

Stanford CS 229, Public Course, Problem Set 1

Dylan Price

December 2, 2016

1

a)

Find the Hessian of the cost function $J(\theta) = \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2$

We know that $H_{jk} = \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k}$

First find $\frac{\partial J(\theta)}{\partial \theta_k}$,

$$\begin{aligned} \frac{\partial J(\theta)}{\partial \theta_k} &= \frac{1}{2} \sum_{i=1}^m \frac{\partial}{\partial \theta_k} (\theta^T x^{(i)} - y^{(i)})^2 \\ &= \frac{1}{2} \sum_{i=1}^m 2(\theta^T x^{(i)} - y^{(i)})(x_k^{(i)}) \\ &= \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})(x_k^{(i)}) \end{aligned}$$

Now find $\frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k}$,

$$\begin{aligned} \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} &= \frac{\partial}{\partial \theta_j} \left(\frac{\partial J(\theta)}{\partial \theta_k} \right) \\ &= \sum_{i=1}^m \frac{\partial}{\partial \theta_j} ((\theta^T x^{(i)} - y^{(i)})(x_k^{(i)})) \\ &= \sum_{i=1}^m x_j^{(i)} x_k^{(i)} \text{ for } 1 \leq j \leq n \text{ and } 1 \leq k \leq n \end{aligned}$$

Therefore

$$\begin{aligned}
H &= \begin{bmatrix} \sum_{i=1}^m x_1^{(i)} x_1^{(i)} & \sum_{i=1}^m x_2^{(i)} x_1^{(i)} & \cdots & \sum_{i=1}^m x_n^{(i)} x_1^{(i)} \\ \sum_{i=1}^m x_1^{(i)} x_2^{(i)} & \sum_{i=1}^m x_2^{(i)} x_2^{(i)} & \cdots & \sum_{i=1}^m x_n^{(i)} x_2^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m x_1^{(i)} x_n^{(i)} & \sum_{i=1}^m x_2^{(i)} x_n^{(i)} & \cdots & \sum_{i=1}^m x_n^{(i)} x_n^{(i)} \end{bmatrix} \\
&= \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(n)} \\ \vdots & \ddots & \vdots \\ x_m^{(1)} & \cdots & x_m^{(n)} \end{bmatrix} \begin{bmatrix} x_1^{(1)} & \cdots & x_m^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \cdots & x_m^{(n)} \end{bmatrix} \\
&= X^T X
\end{aligned}$$

b)

Show that the first iteration of Newton's method gives us $\theta^* = (X^T X)^{-1} X^T \vec{y}$, the solution to our least squares problem.

One iteration of Newton's Method:

$$\theta := \theta - H^{-1} \nabla_{\theta} J(\theta)$$

Therefore,

$$\begin{aligned}
\theta^* &= \theta - (X^T X)^{-1} \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_m} \end{bmatrix} \\
&= \theta - (X^T X)^{-1} \begin{bmatrix} \sum_{i=1}^m (x_1^{(i)} \theta^T x^{(i)} - x_1^{(i)} y^{(i)}) \\ \vdots \\ \sum_{i=1}^m (x_m^{(i)} \theta^T x^{(i)} - x_m^{(i)} y^{(i)}) \end{bmatrix}
\end{aligned}$$

let $\theta = \vec{0}$ (initialize θ)

$$\begin{aligned}
&= -(X^T X)^{-1} \begin{bmatrix} \sum_{i=1}^m (-x_1^{(i)} y^{(i)}) \\ \vdots \\ \sum_{i=1}^m (-x_m^{(i)} y^{(i)}) \end{bmatrix} \\
&= (X^T X)^{-1} \begin{bmatrix} \sum_{i=1}^m (x_1^{(i)} y^{(i)}) \\ \vdots \\ \sum_{i=1}^m (x_m^{(i)} y^{(i)}) \end{bmatrix} \\
&= (X^T X)^{-1} X^T \vec{y}
\end{aligned}$$

2

a

See q2/ folder

b

At low values of τ , the classification boundaries are clustered around the positive training examples. As you increase τ , these boundaries begin to merge into bigger areas, i.e. the classification boundaries look less 'local' to the positive training examples. At high values of τ , the classification boundary is essentially a straight line dividing positive and negative classes.

The decision boundary of unweighted logistic regression would look like the plots with the highest values of τ . This is because as τ approaches infinity, the weight function $w^{(i)}$ goes to 1 for every training example, making the regression unweighted (that is, every training example gets the same weight).

3

a)

$$\begin{aligned}
J(\theta) &= \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^p ((\theta^T x^{(i)})_j - y_j^{(i)})^2 \\
&= \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^T (\theta^T x^{(i)} - y^{(i)})
\end{aligned}$$

Let $\vec{1}$ be a vector of all ones.

$$= \frac{1}{2} \vec{1}^T (X\theta - Y)^T (X\theta - Y) \vec{1}$$

b)

$$\begin{aligned}
J(\theta) &= \frac{1}{2} \vec{1}^T (X\theta - Y)^T (X\theta - Y) \vec{1} \\
&= \frac{1}{2} \vec{1}^T ((X\theta)^T - Y^T) (X\theta - Y) \vec{1} \\
&= \frac{1}{2} \vec{1}^T (\theta^T X^T - Y^T) (X\theta - Y) \vec{1} \\
&= \frac{1}{2} \vec{1}^T (\theta^T X^T X\theta - \theta^T X^T Y - Y^T X\theta + Y^T Y) \vec{1} \\
&= \frac{1}{2} \vec{1}^T \theta^T X^T X\theta \vec{1} - \frac{1}{2} \vec{1}^T \theta^T X^T Y \vec{1} - \frac{1}{2} \vec{1}^T Y^T X\theta \vec{1} + \frac{1}{2} \vec{1}^T Y^T Y \vec{1}
\end{aligned}$$

$$\begin{aligned}
\Delta_{\theta} J(\theta) &= \frac{1}{2} * 2X^T X\theta - \frac{1}{2} X^T Y - \frac{1}{2} (Y^T X)^T \\
&= X^T X\theta - X^T Y
\end{aligned}$$

$$\begin{aligned}
0 &= X^T X \theta - X^T Y \\
X^T X \theta &= X^T Y \\
\theta &= (X^T X)^{-1} X^T Y
\end{aligned}$$

c)

Consider the multivariate model $y^{(i)} = \theta^T x^{(i)}$

$$y^{(i)} = \theta^T x^{(i)} = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,n} \\ \theta_{2,1} & \theta_{2,2} & & \\ \vdots & & \ddots & \\ \theta_{p,1} & & & \theta_{p,n} \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix} = \begin{bmatrix} \theta_{1,1}x_1^{(i)} & \theta_{1,2}x_2^{(i)} & \cdots & \theta_{1,n}x_n^{(i)} \\ \theta_{2,1}x_1^{(i)} & \theta_{2,2}x_2^{(i)} & & \\ \vdots & & \ddots & \\ \theta_{p,1}x_1^{(i)} & & & \theta_{p,n}x_n^{(i)} \end{bmatrix} = \begin{bmatrix} \theta_1^T x^{(i)} \\ \theta_2^T x^{(i)} \\ \vdots \\ \theta_p^T x^{(i)} \end{bmatrix} = \begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ \vdots \\ y_p^{(i)} \end{bmatrix}$$

Therefore each row of this multivariate model corresponds to an equation of the form $y_j^{(i)} = \theta_j^T x^{(i)}$.

4

a)

$$\begin{aligned}
\ell(\varphi) &= \log \prod_{i=1}^m P(x^{(i)}, y^{(i)}; \varphi) \\
&= \log \prod_{i=1}^m P(y^{(i)}) P(x^{(i)} | y^{(i)}) \\
&= \sum_{i=1}^m \log P(y^{(i)}) P(x^{(i)} | y^{(i)}) \\
&= \sum_{i=1}^m [\log P(y^{(i)}) + \log \prod_{j=1}^n (\phi_{j|y^{(i)}})^{x_j^{(i)}} (1 - \phi_{j|y^{(i)}})^{1-x_j^{(i)}}] \\
&= \sum_{i=1}^m [\log P(y^{(i)}) + \sum_{j=1}^n (\log(\phi_{j|y^{(i)}})^{x_j^{(i)}} + \log(1 - \phi_{j|y^{(i)}})^{1-x_j^{(i)}})] \\
&= \sum_{i=1}^m [\log P(y^{(i)}) + \sum_{j=1}^n (x_j^{(i)} \log(\phi_{j|y^{(i)}}) + (1 - x_j^{(i)}) \log(1 - \phi_{j|y^{(i)}}))] \\
&= \sum_{i=1}^m [(y^{(i)} \log \phi_y + (1 - y^{(i)}) \log(1 - \phi_y) + \sum_{j=1}^n (x_j^{(i)} \log(\phi_{j|y^{(i)}}) + (1 - x_j^{(i)}) \log(1 - \phi_{j|y^{(i)}}))]
\end{aligned}$$

b)

First find ϕ_y

$$\nabla_{\phi_y} \ell(\varphi) = \sum_{i=1}^m [y^{(i)} \frac{1}{\phi_y} - (1 - y^{(i)}) \frac{1}{1 - \phi_y}]$$

$$\begin{aligned}
0 &= \sum_{i=1}^m \left[\frac{y^{(i)}}{\phi_y} - \frac{1 - y^{(i)}}{1 - \phi_y} \right] \\
&= \sum_{i=1}^m [y^{(i)}(1 - \phi_y) - (1 - y^{(i)})\phi_y] \\
&= \sum_{i=1}^m [y^{(i)} - y^{(i)}\phi_y - (\phi_y - y^{(i)}\phi_y)] \\
&= \sum_{i=1}^m [y^{(i)} - y^{(i)}\phi_y - \phi_y + y^{(i)}\phi_y] \\
&= \sum_{i=1}^m [y^{(i)} - \phi_y] \\
\sum_{i=1}^m \phi_y &= \sum_{i=1}^m \phi_y \\
m\phi_y &= \sum_{i=1}^m y^{(i)} \\
\phi_y &= \frac{\sum_{i=1}^m y^{(i)}}{m} = \sum_{i=1}^m \frac{1\{y^{(i)} = 1\}}{m}
\end{aligned}$$

Now find $\phi_{j|y=0}$ and $\phi_{j|y=1}$

$$\begin{aligned}
\nabla_{\phi_{j|y^{(i)}}} \ell(\varphi) &= \nabla_{\phi_{j|y^{(i)}}} \sum_{i=1}^m [(y^{(i)} \log \phi_y + (1 - y^{(i)}) \log(1 - \phi_y))] + \\
&\quad \nabla_{\phi_{j|y^{(i)}}} \sum_{i=1}^m \left[\sum_{j=1}^n [(x_j^{(i)} \log(\phi_{j|y^{(i)}}) + (1 - x_j^{(i)}) \log(1 - \phi_{j|y^{(i)}}))] \right]
\end{aligned}$$

We can drop the first term (the one containing references to ϕ_y) since its gradient with respect to $\phi_{j|y^{(i)}}$ is 0.

We can also drop $\sum_{j=1}^n$ from the second term because we are taking the gradient with respect to a single $\phi_{j|y^{(i)}}$.

$$\begin{aligned}
&= \nabla_{\phi_{j|y^{(i)}}} \sum_{i=1}^m [x^{(i)} \log \phi_{j|y^{(i)}} + (1 - x^{(i)}) \log(1 - \phi_{j|y^{(i)}})] \\
&= \sum_{i=1}^m \left[x^{(i)} \frac{1}{\phi_{j|y^{(i)}}} - (1 - x^{(i)}) \frac{1}{1 - \phi_{j|y^{(i)}}} \right]
\end{aligned}$$

$$\begin{aligned}
0 &= \sum_{i=1}^m [x^{(i)} \frac{1}{\phi_{j|y^{(i)}}} - (1 - x^{(i)}) \frac{1}{1 - \phi_{j|y^{(i)}}}] \\
&= \sum_{i=1}^m [x^{(i)}(1 - \phi_{j|y^{(i)}}) - (1 - x^{(i)})\phi_{j|y^{(i)}}] \\
&= \sum_{i=1}^m [x^{(i)} - \phi_{j|y^{(i)}}] \\
\sum_{i=1}^m \phi_{j|y^{(i)}} &= \sum_{i=1}^m x^{(i)} \\
\sum_{i=1}^m \phi_{j|y=0} 1\{y^{(i)} = 0\} &= \sum_{i=1}^m x^{(i)} 1\{y^{(i)} = 0\} \\
\phi_{j|y=0} &= \frac{\sum_{i=1}^m x^{(i)} 1\{y^{(i)} = 0\}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}} = \frac{\sum_{i=1}^m 1\{x^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}}
\end{aligned}$$

Similarly,

$$\phi_{j|y=1} = \frac{\sum_{i=1}^m 1\{x^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}}$$

c)

We predict $y = 1$ when

$$P(y = 1|x) \geq P(y = 0|x)$$

$$P(y = 1|x) - P(y = 0|x) \geq 0$$

$$\frac{P(x|y=1)P(y=1)}{\sum_{a \in Val(y)} P(x|y=a)P(y=a)} - \frac{P(x|y=0)P(y=0)}{\sum_{a \in Val(y)} P(x|y=a)P(y=a)} \geq 0$$

$$\frac{\phi_y \prod_{j=1}^n (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j} - (1 - \phi_y) \prod_{j=1}^n (\phi_{j|y=0})^{x_j} (1 - \phi_{j|y=0})^{1-x_j}}{\phi_y \prod_{j=1}^n (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j} + (1 - \phi_y) \prod_{j=1}^n (\phi_{j|y=0})^{x_j} (1 - \phi_{j|y=0})^{1-x_j}} \geq 0$$

$$\phi_y \prod_{j=1}^n (\phi_{j|y=1})^{x_j} (1 - \phi_{j|y=1})^{1-x_j} \geq (1 - \phi_y) \prod_{j=1}^n (\phi_{j|y=0})^{x_j} (1 - \phi_{j|y=0})^{1-x_j}$$

$$\log \phi_y + \sum_{j=1}^n x_j \log \phi_{j|y=1} + \sum_{j=1}^n (1 - x_j) \log(1 - \phi_{j|y=1}) \geq \log(1 - \phi_y) + \sum_{j=1}^n x_j \log \phi_{j|y=0} + \sum_{j=1}^n (1 - x_j) \log(1 - \phi_{j|y=0})$$

$$\begin{aligned}
&\log \phi_y - \log(1 - \phi_y) + \sum_{j=1}^n x_j \log \phi_{j|y=1} + \sum_{j=1}^n \log(1 - \phi_{j|y=1}) - \sum_{j=1}^n x_j \log(1 - \phi_{j|y=1}) \\
&\quad - \sum_{j=1}^n x_j \log \phi_{j|y=0} - \sum_{j=1}^n \log(1 - \phi_{j|y=0}) + \sum_{j=1}^n x_j \log(1 - \phi_{j|y=0}) \geq 0
\end{aligned}$$