Machine Learning: Generalization

Dimitri Bourilkov
University of Florida

USDA-ARS / UF Machine Learning Training 2019
August 27, 2019

UF UNIVERSITY of FLORIDA

Introduction

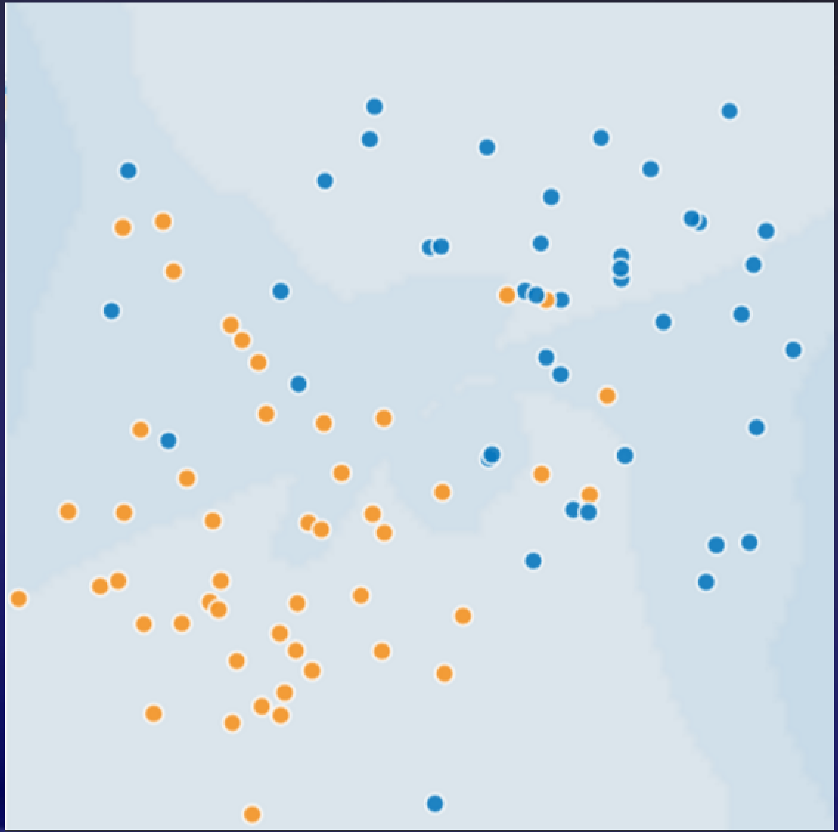
- How will my model perform on new, unseen, data? In other words, how will it generalize?
- The peril of Overfitting (Overtraining)
- How to measure Machine Learning (ML) performance and reduce this peril
- Based on the Google Machine Learning Crash Course; License CC BY 4.0

A Classification Problem

Example: tree positions in a forest:

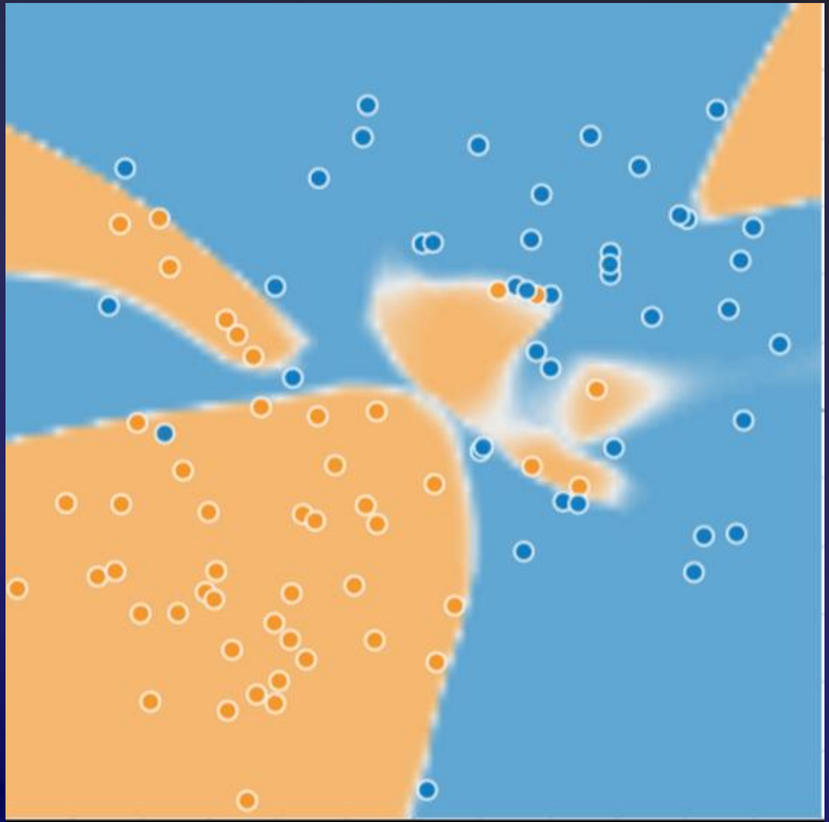
- The blue dots represent sick trees
- The orange dots represent healthy trees

Build a model to separate them



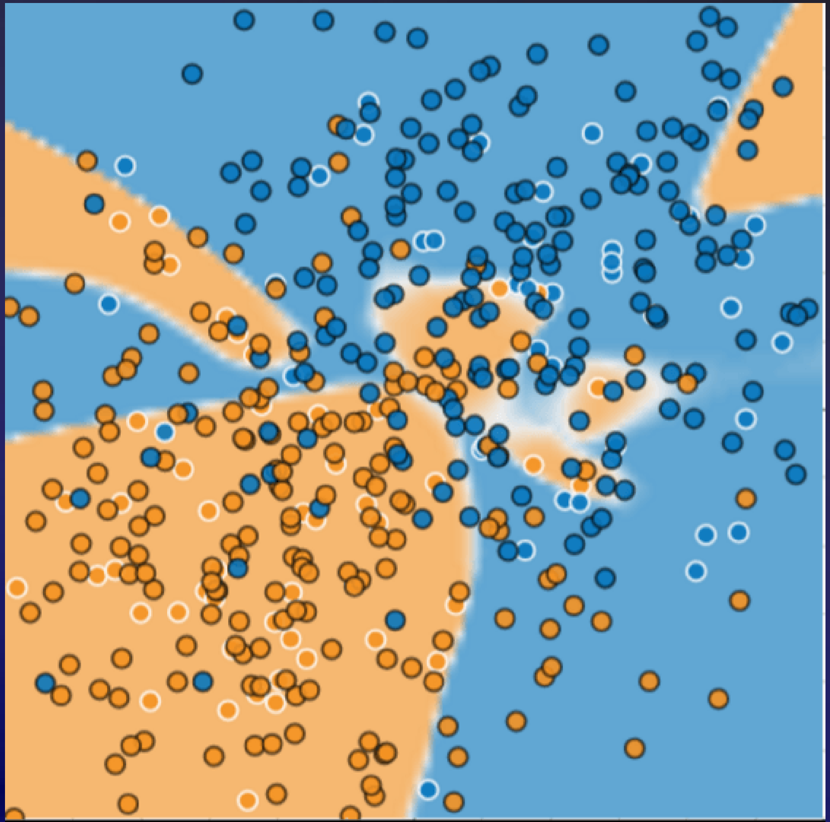
A "Perfect" Model

- A simple, e.g. linear model (line) will not do the job
- We can build complex models which will provide almost perfect separation
- Will they work on new data?



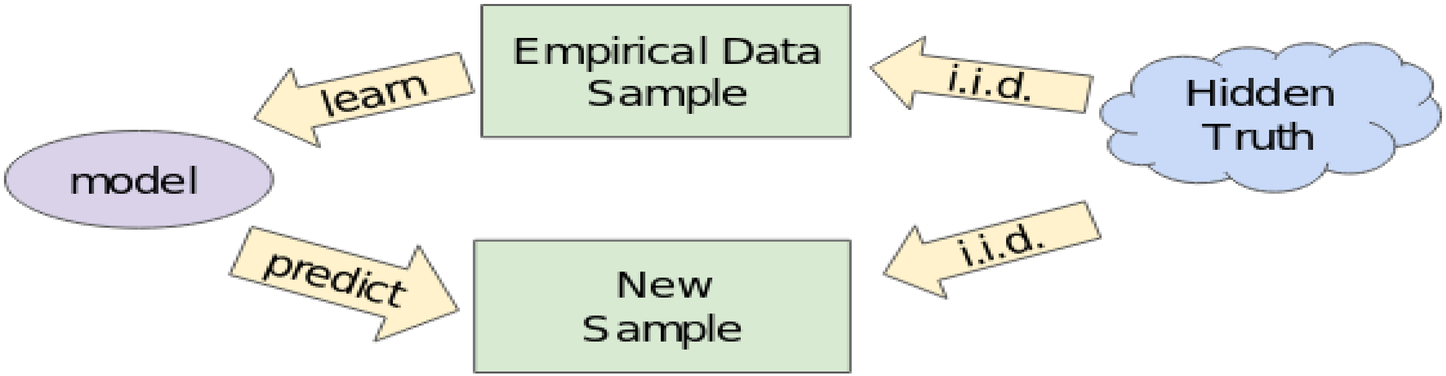
Overfitting

- A resounding NO
- Our model does not generalize well, even if the loss on training data is low
- Our model is tuned to the peculiarities of the training sample and too complex as a result



The Big Picture

The Big Picture



- Goal: predict well on new data drawn from (hidden) true distribution.
- Problem: we don't see the truth.
    - We only get to sample from it.
- If model h fits our current sample well, how can we trust it will predict well on other new samples?

Simplify “enough”

- Ockham's razor: More things should not be used than are necessary.
- Albert Einstein: Everything should be made as simple as possible, but not simpler.

How Do We Know If Our Model Is Good?

- Theoretically:
  - Interesting field: generalization theory
  - Based on ideas of measuring model simplicity / complexity
- Intuition: formalization of Ockham's Razor principle
  - The less complex a model is, the more likely that a good empirical result is not just due to the peculiarities of our sample

Good ML models find a balance between the two

Empirical Tests

How Do We Know If Our Model Is Good?

- Empirically:
  - Asking: will our model do well on a new sample of data?
  - Evaluate: get a new sample of data-call it the test set
  - Good performance on the test set is a useful indicator of good performance on the new data in general:
    - If the test set is large enough
    - If we don't cheat by using the test set over and over

The ML Fine Print

The following three basic assumptions guide generalization:

- We draw examples independently and identically (i.i.d) at random from the distribution; in other words, examples don't influence each other
- The distribution is stationary; that is the distribution doesn't change within the data set
- We always pull from the same distribution (for training, validation and test samples)

In practice, we sometimes violate these assumptions.