Regularization

- Small value of parameter ⇒ simpler hypothesis ⇒ less overfitting

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

$\lambda\sum_{j=1}^{n}\theta_j^2$: regularization term

if λ too large then underfit

high penalty → ∴ as if model don't contain $\theta_1, \theta_2, \theta_3 \ldots$

∴ Now the problem is finding the optimal λ that minimizes $J(\theta)$

cost fn → $J(\theta)$

penalize $\theta_1 \ldots \theta_n$ but not θ_0

(ex) Regularized linear regression:

Gradient descent $\theta_j = \theta_j\left(1 - \alpha\frac{\lambda}{m}\right) - \alpha\frac{1}{m}\sum_{i=1}^{m}\left[h_\theta(x_i) - y_i\right]x_j$

$1 - \alpha\frac{\lambda}{m} < 1$ ex) 0.99

θ_j gets replaced by $0.99\,\theta_j$ ⇒ make θ_j smaller

→ regular gradient descent

→ So basically you are just shrinking the parameter by a bit when you do regularization

Normal Equation

$\theta = \left(X^TX + \lambda\begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}\right)^{-1} X^Ty$ ⇒ yield global optimum

guaranteed to be invertible matrix

(ex) Regularized Logistic Regression

$h_\theta(x) = g(\theta_0 + \theta_1x_1 + \theta_2x^2 + \cdots)$

with regularization, even lots of features $\theta_1 \ldots \theta_n$ is OK