TAMS ARTIFICIAL INTELLIGENCE SOCIETY
2021 - 2022

Overfitting/Underfitting

10/6/2021

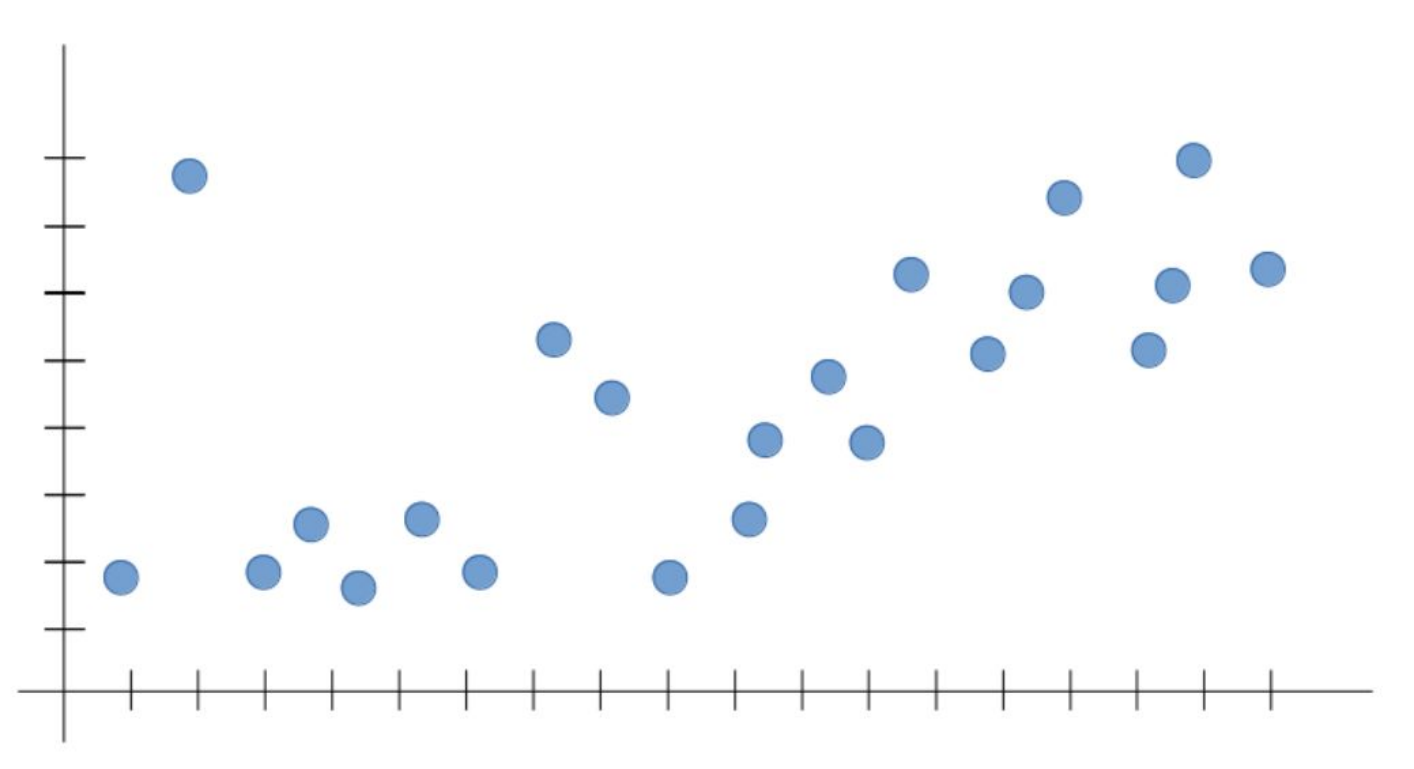
Lesson Plan

- Definition
- Bias/Variance
- How to Address Overfitting/Underfitting
- Cross-validation
- Learning Curve

Definition

- Goal of ML: generalize well to new data
- A model that generalizes well is said to neither overfit or underfit
- Overfitting: model is too complex (fits the training data well but not the testing data)
- Underfitting: model is too simple (does not fit both the training and testing data well)

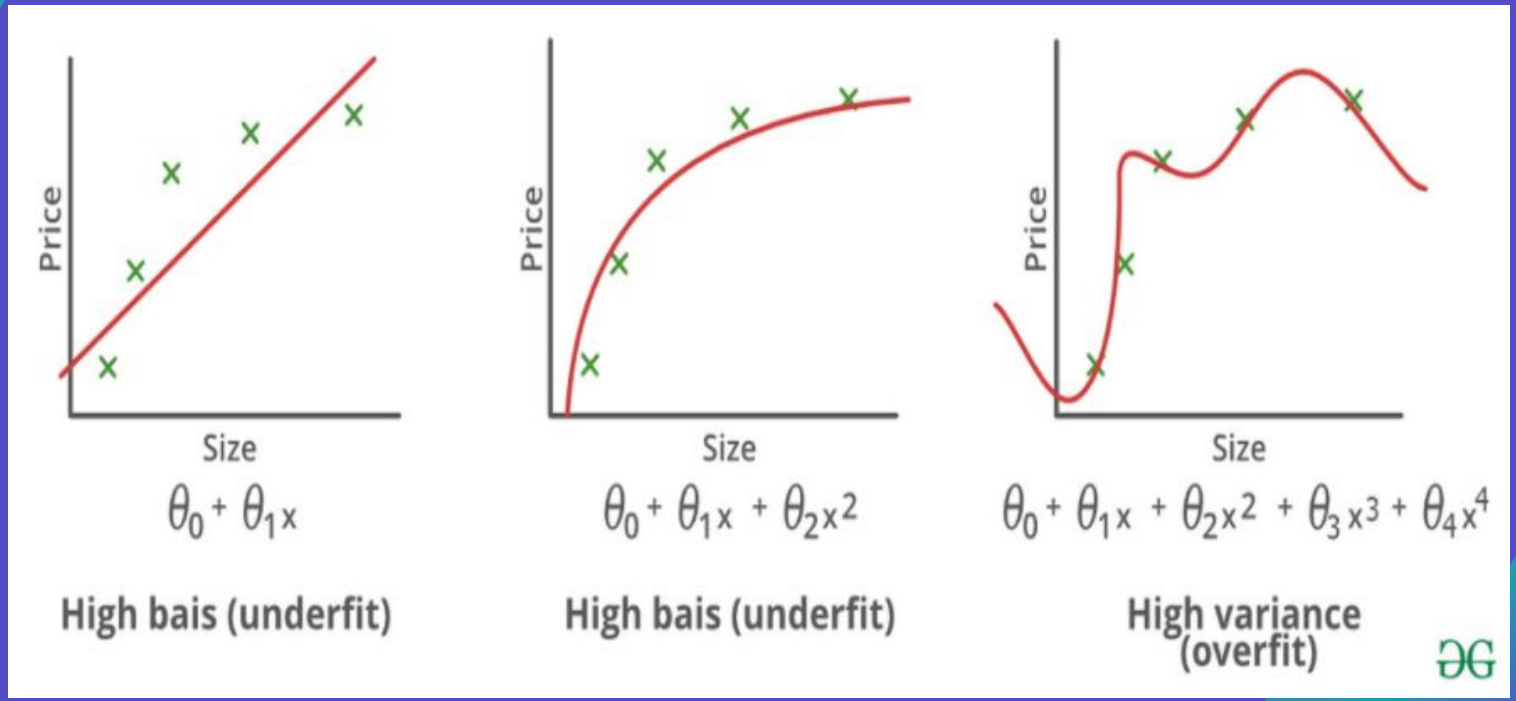
Example (1)



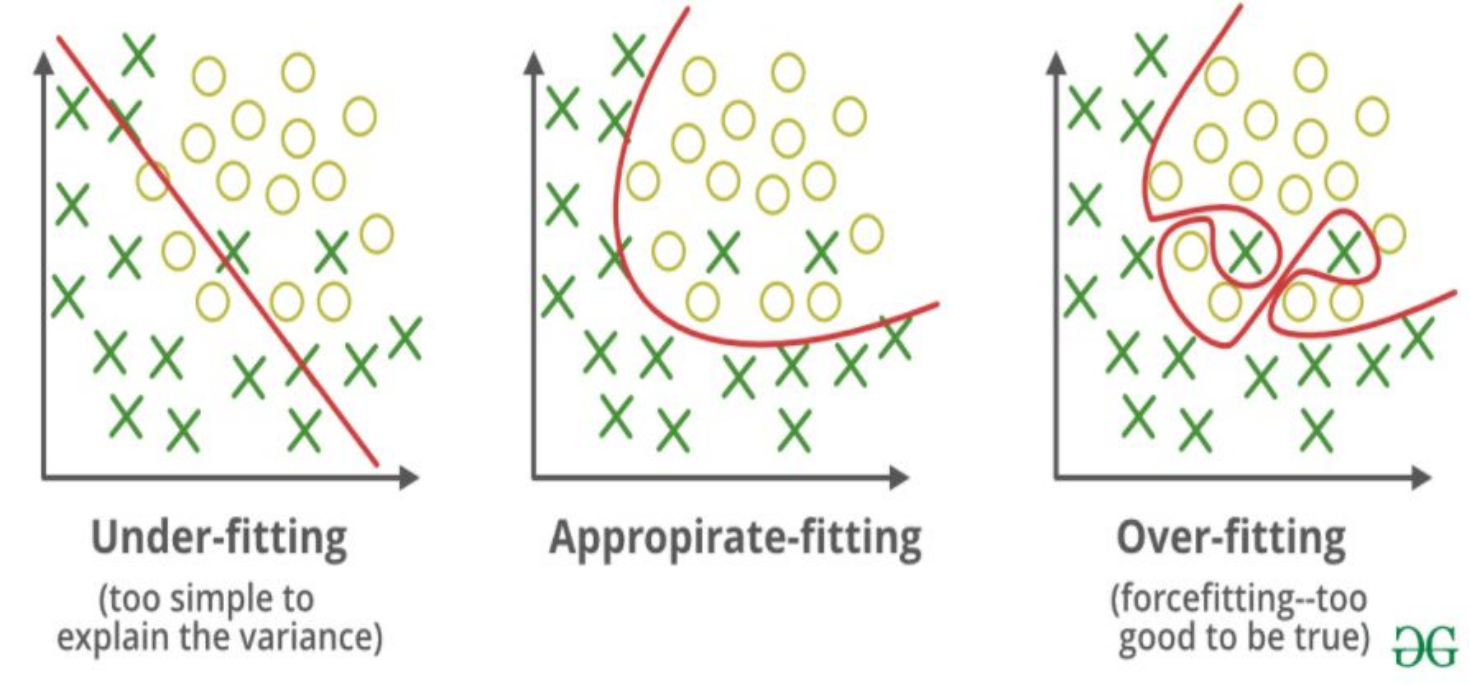
Bias/Variance

- Bias: Simplifying assumptions made by a model to learn the hypothesis function easier
- Variance: The model’s sensitivity to fluctuations in the training data
- High bias, low variance: Underfit
- High variance, low bias: Overfit

Example (2) - Regression



Example (3) - Classification



How to Address Overfitting/Underfitting

- Overfitting
  - Reduce the number of features used
  - Apply regularization (controls the size of the parameters)
  - Cross-validation
- Underfitting
  - Increase the number of features used (increase complexity of the model)
  - Reduce regularization

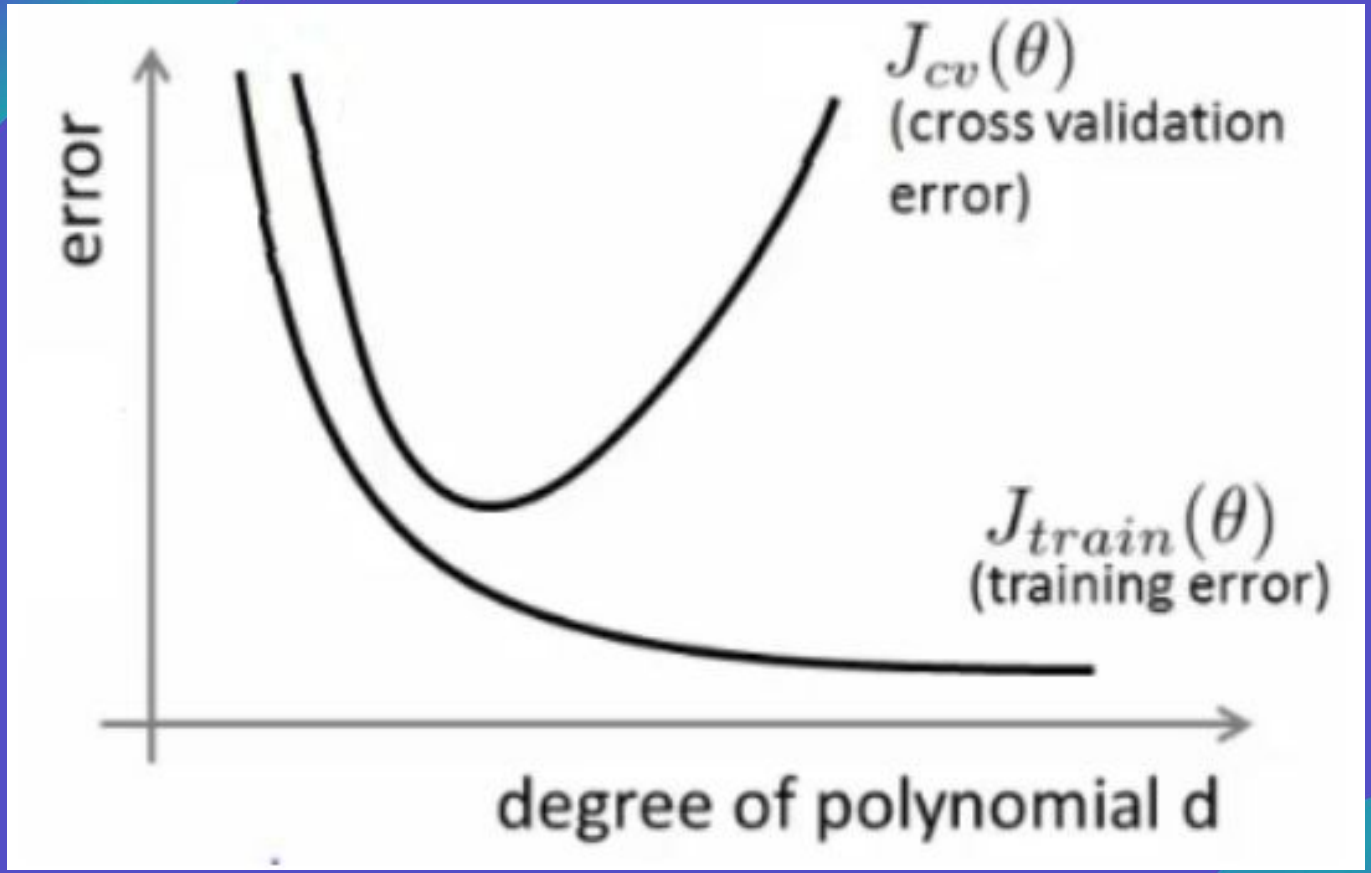
Cross-validation

- Split data into train/cross-validation/test sets
- Find parameter values from train set
- Get cost values for each model from cross-validation set
- Select the model that returns the lowest cost in the cross-validation stage and test in on the test set.

1. $h_\theta(x) = \theta_0 + \theta_1 x$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

$\vdots$

10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$


$J_{cv}(\theta)$
(cross validation error)
$J_{train}(\theta)$
(training error)
error
degree of polynomial d

Learning Curve

- Plot cross validation error and training error against number of training data