\omega}_j' \underset{\sim}{x}_i$ (activation func^n)

Also, we know, the cross-entropy error func^n

$$E(W) = -\sum_{i=1}^{n} \sum_{k=1}^{c} t_{ik} \ln Y_{ik}$$

where, $t_{ik} = 1$ if one-of-c encoding $(\underset{\sim}{x}_i, \underset{\sim}{\omega}_k)$ is $(\underset{\sim}{x}_i, \underset{\sim}{t}_i)$ and $t_{ij} = 0$ if $j \neq k$.

So, $$\frac{\partial E(W)}{\partial \underset{\sim}{\omega}_j} = \sum_i \frac{\partial E(W)}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial \underset{\sim}{\omega}_j}$$ (By chain rule)

Now, $$\frac{\partial a_{ij}}{\partial \underset{\sim}{\omega}_j} = \frac{\partial}{\partial \underset{\sim}{\omega}_j}(\underset{\sim}{\omega}_j' x_i) = \underset{\sim}{x}_i \quad \text{— (1)}$$

and $$\frac{\partial E(W)}{\partial a_{ij}} = -\frac{\partial}{\partial a_{ij}}\left[\sum_{i=}^{n} \sum_{k=1}^{c} t_{ik} \ln Y_{ik}\right]$$

$$= -\sum_{k=1}^{c} \frac{\partial}{\partial a_{ij}} (t_{ik} \ln Y_{ik})$$

$$= -\sum_{k=1}^{c} t_{ik} \frac{1}{Y_{ik}} \frac{\partial Y_{ik}}{\partial a_{ij}}$$

$$= -\sum_{k=1}^{c} t_{ik} \frac{1}{\cancel{Y_{ik}}} \cdot \cancel{Y_{ik}} (\delta_{kj} - Y_{ij})$$ (from Part a)

$$= -\sum_{k=1}^{c} t_{ik} \delta_{kj} + Y_{ij} \sum_{k=1}^{c} t_{ik}$$

Since, if $k=j$ then $\delta_{kj} = 1$ otherwise 0. Similarly for t_{ik}.

$$= Y_{ij} - t_{ij} \quad \text{— (2)}$$

So, we can write gradient equation — from eq 1 & 2

$$\frac{\partial E(W)}{\partial \underset{\sim}{w}_j} = \sum_{i=1}^{n} (Y_{ij} - t_{ij})\, \underset{\sim}{x}_i$$

Thus, Batch gradient descent rule can be written as —

update $\underset{\sim}{w}_j \leftarrow \underset{\sim}{w}_j - \eta \sum_{i=1}^{n} (Y_{ij} - t_{ij})\, \underset{\sim}{x}_i$

and for each update, (single update)

$$\underset{\sim}{w}_j \leftarrow \underset{\sim}{w}_j - \eta (Y_{ij} - t_{ij})\, \underset{\sim}{x}_i$$