Learning Theory

WilliamLiusy

August 27, 2024

1 PAC assumption

- There is a true data distribution D, where the train, valid, test set are all sampled from D.
- All data are I.I.D sampled.

2 Notations

Let h be a hypothesis, $\mathcal{H}$ be a class of hypotheses (that the learning algorithm might focus on).

- Generalization Error $\epsilon(h) := E_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq y\}]$
- Empirical Error $\hat{\epsilon}_S(h) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$, where S denotes the finite dataset

3 Empirical Risk Minimizer

3.1 Definition

Our goal in learning is to find a hypothesis to minimize the generalization error.

A intuitive learning algorithm is the ERM(Empirical Risk Minimizer).

The ERM estimator is to find the hypothesis with the least empirical error.

$$\hat{\epsilon}(h)_{ERM} = \operatorname*{argmin}_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$$

3.2 Uniform Convergence

The next few questions we are interested in are:

- To what extent can we claim about our prediction accuracy through finite data?
- How well can ERM do?

3.2.1 Generalization Error V.S. Empirical Error

In this section, we talk about the first question. We are going to compare $\epsilon(h)$ and $\hat{\epsilon}_S(h)$.

Since the data are I.I.D sampled from D, assign a random variable $Z_i := \mathbb{1}\{h(x^{(i)}) \neq y^{(i)}\}$, and thus $\hat{\epsilon}_S(h) = \frac{1}{m}\sum_{i=1}^{m} Z_i$ Then $\forall i = 1, ..., m, Z_i$ are I.I.D sampled from a Bernoulli distribution $Bern(\epsilon(h))$. By the Hoeffding's Inequality, we have

$$Pr[|\epsilon(h) - \hat{\epsilon}_S(h)| > \gamma] < 2e^{-2\gamma^2 m}$$

Suppose the hypothesis class $\mathcal{H}$ is finite with k elements.Then by union inequality,

$$Pr[\exists h \in \mathcal{H}, |\epsilon(h) - \hat{\epsilon}_S(h)| > \gamma] < 2ke^{-2\gamma^2 m}$$

To summarize,

1° Assume you want to guarantee that
With probability $1 - \delta$, $|\epsilon(h) - \hat{\epsilon}_S(h)| < \gamma$ for all $h \in \mathcal{H}$.
Then m should satisfy $m \geq \frac{1}{2\gamma^2} ln \frac{2k}{\delta}$

2° Assume you want to fix m, then
With probability $1 - \delta$, $|\epsilon(h) - \hat{\epsilon}_S(h)| < \sqrt{\frac{1}{2m} ln \frac{2k}{\delta}}$ for all $h \in \mathcal{H}$.

3.3 ERM hypotheis V.S Best In-Class Hypothesis

In this section, we talk about the second question.

According to the last section, assume with probability $1-\delta$, $|\epsilon(h)-\hat{\epsilon}_S(h)| < \gamma$ for all $h \in \mathcal{H}$

The Best hypothesis h^* in $\mathcal{H}$ is defined by $\underset{h\in\mathcal{H}}{\text{argmin}}\, E_{(x,y)\sim D}[\mathbb{1}\{h(x) \neq y\}]$

So, we have

$$\begin{aligned} \epsilon(h_{ERM}) &< \hat{\epsilon}_S(h_{ERM}) + \gamma \\ &\leq \hat{\epsilon}_S(h^*) + \gamma \\ &< \epsilon(h^*) + 2\gamma \end{aligned}$$

4 VC dimension

Definition 1 (VC dimension) *Let S be a set of unlabelled data $\{x^{(1)}, ..., x^{(k)}\}$. Call $\mathcal{H}$* ***shatters*** *S if for all possible labels $\{y^i\}_{i=1}^k$, there exists a $h \in \mathcal{H}$ s.t. $h(x^{(i)}) = y^{(i)}$ for all i = 1,...,k. The* ***VC dimension*** *(Vapnik-Chervonenkis) of $\mathcal{H}$ is defined by the cardinality of the largest S that can be shattered by $\mathcal{H}$. If the largest S is infinite, $VC(\mathcal{H}) = +\infty$*

e.g. 1 *Suppose n = 1, m = 3. Let $\mathcal{H}$ be all linear boundaries (that linear classifier considers).*