## Road Map

In chapter2, we will consider the initial case about statistic estimate, $\cdots$

## Chapter2: A Gentle Start

This chapter is talking about a general model of machine learning and common error.

### 2.1 Formal Model.

- **The learner's input**

· **Domain set:** An arbitrary set, $\chi$. This is the set of objects that we may wish to label.

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

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

- **The learner's output**

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

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 $\chi$ to $y$.

- **Other tools for ML**

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

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

(a) **True error:** $L_{D,f}(h) = \mathbb{P}_{x\sim D}[h(x) \neq f(x)] = D(\{x \mid h(x) \neq f(x)\})$

(b) **Training error:** $L_S(h) = \dfrac{|\{i \in m \mid h(x_i) \neq y_i\}|}{m}$ where $[m] = \{1, \cdots, m\}$

picture here **(same training error but different true error)**

· We denote the probability of getting a non-representative sample by $\delta$, and call $(1-\delta)$ the **confidence parameter** of our prediction

· The **accuracy parameter**, commonly denoted by $\epsilon$. We interpret the event $L_{(D,f)}(h_s) > \epsilon$ as a failure of the learner

### 2.2 Improve Model.

**Empirical Risk Minimization**

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

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

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

**Overfitting**

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

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

### 2.3 The upper bound of $L_{(D,f)}(h_s)$ in finite hypothsis.

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 $L_S(h^*) = 0$
- **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 $m \geq \dfrac{\log(|H|/\delta)}{\epsilon}$, then $L_{D,f}(h_s) \leq \epsilon$

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

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

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 $L_S(h_S) = 0$)

by definition, we can write

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

We can rewrite $M$ as

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

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

$$D^m(\{\, S|_x \mid L_{(D,f)}(h_s) > \epsilon \,\}) \leq D^m(M) = D^m(\cup_{h \in H_B} \{\, S|_x \mid L_S(h) = 0 \,\})(\star_3)$$

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

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

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

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

**Next, we focus on the bad hypothesis** $h \in H_B \implies L_{(D,f)}(h) > \epsilon$
because the event are i.i.d, we get that

$$D^m(\{\, S|_x \mid L_S(h) = 0 \,\}) = D^m(\{\, S|_x \mid \forall i, h(x_i) = f(x_i) = f(x_i) \,\}) = \prod_{i=1}^{m} D(\{\, x_i \mid h(x_i) = f(x_i) \,\})(\star_5)$$

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

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

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

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

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

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

## Chapter5: The Bias-Complexity Tradeoff

First, we can decomposition the error to following terms:

$$L_D(h_s) = \epsilon_{app} + \epsilon_{est} \text{ where: } \epsilon_{app} = \min_{h \in H} L_D(h),\ \epsilon_{est} = L_D(h_s) - \epsilon_{app}$$

**Error type**

· **The Approximation Error($\epsilon_{app}$)**

· **The Estimation Error($\epsilon_{est}$)**

**Bias-complexity Tradeoff**