Linear Regression

Cost Function

$$egin{align} J(ec{ heta}) &= rac{1}{2m} \sum_i (\overrightarrow{x_i} \cdot ec{ heta} - y_i)^2 \ &= rac{1}{2m} \sum_i ((x_{i,1} heta_1 + x_{i,2} heta_2 + \dots + x_{i,n} heta_n) - y_i)^2 \end{split}$$

To take the gradient of $J(ec{ heta})$, we consider the partial derivative with respect to $heta_k$ for some $k \in 1 \dots n$

$$rac{\partial J}{\partial heta_k} = rac{1}{m} \sum_i (\overrightarrow{x_i} \cdot ec{ heta} - y_i) x_{i,k}$$

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

$$abla J(ec{ heta}) = rac{1}{m} \sum_i (\overrightarrow{x_i} \cdot ec{ heta} - y_i) \overrightarrow{x_i}$$

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

Matrix Version

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

<u>Proposition 1</u>: If $y(\vec{eta}) = \vec{x} \cdot \vec{eta}$, and \vec{x} does not depend on \vec{eta} , then:

$$rac{\partial y}{\partial ec{eta}} = ec{x}$$

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

Proof:

$$y(ec{eta}) = \sum_i x_i eta_i = x_1 eta_1 + x_2 eta_2 + \ldots + x_i eta_i$$

We take the partial derivative with respect to β_j for some $j \in 1 \dots i$:

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

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

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

<u>Proposition 2</u>: Let the scalar lpha be defined by $lpha=ec{y}^T\mathbf{A}ec{x}$ where:

- ullet $ec{y}$ is m imes 1
- ullet \vec{x} is n imes 1
- ullet **A** is m imes n, and independent of $ec{x}$ and $ec{y}$

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

Then:

$$lpha = ec{w}^T ec{x}$$

It follows from Proposition 1 that:

$$\therefore rac{\partial lpha}{\partial ec{x}} = ec{w}^T = ec{y}^T \mathbf{A}.$$

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

$$egin{aligned} lpha &= lpha^T = ec{x}^T \mathbf{A}^T ec{y}, \ & \therefore rac{\partial lpha}{\partial ec{y}} &= ec{x}^T \mathbf{A}^T. \end{aligned}$$

<u>Proposition 3</u>: Let the scalar lpha be defined by $lpha=ec{x}^T\mathbf{A}ec{x}$ where:

- ullet \vec{x} is n imes 1
- ullet ${f A}$ is m imes n and independent of ec x

Then:

$$rac{\partial lpha}{\partial ec{x}} = ec{x}^T (\mathbf{A} + \mathbf{A}^T)$$

Proof:

By definition, we have:

$$egin{aligned} lpha &= \sum_{j=1}^n \sum_{i=1}^n A_{i,j} x_i x_j, ext{ so:} \ &rac{\partial lpha}{\partial x_k} = \sum_{j=1}^n A_{k,j} x_j + \sum_{i=1}^n A_{i,k} x_i ext{ for all } k \ &\therefore rac{\partial lpha}{\partial ec{x}} = ec{x}^T \mathbf{A}^T + ec{x}^T \mathbf{A} = ec{x}^T (\mathbf{A}^T + \mathbf{A}). \end{aligned}$$

Notes:

- When taking the derivative with respect to the x_k th component, $\frac{\partial}{\partial x_k}(A_{i,j}x_ix_j)$ 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 **A** is symmetric,

$$rac{\partial lpha}{\partial ec{x}} = ec{x}^T (\mathbf{A} + \mathbf{A}) = 2 ec{x}^T \mathbf{A}.$$

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
- ullet $ec{ heta}$ is n imes 1
- ullet $ec{y}$ is m imes 1

$$egin{aligned} J(ec{ heta}) &= rac{1}{2n} (\mathbf{X} ec{ heta} - ec{y})^T (\mathbf{X} ec{ heta} - ec{y}) \ &= rac{1}{2n} (ec{ heta}^T \mathbf{X}^T - ec{y}^T) (\mathbf{X} ec{ heta} - ec{y}) \ &= rac{1}{2n} (ec{ heta}^T \mathbf{X}^T \mathbf{X} ec{ heta} - ec{ heta}^T \mathbf{X}^T ec{y} - ec{y}^T \mathbf{X} ec{ heta} + ec{y}^T ec{y}) \end{aligned}$$

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

$$egin{aligned}
abla J(ec{ heta}) &= rac{1}{2n}(2{ec{ heta}}^T\mathbf{X}^T\mathbf{X} - {ec{y}}^T\mathbf{X} - {ec{y}}^T\mathbf{X} + 0) \ &= rac{1}{n}({ec{ heta}}^T\mathbf{X}^T\mathbf{X} - {ec{y}}^T\mathbf{X}) \end{aligned}$$

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:

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

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.