Data Science from Scratch

First Principles with Python

Stephen CUI

March 20, 2023

Contents

Chapter 1

Machine Learning

Data science is mostly turning business problems into data problems and collecting data and understanding data and cleaning data and formatting data, after which machine learning is almost an afterthought.

1.1 Modeling

What is a model? It's simply a specification of a mathematical (or probabilistic) relationship that exists between different variables.

1.2 What Is Machine Learning?

Everyone has her own exact definition, but we'll use machine learning to refer to creating and using models that are learned from data. In other contexts this might be called predictive modeling or data mining, but we will stick with machine learning.

We'll look at both supervised models (in which there is a set of data labeled with the correct answers to learn from) and unsupervised models (in which there are no such labels). There are various other types, like semisupervised (in which only some of the data are labeled), online (in which the model needs to continuously adjust to newly arriving data), and reinforcement (in which, after making a series of predictions, the model gets a signal indicating how well it did) that we won't cover here.

1.3 Overfitting and Underfitting

A common danger in machine learning is **overfitting**—producing a model that performs well on the data you train it on but generalizes poorly to any new data. This could involve learning noise in the data. Or it could involve learning to identify specific inputs rather than whatever factors are actually predictive for the desired output. The other side of this is **underfitting**—producing a model that doesn't perform well even on the training data, although typically when this happens you decide your model isn't good enough and keep looking for a better one.

Clearly, models that are too complex lead to overfitting and don't generalize well beyond the data they were trained on. So how do we make sure our models aren't too complex? The most fundamental approach involves using different data to train the model and to test the model.

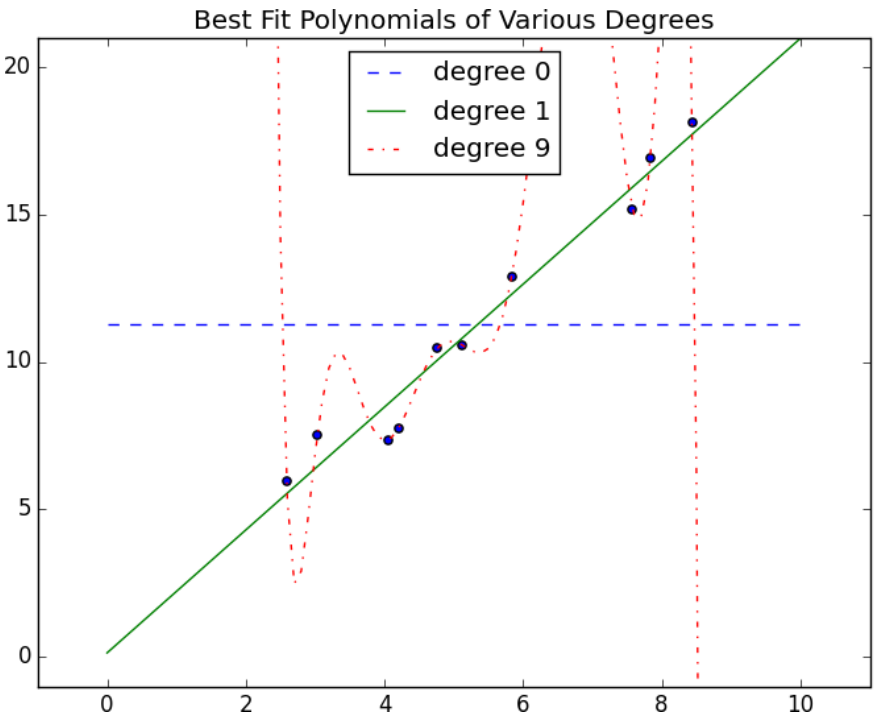

Figure 1.1: Overfitting and underfitting

If the model was overfit to the training data, then it will hopefully perform really poorly on the (completely separate) test data. Said differently, if it performs well on the test data, then you can be more confident that it's fitting rather than overfitting.

1.4 Correctness

Accuracy is defined as the fraction of correct predictions. **Precision** measures how accurate our positive predictions were. **Recall** measures what fraction of the positives our model identified. Sometimes precision and recall are combined into the **F1 score**.

1.5 The Bias-Variance Tradeoff

Another way of thinking about the overfitting problem is as a tradeoff between bias and variance.

Both are measures of what would happen if you were to retrain your model many times on different sets of training data (from the same larger population).

For example, the degree 0 model in Figure 1.1 will make a lot of mistakes for pretty much any training set (drawn from the same population), which means that it has a high bias. However, any two randomly chosen training sets should give pretty similar models (since any two randomly chosen training sets should have pretty similar average values). So we say that it has a low variance. High bias and low variance typically correspond to underfitting.

On the other hand, the degree 9 model fit the training set perfectly. It has very low bias but very high variance (since any two training sets would likely give rise to very different models). This corresponds to overfitting.

If your model has high bias (which means it performs poorly even on your training data), one thing to try is adding more features. If your model has high variance, you can similarly remove features. But another solution is to obtain more data (if you can).

Holding model complexity constant, the more data you have, the harder it is to overfit. On the other hand, more data won't help with bias. If your model doesn't use enough features to capture regularities in the data, throwing more data at it won't help.

1.6 Feature Extraction and Selection

Features are whatever inputs we provide to our model.