

Linear Regression

1. Basics

- Linear hypothesis: $h(x) = \theta_1 x + \theta_0$, $\theta_i (i = 1, 2 \text{ for 2D cases})$.
- cost function:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2, \quad h_{\theta}(x) = \sum_{i=0}^n \theta_i x_i = \theta^T x$$

- best choice for $\theta = \arg \min_{\theta} J(\theta)$

2. Gradient Descent (GD) Algorithm

Algorithm.

1. Given a starting point θ in dom J
2. while convergence criterion is satisfied
 1. Calculate gradient $\nabla J(\theta)$
 2. Update $\theta \leftarrow \theta - \alpha \nabla J(\theta)$

θ Is usually initialized randomly, and α is so-called learning rate.

- For linear regression,

$$\begin{aligned} \theta_j &\leftarrow \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}, \quad \forall j = 0, 1, \dots, n, \quad x_0^{(i)} = 1 \\ \frac{\partial J(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2 \\ &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \sum_{i=1}^m \left(\sum_{j=0}^n \theta_j x_j^{(i)} - y^{(i)} \right)^2 \\ &= \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)}) x_j^{(i)} \end{aligned}$$

- Another commonly used form

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

- where m is introduced to scale the objective function to deal with differently sized training set.

3. Matrix Form

$$\begin{aligned} X &= \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}, \quad Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \\ J(\theta) &= \frac{1}{2} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2 = \frac{1}{2} (X\theta - Y)^T (X\theta - Y) \end{aligned}$$

- Minimize $J(\theta) = \frac{1}{2}(Y - X\theta)^T(Y - X\theta)$

$$\begin{aligned}
 \nabla_{\theta} J(\theta) &= \nabla_{\theta} \frac{1}{2}(Y - X\theta)^T(Y - X\theta) \\
 &= \frac{1}{2} \nabla_{\theta} \text{tr}(Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta) \\
 &= \frac{1}{2} \nabla_{\theta} \text{tr}(\theta^T X^T X\theta) - X^T Y \\
 &= \frac{1}{2} (X^T X\theta + X^T X\theta) - X^T Y \\
 &= X^T X\theta - X^T Y
 \end{aligned}$$

Theorem. Normal Equation

The matrix $A^T A$ is invertible if and only if the columns of A are linearly independent. In this case, there exists only one least-squares solution.

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