Linear Regression

Cost Function

$$
\begin{aligned}
J(\vec{\theta}) &= \frac{1}{2m}\sum_i (\overrightarrow{x_i} \cdot \vec{\theta} - y_i)^2 \\
&= \frac{1}{2m}\sum_i ((x_{i,1}\theta_1 + x_{i,2}\theta_2 + \cdots + x_{i,n}\theta_n) - y_i)^2
\end{aligned}
$$

To take the gradient of $J(\vec{\theta})$, we consider the partial derivative with respect to θ_k for some $k \in 1...n$
.

$$
\frac{\partial J}{\partial \theta_k} = \frac{1}{m}\sum_i (\overrightarrow{x_i} \cdot \vec{\theta} - y_i)x_{i,k}
$$

Since this holds for every $k \in 1...n$, we have:

$$
\nabla J(\vec{\theta}) = \frac{1}{m}\sum_i (\overrightarrow{x_i} \cdot \vec{\theta} - y_i)\overrightarrow{x_i}
$$

Note that $\nabla J(\vec{\theta})$ may also appear as just $J'(\vec{\theta})$ or $\frac{\partial J}{\partial \vec{\theta}}$.

Matrix Version

To derive the matrix version, we require some intermediary lemmas:

Proposition 1: If $y(\vec{\beta}) = \vec{x} \cdot \vec{\beta}$, and $\vec{x}$ does not depend on $\vec{\beta}$, then:

$$
\frac{\partial y}{\partial \vec{\beta}} = \vec{x}
$$

(Note that $y(\vec{\beta})$ is a scalar, but $\frac{\partial y}{\partial \vec{\beta}}$, its derivative/gradient, is a vector!)

Proof:

$$y(\vec{\beta}) = \sum_i x_i\beta_i = x_1\beta_1 + x_2\beta_2 + \ldots + x_i\beta_i$$

$$\text{We take the partial derivative with respect to } \beta_j \text{ for some } j \in 1\ldots i:$$

$$\frac{\partial y}{\partial \beta_j} = x_j.$$

$$\text{Since the } j^{\text{th}} \text{ component of } \frac{\partial y}{\partial \vec{\beta}} \text{ is } x_j \text{ for all } j,$$

$$\therefore \frac{\partial y}{\partial \vec{\beta}} = \vec{x}$$

Proposition 2: Let the scalar α be defined by $\alpha = \vec{y}^T\mathbf{A}\vec{x}$ where:

- $\vec{y}$ is $m \times 1$
- $\vec{x}$ is $n \times 1$
- $\mathbf{A}$ is $m \times n$, and independent of $\vec{x}$ and $\vec{y}$

Then:

$$\frac{\partial \alpha}{\partial \vec{x}} = \vec{y}^T\mathbf{A} \text{ and } \frac{\partial \alpha}{\partial \vec{y}} = \vec{x}^T\mathbf{A}^T.$$

Proof:

Define $\vec{w}^T = \vec{y}^T\mathbf{A}$ (so $\vec{w}$ is an $n \times 1$ column vector).

Then:

$$\alpha = \vec{w}^T\vec{x}$$

It follows from Proposition 1 that:

$$\therefore \frac{\partial \alpha}{\partial \vec{x}} = \vec{w}^T = \vec{y}^T\mathbf{A}.$$

To prove the next part, we note that since α is a scalar:

$$\alpha = \alpha^T = \vec{x}^T\mathbf{A}^T\vec{y},$$

$$\therefore \frac{\partial \alpha}{\partial \vec{y}} = \vec{x}^T\mathbf{A}^T.$$

Proposition 3: Let the scalar α be defined by $\alpha = \vec{x}^T\mathbf{A}\vec{x}$ where:

- $\vec{x}$ is $n \times 1$
- $\mathbf{A}$ is $m \times n$ and independent of $\vec{x}$

Then:

$$\frac{\partial \alpha}{\partial \vec{x}} = \vec{x}^T(\mathbf{A} + \mathbf{A}^T)$$

Proof:

By definition, we have:

$$\alpha = \sum_{j=1}^{n}\sum_{i=1}^{n} A_{i,j}x_i x_j, \text{ so:}$$

$$\frac{\partial \alpha}{\partial x_k} = \sum_{j=1}^{n} A_{k,j}x_j + \sum_{i=1}^{n} A_{i,k}x_i \text{ for all } k$$

$$\therefore \frac{\partial \alpha}{\partial \vec{x}} = \vec{x}^T\mathbf{A}^T + \vec{x}^T\mathbf{A} = \vec{x}^T(\mathbf{A}^T + \mathbf{A}).$$

Notes:

- When taking the derivative with respect to the ${x_k}^{\text{th}}$ component, $\frac{\partial}{\partial x_k}(A_{i,j}x_i x_j)$ is non-zero if and only if $i = k$ or $j = k$.
- To understand $\sum_{j=1}^{n} A_{k,j}x_j$, think that for a fixed k we are "iterating" through the k^{th} **row** of the matrix $\mathbf{A}$ and multiplying the j^{th} element of that row with the j^{th} element of $\vec{x}$, which is the same as $\vec{x}^T\mathbf{A}^T$.
- To understand $\sum_{i=1}^{n} A_{i,k}x_i$, think that for a fixed k we are "iterating" through the k^{th} **column** of the matrix $\mathbf{A}$... so we have the same as $\vec{x}^T\mathbf{A}$.
- If $\mathbf{A}$ is symmetric,

$$\frac{\partial \alpha}{\partial \vec{x}} = \vec{x}^T(\mathbf{A} + \mathbf{A}) = 2\vec{x}^T\mathbf{A}.$$

Back to Linear Regression...

$$J(\vec{\theta}) = \frac{1}{2n}||(\mathbf{X}\vec{\theta} - \vec{y})||$$

If there are m training samples and n features, then:

- $\mathbf{X}$ is $m \times n$. A **row** represents 1 training sample
- $\vec{\theta}$ is $n \times 1$
- $\vec{y}$ is $m \times 1$

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

We wish to calculate $\nabla J(\vec{\theta})$.

$$
\begin{aligned}
\nabla J(\vec{\theta}) &= \frac{1}{2n}(2\vec{\theta}^T\mathbf{X}^T\mathbf{X} - \vec{y}^T\mathbf{X} - \vec{y}^T\mathbf{X} + 0) \\
&= \frac{1}{n}(\vec{\theta}^T\mathbf{X}^T\mathbf{X} - \vec{y}^T\mathbf{X})
\end{aligned}
$$

where we have used the fact that $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix, and is independent of $\vec{\theta}$ and $\vec{y}$. We set $\nabla J(\vec{\theta})$ to 0 to calculate the closed-form solution:

$$
\begin{aligned}
0 &= \frac{1}{n}(\vec{\theta}^T\mathbf{X}^T\mathbf{X} - \vec{y}^T\mathbf{X}) \\
\vec{\theta}^T\mathbf{X}^T\mathbf{X} - \vec{y}^T\mathbf{X} &= 0 \\
\vec{\theta}^T\mathbf{X}^T\mathbf{X} &= \vec{y}^T\mathbf{X} \\
\vec{\theta}^T &= \vec{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
\therefore \vec{\theta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y}
\end{aligned}
$$

where we have used the fact that $\mathbf{X}^T\mathbf{X}$ is invertible, and that the inverse of a symmetric matrix is also symmetric.