NOTES ON
STATISTICAL LEARNING THEORY

JIECAO CHEN
JIECA001@UMN.EDU

ABSTRACT. Most results of statistical learning theory take the form of so-called error bounds. This note contains introduction and brief summary of some key ideas and techniques to obtain those result.

1. INTRODUCTION

The main goal of statistical learning theory is to provide a framework for studying the problem of inference, which includes gaining knowledge, making prediction or constructing models from the date set. As it is put into a statistical framework, that is there are assumptions of statistical nature about the underlying phenomena.

V. Vapnik once described why we need theories:

Nothing is more practical than a good theory.

A theory of interference should be able to give the formal definition of words such as learning, generalization, overfitting and also to characterize the performance of learning algorithms so that it would eventually be able to help us to design better learning algorithms.

Two goals: make things more precise and derive new or improved algorithms

1.1. **Learning and Inference.** The process of inductive inference:

- Observe a phenomenon
- Construct a model of that phenomenon
- Make predictions using this model

The task of Machine Learning is to actually *automate* this process and the goal of Learning Theory is to *formalize* it.

To make our discussion simpler, in this note, we only consider a special case of above process which is the supervised learning framework for pattern recognition. In this special case, data set consists of instance-label pairs, where the value of label $\in \{-1, 1\}$. The task of a learning algorithm is to construct a function which is able to predict the corresponding label given the instance. A "Good" learning algorithm will give less errors when do such prediction.

A model that can exactly fit all instance-label pairs in a training data set may not always be the best model, as we know that the data set may sometime include noise, in such case, a model fitting all the data may give a worse prediction in the unseen data, such phenomenon is usually referred to as **overfitting**. A way to overcome **overfitting** is to look for **regularities**.

If there are many models available, and all of the well fit the data, we normally follow Occam's idea to choose the simplest model, which would more likely to be generalized from the training data to future data. This immediately raises the question of how to measure and qualify simplicity of a model.

There is no universal way of measuring simplicity and the choice of a specific measure inherently depends on the problem at hand.

A formula we believe:

Generalization = Data + Knowledge

Generalization can only come when one adds specific knowledge to the data. Each learning algorithm encodes specific knowledge.

1.2. **Assumptions.** Several more precise assumptions that made by the Statistical Learning Theory Framework:

- Probabilistic model of the phenomenon (or data generation process). Within this model, the relationship between past and future observations is that they both are sampled independently from the same distribution (i.i.d).

2. Formalization

Consider an input space $\mathcal{X}$ and output space $\mathcal{Y}$. As we only consider our specific case (binary classification), we choose $\mathcal{Y} = -1, 1$.

$(X, Y) \in \mathcal{X} \times \mathcal{Y}$ are random variables follow the an unknown distribution P. We observe a sequence of i.i.d pairs (X_i, Y_i) sampled according to P and the goal is to construct a function $g : \mathcal{X} \to \mathcal{Y}$ which can predict Y from X.

A criterion is need to choose the function g. It can be chosen as $P(g(X) \neq Y)$. We thus define the *risk* of g as

$$R(g) = P(g(X) \neq Y) = \mathbb{E}[\mathbb{I}_{g(X)\neq Y}] \tag{1}$$

Note that P can be decomposed as $P_X \times P(Y|X)$. Define the *regression function* $\eta(x) = \mathcal{E}[Y|X = x] = 2\mathcal{P}[Y = 1|X = x]-1$ and the *target function* (or Bayes classifier) $t(x) = sgn\eta(x)$. This function achieves the minimum risk over all possible measurable functions:

$$R(t) = \inf_g R(g) \tag{2}$$

Denote the value $R(t)$ as R^*, called the Bayes risk.

Define the *noise level* as $s(x) = \min(\mathbb{P}[Y = 1|X = x], 1 - \mathbb{P}[Y = 1|X = x]) = (1 - \eta(x))/2$ and this gives $R^* = \mathbb{E}s(X)$.

Our goal is thus to find this function t. but since P is unknown we cannot directly measure the risk. Let's define *empirical risk*:

$$R_n(g) = \frac{1}{b}\sum_{i=1}^{n} \mathbb{I}_{g(X_i)\neq Y_i} \tag{3}$$

It is common to use this quantity as a criterion to select an estimate of t.

2.1. **Algorithms. Overfitting**. When the input space is infinite, one can always construct a function g_n which perfectly predicts the labels of the training data but behaves on the other points as the opposite of the target function t. So one would have minimum empirical risk but maximum risk.

There are essentially two ways to prevent this overfitting situation:

- restrict the class of functions in which the minimization is performed
- modify the criterion to be minimized (e.g. adding a penalty for complicated function)

Empirical Risk Minimization. This algorithm is one of the most straightforward, yet it is usually efficient:

$$g_n = arg \min_{g \in \mathcal{G}} R_n(g) \tag{4}$$

where $\mathcal{G}$ is the function class we perform minimization.

Sometime (actually in most case) the *best* function in $\mathcal{G}$ is not necessary the overall best function. So one may want to enlarge the model as much as possible, while preventing overfitting.

Structural Risk Minimization. An infinite sequence $\{\mathbf{G_d}|d = 1, 2, \ldots\}$ of models of increasing size:

$$g_n = arg \min_{g \in \mathcal{G}_d, d \in \mathcal{N}} R_n(g) + \text{pen}(d, n) \tag{5}$$

The penalty $\text{pen}(d, n)$ gives preference to models where estimation error is small and measures the size of capacity of the model.

Regularization. Another, usually easier to implement approach consists in choosing a large model $\mathcal{G}$ and define on $\mathcal{G}$ a *regularizer*, typically a norm $\|g\|$:

$$g_n = arg \min_{g \in \mathcal{G}} R_n(g) + \lambda \|g\|^2 \tag{6}$$

Most existing (and successful) methods can be thought of as regularization methods.

Normalized Regularization.

2.2. **Bounds.** A learning algorithm takes as input the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and produce a function g_n which depends on this data. We want to estimate the risk of g_n. However, $R(g_n)$ is a random variable and it cannot be computed from the data (since it also depends on the unknow P). Estimates of $R(g_n)$ will take the form of probabilistic bounds.

Suppose the algorithm chooses its output from a model $\mathcal{G}$, let g^* be the best function in $\mathcal{G}$ with $R(g^*) = \inf_{g \in \mathcal{G} R(g)}$, we have

$$R(g_n) - R^* = [R(g^*) - R^*] + [R(g_n) - R(g^*)] \tag{7}$$

The first term on the right hand side is usually called the approximation error, and measures how well can functions in $\mathcal{G}$ could approach the target (if would be zero if $t \in \mathcal{G}$). The second terms, called estimation error is a random quantity (depends on the data) and measures how close is g_n to the best possible choice in $\mathcal{G}$.

The first term is usually difficult to decide, it requires us to make some assumption on the model and data. We will focus on the estimation error

Another possible decomposition of the risk is the following:

$$R(g_n) = R_n(g_n) + [R(g_n) - R_n(g_n)] \tag{8}$$

Summary: three type of results we may be interested in.

- $R(g_n) \leq R_n(g_n) + B(n, \mathcal{G})$
- $R(g_n) \leq R(g^*) + B(n, \mathcal{G})$
- $R(g_n) \leq R^* + B(n, \mathcal{G})$

3. Basic Bounds

In this section, simple error bounds (also called generalization bounds) will be covered.

3.1. **Relationship to Empirical Processes.** TOne way to make a statement about $R(g_n)$ is to say how it relates to an estimate such as the empirical risk $R_n(g_n)$. This relationship can take the form of upper and lower bounds for

$$\mathbb{P}[R(g_n) - R_n(g_n) > \epsilon] \tag{9}$$

Let's denote $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$. given $\mathcal{G}$ define the loss class

$$\mathcal{F} = \{f : (x, y) \mapsto \mathbb{I}_{g(x)\neq y} : g \in \mathcal{G}\} \tag{10}$$

Some shorthand notation: $Pf = \mathbb{E}[f(X, Y)]$ and $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i, Y_i)$. P_n is usually called the empirical measure associated to the training sample. Our interest if the difference between true and empirical risks:

$$Pf_n - P_n f_n \tag{11}$$

An empirical process is a collection of random variable indexed by a class of functions, and such that each random variable is distributed as a sum of i.i.d. random variables:

$$\{Pf - P_n f\}_{f\in\mathcal{F}} \tag{12}$$

One of the most studied quantity associated to empirical process is their supremum:

$$\sup_{f\in\mathcal{F}} Pf - P_n f \tag{13}$$

A upper bound on this quantity will also be a upper bound on (11). This shows that the theory of empirical process is great source of tools and techniques for Statistical Learning Theory.

3.2. **Hoeffding's Inequality.** Let's rewrite $R(g) - R_n(g)$ as the following:

$$R(g) - R_n(g) = \mathcal{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \tag{14}$$

By the law of large numbers, we have:

$$\mathbb{P}[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(Z_i) = \mathbb{E}[f(Z)]] = 1 \tag{15}$$

This indicates that with enough samples, the empirical risk of a function is a good approximation to its true risk.

Theorem 3.1. ***(Hoeffding)****. Let $Z_1, \ldots, Z_n$ be i.i.d random variables with $f(Z) \in [a, b]$. Then for all $\epsilon > 0$, we have*

$$\mathbb{P}[\big|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}[f(Z)]\big| > \epsilon] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \tag{16}$$

Denote the right hand side by δ, solve ϵ, we have

$$\mathbb{P}\big[\big|P_n f - Pf\big| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\big] \leq \delta \tag{17}$$

or with probability at least $1 - \epsilon$, we have

$$\big|P_n f - Pf\big| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}} \tag{18}$$

As $f(Z) \in [0,1]$ in our case, we can conclude that with probability at least $1-\epsilon$

$$R(g) \leq R_n(g) + \sqrt{\frac{\log \frac{2}{\delta}}{2n}} \tag{19}$$

Notice that the function g has to be fixed, if this function depends on the data then above inequality couldn't be applied directly. However, these sets S may be different for different functions. In other words, for the observed sample, only some of the functions in $\mathcal{F}$ will satisfy this inequality.

3.3. **Limitations.** What above result essentially says is that for each (fixed) function $f \in \mathcal{F}$, there is a set S of samples for which $Pf - P_n f \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}}$ (and the set of samples has measure $P[S] \geq 1-\epsilon$)

3.4. **Uniform Deviations.** Before seeing the data, we do not know which function the algorithm will choose. The idea is to consider *uniform* deviations

$$R(f_n) - R_n(f_n) \leq \sup_{f \in \mathcal{F}} (R(f) - R_n(f)) \tag{20}$$

In other words, if we can upper bound the supremum on the right, we are done. For this we need a bound which holds simultaneously for all functions in a class.

Consider two functions f_1, f_2 and define

$$C_i = \{(x_1, y_1), \ldots, (x_n, y_n) : Pf_i - P_n f_i > \epsilon\}$$

by Hoeffding's inequality, for each i

$$\mathbb{P}[C_i] \leq \delta$$

We want to measure how many samples are 'bad' for $i = 1, 2$.

$$\mathbb{P}[C_1 \cup C_2] \leq \mathbb{P}[C_1] + \mathbb{P}[C_2] \leq 2\delta$$

If $|\mathcal{F}| = N$, then

$$\mathbb{P}[C_1 \cup \ldots \cup C_N] \leq \sum_{i=1}^{N} \mathbb{P}[C_i]$$

As a result:

$$\mathbb{P}[\exists f \in \mathcal{F} : Pf - P_n f > \epsilon] \leq \sum_{i=1}^{N} \mathbb{P}[Pf_i - P_n f_i > \epsilon] \leq N \exp(-2n\epsilon^2)$$

Now turn to $\mathcal{G} = g_1, \ldots, g_N$, for all $\delta > 0$ with probability at least $1-\delta$,

$$\forall g \in \mathcal{G}, R(g) \leq R_n(g) + \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

REFERENCES

[1] Montenegro, Ravi, and Prasad Tetali. ***Mathematical aspects of mixing times in Markov chains.*** Now Pub, 2006.

[2] Stein, Elias M., and Rami Shakarchi. ***Real analysis: measure theory, integration, and Hilbert spaces***. Vol. 3. Princeton University Press, 2010.