

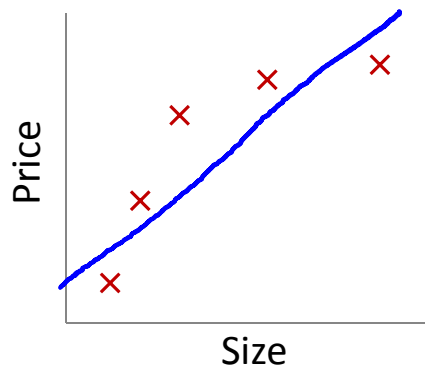


Machine Learning

Regularization

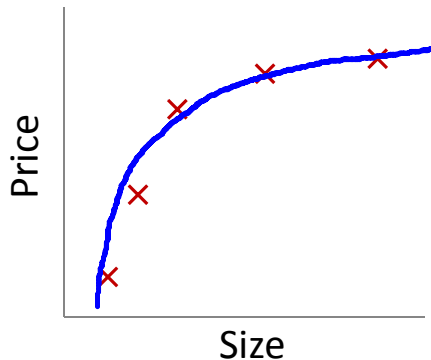
The problem of
overfitting

Example: Linear regression (housing prices)



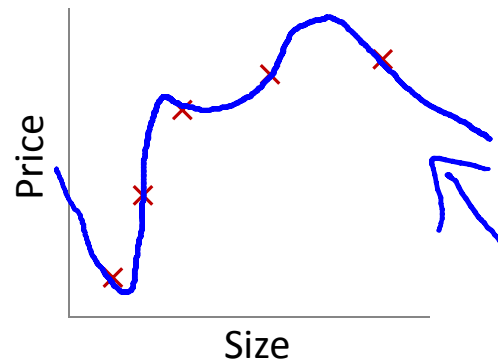
$$\rightarrow \theta_0 + \theta_1 x$$

"Underfit" "High bias"



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

"Just right"

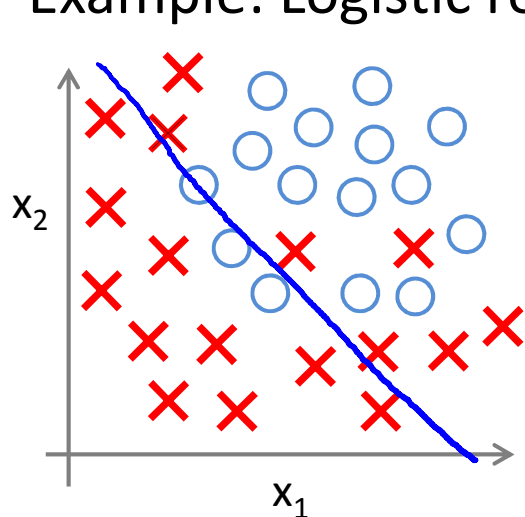


$$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

"Overfit" "High variance"

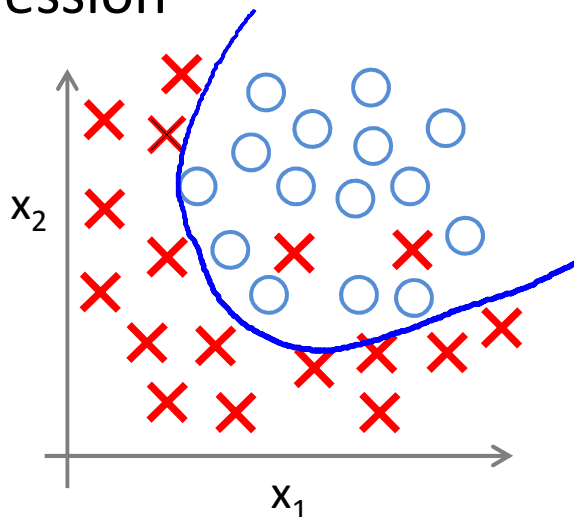
Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).

Example: Logistic regression



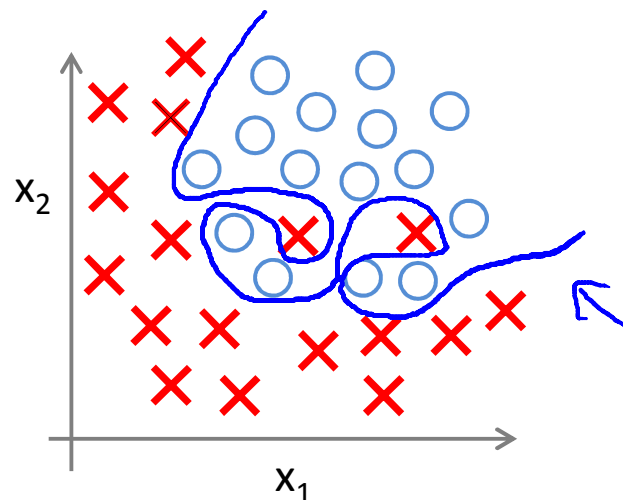
$\rightarrow h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$
 (g = sigmoid function)

↩
 "Underfit"



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

↗



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
 $+ \theta_3 \underline{x_1^2 x_2} + \theta_4 \underline{x_1^2 x_2^2}$
 $+ \theta_5 \underline{x_1^2 x_2^3} + \theta_6 \underline{x_1^3 x_2} + \dots)$

↖
 "Overfit"

Addressing overfitting:

x_1 = size of house

x_2 = no. of bedrooms

x_3 = no. of floors

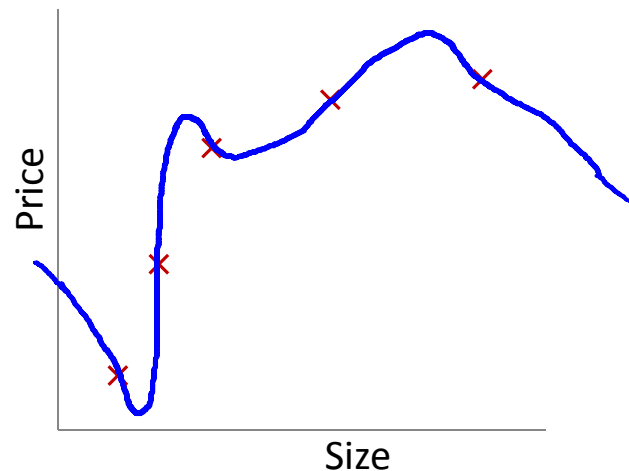
x_4 = age of house

x_5 = average income in neighborhood

x_6 = kitchen size

\vdots

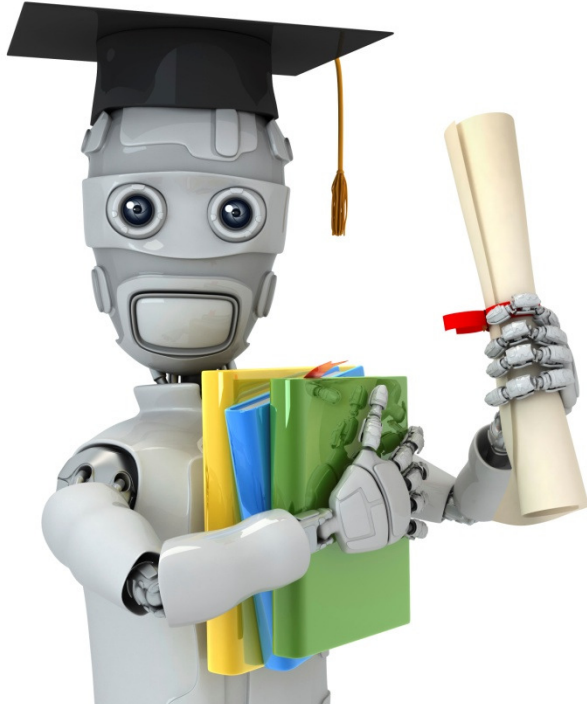
x_{100}



Addressing overfitting:

Options:

1. Reduce number of features.
 - — Manually select which features to keep.
 - — Model selection algorithm (later in course).
2. Regularization.
 - — Keep all the features, but reduce magnitude/values of parameters θ_j .
 - Works well when we have a lot of features, each of which contributes a bit to predicting y .

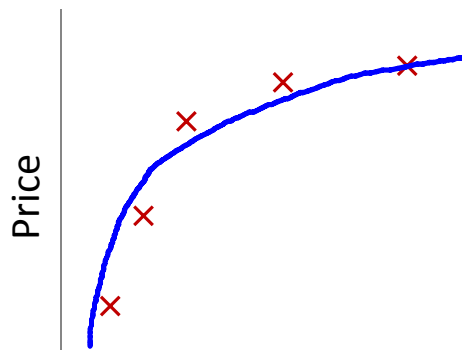


Machine Learning

Regularization

Cost function

Intuition



Size of house

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



Size of house

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

Two pink arrows point upwards from the θ_3 and θ_4 terms, which are crossed out with blue lines.

Suppose we penalize and make θ_3, θ_4 really small.

$$\rightarrow \min_{\theta} \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \underbrace{1000 \theta_3^2}_{\theta_3 \approx 0} + \underbrace{1000 \theta_4^2}_{\theta_4 \approx 0}$$

The entire equation is underlined in blue. The terms $\theta_3 \approx 0$ and $\theta_4 \approx 0$ are written below the corresponding terms in the equation, underlined in pink.

Regularization.

Small values for parameters $\theta_0, \theta_1, \dots, \theta_n$

- “Simpler” hypothesis
- Less prone to overfitting

$$\rightarrow \boxed{\theta_3, \theta_4} \approx 0$$

Housing:

- Features: x_1, x_2, \dots, x_{100}
- Parameters: $\theta_0, \theta_1, \theta_2, \dots, \theta_{100}$

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

~~θ_0~~ $\theta_1, \theta_2, \theta_3, \dots, \theta_{100}$ ~~θ_0~~

Regularization.

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

$\min_{\theta} J(\theta)$

regularization parameter

