

CSE 417T

Introduction to Machine Learning

Lecture 5

Instructor: Chien-Ju (CJ) Ho

Logistics: Office Hours

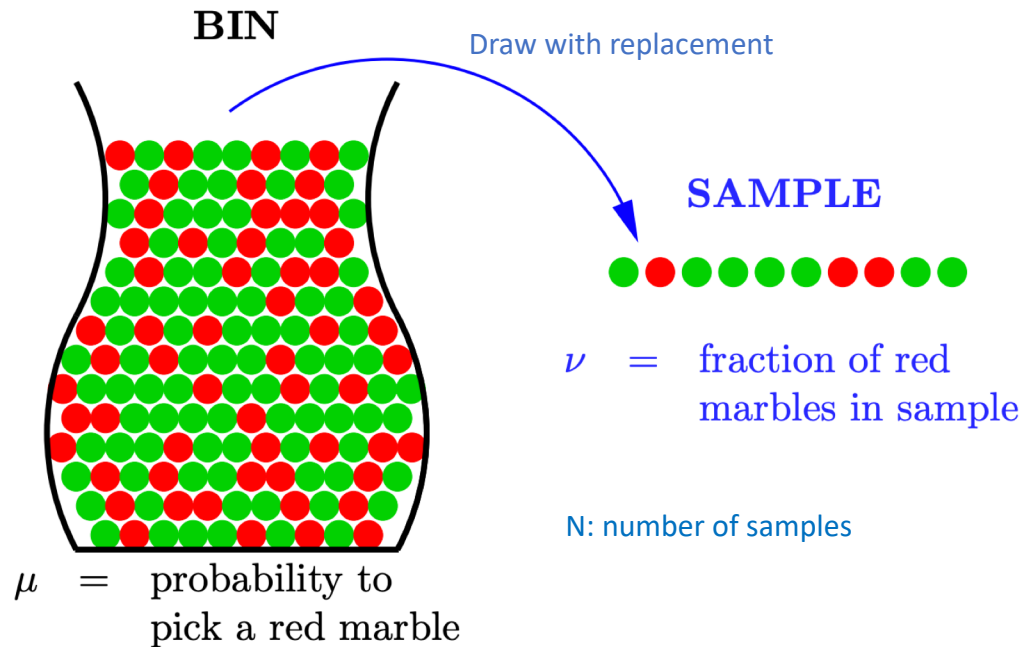
- TA Office Hours

	Office Hours		Location
Mondays	1:30-3:00PM (Flora)	4:30-6:00PM (Heming)	Jolley 431
Tuesdays	9:30-11:00AM (Xinyu)	3:00-4:30PM (Yi)	Jolley 431
Wednesdays	10:00-11:30AM (Ruoyao)	12:30-2:00PM (Ziyang)	Jolley 431
Thursdays	2:30-4:00PM (Tong)	4:00-5:30PM (Connor)	Jolley 431
Fridays	12:30-2:00PM (Brendan)	2:30-4:00PM (Jiahao)	Jolley 309
Sundays	5:00-6:30PM (Ina)		Jolley 408

- Utilize the office hours early to avoid the crowd.

Recap

Hoeffding's Inequality



$$\Pr[|\mu - \nu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Define $\delta = \Pr[|\mu - \nu| > \epsilon]$

- Fix δ , ϵ decreases as N increases
- Fix ϵ , δ decreases as N increases
- Fix N , δ decreases as ϵ increases

Informal intuitions of notations
 N : # sample
 δ : probability of “bad” event
 ϵ : error of estimation

Fixed Hypothesis and Finite Hypothesis Set

- Given dataset $D = \{(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)\}$
 - $E_{in}(h) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{n=1}^N \mathbb{I}[h(\vec{x}_n) \neq f(\vec{x}_n)]$ [In-sample error, analogy to v]
 - $E_{out}(h) \stackrel{\text{def}}{=} \Pr_{\vec{x} \sim P(\vec{x})} [h(\vec{x}) \neq f(\vec{x})]$ [Out-of-sample error, analogy to μ]

- Learning bounds

- Fixed h (verification)

$$\Pr[|E_{out}(h) - E_{in}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- Finite hypothesis set: learn $g \in \{h_1, \dots, h_M\}$

$$\Pr[|E_{out}(g) - E_{in}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

Dealing with Infinite Hypothesis Set: $M \rightarrow \infty$

- Most of the practical cases involve $M \rightarrow \infty$
- Instead of # hypothesis, counting “effective” # hypothesis

- Dichotomy

- Informally, consider it as “data-dependent” hypothesis
 - Characterized by both H and N data points $(\vec{x}_1, \dots, \vec{x}_N)$

$$H(\vec{x}_1, \dots, \vec{x}_N) = \{h(\vec{x}_1), \dots, h(\vec{x}_N) | h \in H\}$$

- The set of possible prediction combinations $h \in H$ can induce on $\vec{x}_1, \dots, \vec{x}_N$

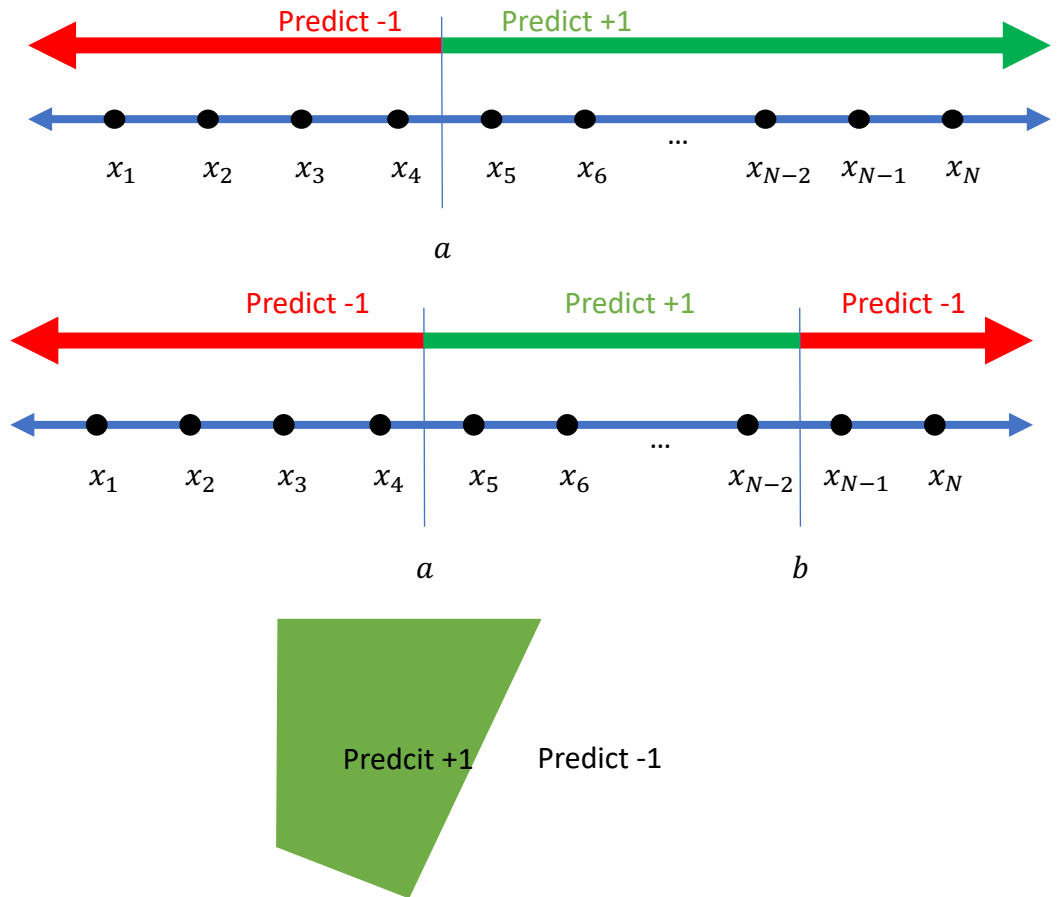
- Growth function

- Largest number of dichotomies H can induce across all possible data sets of size N

$$m_H(N) = \max_{(\vec{x}_1, \dots, \vec{x}_N)} |H(\vec{x}_1, \dots, \vec{x}_N)|$$

Examples on Growth Functions

- H = Positive rays
 - $m_H(N) = N + 1$
- H = Positive intervals
 - $m_H(N) = \binom{N+1}{2} + 1 = \frac{N^2}{2} + \frac{N}{2} + 1$
- H = Convex sets
 - $m_H(N) = 2^N$
- $m_H(N) \leq 2^N$ for all H and for all N



Why Growth Function?

- Growth function $m_H(N)$
 - Largest number of “effective” hypothesis H can induce on N data points
 - A more precise “complexity” measure for H
 - Goal: Replace M in finite-hypothesis analysis with $m_H(N)$
 - With prob at least $1 - \delta$, $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$
- VC Generalization Bound (VC Inequality, 1971)
With prob at least $1 - \delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_H(2N)}{\delta}}$$


Brief Lecture Notes Today

The notes are not intended to be comprehensive. They should be accompanied by lectures and/or textbook.
Let me know if you spot errors.

How to Bound Growth Function

- What we know so far
 - $H =$ Positive rays: $m_H(N) = N + 1$
 - $H =$ Positive intervals: $m_H(N) = \binom{N+1}{2} + 1$
 - $H =$ Convex sets: $m_H(N) = 2^N$
- What about $H =$ 2-D Perceptron?
 - $m_H(3) = 8, m_H(4) = 14, \dots$
 - Generally hard to write down the growth function exactly
- Alternative approach
 - Use the idea of "break point" (defined next) to bound the growth function

Bounding Growth Function

- More definitions....
 - Shatter:
 - H **shatters** $(\vec{x}_1, \dots, \vec{x}_N)$ if $|H(\vec{x}_1, \dots, \vec{x}_N)| = 2^N$
 - H can induce all label combinations for $(\vec{x}_1, \dots, \vec{x}_N)$
 - Break point
 - k is a **break point** for H if no data set of size k can be shattered by H
- Key result:
 - If k is a break point for H , $m_H(N)$ is polynomial in N .
In particular $m_H(N) = O(N^{k-1})$ 
 - If there are no break points for H , $m_H(N) = 2^N$

A bit more accurately:

- $m_H(N) \leq \sum_{i=1}^{k-1} \binom{N}{i}$, or
- $m_H(N) \leq N^{k-1} + 1$

Practice

- What is the break point for
 - Positive rays: $k=2, 3, 4, \dots$ are break points
 - Positive intervals: $k=3, 4, 5, \dots$ are break points
 - Convex sets: there are no break points
 - 2-dimensional perceptron (2-D linear separators)
 - $K=4, 5, \dots$ are break points

Why Break Points?

- VC Generalization Bound

With prob at least $1 - \delta$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4m_H(2N)}{\delta}}$$

- In the following discussion, we treat δ as a constant [i.e., with high probability, the following is true]

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{\frac{1}{N} \ln m_H(N)}\right)$$

[For example, we can set δ to be a small constant, say 0.01. Then every time we wrote the above inequality, we mean that it is true with probability at least 99%.]

Applying Break Points in VC Bound

- If there are no break point ($m_H(N) = 2^N$)

$$E_{out}(g) \leq E_{in}(g) + \text{Constant}$$

(This implies that learning is infeasible even when $N \rightarrow \infty$)

- If k is a break point for H , i.e., $m_H(N) = O(N^{k-1})$

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{(k-1) \frac{\ln N}{N}}\right)$$

VC Dimension

- VC Dimension of H : $d_{vc}(H)$ or just d_{vc}
 - The VC dimension of H is the **largest N such that $m_H(N) = 2^N$** .
 - $d_{vc}(H) = \infty$ if $m_H(N) = 2^N$ for all N .
 - Equivalently,
 - let k^* be the smallest break point for H , the VC dimension of H is $k^* - 1$
- Plug the definition into VC Generalization Bound

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{d_{vc} \frac{\ln N}{N}}\right)$$

Discussion on the VC Theory

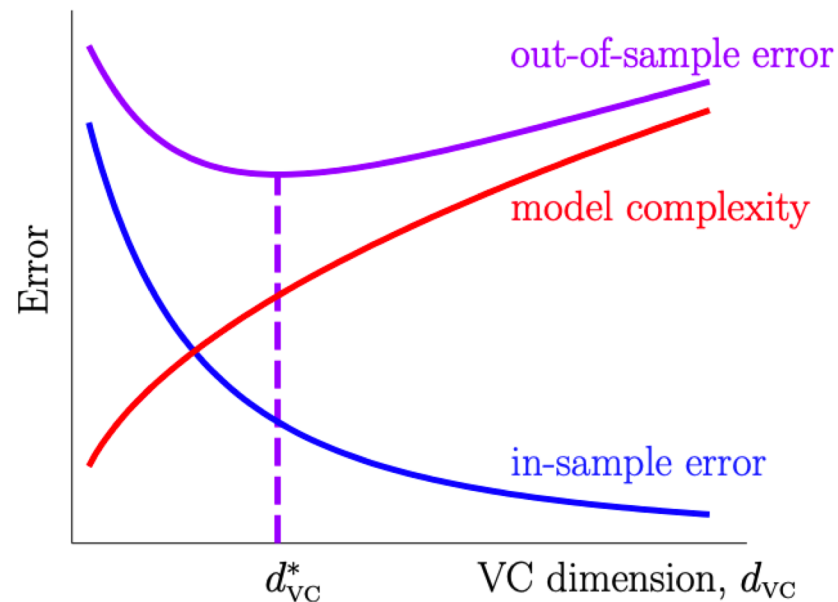
- VC Bound

$$E_{out}(g) \leq E_{in}(g) + O\left(\sqrt{d_{VC} \frac{\ln N}{N}}\right)$$

- The bound is “loose”
 - Depends only on H and N
 - The analysis is loose in many places
- However, it qualitatively characterizes the practice reasonably well
 - (the bound is roughly equally loose for every H)

Discussion on the VC Theory

- It establishes the feasibility of learning for infinite hypothesis set.
- It provides nice intuitions on what's happening underneath ML.
 - A single parameter to characterize complexity of H



Bias-Variance Decomposition

Another theory of generalization

Real-Value Target and Squared Error

- So far, we focus on binary target function and binary error
 - Binary target function $f(\vec{x}) \in \{-1, 1\}$
 - Binary error $e(h(\vec{x}), f(\vec{x})) = \mathbb{I}[h(\vec{x}_n) \neq f(\vec{x}_n)]$
- What about real-value functions [called “**regression**’] and squared error?
 - Real-value target function $f(\vec{x}) \in \mathbb{R}$
 - Squared error $e(h(\vec{x}), f(\vec{x})) = (h(\vec{x}_n) - f(\vec{x}_n))^2$
- What can we say about $E_{out}(g)$?
 - $E_{out}(g) = \mathbb{E}_{\vec{x}}[e(g(\vec{x}), f(\vec{x}))] = \mathbb{E}_{\vec{x}}[(g(\vec{x}) - f(\vec{x}))^2]$