Overfitting

Riheng ZHU
zrh0712@outlook.com

Chapter 1

Introduction

Overfitting is the phenomenon (Figure 1)where fitting the observed facts (data) well no longer indicates that we will get a decent out-of-sample error, and may actually lead to the opposite effects. The model uses its additional degrees of freedom to fit idiosyncrasies in the data (ex. noise). Overfitting can even happen when the hypothesis set contains only functions which are *far simpler* to the target function, and so the plots thickens.

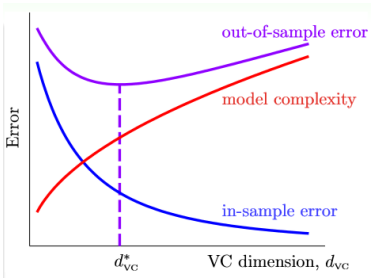

Figure 1. Overfitting

Chapter 2

Hazard of overfitting

Fours reasons for serious overfitting: Stochastic noise, deterministic noise, data size and excessive power.

Stochastic Noise The stochastic noise is the data do not lie in the target function curve.

Deterministic Noise The deterministic noise is the part of the target function which cannot be modeled by the hypothesis. There are two basic difference between stochastic noise and deterministic noise:

1. *depends on H* The same data set will have different deterministic noise depending on which model we choose.
2. *fixed for a given x* If we generate the same data(x values), the deterministic noise would not change, but the stochastic noise would.

The bias-variance decomposition is a useful tool for understanding how noise affects performances:

$$\mathbb{E}_D[E_{out}] = \sigma^2 + bias + vars \tag{1}$$

where

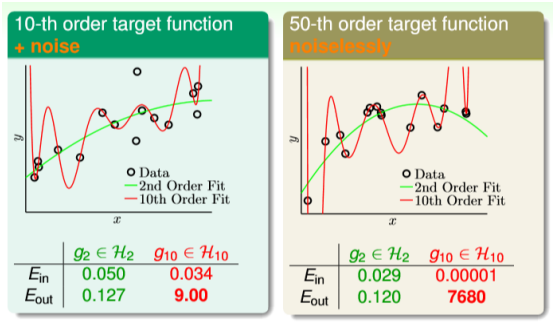

Figure 2. Effects of stochastique noise and deterministic noise

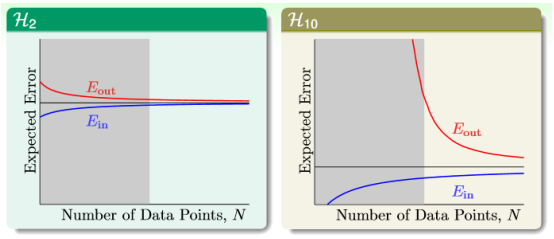

Figure 3. Effects of dataSize for overfitting

- σ^2 is the variance of stochastic noise.
- *bias* is directly related to the deterministic noise in that it captures the model's inability to approximate f.
- *vars* is indirectly impacted by both noise, capturing a model's susceptibility to being led astray by noise.

Data Size Overfitting is occurring for N in the shaded gray region because by choosing H_{10} which is better E_{in}, and you get worse E_{out} as shown in Figure 3.
Excessive power model complexity.

Chapter 3

Regularization

Chapter 4

Validation

Chapter 5

Three Learning Principles