Gradient Computation of Softmax Regression

This note computes the gradient of the cost function of softmax regression in detail.

Suppose there are K classes and the class probabilities the label y of a training sample x is computed as

$$\begin{pmatrix} P(y=1|x) \\ P(y=2|x) \\ \vdots \\ P(y=K|x) \end{pmatrix} = \begin{pmatrix} \frac{\exp\left(\theta^{(1)\top} x\right)}{\sum_{j=1}^{K} \exp\left(\theta^{(j)\top} x\right)} \\ \frac{\exp\left(\theta^{(2)\top} x\right)}{\sum_{j=1}^{K} \exp\left(\theta^{(j)\top} x\right)} \\ \vdots \\ \frac{\exp\left(\theta^{(K)\top} x\right)}{\sum_{j=1}^{K} \exp\left(\theta^{(j)\top} x\right)} \end{pmatrix}.$$

Suppose we have m training samples $\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{m}$, then the cost function is as follows:

$$J(\theta) = -\left[\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \log \frac{\exp\left(\theta^{(k)\top} x^{(i)}\right)}{\sum_{j=1}^{K} \exp\left(\theta^{(j)\top} x^{(i)}\right)}\right],$$

in which $\mathbf{1}\{\cdot\}$ is the indicator function.

Now we show the process of computing $\nabla_{\theta^{(k)}} J(\theta)$ step by step.

$$\begin{aligned}
\nabla_{\theta^{(k)}} J(\theta) &= -\nabla_{\theta^{(k)}}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \log \frac{\exp\left(\theta^{(k)\top} x^{(i)}\right)}{\sum_{j=1}^{K} \exp\left(\theta^{(j)\top} x^{(i)}\right)}\right] \\
&= -\nabla_{\theta^{(k)}}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \left(\theta^{(k)\top} x^{(i)} - \log \sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top} x^{(i)}\right)\right)\right)\right] \\
&= -\nabla_{\theta^{(k)}} \sum_{i=1}^{m}\left[\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \theta^{(k)\top} x^{(i)} - \underbrace{\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\}}_{\text{sums to 1}} \underbrace{\log \sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top} x^{(i)}\right)\right)}_{\text{Without the summation index } k}\right] \\
&= -\nabla_{\theta^{(k)}} \sum_{i=1}^{m}\left[\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \theta^{(k)\top} x^{(i)} - \log \sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top} x^{(i)}\right)\right)\right] \\
&= -\sum_{i=1}^{m}\left[\underbrace{\nabla_{\theta^{(k)}}\left(\sum_{k=1}^{K} \mathbf{1}\left\{y^{(i)}=k\right\} \theta^{(k)\top} x^{(i)}\right)}_{\text{Only one term contains } \theta^{(k)}} - \nabla_{\theta^{(k)}} \log \sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top} x^{(i)}\right)\right)\right] \\
&= -\sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=k\right\} x^{(i)} - \nabla_{\theta^{(k)}} \log \sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top} x^{(i)}\right)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=k\right\}x^{(i)} - \frac{\nabla_{\theta^{(k)}} \underbrace{\sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top}x^{(i)}\right)\right)}_{\text{Only one term contains } \theta^{(k)}}}{\sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top}x^{(i)}\right)\right)}\right] \\
&= -\sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=k\right\}x^{(i)} - \frac{\exp\left(\theta^{(k)\top}x^{(i)}\right)\nabla_{\theta^{(k)}}\left(\theta^{(k)\top}x^{(i)}\right)}{\sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top}x^{(i)}\right)\right)}\right] \\
&= -\sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=k\right\}x^{(i)} - \frac{\exp\left(\theta^{(k)\top}x^{(i)}\right)}{\sum_{j=1}^{K}\left(\exp\left(\theta^{(j)\top}x^{(i)}\right)\right)}x^{(i)}\right] \\
&= -\sum_{i=1}^{m}\left[\mathbf{1}\left\{y^{(i)}=k\right\}x^{(i)} - P\left(y^{(i)}=k|x^{(i)}\right)x^{(i)}\right] \\
&= -\sum_{i=1}^{m}\left[x^{(i)}\left(\mathbf{1}\left\{y^{(i)}=k\right\} - P\left(y^{(i)}=k|x^{(i)}\right)\right)\right].
\end{aligned}
$$