2 Locally weighted linear regression

Part 1

Show that

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2 = (X\theta - y)^T W (X\theta - y).$$

$$A = X\theta - y = \begin{bmatrix} (x^{(1)})^T\theta \\ (x^{(2)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} (x^{(1)})^T\theta - y^{(1)} \\ (x^{(2)})^T\theta - y^{(2)} \\ \vdots \\ (x^{(m)})^T\theta - y^{(m)} \end{bmatrix}$$

$$W = \frac{1}{2}\begin{bmatrix} w^{(1)} \cdots 0 \\ \vdots \ddots \vdots \\ 0 \cdots w^{(m)} \end{bmatrix}$$

$$J(\theta) = A^T W A = \begin{bmatrix} (x^{(1)})^T\theta - y^{(1)} \cdots (x^{(m)})^T\theta - y^{(m)} \end{bmatrix} \frac{1}{2}\begin{bmatrix} w^{(1)} \cdots 0 \\ \vdots \ddots \vdots \\ 0 \cdots w^{(m)} \end{bmatrix} \begin{bmatrix} (x^{(1)})^T\theta - y^{(1)} \\ \vdots \\ (x^{(m)})^T\theta - y^{(m)} \end{bmatrix}$$

$$= \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$$

Part 2

$$\begin{aligned} J(\theta) &= (X\theta - y)^T W (X\theta - y) \\ &= (X^T\theta^T - y^T)(WX\theta - Wy) \\ &= (\theta^T X^T W X\theta - \theta^T X^T W y - y^T W X\theta + y^T W y) \end{aligned}$$

Because $(\theta^T X^T W)$ and y are 1 x m and m x 1, respectively, $\theta^T X^T W y = y^T W X\theta$.

$$\begin{aligned} \frac{dJ(\theta)}{d\theta} &= \frac{d}{d\theta}(\theta^T X^T W X\theta - 2(\theta^T X^T W y) + y^T W y) \\ &= X^T W X\theta - 2X^T W y + X^T W X\theta \\ &= 2X^T W X\theta - 2X^T W y \end{aligned}$$

To find θ which minimizes $J(\theta)$, we set $2X^T W X\theta - 2X^T W y = 0$ and get

$$\theta = (X^T W X)^{-1} X^T W y$$

Part 3

Algorithm 1 Calculating θ by Batch Gradient Descent

Input: Data matrix $X \in m \times d+1$, vector $y \in m \times 1$, learning rate $\alpha \in \mathbb{R}$, input vector $x \in \mathbb{R}^{d+1}$

$w \leftarrow m \times n$ zeros matrix
$\theta \leftarrow d \times 1$ zeros matrix
$grad \leftarrow d \times 1$ zeros matrix

for $j = 0$ to m **do**
    $w_j^{(j)} \leftarrow \frac{(x-X^{(j)})^T(x-X^{(j)})}{2length(x)^2}$
end for

for $j = 0$ to 5000 **do** $\triangleright$ arbitrary number of iterations
    $grad \leftarrow \frac{X^T w(X\theta - y)}{m}$
    $\theta \leftarrow \theta - \alpha * grad$
end for

return θ

Locally weighted linear regression is a non-parametric method.

3 Properties of the linear regression estimator

Part 1

Show that $E(\theta) = \theta^*$.
Normal equation states: $X^T X\theta = X^T y$.

$$\begin{aligned} \therefore (X^T X)^{-1}(X^T X\theta) &= (X^T X)^{-1} X^T y \\ I\theta = \theta &= (X^T X)^{-1} X^T y \\ if (X^T X)^{-1} X^T &= A, \\ E(\theta) &= E(Ay) = AE(y) \end{aligned}$$

And since y is normally distributed, $\epsilon = 0 \qquad \therefore y = \theta^T x$

$$\text{By this definition, } E(y) = X\theta^*$$

$$\begin{aligned} \therefore E(\theta) &= AX\theta^* \\ &= (X^T X)^{-1} X^T X\theta^* \\ &= I\theta^* \\ &= \theta^* \end{aligned}$$

Part 2

Show that $Var(\theta) = (X^T X)^{-1}\sigma^2$.

From Part 1,

$$\begin{aligned}
\theta &= (X^T X)^{-1} X^T y \\
if A &= (X^T X)^{-1} X^T, \\
Var(\theta) &= Var(Ay) = AVar(y)A^T = (X^T X)^{-1} X^T Var(y)((X^T X)^{-1} X^T)^T \\
&= (X^T X)^{-1} X^T \sigma^2 I((X^T X)^{-1} X^T)^T
\end{aligned}$$

$(A^T B^T = (BA)^T)$

$$\begin{aligned}
\therefore &= \sigma^2 I (X^T X)^{-1}((X^T X)^{-1} X^T X)^T \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$