Sid

Overfitting – "High Variance

- fail to generalize to new examples

1) Reduce # of features

2) Regularization
   - keep features but reduce magnitude/values of θ
     - "simpler" hypothesis
     - less prone to overfitting

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \underline{\lambda}\sum_{j=1}^{n}\theta_j^2\right]$$

regularization parameter

Regularized Linear Regression

θ_0 not regularized

Gradient Descent

$$\theta_0 = \theta_0 - \alpha\frac{1}{m}\sum_{i=0}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_0^{(i)}$$

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]$$

$$\theta_j := \underbrace{\theta_j\left(1 - \alpha\frac{\lambda}{m}\right)}_{< 1} - \alpha\frac{1}{m}\sum_{i=1}^{m}h_\theta(x^{(i)}) - y^{(i)})\,x_j^{(i)}$$

Normal Equation : $\theta = \left(x^T X + \lambda\underbrace{\begin{bmatrix}0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1\end{bmatrix}}_{(n+1)\times(n+1)}\right)^{-1} X^T y$

Regularized Logistic Regression

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right]$$

$$+ \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

$$\theta_j := \theta_j - \alpha\left[\frac{1}{m}\sum_{i=1}^{m}\left(\underbrace{h_\theta(x^{(i)})}_{h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}} - y^{(i)}\right)x_j^{(i)} + \frac{\lambda}{m}\theta_j\right]$$