

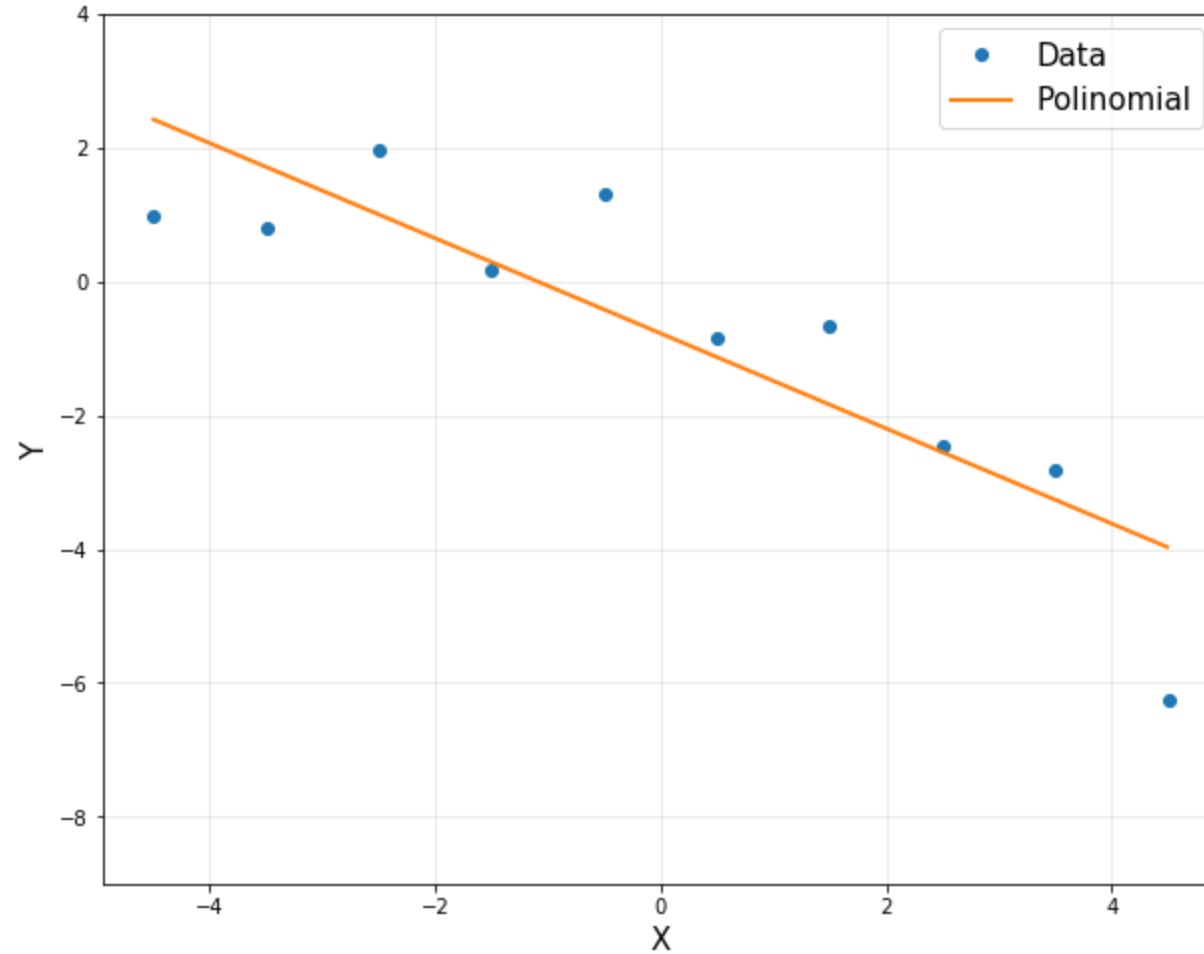


Optimization: Overfitting

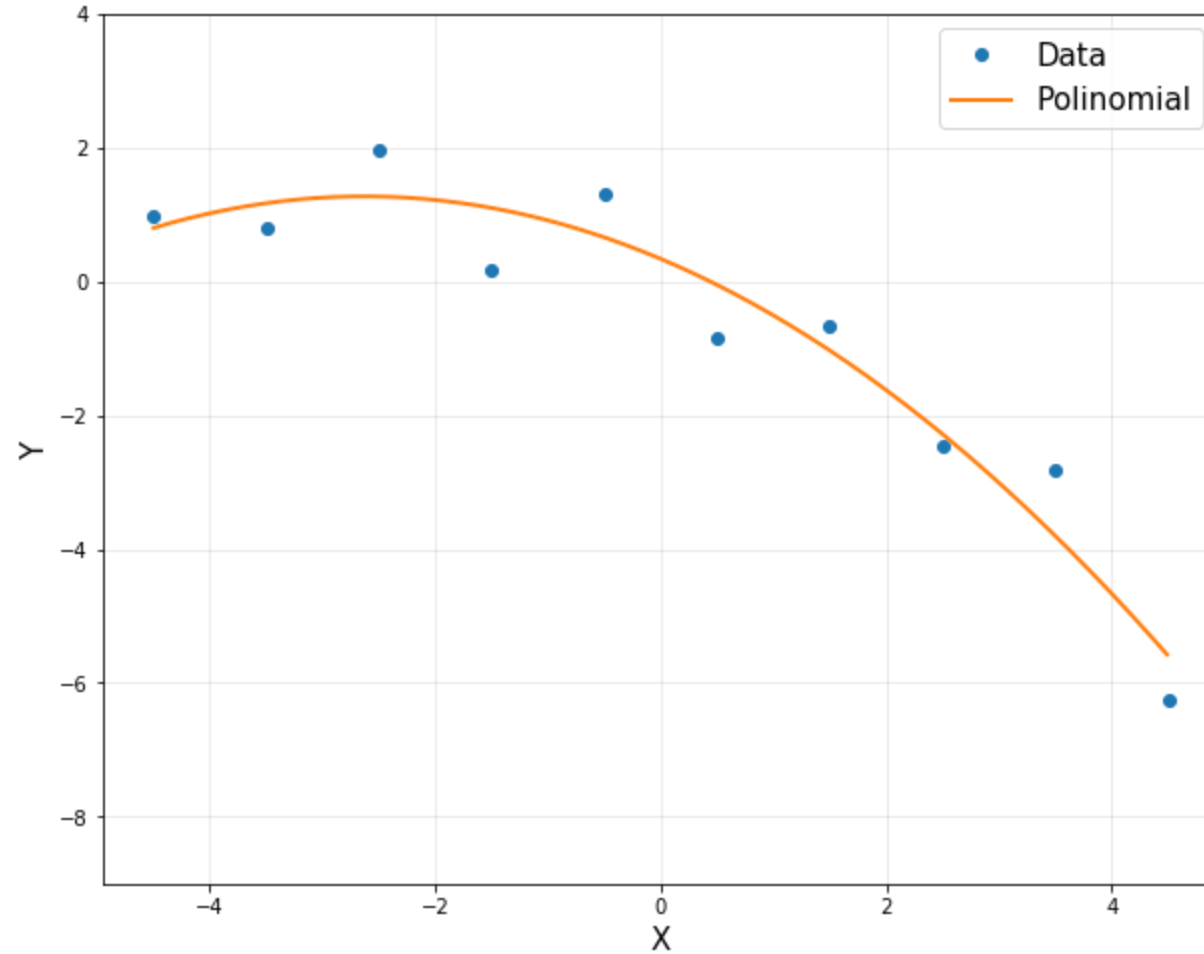
Industrial AI Lab.

Prof. Seungchul Lee

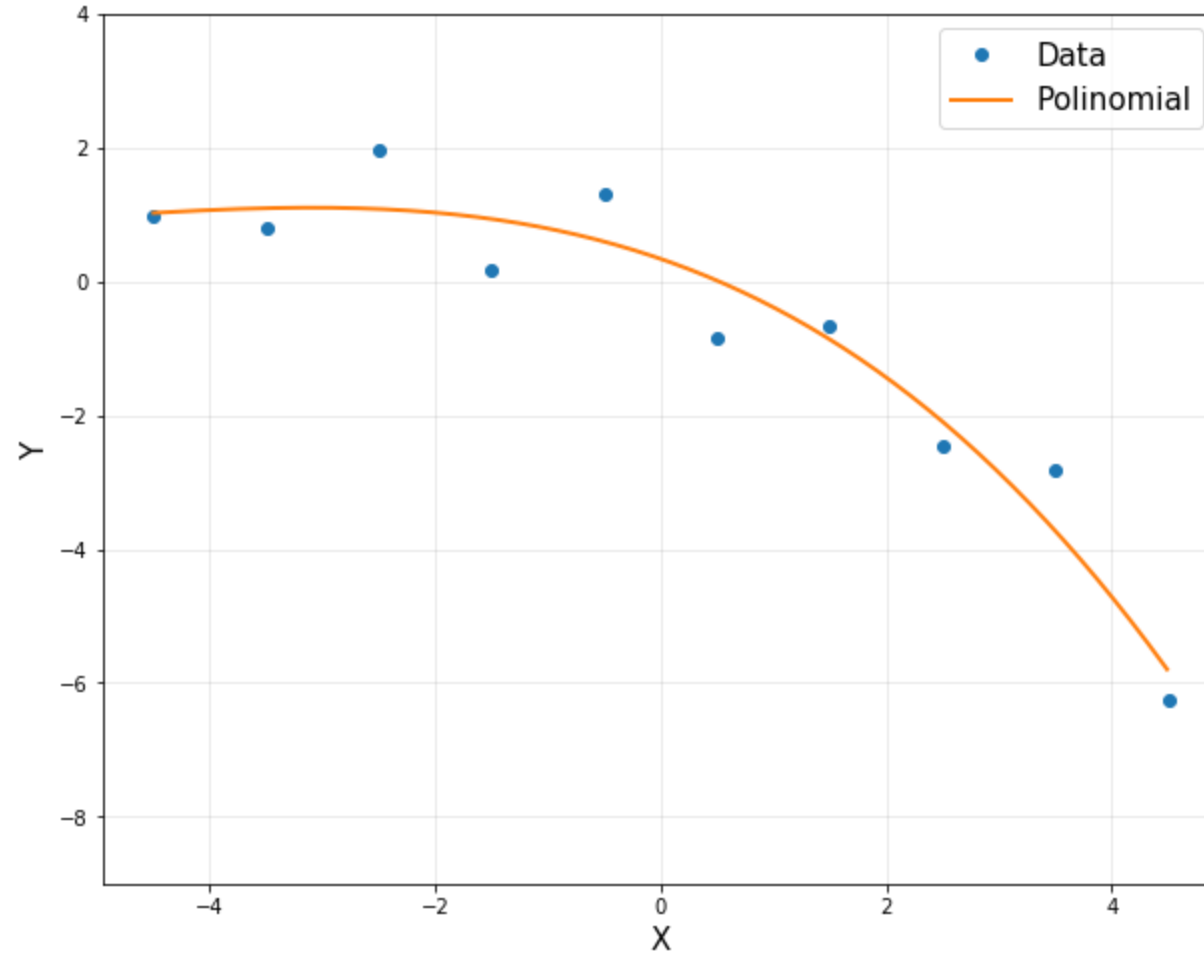
Polynomial Regression ($d = 1$)



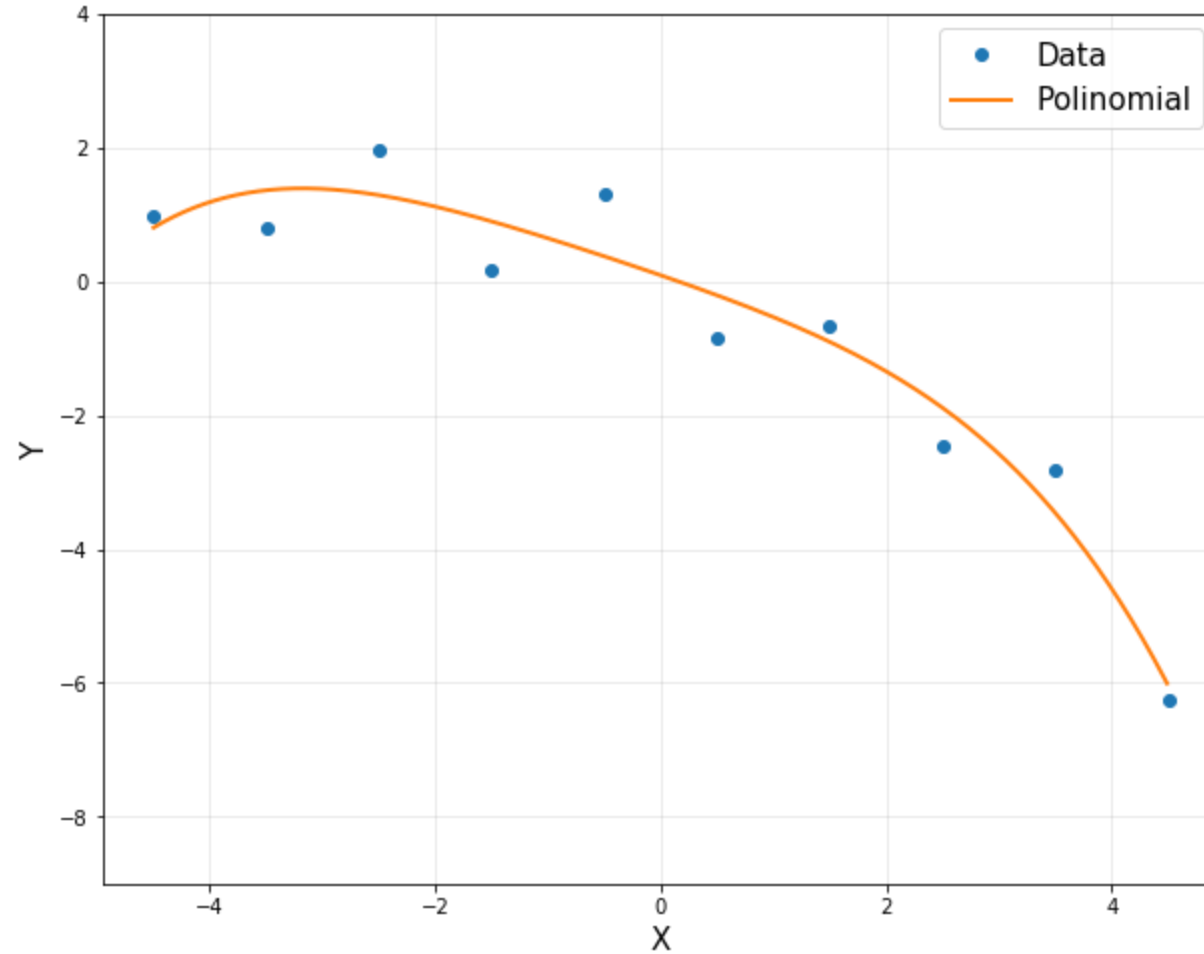
Polynomial Regression ($d = 2$)



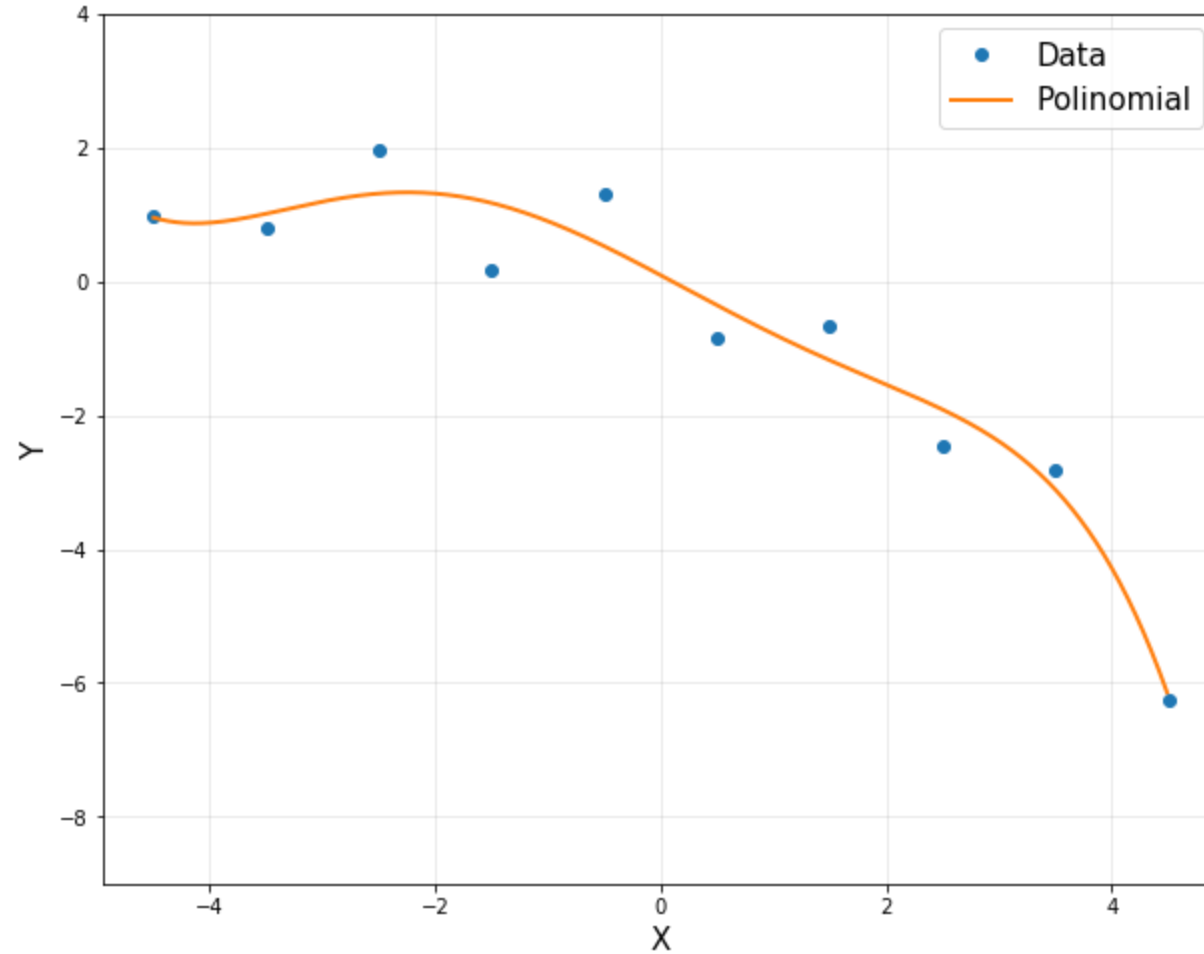
Polynomial Regression ($d = 3$)



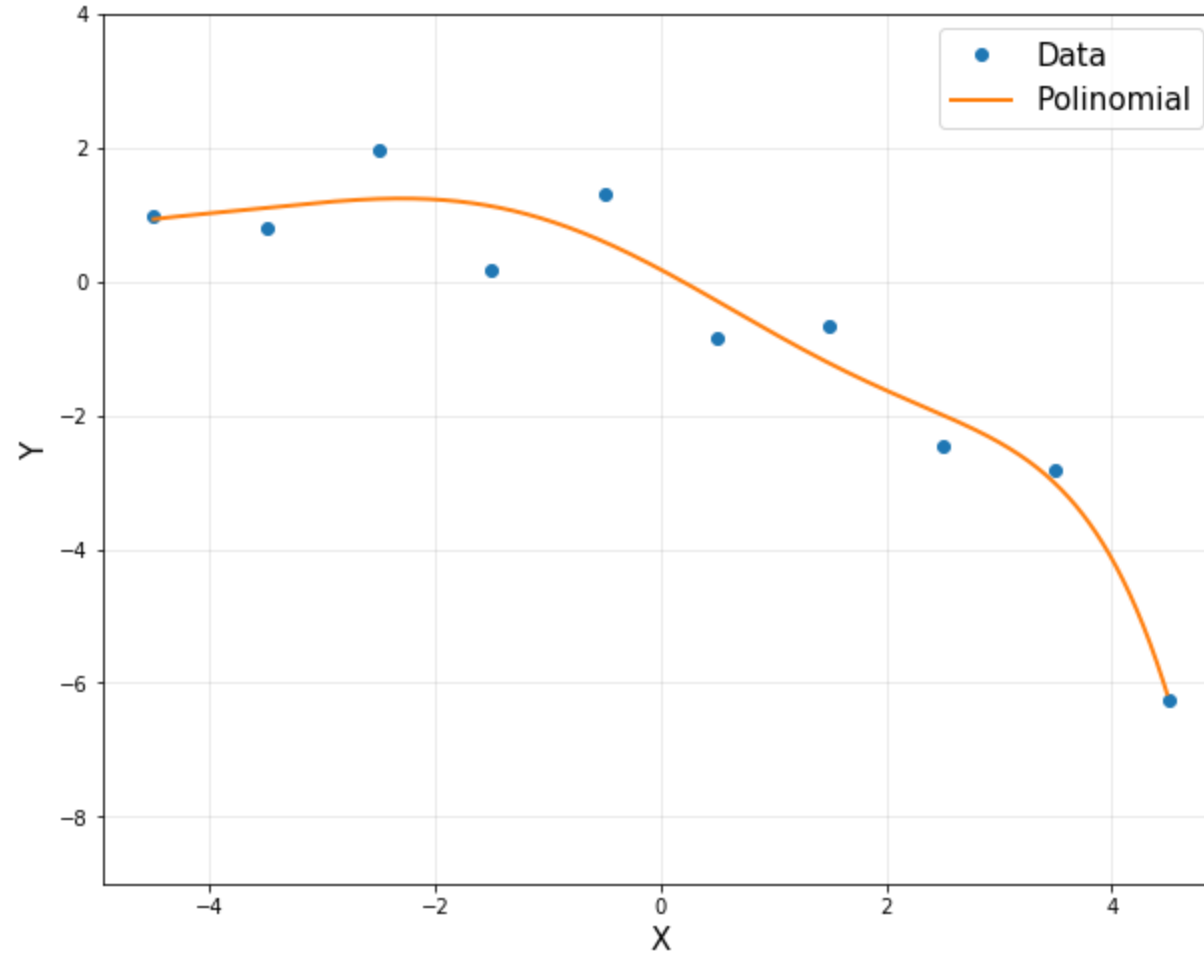
Polynomial Regression ($d = 4$)



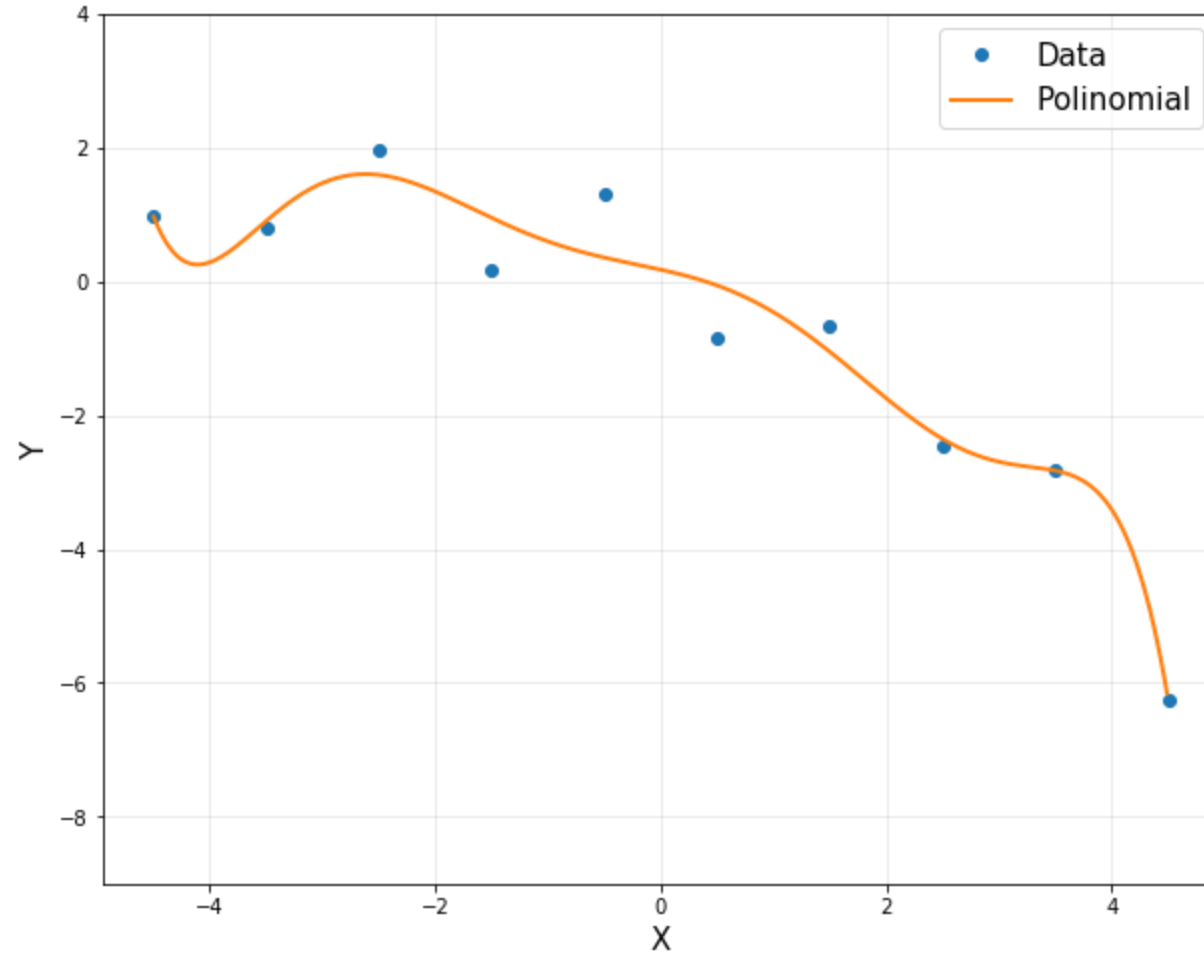
Polynomial Regression ($d = 5$)



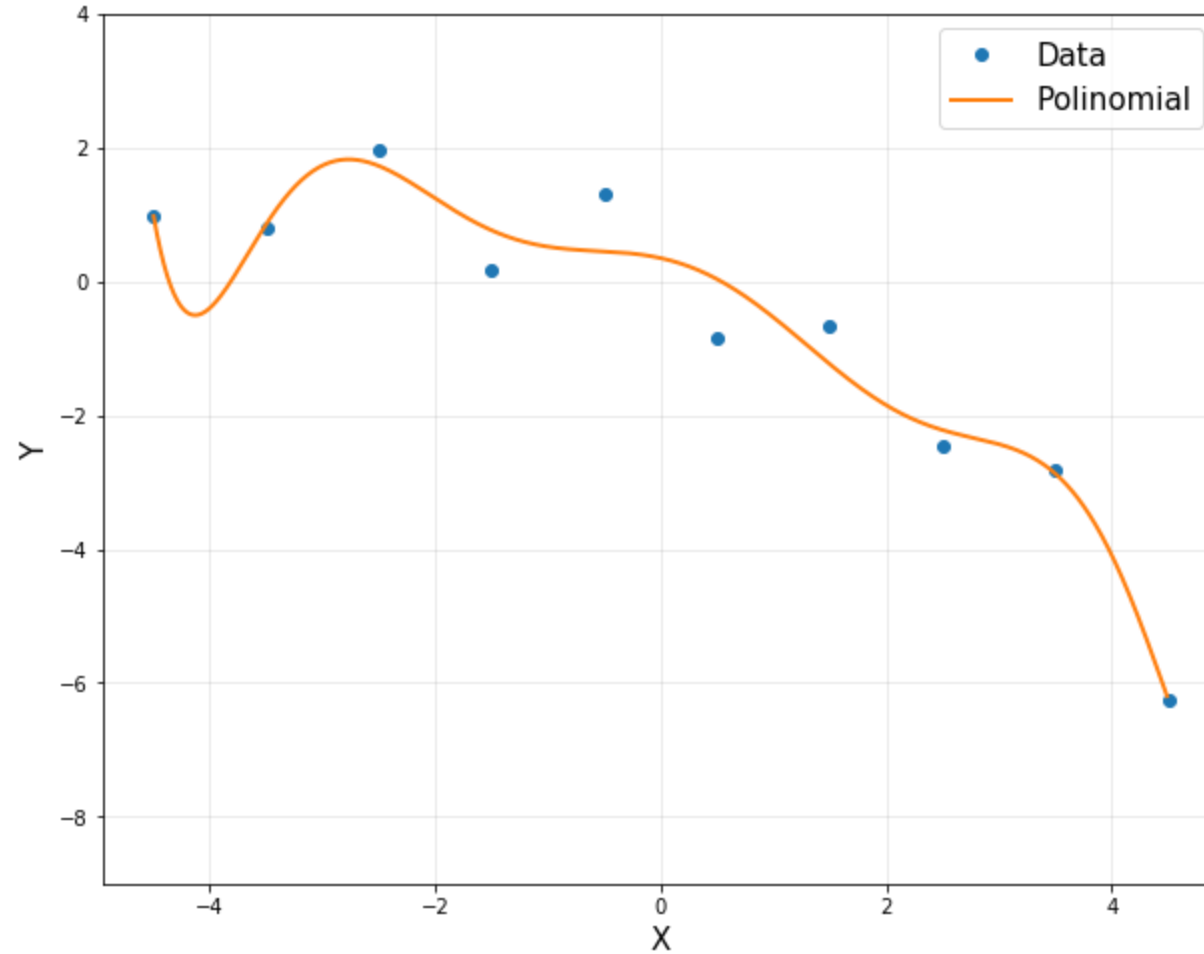
Polynomial Regression ($d = 6$)



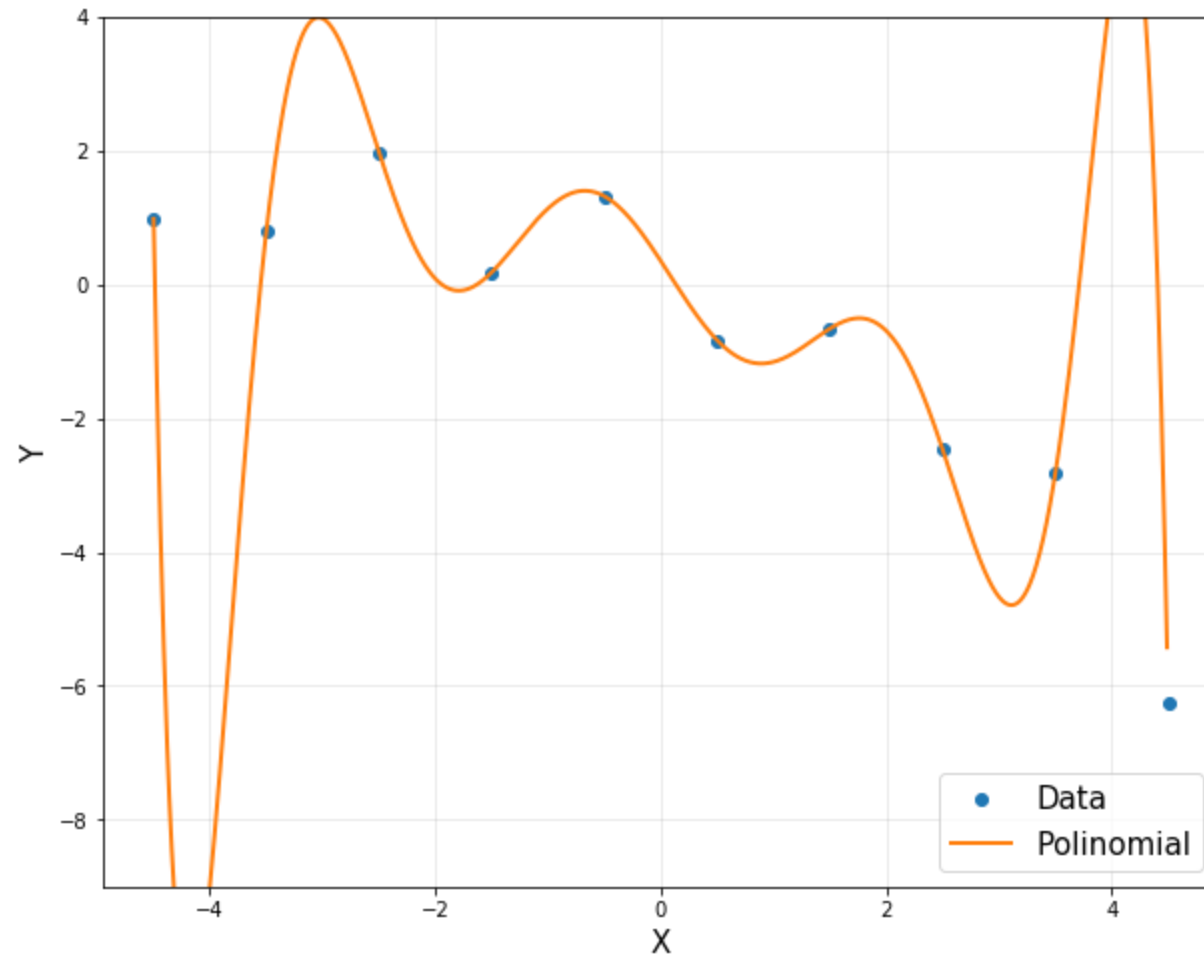
Polynomial Regression ($d = 7$)



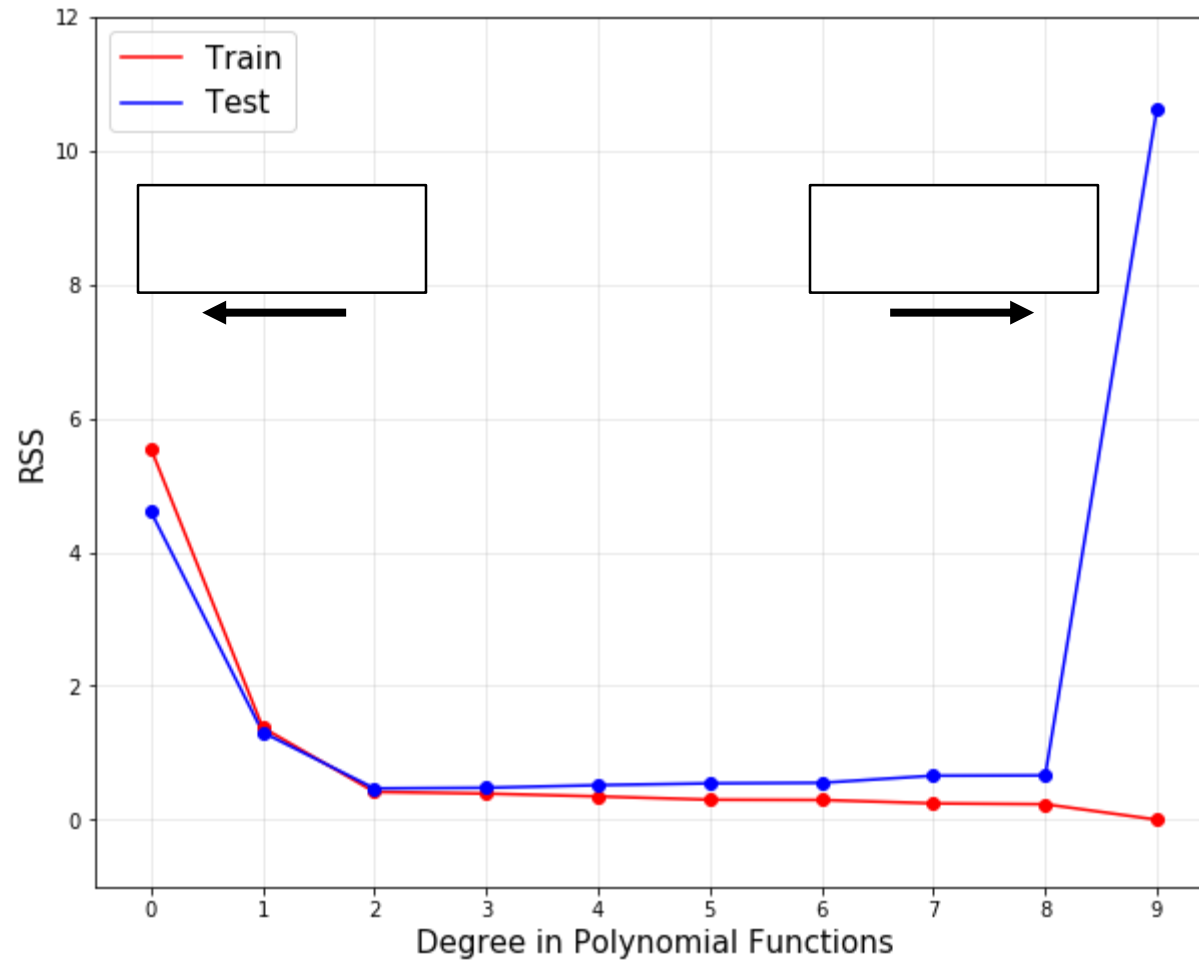
Polynomial Regression ($d = 8$)



Polynomial Regression ($d = 9$)



Errors on Train and Test Datasets

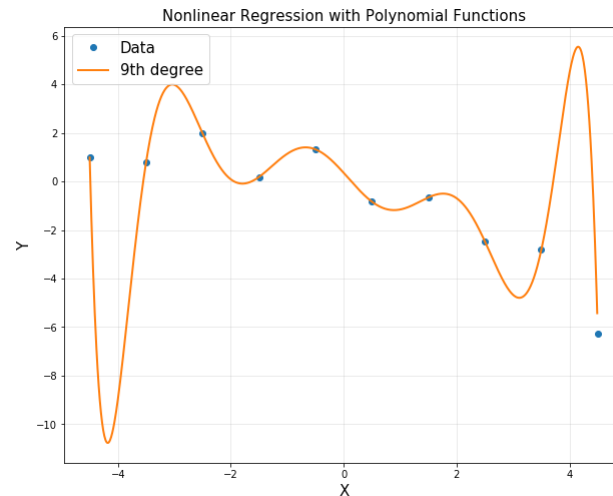
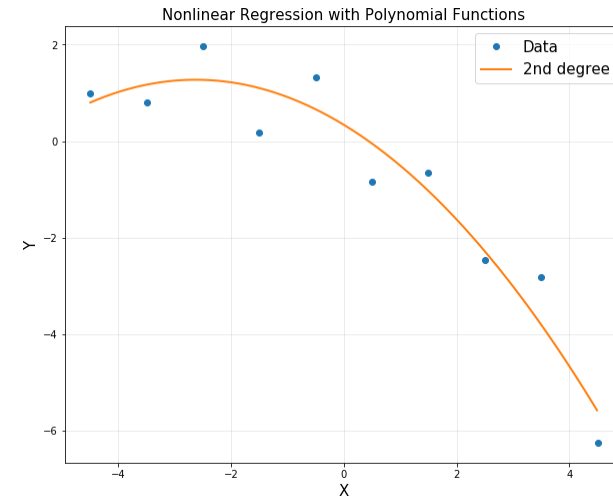
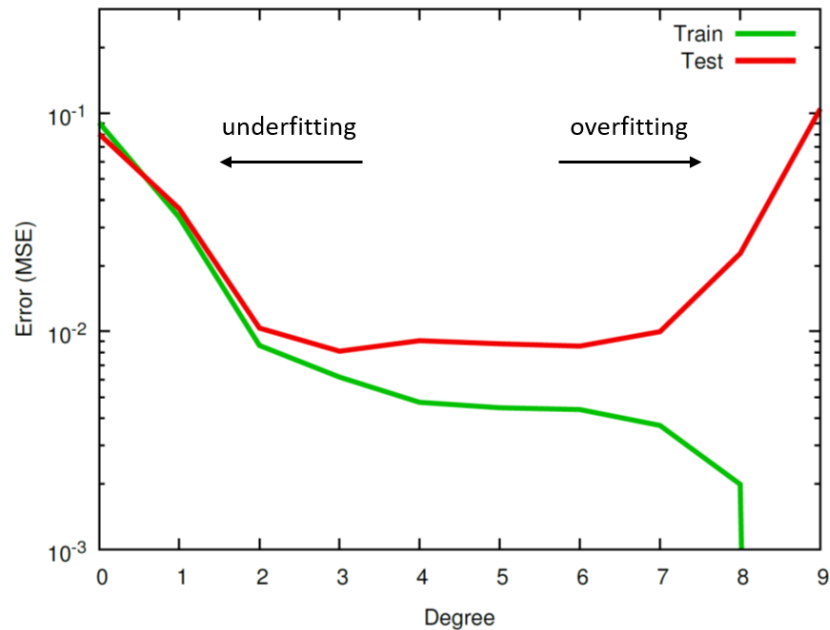


Overfitting Problem

- Have you come across a situation where your model performed exceptionally well on train data, but was not able to predict test data ?
- One of the most common problem data science professionals face is to avoid overfitting.

Issue with Rich Representation

- Low error on input data points, but high error nearby
- Low error on training data, but high error on testing data

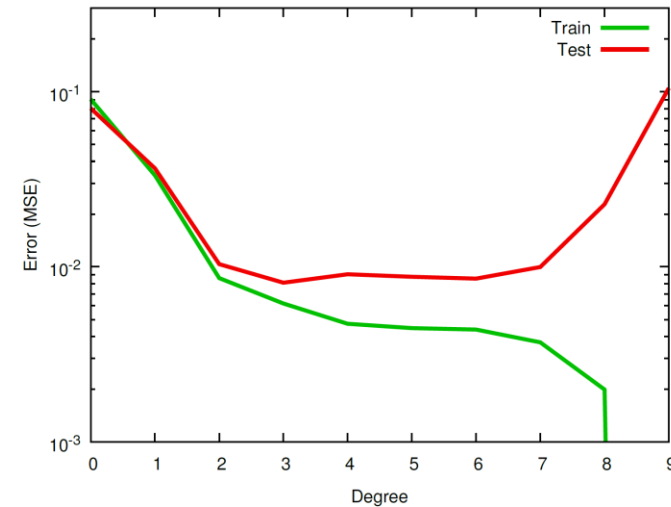
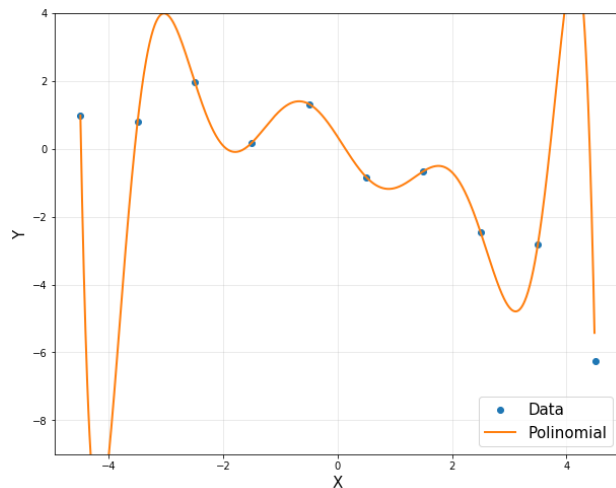


Generalization Error

- Fundamental problem: we are optimizing parameters to solve

$$\min_{\theta} \sum_{i=1}^m \ell(y_i, \hat{y}_i) = \min_{\theta} \sum_{i=1}^m \ell(y_i, \Phi\theta)$$

- But what we really care about is loss of prediction on new data (x, y)
 - also called **generalization error**



- Divide data into training set, and validation (testing) set

Regularization (Shrinkage Methods)

- With many features, prediction function becomes very expressive (model complexity)
 - Choose less expressive function (e.g., lower degree polynomial, fewer RBF centers, larger RBF bandwidth)
 - Keep the magnitude of the parameter small
 - Regularization: penalize large parameters θ

$$\min \|\Phi\theta - y\|_2^2 +$$



- λ : regularization parameter, trades off between low loss and small values of θ

Regularization (Shrinkage Methods)

- Often, overfitting associated with very large estimated parameters
- We want to balance
 - how well function fits data
 - magnitude of coefficients

$$\text{Total cost} = \underbrace{\text{measure of fit}}_{RSS(\theta)} + \lambda \cdot \underbrace{\text{measure of magnitude of coefficients}}_{\lambda \cdot \|\theta\|_2^2}$$

$$\implies \min \|\Phi\theta - y\|_2^2 + \lambda \|\theta\|_2^2$$

- multi-objective optimization
- λ is a tuning parameter

Different Regularization Techniques

- Big Data
- Data augmentation
 - The simplest way to reduce overfitting is to increase the size of the training data.



shift shift shear shift & scale rotate & scale



Different Regularization Techniques

- Early stopping
 - When we see that the performance on the validation set is getting worse, we immediately stop the training on the model.

