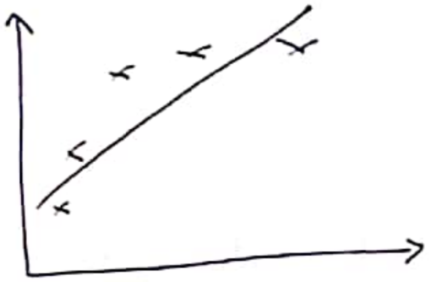
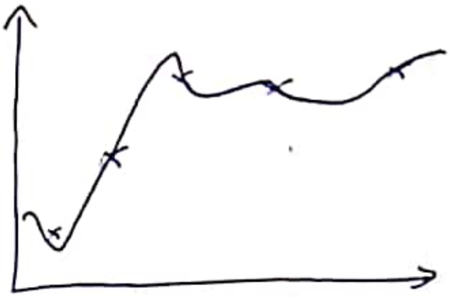
7

Regularization

7.1) Problem of Over fitting





$\theta_0 + \theta_1 x$

$\Rightarrow$ "Under fit"

$\Rightarrow$ It has "high bias"

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$\Rightarrow$ "Overfit"

$\Rightarrow$ It has high variance

Overfitting

↳ If we have too many features, the learned hypothesis may fit the training set very well, but fails to generalize to example.

$\Rightarrow$ Overfitting can also be found in Classification.

* Addressing Overfitting

1. Reduce number of features
   → Manually select which features to keep.
   → Model selection algorithm

2. Regularization

→ Keep all the features, but reduce magnitude of parameters θ_j.

7.2) Cost Function

⇒ Small values for parameters $\theta_0\ \theta_1\ \theta_2 \cdots \theta_n$

→ "Simpler" hypothesis.

→ Less prone to overfitting.

⇒ As we dont know which parameter is less important, so we will penalise all the parameters:-

$$J(\theta) = \frac{1}{2m}\left[\underbrace{\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2}_{\text{Objective :: fit the training data well}} + \underbrace{\lambda}_{\text{regularization Parameter}} \underbrace{\sum_{j=1}^{n}\theta_j^2}_{\text{Objective: Keep the parameter small}}\right]$$

{Objective :: fit the training data well}

regularization Parameter

{Objective: Keep the parameter small}

{It controls the trade off between the two objective}

7.3) Regularized - Normal equation

$$X = \begin{bmatrix} (x^{(1)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix}_{m \times (n+1)} \qquad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}_{m \times 1}$$

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

7.4) Regularized - logistic regression

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