ECE595 / STAT598: Machine Learning I

Lecture 25.3: Generalization Bound - Handling M Hypothesis

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline

- Lecture 25 Generalization
- Lecture 26 Growth Function
- Lecture 27 VC Dimension

Today's Lecture:

- M Hypothesis
  - PAC framework
  - Guarantee and Possibility
  - The M factor
- Generalization Bound
  - $\mathcal{H}$
  - f
  - Lower and upper limits
- Handling M hypothesis
  - A preview

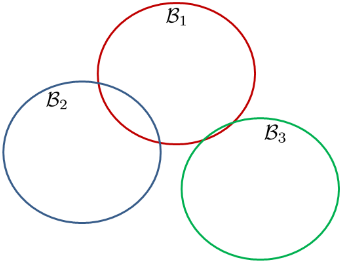
Overcoming the M Factor

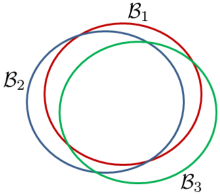
- The $\mathcal{B}$ad events $\mathcal{B}_m$ are

$$\mathcal{B}_m = \{|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\}$$

- The factor M is here because of the Union bound:

$$\mathbb{P}[\mathcal{B}_1 \textbf{ or } \ldots \textbf{ or } \mathcal{B}_M] \leq \mathbb{P}[\mathcal{B}_1] + \ldots + \mathbb{P}[\mathcal{B}_M].$$

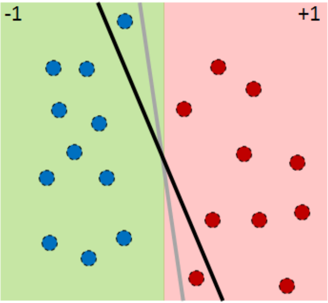




Counting the Overlapping Area

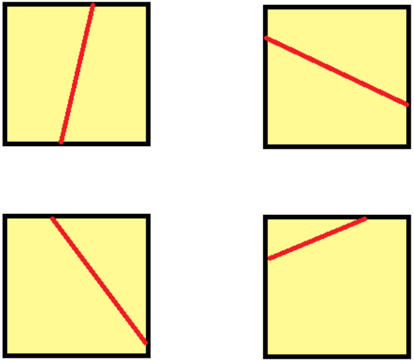
- ΔE_{out} = change in the $+1$ and -1 area
- Example below: Change a little bit
- ΔE_{in} = change in labels of the training samples
- Example below: Change a little bit, too
- So we should expect the probabilities

$$\mathbb{P}[|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon] \approx \mathbb{P}[|E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon].$$



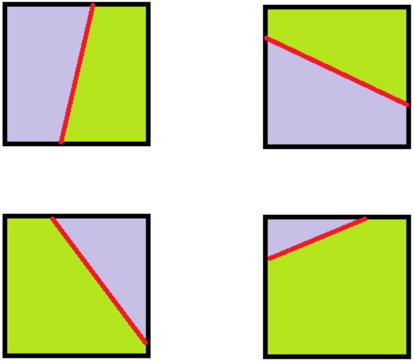
Looking at the Training Samples Only

- Here is a our goal: Find something to replace M.
- But M is big because the whole input space is big.
- Let us look at the input space.



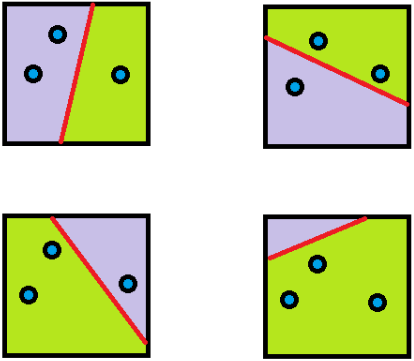
Looking at the Training Samples Only

- If you move the hypothesis a little, you get a different partition
- Literally there are infinitely many hypotheses
- This is M



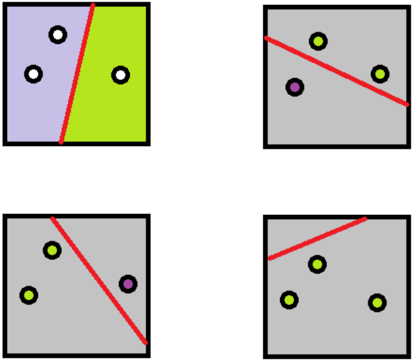
Looking at the Training Samples Only

- Here is a our goal: Find something to replace M
- But M is big because the whole input space is big
- Can we restrict ourselves to just the training sets?



Looking at the Training Samples Only

- The idea is: Just look at the training samples
- Put a mask on your dataset
- Don't care until a training sample flips its sign



Reading List

- Learning from Data, chapter 2
- Martin Wainwright, High Dimensional Statistics, Cambridge University Press 2019. (Chapter 2)
- CMU Note https://www.cs.cmu.edu/~mgormley/courses/10601-s17/slides/lecture28-pac.pdf
- Stanford Note http://cs229.stanford.edu/notes/cs229-notes4.pdf