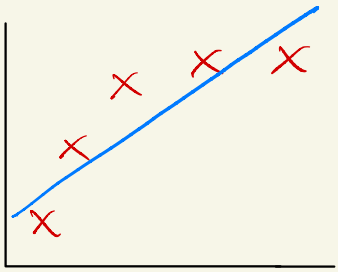
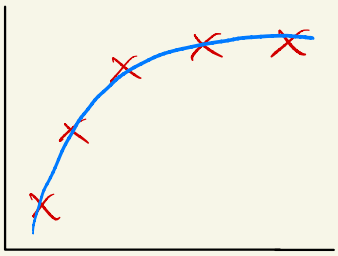
May 16, 2021

e.g. Linear Regression

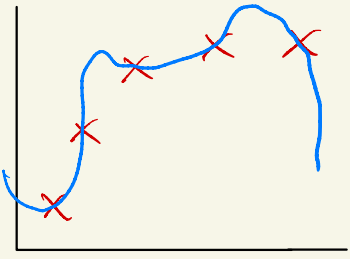


$\theta_0 + \theta_1 x$

↳ underfit / high bias



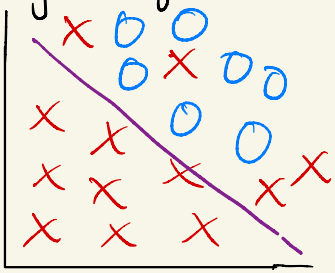
$\theta_0 + \theta_1 x + \theta_2 x^2$



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

↳ overfit / high variance

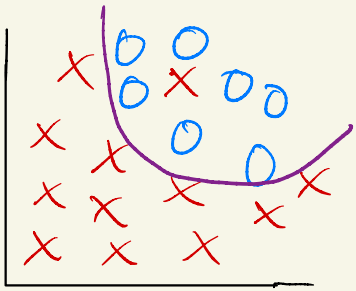
e.g. Logistic Regression



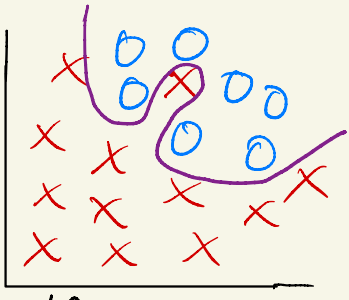
$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

↳ sigmoid

↳ underfit



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1^2 + x_2^2 + \theta_6 x_1^2 + x_2^3 \ldots)$

↳ overfit

Overfitting occurs when the hypothesis has too many features and fits the training data well, but fails to generalize to new examples.

To Address Overfitting:

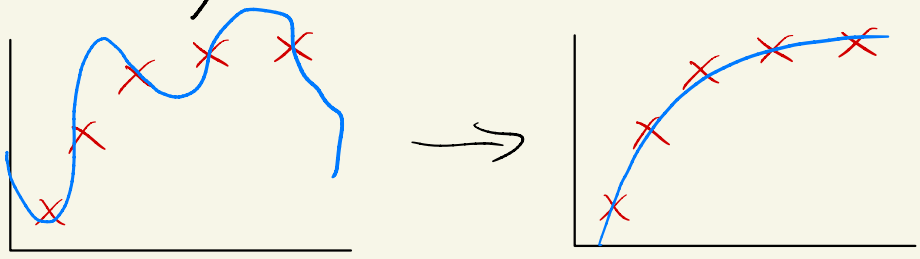
1. Reduce number of features
   - ↳ manually select which features to keep
     - → drawback is you lose some of the data
   - ↳ model selection algorithm

2. Regularization
   - ↳ keep all the features, but reduce the magnitude/values of parameters θ_j
     - → works well when we have a lot of features contributing to predicting y

Regularization:

eg: We want to make $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ more quadratic.

↳ We modify our cost function



We do this by making $\theta_3 x^3 + \theta_4 x^4$ really small.

↳ $\min \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + 1000\,\theta_3^2 + 1000\,\theta_4^2$

↳ by inflating the cost of θ_3 and θ_4, the cost function reduces their values to get to zero

i.e. $\theta_3 \approx 0$, $\theta_4 \approx 0$

✱ Having smaller values for the parameters results in a simpler hypothesis and decreases chances of overfitting.

Regularized Linear Regression:

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

↳ ✱ Doesn't affect θ_0

λ - regularization parameter (a scalar)

✱ However, if λ is set to be too large, then $\theta_1 \approx 0$, $\theta_2 \approx 0$... and $h_\theta(x) = \theta_0$, underfits.

↳ flat line —

Regularized Cost Function:

※ For linear regression

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

↳ gradient descent now solves for the min of this

※ Adding $\lambda\sum_{j=1}^{n}\theta_j^2$ to the loss is called

L2 Regularization

Gradient Descent:

repeat {

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

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

}

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

$\left(1 - \alpha\frac{\lambda}{m}\right) < 1$ since $\alpha > 1, \lambda > 1$

$\alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$ original g.d. update

Normal Equation:

Previously: $\theta = (X^T X)^{-1} X^T y$

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

※ Always invertible!

※ Regularization also takes care of when $m \leq n$ makes $(X^T X)^{-1} X^T y$ non-invertible.

Regularized Logistic Regression:

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

Gradient Descent:

Same as for regularized linear regression, but $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$.