Statistics Learning Theory

Reference

Understanding Machine Learning: From Theory to Algorithms

Example

Imagine that we arrived at a small island. This island is rich in coconut. There are two types of coconut, sweet coconut and unsweetened coconut.
When we buy coconut, we can judge the sweetness from the color and hardness of coconut. Suppose we have a series of trading experience, we may sometimes buy sweet coconuts, and sometimes unsweetened coconuts.
How do we summarize this series of experience to learn how to judge whether coconut is sweet or not? This is a main issue in the statistics learning theory.

The learner's input

- Feature Set($\mathcal{X}$): An arbitary set, $\mathcal{X}$ which is the collections associated with the corresponding features in each coconut.
- Label set($\mathcal{Y}$):In our setting, we can set any two element set as our label set. In general, we use $\{0, 1\}$ or $\{-1, 1\}$.

Assocated with feature set and label set, the learner can formalize the set of training data $\mathcal{S}$. That is,
$\mathcal{S} = ((x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)) \subset \mathcal{X} \times \mathcal{Y}$.

The learner's output

The learner try to use the training data to produce the prediction rule(classifier) $h : \mathcal{X} \to \mathcal{Y}$. When learner encount the new coconut, she/he can use the h to predict the sweetness of coconut,

Environment Setting

There is a natural problem: how to measure the success of our prediction rule? To answer this problem, we need to give the environment setting:

- Data generate process over $\mathcal{X}$: we assume that the distribution $\mathcal{D}$ over $\mathcal{X}$ generates the training sample.
- Label function f: To simply our current problem, we assume there exist the perfect prediction rule(labeling function) such that $f : \mathcal{X} \to \mathcal{Y}$ and that $y_i = f(x_i)$. This assumption will be relaxed in the future.

Measures of success

Naturally, we can define the error of prediction rule(w.r.t $\mathcal{D}$ and f) as follows:

$$L_{D,f}(h) = P_{x\sim D}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\})$$

The next question turns out that how do the learner generate their prediction rule?

Empirical Risk Minimization

The natural way that the learner 'learn' the prediction rule is to choose those rule minimize the training sample error. We define the training error as follows:

$$L_S(h) = \frac{|\{i \in \{1, \cdots, m\} : h(x_i) \neq y_i\}|}{m}$$

We also call the training error as empirical error or empirical risk.

Empirical Risk Minimization

In this setting, the learner can choose $\hat{h}$ which satifying

$$\hat{h} = arg \min_{h:\mathcal{X}\to\mathcal{Y}} L_S(h)$$

as their prediction rule. Notice that:

- we call this selection paradigm as empirical risk minimization(ERM)
- The $\hat{h}$ may not be unique. That is, it is possible to find several $\hat{h}$ to satisfy ERM rule.

The ERM rule looks reasonable. However, it causes a important issue, that is, overfitting!

Overfitting

To explain what is overfitting, imaging our coconut example: suppose that, in the island, half of coconut is sweet and another half is unsweetened. Furthermore, the hardness completely determine the sweetness of coconut and The hard degree is converted to a value between 0 and 1 (the softest is 0 and the hardest is 1).
Assuming that the coconut distribution $\mathcal{D}$ is uniform distribution over [0, 1], and in this island, When the hardness is greater than or equal to one-half, the coconut must be sweet, and if it is less than one-half, it must be unsweetened(That is $f = I(x \geq \frac{1}{2})$).

Overfitting

The learner proposes the following ERM rule:

$$h_S(x) = \begin{cases} y_i & \textit{if} \quad \exists i \in \{1, \cdots, m\} \quad s.t. \quad x_i = x \\ 0 & \textit{otherwise} \end{cases}$$

It is celarly that the h_S indeed minimize the empirical risk($L_S(h_S) = 0$). However, the true risk $L_{D,f}(h_S) = \frac{1}{2}$
That is, ERM prediction rule works extremely well in the training sample but fails in the out-sample. The natureal question arise: How to correct the overfitting?

Correction of overfitting

One way to correct the overfitting problem is to restrict the possbile prediction rule to avoid to select the 'bad' prediction rule like the above example.
To be more specific, we only select the prediction rule in the space $\mathcal{H}$ which is called the hypothesis class.
The important issue in learning theory is about what kind of $\mathcal{H}$ can avoid the overfitting?
Intuitively, choosing a more restricted $\mathcal{H}$ better avoid the overfitting. However, it may lead to more bias. Therefore, there is a tradeoff between bias and overfitting!
In the next, we will prove in the finite $\mathcal{H}$ and some strong assumptions, the overfitting can be avoided.

Finite Hypothesis Classes- Assumptions

Assume that

- Realizability Assumption:There exist $h^* \in \mathcal{H}$ such that $L_{D,f}(h^*) = 0$
- i.i.d assumption: we assume that the training data $S = \{x_1, x_2, \cdots, x_m\}$ are sampled i,i,d from $\mathcal{D}$. Therefore, the distribution of S is D^m
- $|\mathcal{H}|$ is finite

Finite Hypothesis Classes - Theorem

Under above three assumptions, let $\delta \in (0, 1)$ and $\varepsilon > 0$ and let m be an interger that satisfies

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

Then, for any labeling function f, and for any distribution, $\mathcal{D}$, with probability of at least $1 - \delta$ over the choice of an sample S of size m, we have every ERM prediction rule h_S, it holds that

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

Finite Hypothesis Classes -Proof

We try to uppper bound

$$\mathcal{D}^m(\{S : L_{D,f}(h_S) > \varepsilon\})$$

Let

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

$$M = \{S : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

Where $\mathcal{H}_B$ collect those 'bad' prediction rule and M collect those sample to mislead the learner.

Finite Hypothesis Classes–Proof

Then we have

$$\mathcal{D}^m(\{S : L_{D,f}(h_S) > \varepsilon\}) \leq \mathcal{D}^m(M) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S : L_S(h) = 0\})$$

since

$$M = \cup_{h \in \mathcal{H}_B} \{S : L_S(h) = 0\}$$

and

$$\mathcal{D}^m(\{S : L_S(h) = 0\}) = \mathcal{D}^m(\{S : \forall i, h(x_i) = f(x_i))$$

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

and we know that $\forall h \in \mathcal{H}_B$

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

Finite Hypothesis Classes–Proof

Since $1 - \varepsilon \leq e^{-\varepsilon}$, we have

$$\mathcal{D}^m(\{S : L_S(h) = 0\}) \leq (1 - \varepsilon)^m \leq e^{-\varepsilon m}$$

Therefore,

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

Therefore, given (δ, ε), we can bound the above set when sample larege than

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