Stanford CS 229, Public Course, Problem Set 1

Dylan Price

October 11, 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 $H_{jk} = \frac{\partial^2 J(\theta)}{\partial\theta_j\theta_k}$

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

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

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,

$$
\begin{aligned}
\theta^* &= \theta - (X^T X)^{-1} \begin{bmatrix}
\dfrac{\partial J(\theta)}{\partial \theta_1} \\
\vdots \\
\dfrac{\partial J(\theta)}{\partial \theta_m}
\end{bmatrix} \\
&= \theta - (X^T X)^{-1} \begin{bmatrix}
\sum_{i=1}^{m} (x_1^{(i)} \theta^T x^{(i)} - x_1^{(i)} y^{(i)}) \\
\vdots \\
\sum_{i=1}^{m} (x_m^{(i)} \theta^T x^{(i)} - x_m^{(i)} y^{(i)})
\end{bmatrix} \\
\text{let } \theta = \vec{0} \text{ (initialize } \theta\text{)} \\
&= -(X^T X)^{-1} \begin{bmatrix}
\sum_{i=1}^{m} (-x_1^{(i)} y^{(i)}) \\
\vdots \\
\sum_{i=1}^{m} (-x_m^{(i)} y^{(i)})
\end{bmatrix} \\
&= (X^T X)^{-1} \begin{bmatrix}
\sum_{i=1}^{m} (x_1^{(i)} y^{(i)}) \\
\vdots \\
\sum_{i=1}^{m} (x_m^{(i)} y^{(i)})
\end{bmatrix} \\
&= (X^T X)^{-1} X^T \vec{y}
\end{aligned}
$$

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).