# 1.

September 8, 2019 4:46 PM

## Multiple Linear Regression with Matrices

Given a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, where each $x^{(i)} = [x_0^{(i)} \cdots x_{m-1}^{(i)}] \in \mathbb{R}^m$ is a vector of features and $y^{(i)} \in \mathbb{R}$ is the response variable, we define $X \in \mathbb{R}^{n \times m}$ as the matrix of features such that

$\boxed{X_{i,j} \equiv x_j^{(i)}}$, $\forall\ i = 1, \dots, n$, $j = 0, \dots, m-1$ (i.e. $n$ observations and $m$ features)

$y = [y_1 \cdots y_n]^T \in \mathbb{R}^n$ is the response vector, and $\theta = [\theta^{(0)} \dots \theta^{(m-1)}]^T \in \mathbb{R}^m$ is the parameter vector or weight vector.

We also let $\epsilon = [\epsilon_1 \cdots \epsilon_n] \in \mathbb{R}^n$ be the error vector, and sometimes assume $\epsilon \sim \mathcal{N}(0, I)$, where $I$ is the identity matrix.

(Note that there is an intercept feature $x_0$, and the convention is that $x_0^{(i)} \equiv 1\ \forall i$).

Then, we assume that the true model is

$$\boxed{\underbrace{y}_{n\times 1} = \underbrace{X}_{n\times m}\underbrace{\theta}_{m\times 1} + \underbrace{\epsilon}_{n\times 1}}$$

In order to estimate $\hat{\theta}$, we want $y$ close to $X\hat{\theta}$, so that $(y - X\hat{\theta}) = 0$.

We define $\boxed{e = (y - X\hat{\theta})}$ to be the estimation error or residuals, intuitively, want these to be small. Since we don't care about the sign, we want to minimize, for each residual $\boxed{e_i = (y_i - (x^{(i)})^T \cdot \theta^{(i)})}$ want to minimize the square residual

$$\boxed{e_i^2 = (y_i - (x^{(i)})^T \cdot \theta^{(i)})^2}$$

Define our model function $h_\theta(\cdot)$

$$\boxed{h_\theta(x) = \theta^T x = x^T \theta = \sum_{j=0}^{m-1} x_j \theta_j}$$

$$\Rightarrow \boxed{h_\theta(X) = X\theta}$$

Informally, we want to minimize the Sum of Square Errors (SSE), i.e.

$$\boxed{\min_{\theta_0, \dots, \theta_{m-1}} SSE(h_\theta)}$$

where,

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)}))^2$$
$$= \sum_{i=1}^{n} (y^{(i)} - (x^{(i)})^T \theta^{(i)})^2$$

More formally, we want to minimize the Euclidean distance between our model and the truth model, i.e.

$$\boxed{\min_{\theta} \|y - X\theta\|_2^2}$$

where $\|\cdot\|_2$ is the L2-Norm (Euclidean Norm)

Note that $\boxed{\|x\| = \sqrt{\langle x, x \rangle}}$, where $\langle \cdot, \cdot \rangle$ is an inner product, e.g. here $\boxed{\langle x, x \rangle \simeq x^T \cdot x}$.

It follows that $\|x\|^2 = \langle x, x \rangle \simeq x^T \cdot x$, and so it our optimization objective becomes:

$$\boxed{\min_{\theta} \|e\|_2^2 = \min_{\theta} e^T e = \min_{\theta} (y - X\theta)^T (y - X\theta)}$$

(Note that $SSE = \sum_{i=1}^{n} e_i^2 = e^T e$)

Turn the page

2.

September 8, 2019 5:44 PM

Before optimizing, recall that for two matrices $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $AB = C \in \mathbb{R}^{m \times p}$, where

$$\boxed{c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}}, \text{ where } A_{i,j} = a_{ij} \begin{pmatrix} i=1,\dots,m \\ j=1,\dots,n \end{pmatrix}, \quad B_{i,j} = b_{i,j} \begin{pmatrix} i=1,\dots,m \\ j=1,\dots,p \end{pmatrix}$$

Now, recall we want to find $\hat{\theta}$ s.t.

$$\boxed{\hat{\theta} = \min_{\theta} \|y - X\theta\|_2^2}$$

so want to find a minimum by taking derivatives and setting the gradient to 0:

$$\boxed{\nabla_\theta \|y - X\theta\|_2^2 = \frac{\partial}{\partial \theta} \|y - X\theta\|_2^2 := 0}$$

and solving for $\theta$. Since this is a convex objective, we will find and absolute minimum. First note that

$$\|y - X\theta\|_2^2 = (y - X\theta)^T (y - X\theta)$$
$$= (y^T - \theta^T X^T)(y - X\theta)$$
$$= y^T y - y^T X\theta - \theta^T X^T y + \theta^T X^T X\theta \quad (1)$$

But,

$$\underbrace{\underbrace{y^T}_{1 \times n} \underbrace{X}_{n \times m} \underbrace{\theta}_{m \times 1}}_{1 \times 1} + \underbrace{\underbrace{\theta^T}_{1 \times m} \underbrace{X^T}_{m \times n} \underbrace{y}_{n \times 1}}_{1 \times 1}$$

$$= \sum_{j=0}^{m-1} \sum_{i=1}^{n} y_i x_{i,j} \theta_j + \sum_{i=1}^{n} \sum_{j=0}^{m-1} \theta_j x_{i,j} y_i$$

$$= 2 \sum_{i=1}^{n} \sum_{j=0}^{m-1} y_i x_{i,j} \theta_j \quad (2)$$

$$= 2 y^T X \theta = 2 \theta^T X^T y$$

so (1) becomes

$$\|y - X\theta\|_2^2 = y^T y - 2 y^T X\theta + \theta^T X^T X\theta \quad (3)$$

Then we have

$$\frac{\partial}{\partial \theta} \|y - X\theta\|^2$$
$$= \frac{\partial}{\partial \theta} (y^T y - 2 y^T X\theta + \theta^T X^T X \theta)$$

Now note

- $\frac{\partial}{\partial \theta} y^T y = 0 \quad (4)$
- From (2)

$$\frac{\partial}{\partial \theta_\ell} \sum_{i=1}^{n} \sum_{j=0}^{m-1} \theta_j x_{i,j} y_i$$
$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_\ell} \left( \theta_\ell x_{i,\ell} y_i + \sum_{j \neq \ell} \theta_j x_{i,j} y_i \right)$$
$$= \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta_\ell} \theta_\ell x_{i,\ell} y_i + 0 \right)$$
$$= \sum_{i=1}^{n} x_{i,\ell} y_i = (x^{(\ell)})^T \cdot y \quad (\ell = 0, \dots, m-1)$$

$$\Rightarrow \frac{\partial}{\partial \theta} (-2 y^T X \theta) = -2 X^T y \quad (5)$$

- $$\underbrace{\theta^T X^T X \theta}_{1 \times m \; m \times n \; n \times m \; m \times 1} = \sum_{i=1}^{n} \sum_{j=0}^{m-1} \sum_{\ell=0}^{m-1} \theta_j x_{i,j} x_{\ell,i} \theta_\ell$$

$$\Rightarrow \frac{\partial}{\partial \theta_k} \theta^T X^T X \theta$$
$$= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \sum_{j=0}^{m-1} \theta_j x_{i,j} x_{j,i} \theta_j$$
$$= \sum_{i=1}^{n} 2 \sum_{j=0}^{m-1} x_{i,j} x_{j,i} \theta_j, \quad (k = 0, \dots, m-1)$$

$$\Rightarrow \frac{\partial}{\partial \theta} \theta^T X^T X \theta = \boxed{2 X^T X \theta} \quad (6)$$

So from (1), (4), (5), (6) we get

Assuming $X^T X$ is invertible.

$$\frac{\partial}{\partial \theta} \|y - \theta X\|_2^2 = 0 - 2 X^T y + 2 X^T X \theta := 0$$

$$\Rightarrow 2 (X^T X) \theta = 2 X^T y \Rightarrow \boxed{\hat{\theta} = (X^T X)^{-1} X^T y}$$

which is the Least Squares Solution. (OLS).