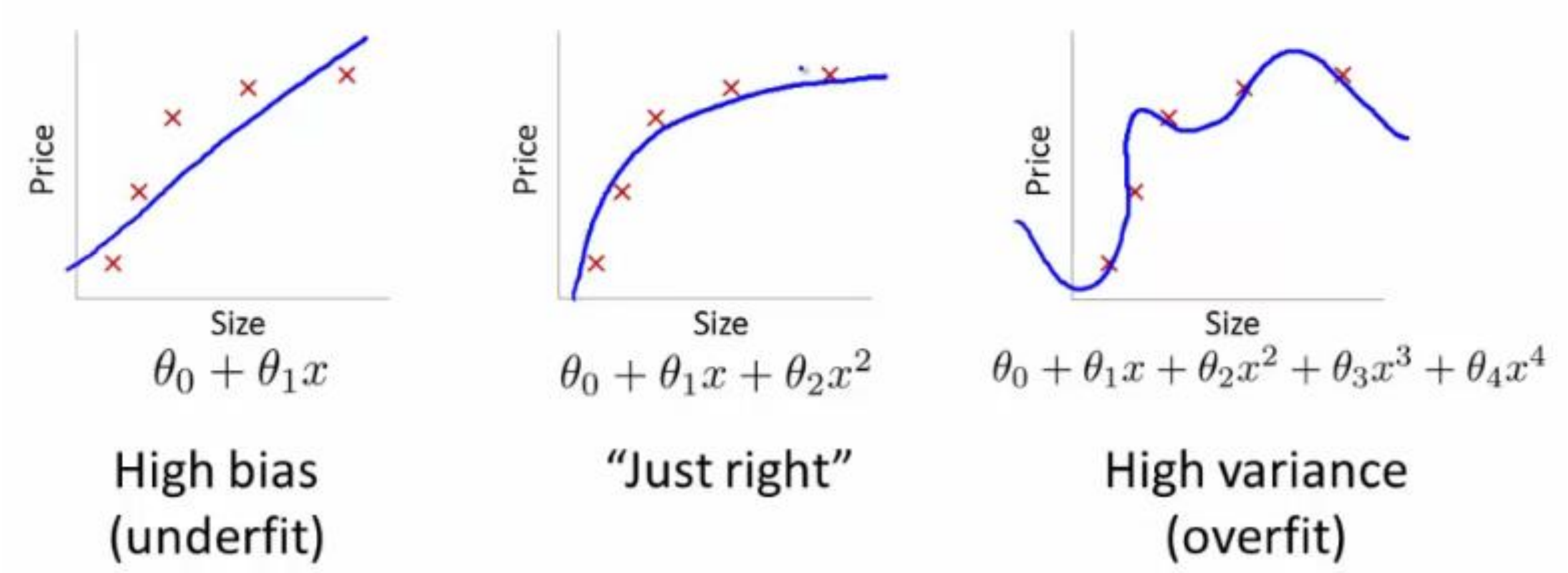
Regularized Regression Models

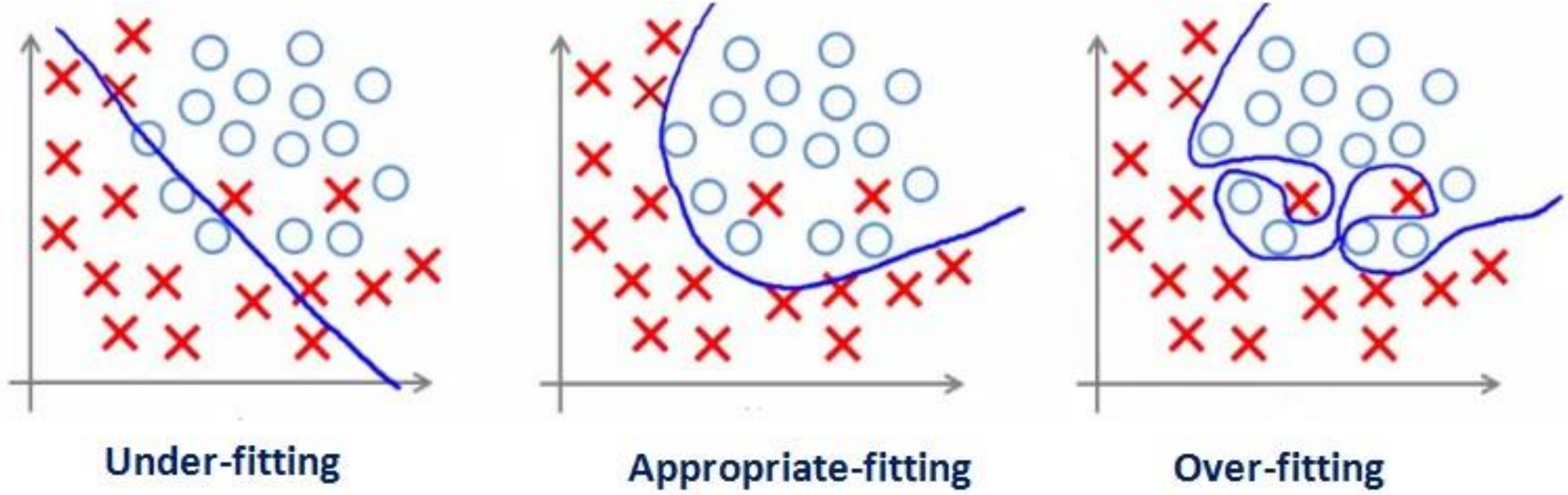
datasciencedojo
unleash the data scientist in you

Linear regression fitting example



datasciencedojo

Logistic regression fitting example



(too simple to explain the variance)

(forcefitting -- too good to be true)

Overfitting

Overfitting when……

- Complex model, too many features, not enough training samples.

How to address overfitting??

- Go through each features to decide which to keep.

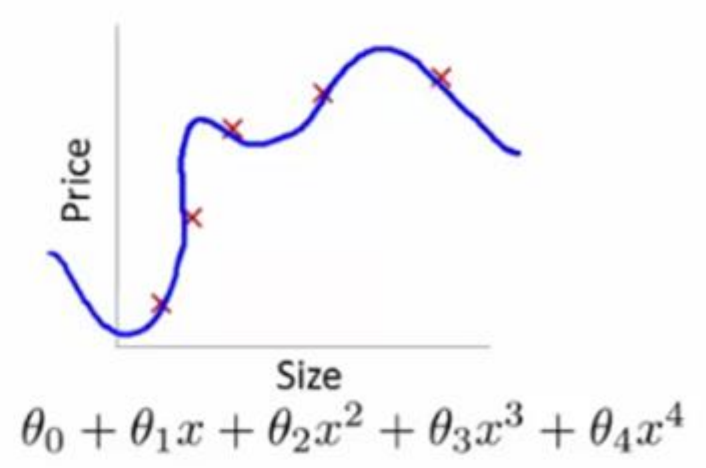
- Use model selection algorithm to automatically choose features.

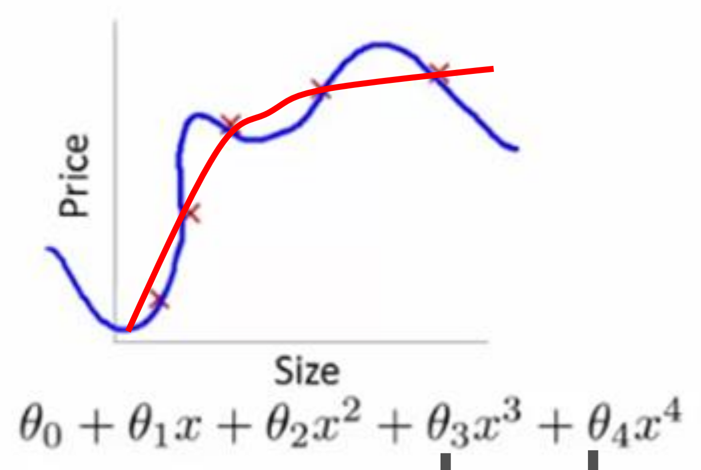
Idea of Regularization

- Keep all the features, but reducing their magnitude of parameter effects in model.
- Shrink θ_j parameters

datasciencedojo

Regularized regression intuition





Close to zero

- Goal: To minimize cost function θ_j

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

- Suppose we penalize and make θ_3 *and* θ_4 very small

datasciencedojo

Regularization

- Two common types of regularization in linear regression

- L2 regularization (a.k.a ridge regression)

$$\sum_{j=1}^{N}(y_j - \sum_{i=0}^{d}\theta_i \cdot x_i)^2 + \lambda \sum_{i=1}^{d}\theta_i^2$$

- L1 regularization (a.k.a lasso regression)

$$\sum_{j=1}^{N}(y_j - \sum_{i=0}^{d}\theta_i \cdot x_i)^2 + \lambda \sum_{i=1}^{d}|\theta_i|$$

Regularized-Ridge regression

Regularization by shrink θ_j smaller values, as a result

- "less complex" hypothesis function without eliminating features
- More protection from overfitting.

L2: Ridge regression

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

$$\min_\theta J(\theta)$$

Regularized-Ridge regression

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

$$\min_\theta J(\theta)$$

- Goal 1: find the best fit
- Goal 2: keep parameter θ_j small
- λ is regularization parameter to controls a trade off

datasciencedojo

Regularized-Ridge regression

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

$$\min_\theta J(\theta)$$

- If λ is too large, θ_j become too small, as if features have no effect in predicting response.

- If λ is too small, θ_j are not regularized.

Questions?