CS 251/2: Data Analysis and Visualization

Lecture 15: Overfitting

Oliver W. Layton

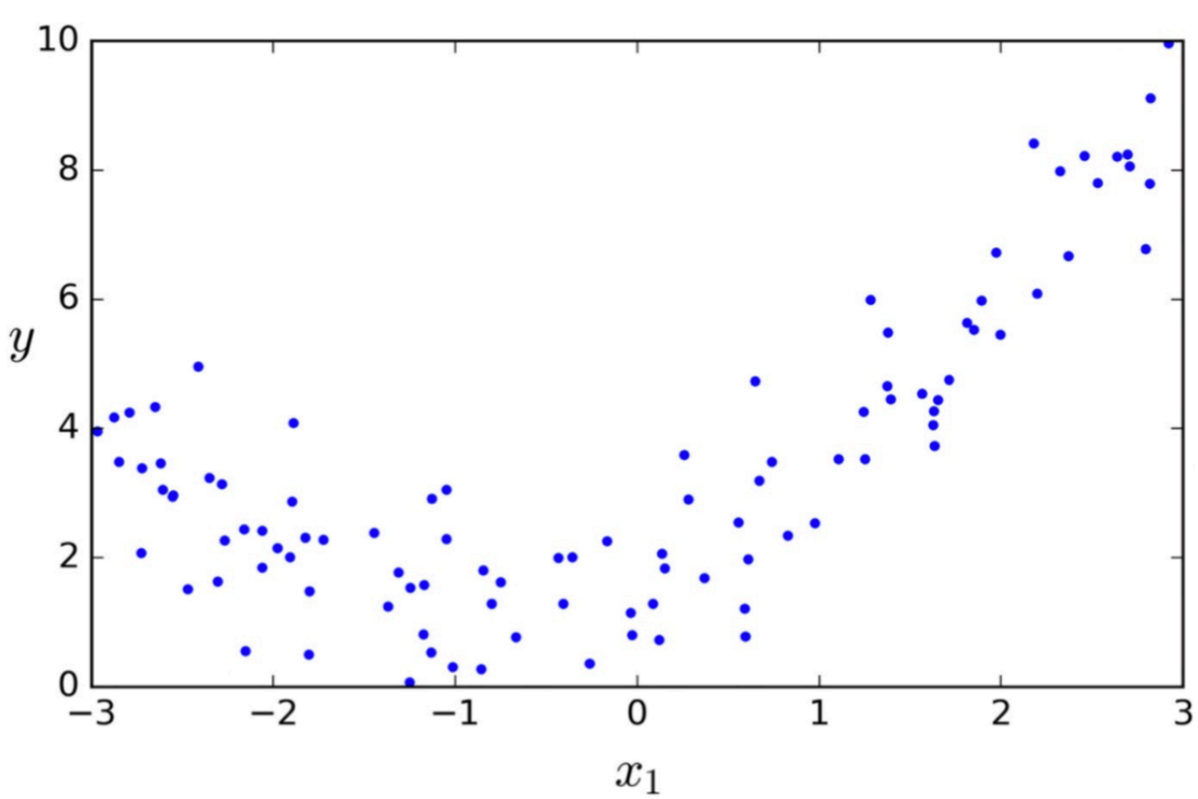
Spring 2021

Overfitting

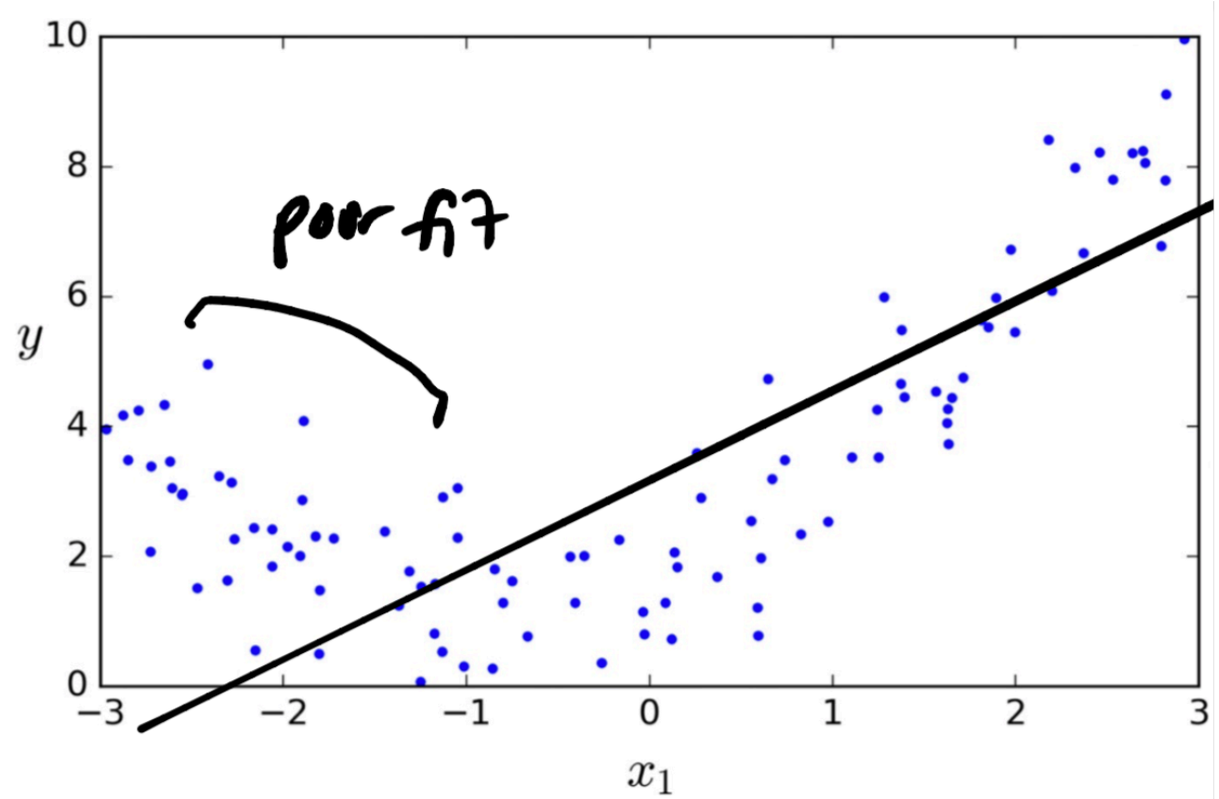
- Fitting data with high degree polynomials can be tempting (e.g. to get better R^2 value), but often can lead to overfitting.
- **Overfitting**: “memorizing” data used to fit linear regression model — model follows the data too closely.

Polynomial regression

$$y = ax_1 + b$$



Truth: quadratic with noise



Linear equation: **underfit**

Quadratic would clearly be better! $y = ax_2^2 + bx_1 + c$

More is better!

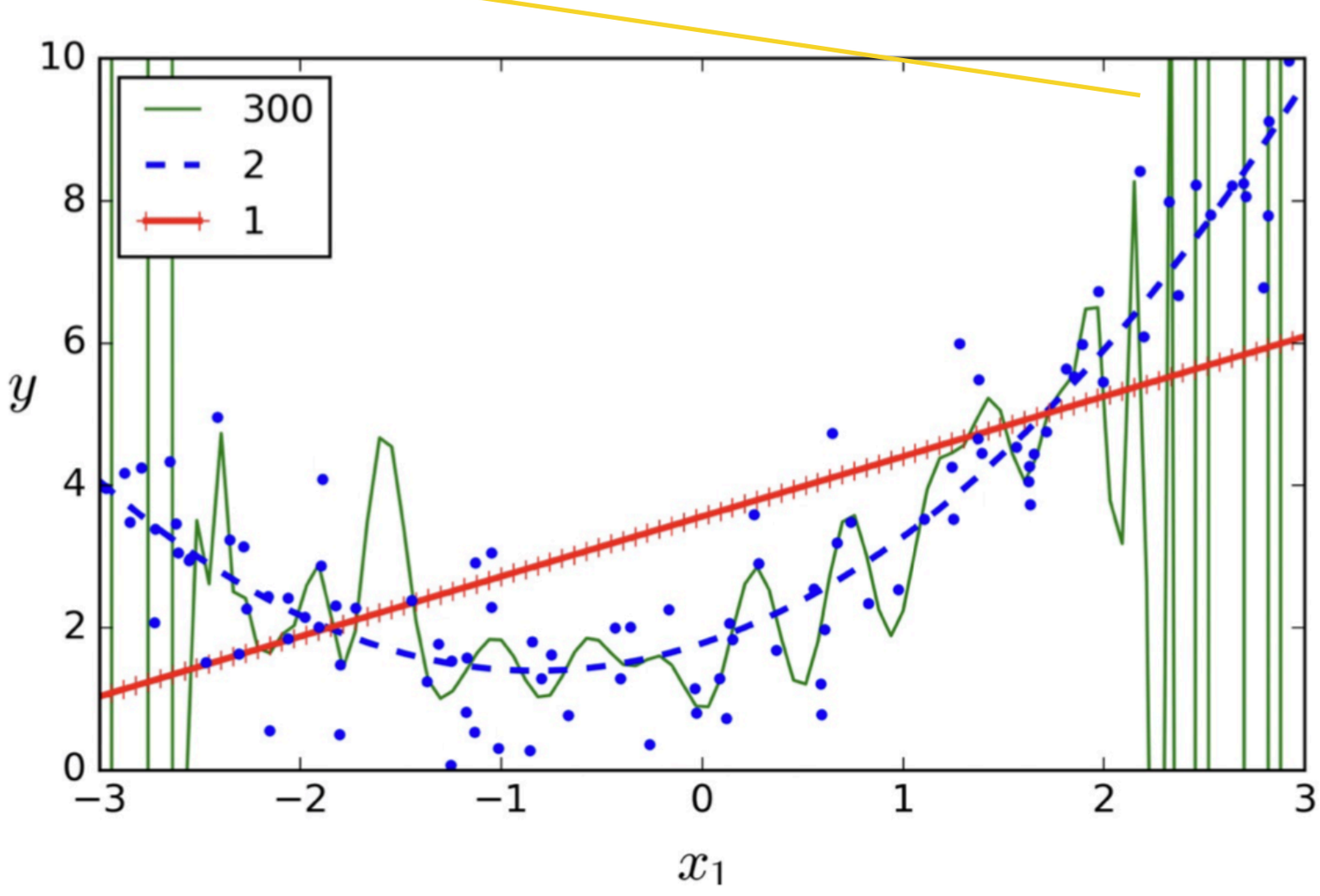
Let's keep going keep increasing the polynomial degree (adding model complexity).

Polynomial regression

…300 degree polynomial…is not good :(

Overfit!

$$y = ax_{300}^{300} + bx_{299}^{299} + \dots$$

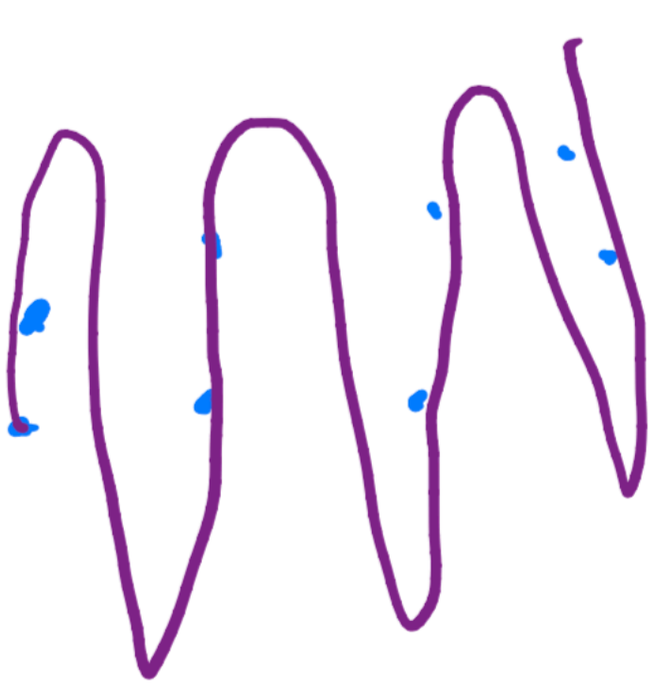


Regression model works hard to pass through as many data points as possible (fit as much detail as possible) — even noise.

Leads to erratic zig-zag pattern.

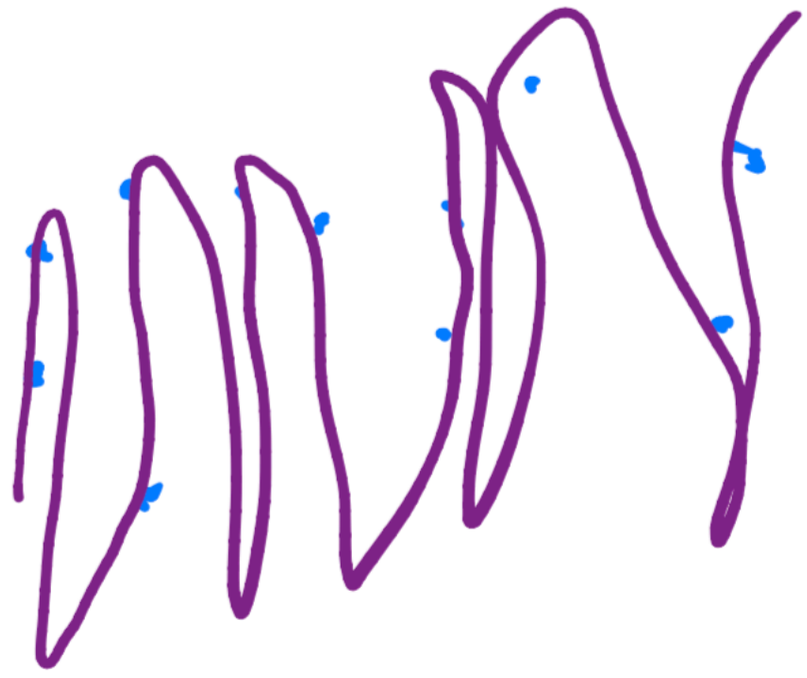
Prediction problem

- **Fit step**: Fit data with noise.
- **Predict step**: Plug in new data generated with new random noise (with the same characteristics as that used to fit data).
- **Problem**: Fit does not resemble new at all!

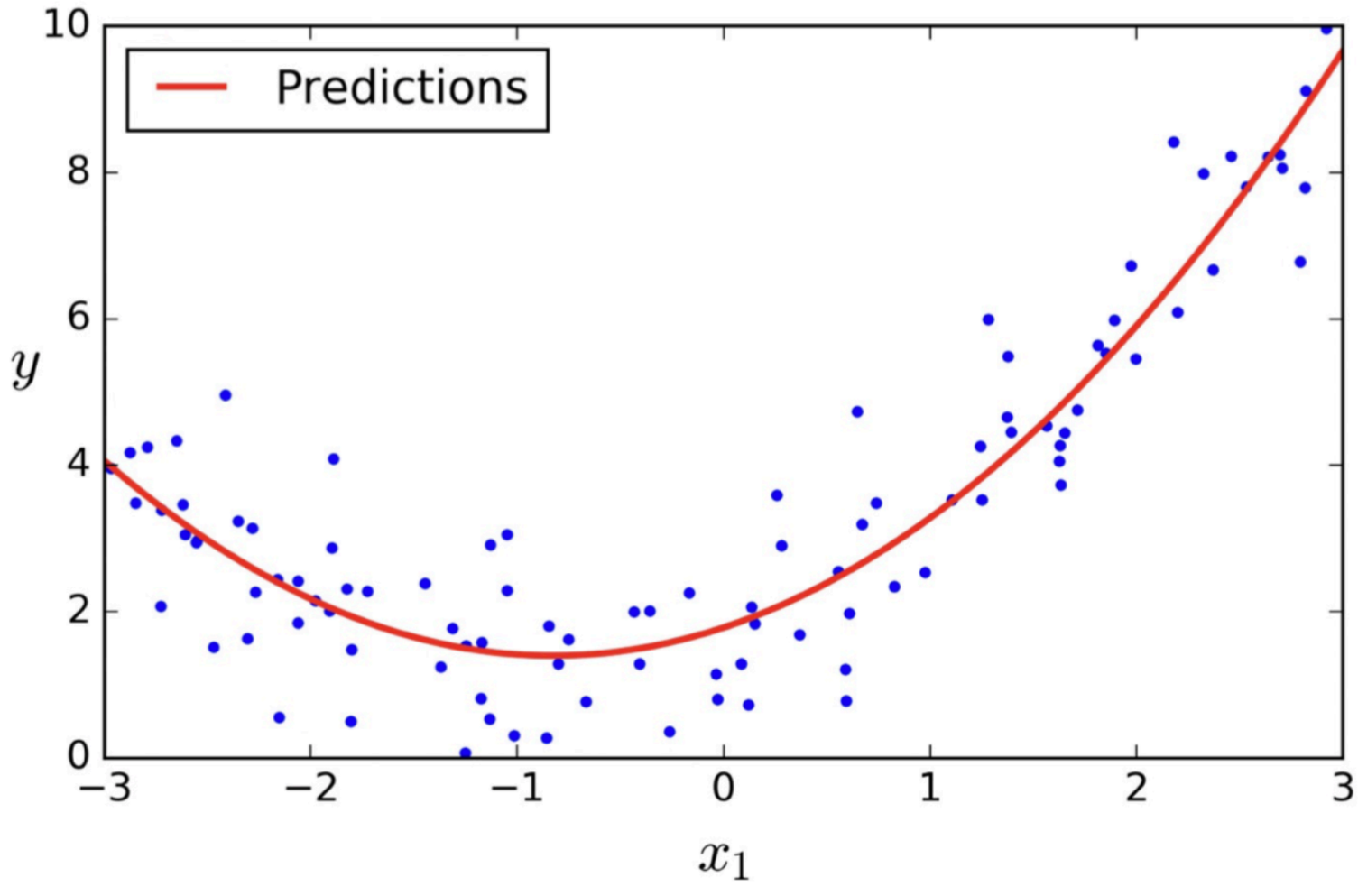


Same data generation process

Very different fits!



Polynomial regression



Quadratic fit (3 coefficients) gives us **HIGHER** error than 300 degree model (301 coefficients), but **MUCH** better generalization.

Simpler fit not sensitive to re-generating data with different noise.