

Linear Regression

Basics

- Linear hypothesis: $h(x) = \theta_1 x + \theta_0$, θ_i ($i = 1, 2$ 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 $\theta = \arg \min_{\theta} J(\theta)$

Gradient Descent (GD) Algorithm

Algorithm.

```
Given a starting point \theta in dom J
while convergence criterion is satisfied
    Calculate gradient \nabla J(\theta)
    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 $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$.
- m is introduced to scale the objective function to deal with differently sized training set.

Matrix Form

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

- 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.

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