Introduction to Machine Learning

Day 2: Regression as a Lab Bench

Joanna Bieri DATA301

Day 2

Important Information

- Email: joanna_bieri@redlands.edu
- Office Hours: Duke 209, Click Here for Joanna's Schedule
- Class Website

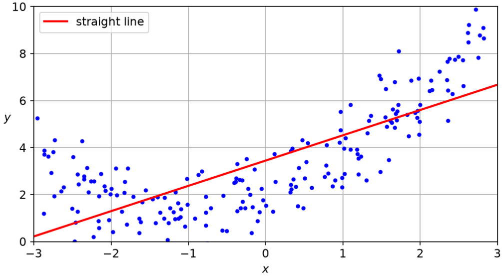
The Three Things From Today

1. **Underfitting** is a model too simple to follow the data. More data will not help.
2. **Overfitting** is a model flexible enough to memorize the noise. It looks fantastic on the training data.
3. A **learning curve** is the fastest way to tell which one you have.

Get out your handwritten notes!

Underfitting

A Straight Line Through A Curve



We built the data ourselves from $y = 0.5x^2 + x + 2$ plus noise, so we know the right answer. A straight line cannot bend, and our data bends.

Discuss

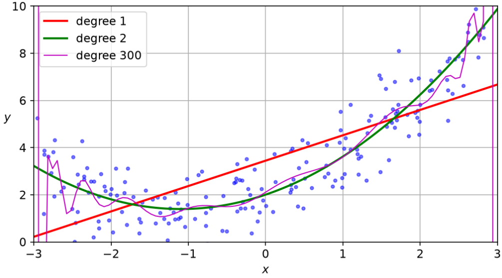
Look at where the errors are. On the left the line sits too high, in the middle too low, on the right too high again. The errors come in **runs**, not scattered randomly.

The straight line's RMSE is **1.619**. The noise we added on purpose has a standard deviation of **1.0**, so a perfect model would still score about 1.0.

- **Q2.** Does this model match the data? Why, in your own words?
- If you collected 10,000 more points and refit the straight line, what happens to the error?

Overfitting

Degree 300 Is A Disaster That Looks Fantastic



Scored on the training data it beats everything else we tried. Its training RMSE is **0.979**.

Discuss

That curve passes closer to the training points than either of the other two models. But look at what it does **between** the points.

- **Q5.** Estimate the residual, just by looking, if you used the degree 300 curve at $x = 2.5$.
- Degree 300 has the lowest training error of anything we fit. Why is that not good news?

The Sneaky Part

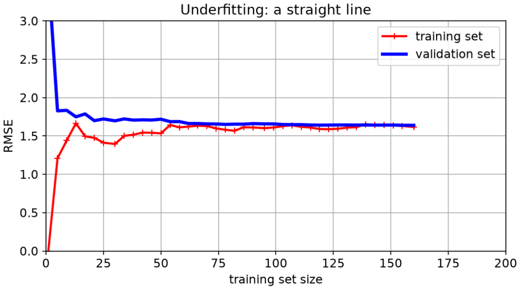
We did **not** switch to a fancier model. It is still `LinearRegression`, the same one that drew the straight line.

We just handed it a new feature, x^2.

The model is still linear **in the features we gave it**. A lot of machine learning is choosing what to feed a model rather than building a cleverer one.

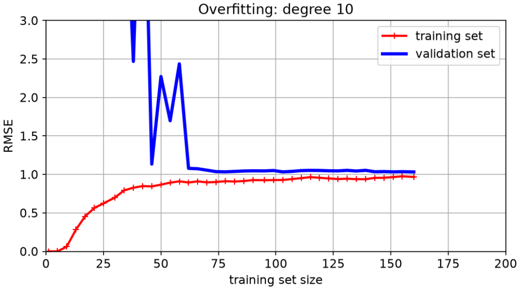
Learning Curves

Underfitting Looks Like This



Both curves climb, meet, and flatten out **high**. Training error bad, validation error bad, close together.

Overfitting Looks Like This



Training error stays **low**. Validation error sits well **above** it. That **gap** is the tell.

Discuss

The rough rule: **high and together means underfitting, low with a gap means overfitting.**

- Your training error is 0.2 and your validation error is 2.5. Which problem do you have, and what do you do?
- Both errors are 1.8 and flat. Would collecting more data help? Why or why not?

Wrap-Up

What To Take Away

- Underfitting: the model is **wrong**. Better model or better features.
- Overfitting: the model is **memorizing**. Simplify, regularize, or get more data.
- **RMSE** puts one number on how wrong you are, in the units of y.
- A score on data the model has already seen tells you almost nothing.

Before Next Class

1. **Weekly Homework 1, due Sunday 9/6 at 11:59pm.** Covers Day 1 and Day 2.
2. Read Geron chapter 4, **Regularized Linear Models**.
3. Watch the Day 3 video.

Day 3

Today: overfitting comes from a model with too much freedom.

Day 3: what happens when you deliberately take some of that freedom away.

That is **regularization**, and it is one of the most useful tools in this class.