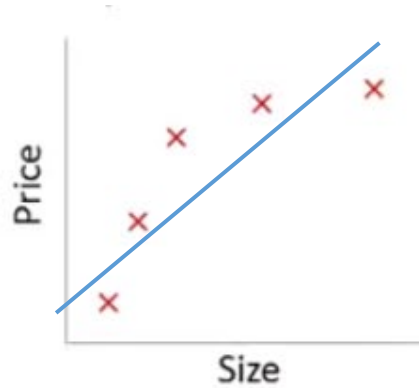


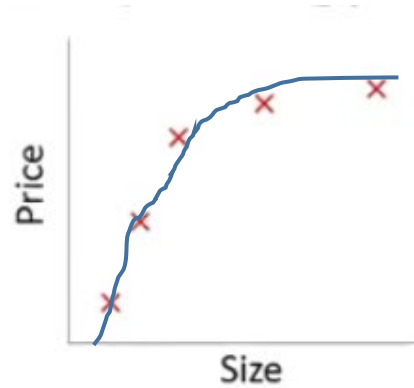
Regularization

Recall the prediction problem using regression models (price of house)

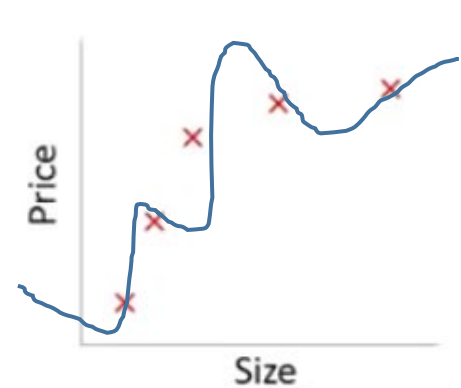


$$\theta_0 + \theta_1 x$$

“Underfit”
“High bias”



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



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

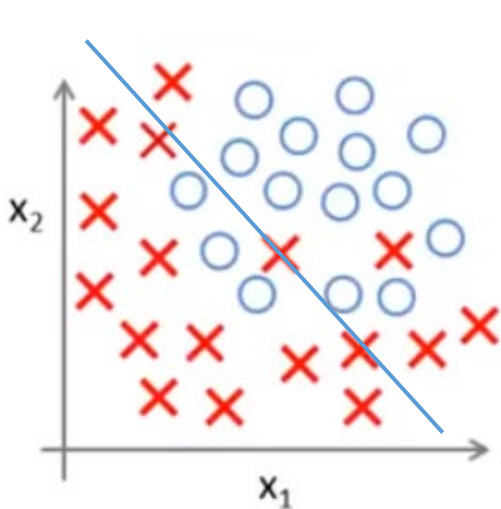
“Overfit”
“High variance”

Overfitting: if we have too many features, the learned hypothesis may fit the training set very well, i.e.,

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \approx 0$$

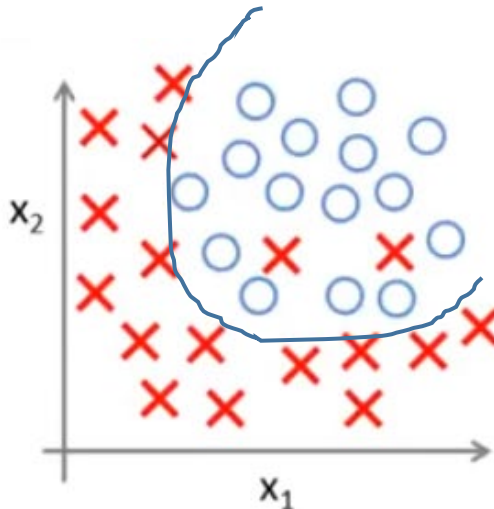
but fail to generalize to new examples, i.e., not able to make good predictions on new examples

Example: predict class labels using logistic regression

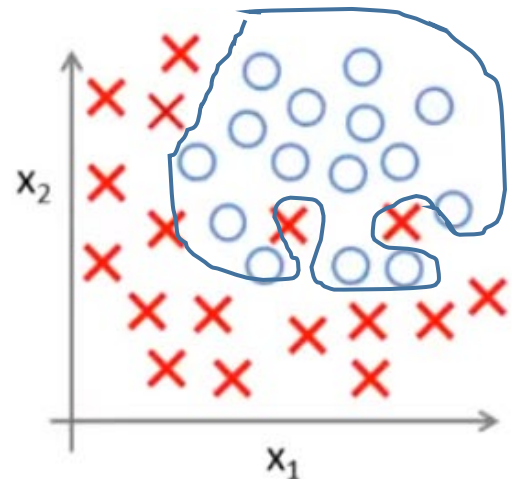


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

underfit



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



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

overfit

Ideas to overcome overfitting

1. Reduce the number of features
 - Manually select useful features to keep
 - Use model selection algorithms to determine model complexity
2. Regularization (error loss + regularization term)
 - Use all features but reduce magnitudes/values of some parameters θ . This works well for a large number of features so that each feature contributes a bit to predicting y
3. Drop out, early stopping, augmenting data...

Mathematical Insights on Regularization

- Regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting.
- Regularization can be applied to objective functions in ill-posed optimization problems.
- The regularization term, or penalty, imposes a cost on the optimization function.

Regularization:

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

- “simpler” hypothesis
- less prone to overfitting

Housing example

- Features: $x_0 x_1 \dots x_{100}$
- Parameters: $\theta_0 \theta_1 \dots \theta_{100}$

Introduce a new cost term into the cost function: the larger the θ'_j s, the higher the cost

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

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

Where for real number $p \geq 1$, the p -norm for vector θ

$$\|\theta\|_p = \left(\sum_{i=1}^n |\theta_i|^p \right)^{1/p}$$

$p = 1$, \rightarrow Lasso regression
(L_1 regularization)

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

$p = 2$, \rightarrow Ridge regression
(L_2 regularization)

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

For $0 < p < 1$, quasi-norm for vector θ , causes more elements of θ to be zeroed out

L2 regularization:

Introduce a new cost term into the cost function: the larger the θ'_j s, the higher the cost

Parameters θ are determined such that $J(\theta)$ is minimized

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

Recall the linear regression problem with hypothesis below,

$$h_{\theta}(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

The parameters $\theta_0 \theta_1 \dots \theta_n$ are determined from minimizing the following cost function,

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

Solution:

Update parameters $\theta_0 \theta_1 \dots \theta_n$ using gradient descent until convergence

Simultaneously update θ_j , $j = 0, 1, \dots, n$, according to

$$\theta_j := \theta_j - \eta \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1, \dots, \theta_n)$$

η is the learning rate

Simultaneously update θ_0 and $\theta_j, j = 1, 2, \dots, n$, according to

$$\theta_0 := \theta_0 - \eta \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$\theta_j := \theta_j - \eta \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Now consider regularized linear regression –
how to determine the parameters in a given
hypothesis?

With the regularization term,
simultaneously update θ_0 and $\theta_j, j = 1, 2, \dots, n$, according to

$$\theta_0 := \theta_0 - \eta \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)}$$
$$\theta_j := \theta_j - \eta \left[\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right]$$

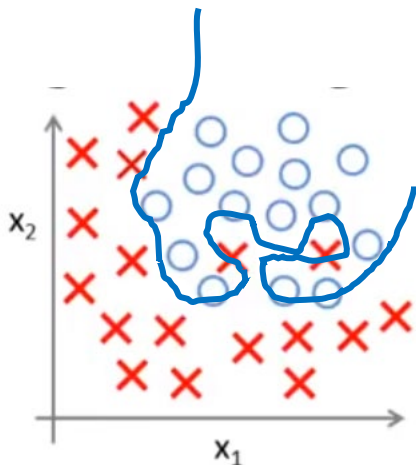
Notice that:

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

$\left(1 - \eta \frac{\lambda}{m} \right)$ is usually less than 1 (weight decay) \rightarrow smaller θ_j

The idea of regularization can be used to regulate other regression models

Regularized logistic regression



Consider a logistic regression model,

$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \dots)$$

To regulate parameters $\theta_1 \dots \theta_n$, minimize the following cost function

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

With the regularization term,
simultaneously update θ_0 and $\theta_j, j = 1, 2, \dots, n$, according to

$$\theta_0 := \theta_0 - \eta \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_0^{(i)}$$
$$\theta_j := \theta_j - \eta \left[\frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right]$$

Note that this appears the same as in the case of linear regression, but they are different. Why?

Consider a regularized (generalized) linear regression problem,

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

What if λ is very large, as large as say $\lambda = 10^{10}$?

