Lecture Pattern Analysis

Part 03: Bias and Variance

Christian Riess

IT Security Infrastructures Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

April 27, 2021

Introduction

- We search in the exercise for good kernel parameters via cross validation
- This model selection task touches the question of **generalization** to new data:
  - Too large kernel: smears out the structure of the training data, but covers new samples
  - Too small kernel: closely follows the training data, but might miss new samples
  - "Right" kernel size: represents the structure of the training data and also covers new samples
    (to the extent possible with the given method and data)
- This is an instance of the **bias-variance tradeoff**[1]

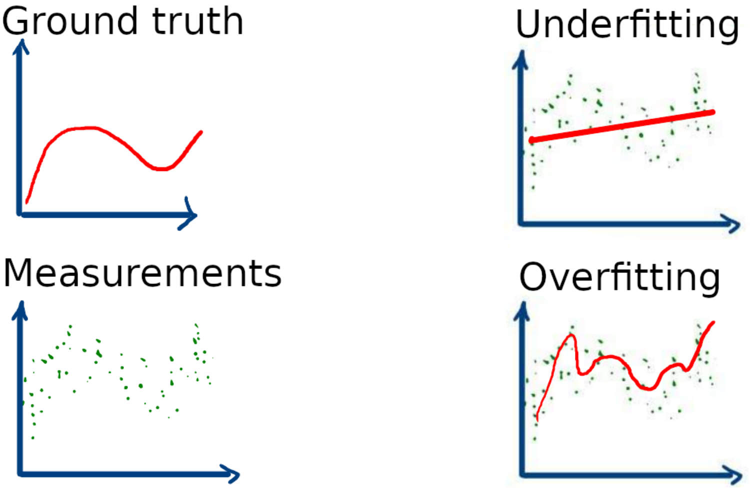
[1] See PR lecture or Hastie/Tibshirani/Friedman Sec. 7-7.3 if more details are desired

Bias and Variance in Regression

- **Bias** is the square of the *average deviation* of an estimator from the ground truth
- **Variance** denotes is the *variance of the estimates*, i.e., the expected squared deviation from the estimated mean[2]
- Informal interpretation:
  - High bias indicates **model undercomplexity**: we obtain a poor fit to the data
  - High variance indicates **model overcomplexity**: the fit also models not just the structure of the data, but also its noise
- Higher model complexity (= more model parameters) tends to lower bias and higher variance
- We will usually not be able to get bias and variance simultaneously to 0
- Regularization increases bias and lowers variance 防过拟合

[2] See Hastie/Tibshirani/Friedman Sec. 7.3 Eqn. (7.9) for a detailed derivation

Sketches for Model Undercomplexity and Overcomplexity



Transferring Bias and Variance to our Density Estimators

- Our kernel framework can directly replicate these investigations by retargeting our kernels to regression or classification:
- Regression:
  - Estimate $f(\mathbf{x})$ at position $\mathbf{x}$ as a kernel-weighted sum of the neighbors or
  - as a k-NN mean of k neighbors
- Classification:
  - Estimate for classes c_1 and c_2 individual densities, evaluate $p_{c_1}(\mathbf{x})$ and $p_{c_2}(\mathbf{x})$, and select the class with higher probability or
  - Select the majority class within k nearest neighbors
- We will then observe that
  - Larger kernel support / larger k increases bias and lowers variance
  - Smaller kernel support / smaller k lowers bias and increases variance
- Analogously, we can use the notion of bias/variance also in our exercise task of unsupervised density estimation

Bonus Slide: Current Research

- Just if you are curious — this is **strictly voluntary**:

- Here is a video on the recent ICML paper "Maximum Likelihood With Bias-Corrected Calibration is Hard-To-Beat at Label Shift Adaptation" by Avanti Shrikumar from Stanford University:
  `https://www.youtube.com/watch?v=ZBXjE9QTruE`

- It adapts the prior of a deep learning model without retraining the model, using tools from our lecture
- I find the talk also generally instructive as an academic presentation

- Enjoy!