Homework 2

Lidian Lin

January 2020

1 Newton's Method

According to Newton's Method,

$$J(w) = J(w^{(k)}) + \nabla_w J(w^{(k)})(w - w^{(k)}) + \frac{1}{2}(w - w^{(k)})H(w - w^{(k)}) \tag{1}$$

$$\begin{aligned}
\nabla_w J(w) &= \nabla_w J(w^{(k)}) + \frac{1}{2}\nabla_w (w^T H w - w^T H w^{(k)} - {w^{(k)}}^T H w + {w^{(k)}}^T H w^{(k)}) \\
&= \nabla_w J(w^{(k)}) + Hw - \frac{1}{2}Hw^{(k)} - \frac{1}{2}Hw^{(k)} \\
&= \nabla_w J(w^{(k)}) + Hw - Hw^{(k)}
\end{aligned} \tag{2}$$

Let $\nabla_w J(w) = 0$:

$$\begin{aligned}
\nabla_w J(w^{(k)}) + Hw - Hw^{(k)} &= 0 \\
H(w) &= Hw^{(k)} - \nabla_w J(w^{(k)}) \\
w^{(k+1)} &= w^{(k)} - H^{-1}\nabla_w J(w^{(k)})
\end{aligned} \tag{3}$$

$$\begin{aligned}
J(w) &= \frac{1}{2n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2 = \frac{1}{2n}(x^T w - y)^T (x^T w - y) \\
\because \quad H(w) &= \frac{1}{n}xx^T \\
\nabla_w J(w^{(k)}) &= \frac{1}{n}x(x^T w^{(k)} - y)
\end{aligned} \tag{4}$$

$$\begin{aligned}
w^{(k+1)} &= w^{(k)} - H^{-1}\nabla_w J(w^{(k)}) \\
\therefore \quad &= w^{(k)} - (\frac{1}{n}xx^T)^{-1}\frac{1}{n}(xx^T w^{(k)} - xy) \\
&= w^{(k)} - (w^{(k)} - (xx^T)^{-1}xy) \\
&= (xx^T)^{-1}xy
\end{aligned} \tag{5}$$

This equation shows that, $w^{(k+1)}$ has nothing to do with $w^{(k)}$. In another word, from whichever $w^{(k)}$ we start, we will come to the converge $(xx^T)^{-1}xy$ with only one iteration.

2 Derivation of Softmax Regression Gradient Updates

2.1 $\nabla_{w^{(l)}}\hat{y}_k^{(i)} =$

For $l = k$:

$$\begin{aligned}
\nabla_{w^{(l)}}\hat{y}_k^{(i)} &= \frac{\partial(\frac{e^{z_l^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}})}{\partial(z_l^{(i)})}\frac{\partial(z_l^{(i)})}{\partial(w^{(l)})} \\
&= \frac{e^{z_l^{(i)}}(\sum_{k'=1}^{c} e^{z_{k'}^{(i)}} - e^{z_l^{(i)}})}{(\sum_{k'=1}^{c} e^{z_{k'}^{(i)}})^2}x^{(i)} \\
&= \frac{e^{z_l^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}} \times \frac{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}} - e^{z_l^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}}x^{(i)} \\
&= \hat{y}_l^{(i)}(1 - \hat{y}_l^{(i)})x^{(i)}
\end{aligned} \tag{6}$$

For $l \neq k$:

$$\begin{aligned}
\nabla_{w^{(l)}}\hat{y}_k^{(i)} &= \frac{\partial(z_l^{(i)})}{\partial(w^{(l)})}\frac{\partial(\frac{e^{z_k^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}})}{\partial(z_l^{(i)})} \\
&= x^{(i)}\frac{e^{z_k^{(i)}}(0 - e^{z_l^{(i)}})}{(\sum_{k'=1}^{c} e^{z_{k'}^{(i)}})^2} \\
&= -x^{(i)}\frac{e^{z_k^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}} \times \frac{e^{z_l^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}} \\
&= -x^{(i)}\hat{y}_k^{(i)}\hat{y}_l^{(i)}
\end{aligned} \tag{7}$$

2.2 $\nabla_{w^{(l)}} f_{CE}(W, b) =$

$$
\begin{aligned}
\nabla_{w^{(l)}} f_{CE}(W, b) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \nabla_{w^{(l)}} \log \hat{y}_k^{(i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left(y_l^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_l^{(i)}}{\hat{y}_l^{(i)}}\right) + \sum_{k \neq l} y_k^{(i)} \left(\frac{\nabla_{w^{(l)}} \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}\right)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left(y_l^{(i)} x^{(i)} (1 - \hat{y}_l^{(i)}) + \sum_{k \neq l} y_k^{(i)} \times (-x^{(i)} \hat{y}_l^{(i)})\right) \qquad (8) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left(y_l^{(i)} x^{(i)} (1 - \hat{y}_l^{(i)}) + (1 - y_l^{(i)}) \times (-x^{(i)} \hat{y}_l^{(i)})\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left(y_l^{(i)} x^{(i)} - y_l^{(i)} x^{(i)} \hat{y}_l^{(i)} - x^{(i)} \hat{y}_l^{(i)} + y_l^{(i)} x^{(i)} \hat{y}_l^{(i)}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} x^{(i)} (y_l^{(i)} - \hat{y}_l^{(i)})
\end{aligned}
$$

2.3 $\nabla_b f_{CE}(W, b) =$

Similarly, we get:
For $l = k$:

$$
\begin{aligned}
\nabla_b \hat{y}_k^{(i)} &= \frac{\partial\left(\frac{e^{z_l^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}}\right)}{\partial(z_l^{(i)})} \frac{\partial(z_l^{(i)})}{\partial b} \\
&= \hat{y}_l^{(i)} (1 - \hat{y}_l^{(i)})
\end{aligned} \qquad (9)
$$

For $l \neq k$:

$$
\begin{aligned}
\nabla_b \hat{y}_k^{(i)} &= \frac{\partial\left(\frac{e^{Z_k^{(i)}}}{\sum_{k'=1}^{c} e^{z_{k'}^{(i)}}}\right)}{\partial(z_l^{(i)})} \frac{\partial(z_l^{(i)})}{\partial b} \\
&= -\hat{y}_k^{(i)} \hat{y}_l^{(i)}
\end{aligned} \qquad (10)
$$

$$
\begin{aligned}
\nabla_b f_{CE}(W,b) &= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} \nabla_b \log \hat{y}_k^{(i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} y_k^{(i)} (\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}}) \\
&= -\frac{1}{n}\sum_{i=1}^{n} (y_l^{(i)} (\frac{\nabla_b \hat{y}_l^{(i)}}{\hat{y}_l^{(i)}}) + \sum_{k \neq l} y_k^{(i)} (\frac{\nabla_b \hat{y}_k^{(i)}}{\hat{y}_k^{(i)}})) \\
\therefore \qquad &= -\frac{1}{n}\sum_{i=1}^{n} (y_l^{(i)} (1-\hat{y}_l^{(i)}) + \sum_{k \neq l} y_k^{(i)} \times (-\hat{y}_l^{(i)})) \\
&= -\frac{1}{n}\sum_{i=1}^{n} (y_l^{(i)} (1-\hat{y}_l^{(i)}) + (1-y_l^{(i)}) \times (-\hat{y}_l^{(i)})) \\
&= -\frac{1}{n}\sum_{i=1}^{n} (y_l^{(i)} - y_l^{(i)}\hat{y}_l^{(i)} - \hat{y}_l^{(i)} + y_l^{(i)}\hat{y}_l^{(i)}) \\
&= -\frac{1}{n}\sum_{i=1}^{n} (y_l^{(i)} - \hat{y}_l^{(i)})
\end{aligned} \tag{11}
$$

3 Implementation of Softmax Regression

After so many trials on hyperparameter sets, I found one set with relatively higher prediction accuracy. Here are the detailed hyperparameters and results.
Batch size: 2500, epoch: 300, learning rate: 1, regularization strength: 0.0001
The prediction accuracy on test data is 92.48%
MSE on test data is 0.2665376591887616