Stanford University

CS 329P : Practical Machine Learning (2021 Fall)

4.2 Underfiting & Overfitting

Qingqing Huang, Mu Li, Alex Smola

https://c.d2l.ai/stanford-cs329p

Who will Repay Their Loans?

- A lender hires you to investigate who will repay their loans
  - You are given all information about the 100 applicants
  - 5 defaulted within 3 years
- A Surprising Finding?!
  - All 5 people who defaulted wore blue shirts during interviews
  - Your model leverages this strong signal as well

Underfitting and Overfitting

- Training error: model error on the training data
- Generalization error: model error on new data

Generalization error \ Training error	Low	High
Low	Good	Bug?
High	Overfitting	Underfitting

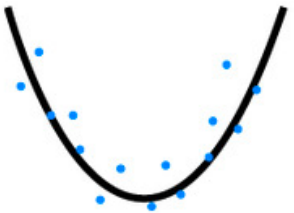
Data and Model Complexity

Data complexity

Model complexity	Low	High
Low	Normal	Underfitting
High	Overfitting	Normal

Underfitting

Normal



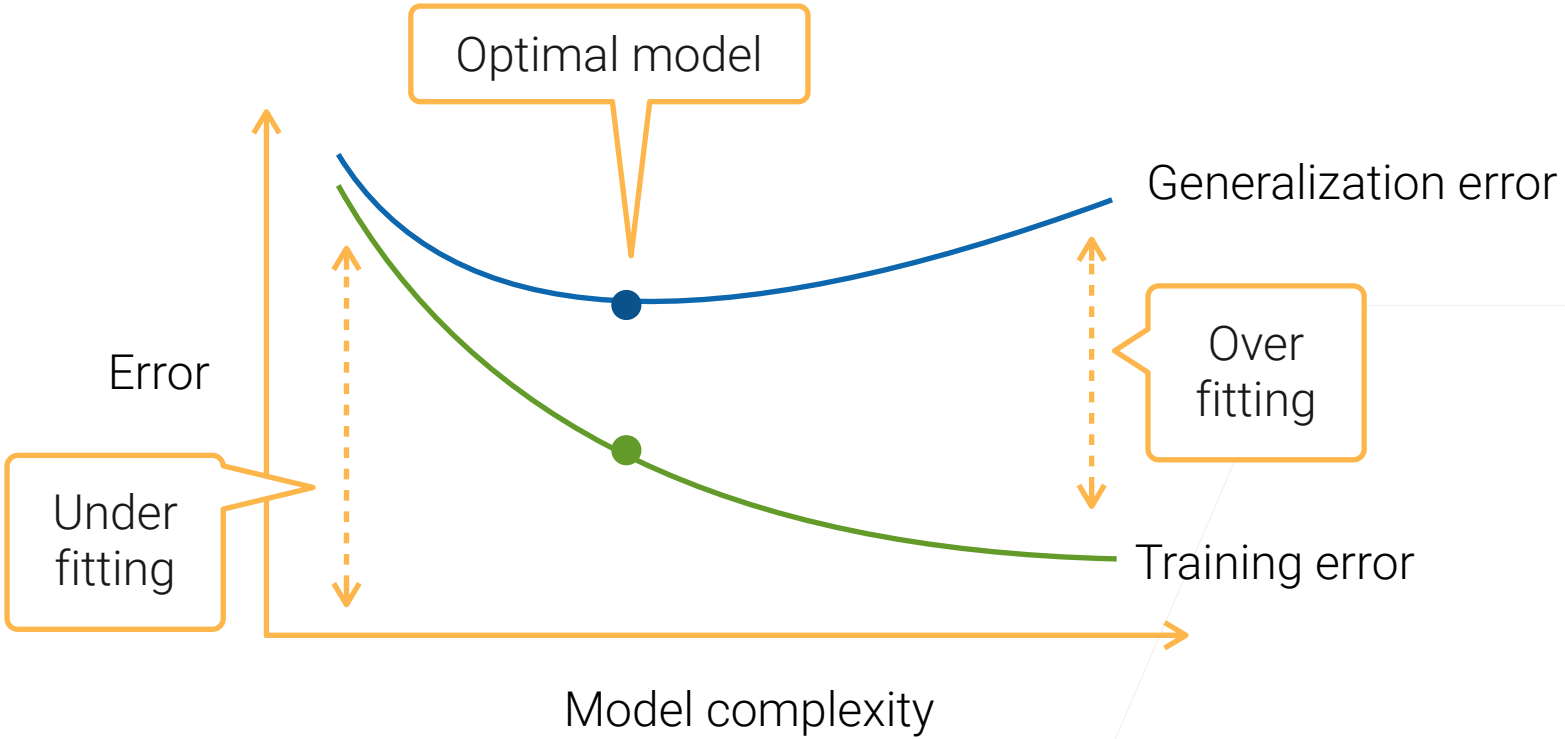
Overfitting

Image credit: hackernoon.com

Model Complexity

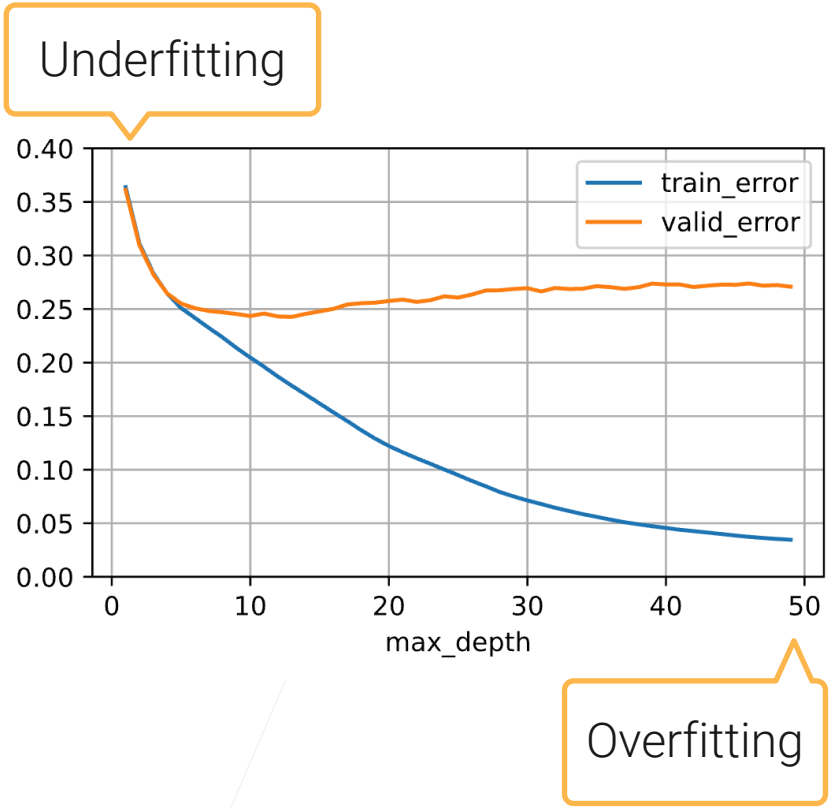
- The capacity of a set of function to fit data points
- In ML, model complexity usually refers to:
  - The number of learnable parameters
  - The value range for those parameters
- It's hard to compare between different types of ML models
  - E.g. trees vs neural network
- More precisely measure of complexity: VC dimension
  - VC dim for classification model:
    the maximum number of examples the model can shatter

Model Complexity



Model Complexity Example: Decision Tree

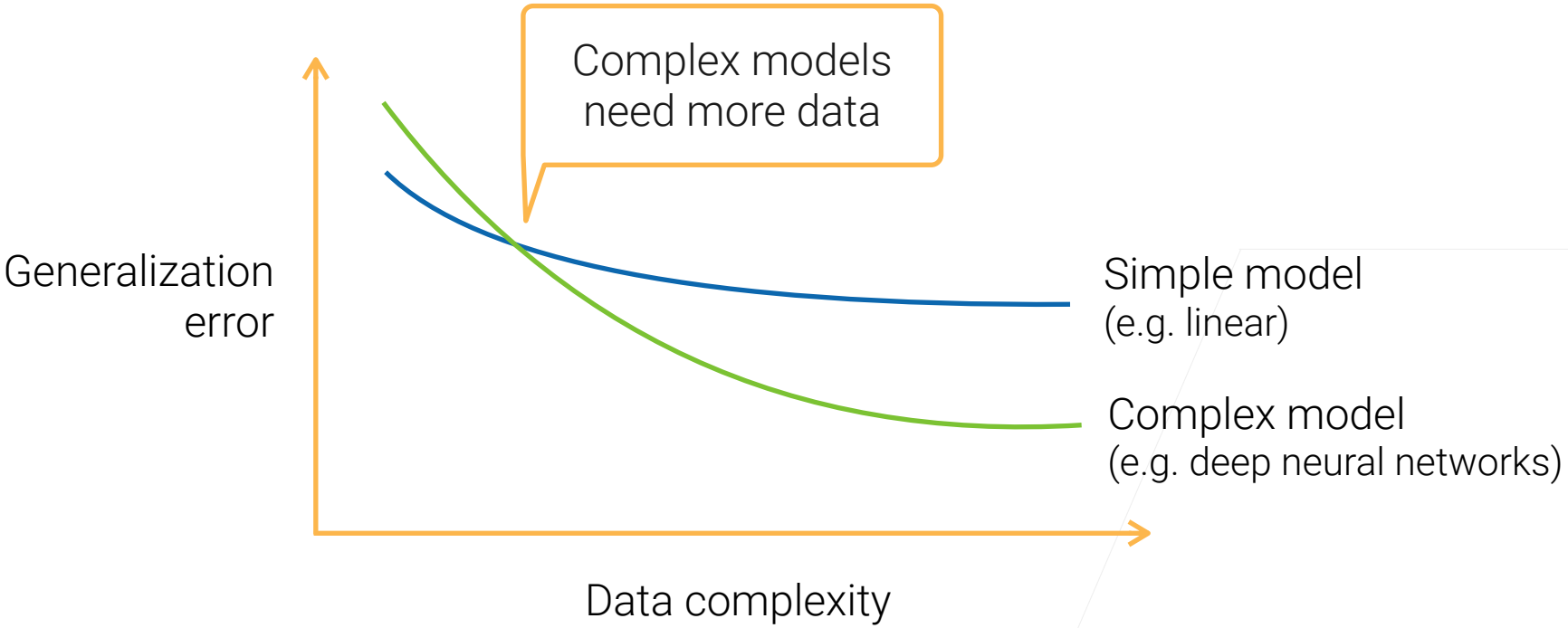
- The tree size can be controlled by the number of levels
- Use scikit-learn `DecisionTreeRegressor(max_depth=n)` on house sales data



Data Complexity

- Multiple factors matters
  - # of examples
  - # of features in each example
  - the separability of the classes
- Again, hard to compare among very different data
  - E.g a char vs a pixel
- More precisely, Kolmogorov complexity
  - A data is simple if it can be generated by a short program

Model Complexity vs Data Complexity



Generalization error

- Generalization error bound (an informal statement)

$$\left|\text{errror on unseen data} - \text{training error}\right| \leq \sqrt{\frac{D}{N}\left(\log\left(\frac{2N}{D}\right)+1\right)}$$

  - D: VC-dim, M: number of training examples
- Generalization error also depends on the training algorithm
  - Adding regularization can penalize complex models
  - Model trained with stochastic gradient methods generalizes better

Model Selection

- Pick a model with a proper complexity for your data
  - Minimize the generalization error
  - Also consider business metrics
- Pick up a model family, then select proper hyper-parameters
  - Trees: #trees, maximal depths
  - Neural networks: architecture, depth (#layers), width (#hidden units), regularizations

Summary

- We care about generalization error
- Model complexity: the ability to fit various functions
- Data complexity: the richness of information
- Model selection: match model and data complexities