Linear Regression Recap

- Training set = $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$
  - $x^{(i)}$: features $\in \mathbb{R}^d$
  - $y^{(i)}$: output $\in \mathbb{R}$
- Goal: Compute hypothesis $h_\theta : \mathbb{R}^d \to \mathbb{R}$ where
  $$h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots \theta_d x_d$$
  such that $h_\theta(x)$ is a good prediction for new feature vectors x. (i.e $x \neq x^{(i)} \; \forall i$)
- Under the assumption that the training set is "representative" of future feature vectors, it is reasonable to minimize the cost of prediction of the hypothesis on training set.

Linear Regression under squared loss is

Given Training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ find $\theta = (\theta_0, \ldots \theta_n)$ such that

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{2}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}_{\text{cost/loss on a single Traing data}}$$ is minimized

(Total cost on Traing)

$$= \frac{1}{2n} \sum_{i=1}^{n} \left(\left(\theta_0 + \sum_{j=1}^{d} \theta_j x_j^{(i)} \right) - y^{(i)} \right)^2$$

Math Recap

Convention: Vectors in $\mathbb{R}^d$ thought of as $(d \times 1)$ matrices (column vectors)

$$\|x\|_2^2 = \sum_{i=1}^{d} x_i^2 = x^T x$$

Partial Derivative $f : \mathbb{R}^d \to \mathbb{R}$ i.e $f(\theta_1 \dots \theta_d) \in \mathbb{R}$

$\frac{\partial f}{\partial \theta_i}$ – measures how f changes as θ_i changes (while $\theta_1, \theta_2 \dots \theta_{i-1}, \theta_{i+1} \dots \theta_n$ are fixed).

Gradient: $\nabla_\theta f = \begin{bmatrix} \frac{\partial f}{\partial \theta_1} \\ \frac{\partial f}{\partial \theta_2} \\ \vdots \\ \frac{\partial f}{\partial \theta_d} \end{bmatrix}$

Standard Gradients:

- For $f(\theta_1 \dots \theta_d) = \sum_{i=1}^{d} \theta_i x_i = x^T \theta = \theta^T x$,

  $\nabla_\theta f = x$

- For $f(\theta) = \theta^T A \theta$ where A is $d \times d$,

  $\nabla_\theta f = (A + A^T)\theta$.

$$X = \begin{bmatrix} | & & | \\ — & (x^{(1)})^T & — \\ & \vdots & \\ — & (x^{(i)})^T & — \\ | & & | \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \end{matrix} \Big\} n \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \Big\} d+1$$

$$X\theta = \begin{bmatrix} \theta_0 + \theta_1 x_1^{(1)} + \cdots \theta_d x_d^{(1)} \\ \vdots \\ \theta_0 + \theta_1 x_1^{(i)} + \cdots \theta_d x_d^{(i)} \\ \vdots \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(n)}) \end{bmatrix}$$

$$X\theta - y = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h_\theta(x^{(n)}) - y^{(n)} \end{bmatrix} \quad \text{where} \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(i)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

Linear Regression: Want to find $\theta \in \mathbb{R}^{d+1}$ s.t.

minimize $J(\theta) = \frac{1}{2n} \|X\theta - y\|_2^2$

$$= \frac{1}{2n} (X\theta - y)^T (X\theta - y)$$

Find θ s.t. $\nabla_\theta J(\theta) = 0$.

$$\nabla_\theta J(\theta) = \frac{1}{2n} \left[\nabla_\theta (X\theta - y)^T (X\theta - y) \right]$$

$$= \frac{1}{2n} \left[\nabla_\theta \{ \theta^T X^T X\theta - \underbrace{\theta^T X^T y}_{(X\theta)^T y} - \underbrace{y^T X\theta}_{y^T(X\theta)} + y^T y \} \right]$$

$$\left[\begin{array}{l} (A+B)^T = A^T + B^T \\ (AB)^T = B^T A^T \end{array} \right]$$

$$= \frac{1}{2n} \left[\nabla_\theta \{ \theta^T X^T X\theta - 2\theta^T X^T y + y^T y \} \right] \quad (x^T\theta = \theta^T x)$$

$$= \frac{1}{2n} \left[(X^T X + (X^T X)^T)\theta - 2X^T y \right]$$

$$= \frac{1}{n} \left[X^T X\theta - X^T y \right] = \frac{1}{n} X^T (X\theta - y)$$

$(X^T X)^T = X^T (X^T)^T = X^T X.$

Normal Equations:

$$\nabla_\theta J(\theta) = 0.$$

$$X^T X\theta - X^T y = 0$$

Find θ s.t. $X^T X\theta = X^T y$. (normal equations)

When $(X^T X)^{-1}$ is defined,

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

Moore-Penrose pseudo inverse: For matrix A, A^+

Solution $\theta = (X^T X)^+ X^T y$.

Proposition: Any θ^* s.t. $X^T X\theta^* = X^T y$. has the property

$$J(\theta^*) \leq J(\theta) \quad \forall \theta.$$

Proof: Let θ be any values to the parameters.

$$J(\theta) = \|X\theta - y\|_2^2 \geq \|X\theta^* - y\|_2^2 = J(\theta^*)$$

$$= \|(X\theta - X\theta^*) + (X\theta^* - y)\|_2^2$$

$$= [(X\theta - X\theta^*) + (X\theta^* - y)]^T [(X\theta - X\theta^*) + (X\theta^* - y)]$$

$$= \underbrace{\|X\theta - X\theta^*\|_2^2}_{\geq 0} + (X\theta - X\theta^*)^T (X\theta^* - y) + (X\theta^* - y)^T (X\theta - X\theta^*) + \underbrace{\|X\theta^* - y\|_2^2}_{J(\theta^*)}$$

$$= \|X\theta - X\theta^*\|_2^2 + 2\underbrace{(X\theta - X\theta^*)^T (X\theta^* - y)} + \|X\theta^* - y\|_2^2$$

$$(X\theta - X\theta^*)^T (X\theta^* - y) = \theta^T X^T X\theta^* - \theta^{*T} X^T X\theta^* - \theta^T X^T y + \theta^{*T} X^T y.$$

$$= \theta^T [X^T X\theta^* - X^T y] - \theta^{*T} \underbrace{(X^T X\theta^* - X^T y)}_{=0 \text{ because normal equations}}.$$

$$= 0$$

$$\Rightarrow J(\theta) \geq J(\theta^*)$$

minimize $J(\theta) = \| X\theta - y \|_2^2$

Each θ gives you a linear combination $a^{(j)}$.

$X\theta$ – Linear span of $a^{(j)}$.

$$X = \begin{bmatrix} 1 & - & x^{(1)T} & - \\ & & \vdots & \\ & & x^{(i)T} & \\ & & \vdots & \end{bmatrix}$$

$a^{(j)}$ – jth column.