Machine Learning

Xuzhe Xia

January 1, 2023

Contents

Chapter 0

Introduction

0.1 Basic concepts in Machine Learning

Definition 1 (Dataset)**.**

1. **Training dataset**:
2. **Testing dataset**:
3. **Validation dataset**:

Definition 2 (Supervised and unsupervised learning)**.**

1. **Supervised learning**:
2. **Unsupervised learning**:
3. **Reinforcement learning**:

Definition 3 (Online and Batch Learning)**.**

- **Online Learning**:
- **Batch Learning**:

Definition 4 (Parametric and non-parametric models)**.**

- **Parametric model**: The model have a fixed number of parameters.
- **Non-parametric model**: The number of parameters grows with the amount of training data.

0.2 Notations

Chapter 1

Foundations

1.1 Definitions and Finite hypothesis classes

1.1.1 A Formal Model

Definition 5 (Domain set). An arbitrary set $\mathcal{X}$, the set of objects that we may wish to label.

Definition 6 (Label set). (temporary) A two-element set, usually $\{0, 1\}$, let $\mathcal{Y}$ denote the **set of possible labels**.

The learner's input:

Definition 7 (Training data). $S = ((x_1, y_1) \cdots (x_m, y_m))$ is a finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$, that is, a sequence of labeled domain points.

A simple data-generation model:

1. First we have some probability distribution over $\mathcal{X}$ denoted by $\mathcal{D}$.
2. (temporary) Also we assume there exist some true labeling function, $f : \mathcal{X} \to \mathcal{Y}$, such that $y_i = f(x_i)$ for all i, and it is unknown to the learner.
3. Each pair in the training dataset is generated by first sampling a point x_i according to $\mathcal{D}$ and the labeling it by the true labeling function f.

The learner's output:

Definition 8 (Hypothesis). Hypothesis: $h : \mathcal{X} \to \mathcal{Y}$.

Measure of success:

Definition 9 (True Risk).

The probability that it does not predict the correct label on a random data point generated by the aforementioned underlying distribution $\mathcal{D}$:

$$L_{\mathcal{D},f}(h) := \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[h(x) \neq f(x)] := \mathcal{D}(\{x : h(x) \neq f(x)\})$$

Also means that the error is the probability of randomly choosing an example x for which $h(x) \neq f(x)$.

Notation. Denote $A(S)$ as the hypothesis that a learning algorithm A returns upon receiving the training data S.

Summary. Begin with training data, feed into the learning algorithm $A(S)$, and output the hypothesis h used for predicting. Then measure the success of h using loss function $L_{\mathcal{D},f}(h)$ which depends on the distribution $\mathcal{D}$ and the ture label function f, which tells us the probability of prediction a wrong label on a random data point.

1.1.2 Empirical Risk Minimization

Definition 10 (Empirical Risk; Training error).

$$L_S(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

where $[m] = \{1, \cdots, m\}$.

Intuition. Training error is the probability making wrong prediction within the training dataset.

Definition 11 (Empirical Risk Minimization). The process of finding h that minimizing $L_S(h)$ is called **Empirical Risk Minimization**, and we such h the **ERM hypothesis**.

1.1.3 Empirical Risk Minimization with Inductive Bias

Definition 12 (Hypothesis class). **Hypothesis class** ($\mathcal{H}$) is the set of hypothesis that the learner choose in advance without have any knowledge about the training set, but might depend on some prior knowledge about the problem to be learned.

Definition 13 (Inductive bias). **Inductive bias** is the restriction that we only allow the learner to choose hypothesis from $\mathcal{H}$.

Definition 14 (ERM with Inductive Bias)**.** $\text{ERM}_{\mathcal{H}}$ is the process of finding

$$h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$$

Notation. Let h_S denote a result of applying $\text{ERM}_{\mathcal{H}}$ to S.

Remark.

- A **fundamental questions** is: over which hypothesis classes $\text{ERM}_{\mathcal{H}}$ learning will not result in overfitting?
- A **fundamental trade-off** is: choosing a more restricted hypothesis class better protects us against overfitting but at the same time might cause us a stronger inductive bias.

In this chapter we make a very strong assumption:

Definition 15 (The Realizability Assumption)**.** We assume that for the hypothesis class $\mathcal{H}$ the learner have chosen, there exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$, intuitively means there exists a perfect hypothesis in this hypothesis class.

The realizability assumption implies that:

Corollary. For any training set S sampled according to $\mathcal{D}$ and labeled by f, we have that $L_S(h_S) = 0$.

Proof. $L_{(\mathcal{D},f)}(h^*) = 0 \Rightarrow L_S(h^*) = 0$, which means that $\min_{h \in \mathcal{H}} L_S(h) = 0$, so $L_S(h_S) = 0$. Note that h_S is not necessarily equal to h^*. □

Observe. Note that when the learning algorithm have only accessed to train set S, any error with respect to the distribution $\mathcal{D}$ should depend on the relation between $\mathcal{D}$ and S. Therefore, we make the **i.i.d.** assumption:

Definition 16 (The i.i.d. assumption)**.** The sample in the training set are independently and identically distributed (i.i.d.) according to the distribution $\mathcal{D}$. We denote this assumption by $S \sim \mathcal{D}^m$.

Observe. Even with the i.i.d. assumption, there is still a small chance that the training set is nonrepresentative.

Definition 17 (Confidence parameter)**.** Denote the probability of getting a nonrepresentative sample by δ, and call $(1 - \delta)$ the **confidence parameter** of our prediction.

Definition 18 (Accuracy parameter).
We introduce **accuracy parameter** ε such that we interpret

- the event $L_{(\mathcal{D},f)}(h_S) > \varepsilon$ as a **failure of the learner**,
- and if $L_{(\mathcal{D},f)}(h_S) \leq \varepsilon$, we view the output of the algorithm as an **approximately correct predictor**.

Lemma 1. The upper bounding of the probability of sampling m training data that leads to the failure of the learner is bounded by $|\mathcal{H}|e^{-\varepsilon m}$, in another word:

$$D^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}|e^{-\varepsilon m}$$

where $S|_x = (x_1, \cdots, x_m)$ denotes the instances of the training set.

First define:

- $\mathcal{H}_B := \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \varepsilon\}$ be the **set of bad hypotheses**,
- and $M := \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\} = \bigcup_{h\in\mathcal{H}_B}\{S|_x : L_S(h) = 0\}$ be the **set of misleading samples**.

Intuition. A misleading sample is the training set $S|_x$ that there exists a bad hypothesis h perfectly fits $S|_x$ but does not perform well in general, that is, an easily-overfitted training set, or non-representative training set.

Remark. Every training data instance that leads to the failure of the learner is a misleading sample:

$$\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\} \subseteq M$$

Proof. For any $S|_x \in \{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}$, by the realizability assumption, we have that $L_S(h_S) = 0$. Also since $L_{(\mathcal{D},f)}(h_S) > \varepsilon$, so h_S is a bad hypothesis and $h_S \in \mathcal{H}_B$. We have find $h_S \in \mathcal{H}_B$ s.t. $L_S(h_S) = 0$, so $S|_x \in M$.

Now we derive:

$$\begin{aligned} D^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq D^m(M) &= D^m\Big(\bigcup_{h\in\mathcal{H}_B}\{S|_x : L_S(h) = 0\}\Big) \\ &\leq \sum_{h\in\mathcal{H}_B} D^m(\{S|_x : L_S(h) = 0\}) \\ &= \sum_{h\in\mathcal{H}_B} D^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \end{aligned}$$

By the i.i.d. assuptiom:

$$= \sum_{h\in\mathcal{H}_B} \left[\prod_{i=1}^{m} D(\{x_i : h(x_i) = f(x_i)\})\right]$$

Since $D(\{x_i : h(x_i) = f(x_i)\}) = 1 - L_{(\mathcal{D},f)}(h)$:

$$= \sum_{h \in \mathcal{H}_B} (1 - L_{(\mathcal{D},f)}(h))^m$$

Since $h \in \mathcal{H}_B$ so $L_{(\mathcal{D},f)} > \varepsilon$:

$$\leq |\mathcal{H}_B|(1-\varepsilon)^m \leq |\mathcal{H}_B| e^{-\varepsilon m} \leq |\mathcal{H}| e^{-\varepsilon m}$$

This complete the proof of lemma 1.

Corollary. Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0,1)$ and $\varepsilon > 0$, and let m be an integer hat satisfies

$$m \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$

Then for any labeling function f, and for any distribution $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}$, $L_{(\mathcal{D},f)}(h) = 0$), with probability of at least $1 - \delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, h_S, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \varepsilon.$$

Proof. By the lemma above,

$$D^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \varepsilon\}) \leq |\mathcal{H}| e^{-\varepsilon m}$$
$$\leq |\mathcal{H}| e^{-\varepsilon \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}} = \delta,$$

which completes the proof. □

Intuition. This tells us that for a sufficiently large training set, we are $(1 - \delta)$ confident that the probability for our $\text{ERM}_{\mathcal{H}}$ hypothesis making errors on a randomly chosen data is no larger than ε.

1.1.4 Exercise

1. Let $\mathcal{H}$ be a class of binary classifier over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. Show that the expected value of $L_S(h)$ over the choice of $S|_x$ equals $L_{(\mathcal{D},f)}(h)$, namely,

$$\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_S(h)] = L_{(\mathcal{D},f)}(h).$$

Solution:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S|_x \sim \mathcal{D}^m}[L_S(h)] &= \sum_{i=1}^{m} \frac{i}{m} \binom{m}{i} \mathcal{D}(\{x : h(x) = f(x)\})^{m-i} \mathcal{D}(\{x : h(x) \neq f(x)\})^i \\
&= \sum_{i=1}^{m} \frac{(m-1)!}{(i-1)!(m-i)!} (1 - L_{(\mathcal{D},f)}(h))^{m-i} L_{(\mathcal{D},f)}(h)^i \\
&= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} (1 - L_{(\mathcal{D},f)}(h))^{n-k} L_{(\mathcal{D},f)}(h)^{k+1} = L_{(\mathcal{D},f)}(h)
\end{aligned}
$$

1.2 PAC learning model

Definition 19 (PAC Learnability). A hypothesis class $\mathcal{H}$ is **Probably Approximately Correct (PAC) Learnable** if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm such that:

- For every $\varepsilon, \delta \in (0,1)$,
- for every distribution $\mathcal{D}$ over $\mathcal{X}$,
- and for every labeling function $f : \mathcal{X} \to \{0,1\}$,

and if the realizable assumption holds with respect to $\mathcal{H}$, $\mathcal{D}$, f,
then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h s.t. with probability of at least $1 - \delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \varepsilon$, and such hypothesis h is a **probably approximately correct solution**.

Definition 20 (Sample Complexity). **Sample complexity** $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ is the function of ε and δ that returns the minimal size of training set needed to guarantee a probably approximately correct solution.

Remark. $m_{\mathcal{H}}$ also depend on $\mathcal{H}$, for example when the hypothesis class is finite, $m_{\mathcal{H}}$ is proportional to $\log |\mathcal{H}|$.

Corollary. Every finite hypothesis class is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$$

Intuition. Since we know that when the sample size is larger that $\left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$, the resulting hypothesis is a probably approximately correct solution, so the minimal sample size needed is no larger than this value.

1.2.1 Agnostic PAC learning