Overfitting and Generalization Performance

October 11, 2020

Introduction

General Aim

Given training sample

$$(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$$

learn a predictor $h_n : \mathbb{R}^d \to \mathbb{R}$ that predicts y given new x.

Empirical Risk Minimization (ERM)

Minimize training risk: $\frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$ given a loss function ℓ.

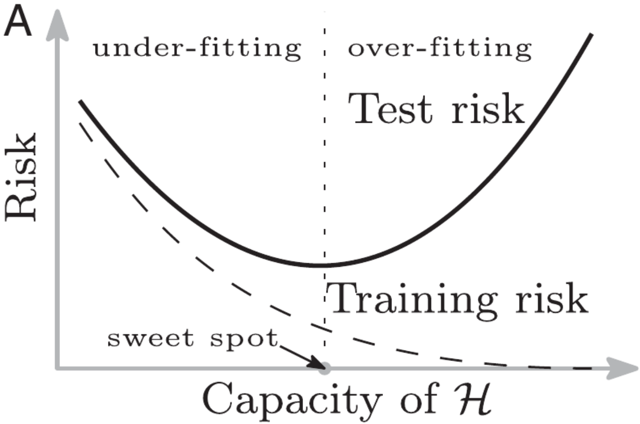
Generalization

- Find h_n that performs well on unseen data.
- Minimize true risk: $E[\ell(h(x), y)]$ where (x, y) drawn independently from P.

"Classical" thinking

- Finding a balance between underfitting and overfitting.
- "Bias-Variance Tradeoff"
- 0 training error does not tend to generalize well.
- Control function class $\mathcal{H}$ implicitly or explicitly.

Generalization of performance



Classical curve from bias variance tradeoff.

Modern practice

- Modern ML methods such as large neural networks and other non-linear predictors have very low to no training risk
- NN architectures chosen such that interpolation can be achieved.
- Works even when training data have high levels of noise.

Examples Even when levels of noise

"Double Descent"

Belkin proposed curve the extends beyond the poiny of interpolation
Observed empirically in a range of datasets

Double Descent

Graph picture, explain points of the graph.

Double Descent

Possible explanation by inductive bias and Occam's razor.

Empirical Evidence

RFFs. Might wanna explain more about RFFs approximating RKHS.

Empirical Evidence

Neural Networks. (Might be hard to explain why SGD is the inductive bias.)

Historical absence

Appendix on Approximation Theorem

On why they choose a function with a smaller norm in RKHS.