HW1- Theory

* Vectors are denoted by **boldfaced** characters, matrices by **BOLDFACED CAPITAL** letters.

1. Prove Normal Equations:
   Given a training set $\mathcal{S} = \{\mathbf{X}, \mathbf{y}\}$, a linear hypothesis class $\{h_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j=1}^{N} \theta_j x_j\}$ and the mean squared error loss function:

$$\mathcal{L} = \frac{1}{2M} \sum_{i=1}^{M} \left(h_{\boldsymbol{\theta}}(\mathbf{x_i}) - y_i\right)^2$$

   prove that $\boldsymbol{\theta}$ that minimizes $\mathcal{L}$ satisfies:

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$$

   where: $\quad \mathbf{x_i}, \boldsymbol{\theta} \in \mathbb{R}^N, \quad \mathbf{y} \in \mathbb{R}^M, \quad \mathbf{X} = \begin{bmatrix} — \ \mathbf{x_1}^T \ — \\ — \ \mathbf{x_2}^T \ — \\ \vdots \\ — \ \mathbf{x_M}^T \ — \end{bmatrix}, \quad M \geq N$

2. Unique solution:
   Show that a unique solution for linear regression exists iff the features are not linearly dependent. Namely, show that a unique solution:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \mathcal{L} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$$

   exists iff $\mathbf{X}$ has full column rank.

solution:

Guy Ilan

1. $S = \{X, y\}$

hypothesis: $h_\theta(x) = \sum_{j=1}^{N} \theta_j x_j = x^T\theta$

Loss function: $L = \frac{1}{2M}\sum_{i=1}^{M}\left(h_\theta(x_i) - y_i\right)^2$

Proving that minimizing the MSE loss leads to $X^T X\theta = X^T y$:

Replace $h_\theta(x) = x^T\theta$:

$$L = \frac{1}{2M}\sum_{i=1}^{M}\left(x_i^T\theta - y_i\right)^2 \longrightarrow L = \frac{1}{2M}(X\theta - y)^T(X\theta - y)$$

$\rightarrow$ Minimizing the loss: $\frac{\partial L}{\partial \theta} = 0$

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

$$\rightarrow \frac{1}{2M}\left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y\right)$$

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

$$\rightarrow \frac{\partial L}{\partial \theta} = \frac{1}{2M}(2X^T X\theta - 2X^T y) \rightarrow \frac{\partial L}{\partial \theta} = \frac{1}{M}(X^T X\theta - X^T y)$$

$\downarrow$

finding min by setting gradient equal to zero:

$$\frac{1}{M}\left(X^T X\theta - X^T y\right) = 0 \quad \cdot M \quad \rightarrow \left(X^T X\theta - X^T y\right) = 0$$

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

thus minimizing θ satisfies: $X^T X\theta = X^T y$

2. Show unique solution for normal equation:

$X^T X\theta = X^T y \rightarrow$ assuming that inverse of $X^T X$ exists: $X^T X\theta = X^T y \quad \cdot (X^T X)^{-1}$

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

for $(X^T X)^{-1}$ to exist it must be invertible:

$X^T X$ is a $N \times N$ matrix and is invertible if and only if it has full rank.

By rank property $\text{rank}(X^T X) = \text{rank}(X)$.

Thus $X^T X$ is invertible if X has full rank.

X will have a full rank if all of its features are linearly independant.

$\text{rank}(X) = N$, implying $\text{rank}(X^T X) = N$

If X do not have a full rank, then $X^T X$ is not invertible and the normal equation will not have a unique solution:

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