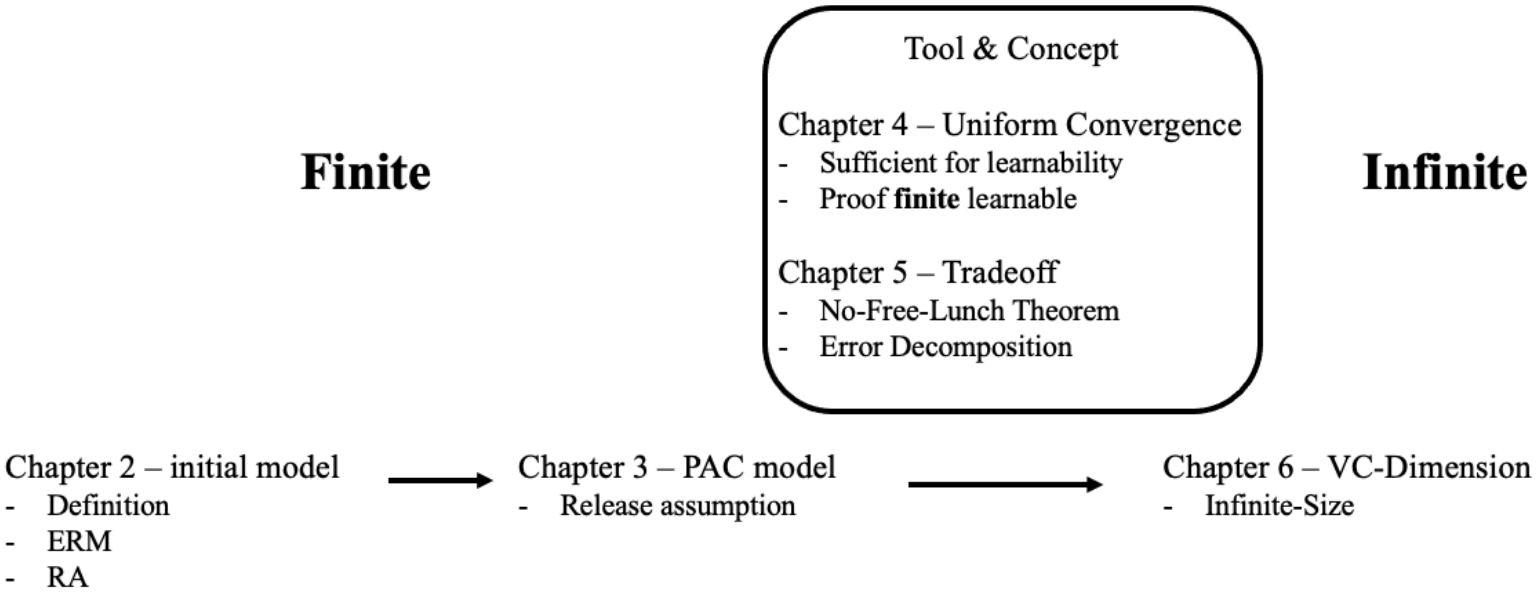
Understand Machine Learing

July 27, 2022

Road Map


Tool & Concept
Chapter 4 – Uniform Convergence
- Sufficient for learnability
- Proof **finite** learnable
Chapter 5 – Tradeoff
- No-Free-Lunch Theorem
- Error Decomposition
Finite
Infinite
Chapter 2 – initial model
- Definition
- ERM
- RA
Chapter 3 – PAC model
- Release assumption
Chapter 6 – VC-Dimension
- Infinite-Size

Definition

Domain set: An arbitrary set, χ. This is the set of objects that we may wish to label.

Definition

Label set: The Answer of the Domain set, usually $\{0, 1\}$ or $\{-1, +1\}$

Definition

Training data: $S = ((x_1, y_1), \cdots, (x_m, y_m))$ is a sequence of labeled domain points.

Definition

$h : \chi \rightarrow y$, a prediction function, also called a predictor, hypothesis, classifier.

Definition

Formally, the learner should choose tin advance a set of predictors. This set is called a hypothesis class and is denoted by H. Each $h \in H$ is a function mapping from χ to y.

Definition

A data-generation model: We now explain how the training data is generated by som probability distribution. Let us denote that probability distribution over χ by D.

Definition

Measure of Success: To know is the output is good or not, we define the loss function to check it

Definition

1. **True error:**
   $L_{D,f}(h) = \mathbb{P}_{x\sim D}[h(x) \neq f(x)] = D(\{x \mid h(x) \neq f(x)\})$
2. **Training error:**
   $L_S(h) = \frac{|\{i \in m \mid h(x_i) \neq y_i\}|}{m}$ where $[m] = \{1, \cdots, m\}$

Definition

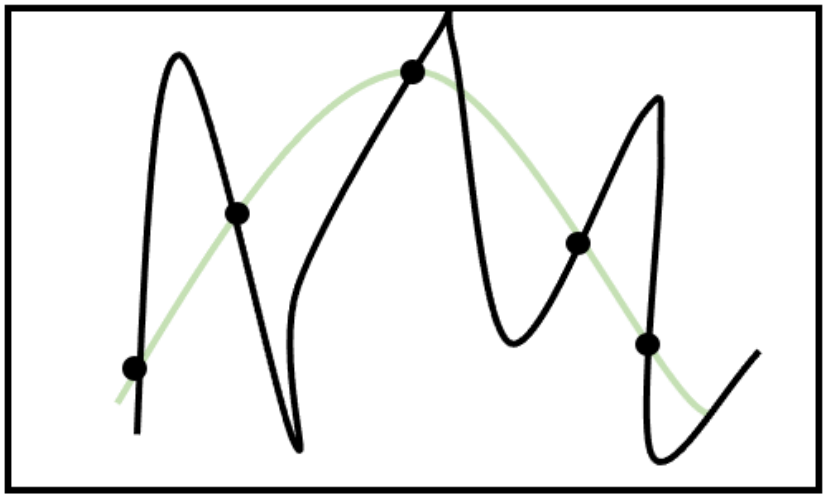


Figure 2-1, training error is 0 but true error is bad

Definition

- We denote the probability of getting a non-representative sample by δ, and call $(1-\delta)$ the **confidence parameter** of our prediction
- The **accuracy parameter**, commonly denoted by ϵ. We interpret the event $L_{(D,f)}(h_s) > \epsilon$ as a failure of the learner

Empirical Risk Minimization

The method to proof the model is to minimize the loss function by using training data, i.e. we check $L_S(h)$

Empirical Risk Minimization

The ERM_H learner uses the ERM rule to choose a predictor $h \in H$, with the lowest possible error over S. Formally

$$h_S = \text{ERM}_H(S) \in \arg\min_{h \in H} L_S(h)$$

Overfitting

cause by ERM, the data is too fit the training set

example: $h_s(x) = \begin{cases} y_i \text{ if } \exists\ i \in [m] \text{ s.t. } x_i = x \\ 0 \text{ otherwise} \end{cases}$

Inductive bias

Restricting the learner to choosing a predictor from H are often called an **inductive bias**. In following statement, we will proof that when we have some assumption, H is a finite set and have enough quantity of data set, then we can avoid the overfitting problem.

Some assumption

The Realizability Assumption: There exists $h^* \in H$ s.t. $L_{(D,f)}(h^*) = 0$. Note that this assumption implies that with probability 1 over random samples, S where the instances of S are sampled according to D and are labeled by f, we have $L_S(h^*) = 0$

Some assumption

The i.i.d. assumption : The examples in the training set are independently an identically distributed (i.i.d) according to the distribution D. We denote this assumption by $S \sim D^m$ where m is the size of S, and D^m denotes the probability over m-tuples induced by applying D to pick each element of the tuple independently of the other members of the tuple.

Proof Section

The goal is to proof when $m \geq \frac{\log(|H|/\delta)}{\epsilon}$, then $L_{D,f}(h_s) \leq \epsilon$

Proof Section

Let H_B be the set of "bad" hypotheses, that is,

$$H_B = \{h \in H \mid L_{D,f}(h) > \epsilon\}$$

Proof Section

In addition, let

$$M = \{S|_x \mid \exists\ h \in H_B, L_S(h) = 0\}$$

be the set of **misleading sample**(they are bad but $L_S(h_S) = 0$)

Proof Section

by definition, we can write

$$\left\{S|_x \mid L_{(D,f)}(h_S) > \epsilon\right\} \subseteq M (\star_1)$$

We can rewrite M as (thought it's intersection was not empty)

$$M = \cup_{h \in H_B} \left\{S|_x \mid L_S(h) = 0\right\} (\star_2)$$

Proof Section

by $(\star_1), (\star_2)$,

$$D^m(\{S|_x \mid L_{(D,f)}(h_s) > \epsilon\}) \leq D^m(M)$$

$$= D^m(\cup_{h \in H_B} \{S|_x \mid L_S(h) = 0\})(\star_3)$$

Proof Section

LEMMA(Union Bound) For any two sets A, B and a distribution D we have

$$D(A \cup B) \leq D(A) + D(B)$$

Proof Section

and the $(\star_3)$ can be bound like this

$$D^m(\{S|_x \mid L_{(D,f)}(h_S) > \epsilon\}) \leq \sum_{h \in H_B} D^m(\{S|_x \mid L_S(h) = 0\})(\star_4)$$

Proof Section

Next, we fix h_B on the bad hypothesis $h_B \in H_B$

$\Longrightarrow \; L_{(D,f)}(h) > \epsilon$

Proof Section

because the event are i.i.d, we get that

$$D^m(\{S|_x \mid L_S(h) = 0\}) = D^m(\{S|_x \mid \forall i, h(x_i) = f(x_i)\})$$

$$= \prod_{i=1}^{m} D(\{x_i \mid h(x_i) = f(x_i)\})(\star_5)$$

Proof Section

check the each individual sampling of an element of the training set, we have

$$D(\{x_i \mid h_B(x_i) = y_i\}) = 1 - D(\{x \mid h_B(x) \neq f(x)\})$$
$$= 1 - L_{D,f}(h_B) \leq 1 - \epsilon$$

Proof Section

put it to the $(\star_5)$ and use the inequality $1 - \epsilon \leq e^{-\epsilon}$

$$D^m(\{S|_x \mid L_S(h_B) = 0\}) \leq (1-\epsilon)^m \leq e^{-\epsilon m}(\star_6)$$

Proof Section

$$D^m(\{S|_x \mid L_{(D,f)}(h_s) > \epsilon\}) = |H_B| D^m(\{S|_x \mid L_S(h_B) = 0\})$$

$|H_B|$ means the cardinality(element number) of H_B

Proof Section

put $(\star_6)$ back to $(\star_4)$

$$D^m(\{S|_x \mid L_{(D,f)}(h_S) > \epsilon\}) \leq |H_B|e^{-\epsilon m} \leq |H|e^{-\epsilon m}$$

Proof Section

In this equation, we can know that when m increase, the overfitting hypothesis's probability (where the hypothesis of L_S is small but $L_{(D,f)}$ is big, i.e. $D^m(\{S|_x \mid L_{(D,f)}(h_S) > \epsilon\})$ will decrease.

Since $D^m(\{S|_x \mid L_{(D,f)}(h_S) > \epsilon\})$ is δ, and have a nature log, we get

$$m \geq \frac{\log(|H|/\delta)}{\epsilon}$$