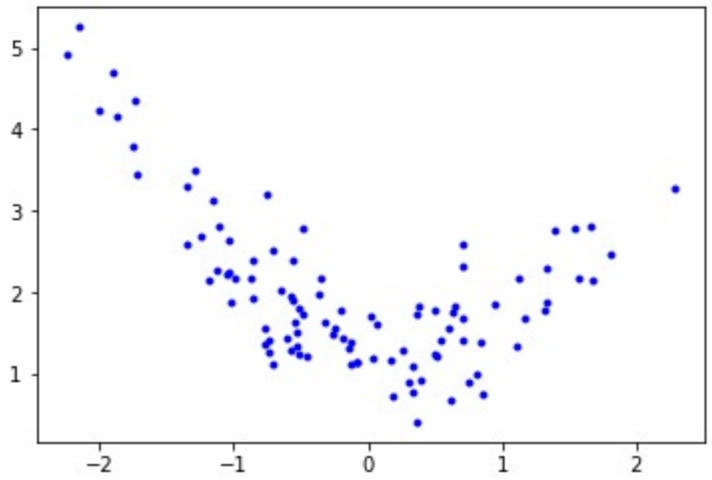
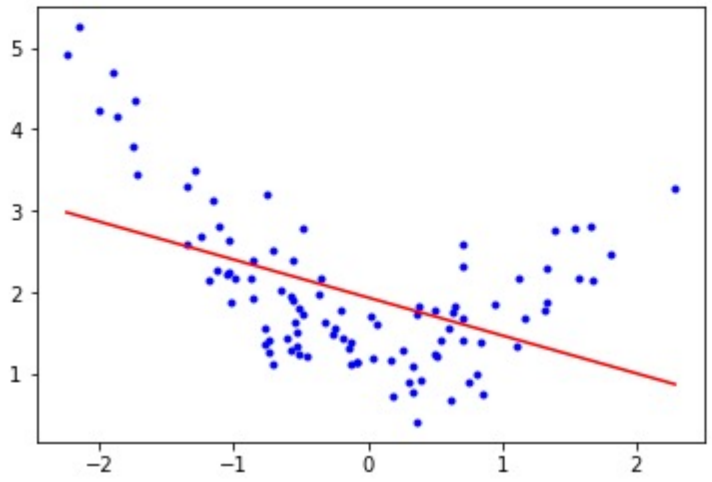
Polynomial Regression & Overfitting

Outline

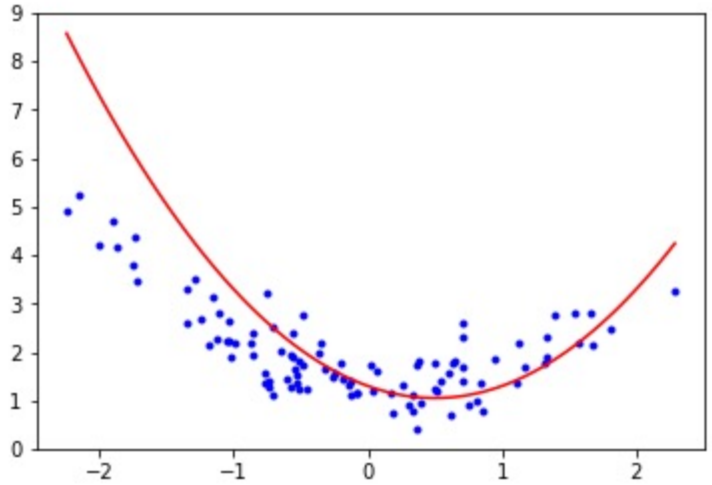
1. Recall **Linear Regression**
2. **Polynomial Regression**: What is it and how is it different (or not so different)?
3. Caution! **Dangers** of polynomial regression!



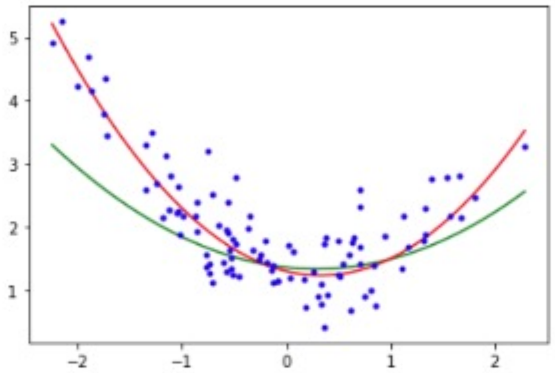
5
4
3
2
1
−2
−1
0
1
2



5
4
3
2
1
−2
−1
0
1
2

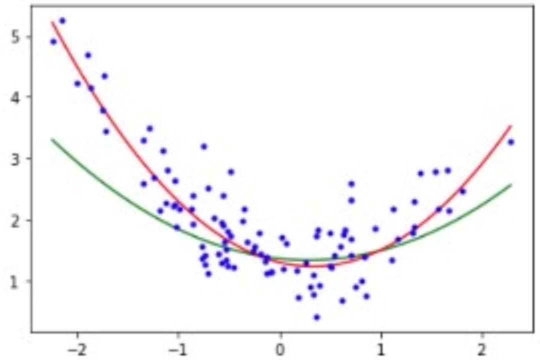

9
8
7
6
5
4
3
2
1
0
−2
−1
0
1
2

$$y = ax + b$$
$$y = ax^2 + bx + c$$
$$y = ax^3 + bx^2 + cx + d$$
$$y = ax^4 + bx^3 + cx^2 + dx + e$$



Which fit looks better?
Why?

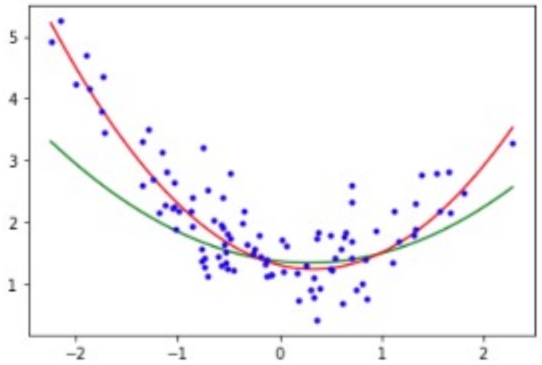
What degree polynomial does this appear to be?



Both are ***quadratics***

$$y = ax^2 + bx + c$$

They have different *polynomial coefficients: a, b, & c*



Polynomial Regression:

For a polynomial of a given degree find the coefficients that reduce the distance/error between the training data and the polynomial line.

Polynomial Regression = Linear Regression

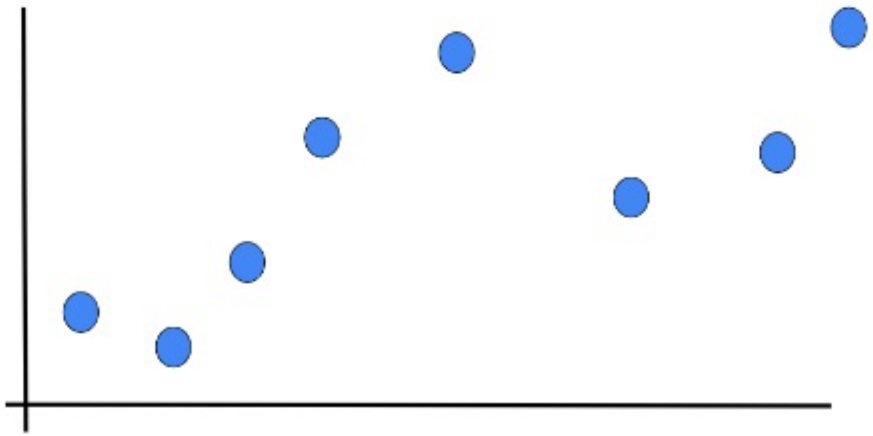
Turn the original polynomial regression problem into a polynomial regression problem with multiple features

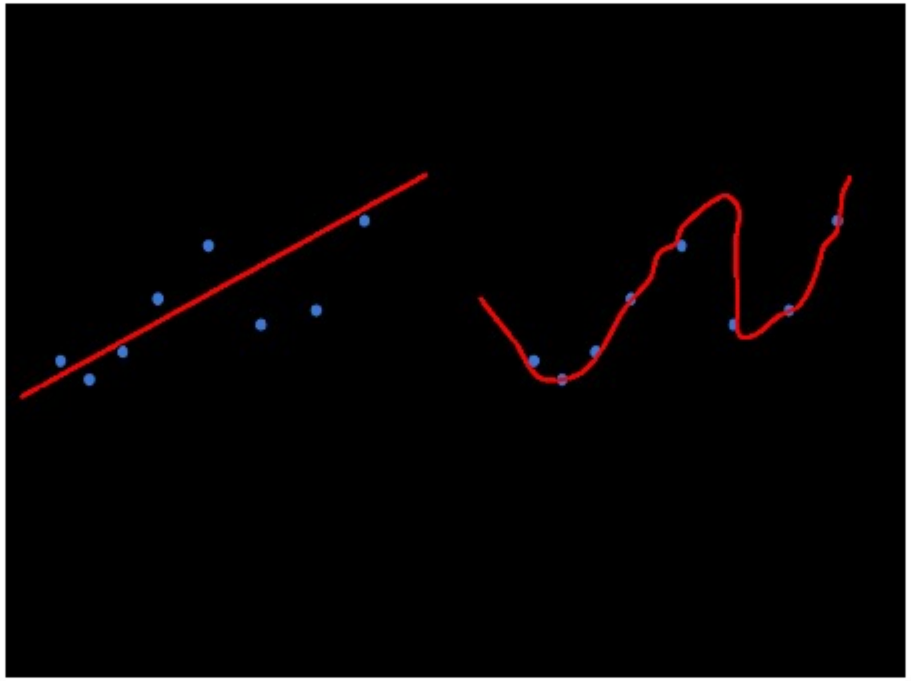
Polynomial $y = \theta_2 * x^2 + \theta_1 * x + \theta_0$

Linear $y = \theta_2 * x_2 + \theta_1 * x_1 + \theta_0$

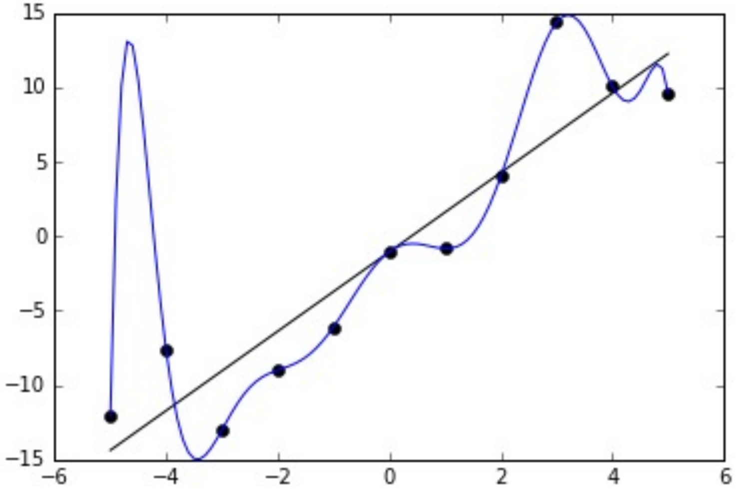
Overfitting

How would you fit this data?






MR. TASTY FRIES



15
10
5
0
−5
−10
−15
−6
−4
−2
0
2
4
6

How Do We Avoid This?

Avoiding Overfitting

- Simpler polynomial
- More training data
- Dropping out some training data
- Overfitting penalties (regularization)

Regularization

Recall: Mean Squared Error

$$MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_{true} - y_{pred})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_{true} - \theta^T X)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_{true} - (\theta_0 + \sum_{j=1}^{p} \theta_j x_i))^2$$

Lasso (L1)

$$MSE(\theta) + \lambda \sum_{j=1}^{p} |\theta_j|$$

Ridge (L2)

$$MSE(\theta) + \lambda\Sigma_{j=1}^{p}\theta_j^2$$

ElasticNet (L1 + L2)

Which Regularization Is Best?

Your Turn