INT 305 Assignment 1

(The deadline is 31st of Oct.)

1. Please write down the whole derivation process to obtain the gradient for logistic regression. (30%)

Solution:

The logistic model is: $z = w^{\top}x$.

and the activation function is: $y = \frac{1}{(1+e^{-z})}$,

the loss function is: $\mathcal{L}_{CE}(y,t) = -t\log(y) - (1-t)\log(1-y)$

To optimize the model, we should update the weight w by using gradient descent.

Therefore, by chain rule:

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

1), because $\mathcal{L}_{CE}(y,t) = -t\log(y) - (1-t)\log(1-y)$,

$$\begin{aligned}\frac{\partial \mathcal{L}_{CE}}{\partial y} &= \frac{\partial[-t\log(y)]}{\partial y} - \frac{\partial[(1-t)\log(1-y)]}{\partial y} \\ &= \left(-t \cdot \frac{1}{y}\right) - \frac{\partial[(1-t)\log(1-y)]}{\partial(1-y)} \cdot \frac{\partial(1-y)}{y} \\ &= \left(-t \cdot \frac{1}{y}\right) - \left[(1-t) \cdot \frac{1}{(1-y)}\right] \cdot (-1) \\ &= \left(-t \cdot \frac{1}{y}\right) - \left[-\frac{(1-t)}{(1-y)}\right] \\ &= \left(-\frac{t}{y} + \frac{1-t}{1-y}\right)\end{aligned}$$

2), because $y = \frac{1}{(1+e^{-z})}$,

$$\begin{aligned}\frac{\partial y}{\partial z} &= \frac{\partial\left[\frac{1}{(1+e^{-z})}\right]}{\partial(1+e^{-z})} \cdot \frac{\partial(1+e^{-z})}{\partial z} \\ &= -(1+e^{-z})^{-2} \cdot (-e^{-z}) \\ &= \frac{e^{-z}}{(1+e^{-z})(1+e^{-z})} \\ &= \frac{1}{(1+e^{-z})} \cdot \frac{(1+e^{-z})-1}{(1+e^{-z})} \Rightarrow y(1-y)\end{aligned}$$

3), because $z = w^{\top}x$,

$$\frac{\partial z}{\partial w_j} = \frac{\partial\left(w_j x_j\right)}{\partial w_j} = x_j$$

Therefore,

$$\begin{aligned}
\frac{\partial \mathcal{L}_{CE}}{\partial w_j} &= \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_j} \\
&= \left(-\frac{t}{y} + \frac{1-t}{1-y}\right) \cdot y(1-y) \cdot x_j \\
&= [-t(1-y) + (1-t) \cdot y] \cdot x_j \\
&= (-t + ty + y - ty) \cdot x_j \\
&= (-t + y) \cdot x_j
\end{aligned}$$

Now, we can get the gradient of w, then we need to update weight w.

$$\begin{aligned}
w_j &\leftarrow w_j - \alpha \frac{\partial \mathcal{J}}{\partial w_j} \\
\text{Because } \mathcal{J} &= \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{CE} \\
\text{thus}, w_j &\leftarrow w_j - \alpha \frac{\partial \mathcal{J}}{\partial w_j} \\
&= w_j - \frac{\alpha}{N} \sum_{i=1}^{N} \left(y^{(i)} - t^{(i)}\right) x_j^{(i)}
\end{aligned}$$

2. Please write down the whole derivation process to obtain the gradient for multiclass classification with softmax. (40%)

Solution:

The model function is: $z = Wx$,

$$\Rightarrow z_k = w_k \cdot x$$

and the softmax function is: $y = \mathrm{softmax}(z)$

$$\Rightarrow y_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}},$$

and the cross-entropy loss function is: $\mathcal{L}_{CE}(y, t) = -\sum_{k=1}^{K} t_k \log y_k$

$$= -t^{\top}(\log y)$$

To optimize the model, we should update the weight w by using gradient descent.

Therefore, by chain rule:

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_k} = \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_k}$$

1), because $\mathcal{L}_{CE} = -t^{\top}(\log y)$,

$$\frac{\partial \mathcal{L}_{CE}}{\partial y} = \frac{\partial[-t^{\top}(\log y)]}{\partial y}$$

Since after one-hot encoding, for classification task with K classes, t_k is a K-dimensional vector. And only the position of the correct class is 1, and the rest is 0.

So, there are two cases: for correct class T, $t_T = 1$, $t_{m\neq T} = 0$. However, according to the function $\mathcal{L}_{CE}(y,t) = -\sum_{k=1}^{K} t_k \log y_k$, we only need to consider the condition of t_T, because the rest condition will get 0. For the same reason, we only need to consider y_T as well.

Therefore, $\mathcal{L}_{CE} = -t_T(\log y_T) = -\log y_T$,

so,

$$\frac{\partial \mathcal{L}_{CE}}{\partial y_T} = \frac{\partial(-\log y_T)}{\partial y_T} = -\frac{1}{y_T}$$

2), because $y_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$,

$$\frac{\partial y}{\partial z_k} = \frac{\partial y_T}{\partial z_k} = \frac{\partial y_k}{\partial z_k} = \frac{\partial\left(\frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}\right)}{\partial z_k}$$

Since mentioned before, there are also two cases.

i), For $T = k$:

$$\frac{\partial y_T}{\partial z_k} \quad = \frac{\partial y_k}{\partial z_k} = \frac{\partial\left(\frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}\right)}{\partial z_k}$$

$$= \frac{e^{z_k} \cdot \sum_{k'} e^{z_{k'}} - e^{z_k} \cdot \frac{\partial(\sum_{k'} e^{z_{k'}})}{\partial z_k}}{(\Sigma_{k'} e^{z_{k'}})^2}$$

because z_k is one of $z_{k'}$,

so, $\frac{\partial(\sum_{k'} e^{z_{k'}})}{\partial z_k} = \frac{\partial(e^{z_1}+\cdots+e^{z_k}+\cdots+e^{z_n})}{\partial z_k} = e^{z_k}$

Thus,

$$\begin{aligned}\frac{\partial y_T}{\partial z_k} = \frac{\partial y_k}{\partial z_k} &= \frac{e^{z_k} \cdot \sum_{k'} e^{z_{k'}} - e^{z_k} \cdot e^{z_k}}{(\sum_{k'} e^{z_{k'}})^2} \\ &= \frac{e^{z_k}(\sum_{k'} e^{z_{k'}} - e^{z_k})}{\sum_{k'} e^{z_{k'}} \cdot \sum_{k'} e^{z_{k'}}} \\ &= y_k(1 - y_k)\end{aligned}$$

ii), For $T \neq k$:

$$\begin{aligned}\frac{\partial y_T}{\partial z_k} &= \frac{\partial \left(\frac{e^{z_T}}{\sum_{k'} e^{z_{k'}}} \right)}{\partial z_k} \\ &= \frac{0 - e^{z_T} \cdot e^{z_k}}{(\sum_{k'} e^{z_{k'}})^2} \\ &= \frac{-e^{z_T} \cdot e^{z_k}}{\sum_{k'} e^{z_{k'}} \cdot \sum_{k'} e^{z_{k'}}} \\ &= -y_T \cdot y_k\end{aligned}$$

3), because $z_k = w_k \cdot x$,

$$\frac{\partial z_k}{\partial w_k} = \frac{\partial (w_k \cdot x)}{\partial w_k} = x$$

4), because we have two cases of $\frac{\partial y}{\partial z_k}$, so we should consider them both when calculate $\frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z_k}$.

i), For $T = k \ (or \ t_k = 1)$,

$$\begin{aligned}\frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z_k} = \frac{\partial \mathcal{L}_{CE}}{\partial z_k} \quad &= \frac{\partial \mathcal{L}_{CE}}{\partial y_T} \cdot \frac{\partial y_T}{\partial z_k} \\ &= -\frac{1}{y_k} \cdot y_k(1 - y_k) \\ &= y_k - 1 \\ &= y_k - t_k\end{aligned}$$

ii), For $T \neq k \ (or \ t_k = 0)$,

$$\begin{aligned}\frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z_k} = \frac{\partial \mathcal{L}_{CE}}{\partial z_k} \quad &= \frac{\partial \mathcal{L}_{CE}}{\partial y_T} \cdot \frac{\partial y_T}{\partial z_k} \\ &= -\frac{1}{y_T} \cdot (-y_T \cdot y_k) \\ &= y_k \\ &= y_k - 0 \\ &= y_k - t_k\end{aligned}$$

Therefore, the result of $\frac{\partial \mathcal{L}_{CE}}{\partial z_k}$ are both $y_k - t_k$.

5),

$$\begin{aligned}\frac{\partial \mathcal{L}_{CE}}{\partial w_k} &= \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_k} \\ &= \frac{\partial \mathcal{L}_{CE}}{\partial z_k} \cdot \frac{\partial z_k}{\partial w_k} \\ &= (y_k - t_k) \cdot x\end{aligned}$$

Now, we can get the gradient of w, then we need to update weight w.

$$\begin{aligned} w_k &\leftarrow w_k - \alpha \frac{\partial \mathcal{J}}{\partial w_k} \\ \text{Because } \mathcal{J} &= \frac{1}{N}\sum_{k=1}^{N} \mathcal{L}_{CE} \\ \text{thus,} w_k &\leftarrow w_k - \alpha \frac{\partial \mathcal{J}}{\partial w_k} \\ &= w_k - \frac{\alpha}{N}\sum_{i=1}^{N}\left(y_k^{(i)} - t_k^{(i)}\right) x^{(i)}\end{aligned}$$

3. Please compare the SVM loss and Softmax loss for multiclass classification, please explain which one is better? (30%)

Solution:

The softmax loss is better for multiclass classification.

For example, now there is a multiclass classification task with three classes. And we sample three training examples, their scores are shown below (the bold type class is label):

	Class 1	Class 2	Class 3
Example 1	**10**	-2	3
Example 2	**10**	9	9
Example 3	**10**	-100	-100

Now, we calculate SVM loss and Softmax loss for these three examples respectively.

The formula of softmax loss is:

$$L_i = -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

The formula of SVM loss is:

$$L_i = \sum_{j \neq y_i} \max\left(0, s_j - s_{y_i} + 1\right)$$

The results are shown below:

	Class 1	Class 2	Class 3	SVM	Softmax
Example 1	**10**	-2	3	0	0.4E-3
Example 2	**10**	9	9	0	0.24
Example 3	**10**	-100	-100	0	0

It can be seen from the results that SVM loss cannot reflect the degree of model optimization precisely. When using SVM Loss to optimize the model, we may only find the local optimal solution, because after obtaining a solution, SVM Loss will become 0.

However, Softmax loss doesn't have that problem, it is a good reflection of the current model. Moreover, even if we find a solution, we can continue to optimize until we find the optimal solution.

Therefore, the Softmax loss is better than SVM loss for multiclass classification.