Math Camp for Machine Learning

Synferlo

Mar. 22, 2021

1 Statistical Learning

1.1 Elements in ML

1. Instance/example:

x, $x \in X$

2. Instance space/domain:

X (where the instance comes from).

3. Label:

Each instance has a label, or class. Label can be $0/1$ or $+/$-.

4. Concept:

There's a function c, called concept, that tells the **true** relationship between instance and label.

$$\text{concept } c : X \rightarrow \{0, 1\}$$

Each instance x is labeled by $c(x)$. Our goal is to find this $c(\cdot)$.

5. Hypothesis:

Note, this is NOT the same one as we say in econometrics. Hypothesis, $h(\cdot)$, is a function that the machine use to do the prediction given an instance x.

$$h : X \rightarrow \{0, 1\}$$

6. Concept VS Hypothesis:

Concept is the <u>TRUE</u> relationship between x and label.

Hypothesis is the <u>GUESS</u> of our machine given the training data.

7. Concept class:

C is where concept c comes from, $c \in C$.

8. Distribution:

All instances are generated from a particular distribution D. We call it <u>target distribution</u> or distribution for short.
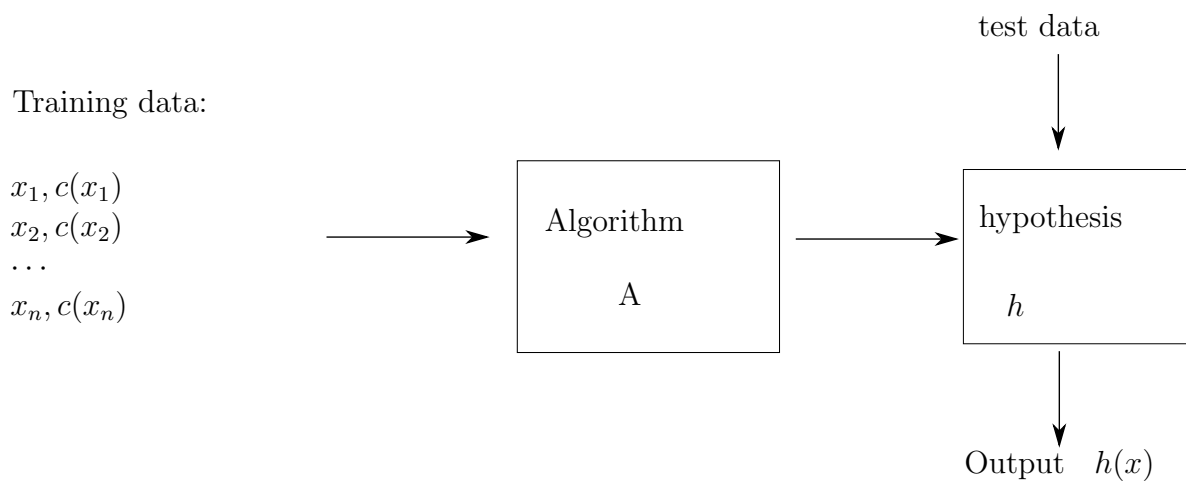
$$x_i \sim D, i.i.d.$$

Hypothesis class:

It tells where the hypothesis comes from. We allow h and c come from different classes.

$$h \in \mathcal{H}$$

1.2 ML Process

Training data:

$x_1, c(x_1)$
$x_2, c(x_2)$
$\ldots$
$x_n, c(x_n)$

$$x_i \in X, \quad x_i \sim D, \quad c \in C, \quad h \in \mathcal{H}$$

Figure 1: ML process

1.3 PAC Learning

We want to see $h(x) = c(x)$

We DO NOT want to see $h(x) \neq c(x)$

1.3.1 How we measure error:

$$err_D(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right]$$

We want this,

$$err_D(h) \leq \varepsilon,$$

where ε is a small positive number.

To guarantee the machine work well, we require the following condition,

$$Pr\left(err_D(h) \leq \varepsilon\right) \geq 1 - \delta$$

where δ is a small positive number.

Hence, $err_D \leq \varepsilon$ requires algorithm to be more accurate. $Pr(err \leq \varepsilon) \geq 1 - \delta$ requires the probability of this correction to be <u>high</u>.

This method is called <u>Probability approximately correct</u>, or PAC for short.

===

We say concept space C is PAC-learnable by $\mathcal{H}$,

if there exist an algorithm (alg.) A, $\forall c \in C$, $\forall$ distribution D, $\forall \varepsilon > 0, \delta > 0$,

A takes $m = poly(\frac{1}{\varepsilon}, \frac{1}{\delta}, \cdots)$ random examples $x_i \sim D$,

that it makes output hypothesis $h \in \mathcal{H}$ $s.t.$ $Pr(err_D(h) \leq \varepsilon) \geq 1 - \delta$.

NB: m is sample size. The more data we have, the higher accuracy that $h(\cdot)$ will be. Hence, m is <u>negative correlated</u> with ε and δ.

===

Here's an example: For $X \in \mathbb{R}$
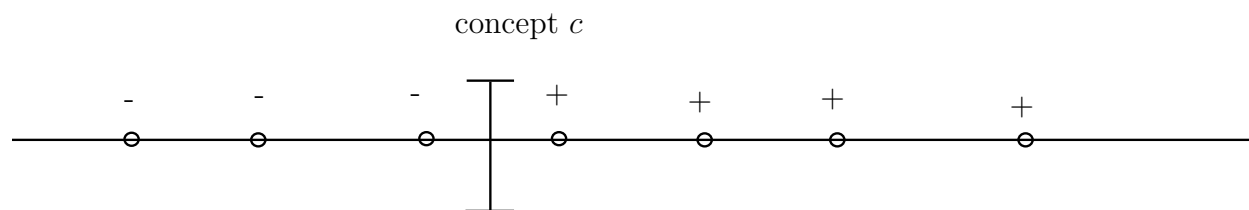
concept c

Figure 2: Threshold

There are many instances on the real line. Concept c is a threshold Function.

All instances on the right are labeled by + (True)

All instances on the left are labeled by - (False)

What we are doing is like this,

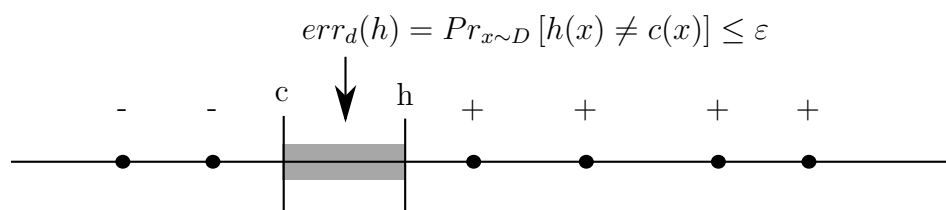$$err_d(h) = Pr_{x \sim D}\left[h(x) \neq c(x)\right] \leq \varepsilon$$

Figure 3: hyp. and concept

So, output $h(x)$ would be

$$h(x) = \begin{cases} + & \text{if } x \geq b \\ - & \text{O.W.} \end{cases}$$

In this case, we say $\mathcal{H} = C$.
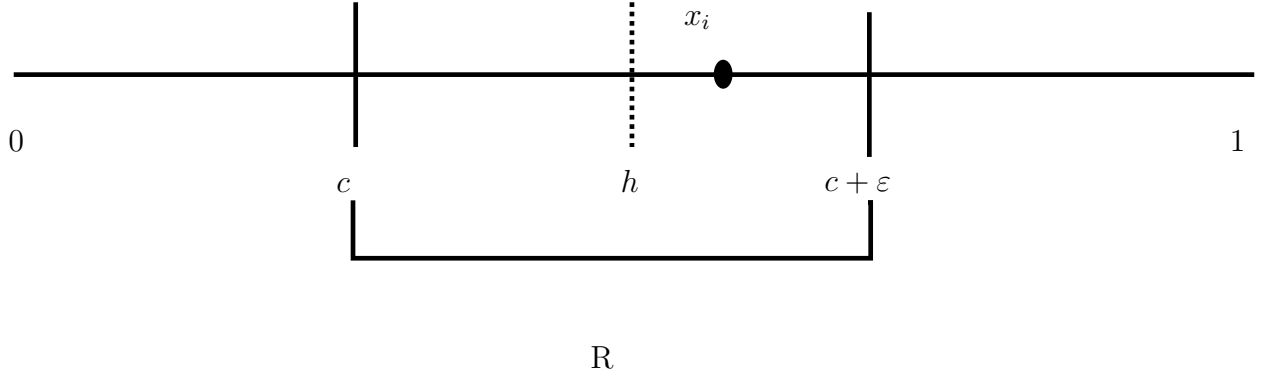
Now, let's consider a general case.

Figure 4: General case

If a training data example x_i falls in $(c, c + \varepsilon)$, region R, we have to shift $h(\cdot)$ to the left of x_i, so that

$$Pr[\quad err_D(h) > \varepsilon \quad] \leq Pr(\quad \text{no } x_i \text{ in } R \quad) = Pr(x_1 \notin R, x_2 \notin R, \cdots x_m, \notin R)$$

Since x_i is $i.i.d.$, we can write

$$Pr(x_1 \notin R, \cdots, x_m \notin R) = \prod_{i=1}^{m} Pr(x_i \notin R) = \prod_{i=1}^{m} (1 - \varepsilon) = (1 - \varepsilon)^m$$

Recall from basic calculus, $1 + x \leq e^x$, so we can rewrite

$$(1 - \varepsilon)^m \leq (e^{-\varepsilon})^m = e^{-\varepsilon m}$$

Also, since

$$Pr(err_D(h) \leq \varepsilon) \geq 1 - \delta$$

we have

$$Pr(err_D(h) > \varepsilon) \leq \delta$$

7

So, we end up with

$$Pr(err_D(h) > \varepsilon) \leq e^{-\varepsilon m}$$

$$\leq \delta$$

We need to solve for m, so that we can know the condition for sample size. Because we need to make sure this upper bound, $e^{-\varepsilon m}$, no greater than δ, we can write this,

$$e^{-\varepsilon m} \leq \delta$$

$$-\varepsilon m \leq \ln \delta$$

$$m \geq -\frac{\ln \delta}{\varepsilon}$$

$$m \geq \frac{\ln \frac{1}{\delta}}{\varepsilon}$$

It says at least you should have sample size greater than this lower bound to guarantee $Pr(err_D(h) \leq \varepsilon) \geq 1 - \delta$.

1.4 Finite Hypothesis Space

Let's start with finite hypothesis space case, $|\mathcal{H}| < \infty$.

==

Theorem 1

Suppose hypothesis space $\mathcal{H}$ is finite, Algorithm A finds hypothesis $h_A \in \mathcal{H}$ is consistent with m random training examples where

$$m \geq \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta}) \tag{1}$$

Then we can say

$$Pr\left[err_D(h_A) > \varepsilon\right] \leq \delta \tag{2}$$

or it is PAC learnable.

Equivalently, we can write,

with $prob. \geq 1 - \delta,$

$$err_D(h) \leq \frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m} = \varepsilon$$

where h is a random variable.

================================

Notice, <u>consistent</u> here is NOT what we mean in Econometrics!!

Here, consistent means <u>hyp.</u> makes NO mistake in training examples.

================================

In last section we have

$$m \geq \frac{\ln\frac{1}{\delta}}{\varepsilon}.$$

Now, we add $\ln|\mathcal{H}|$ to the RHS,

$$m \geq \frac{1}{\varepsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$$

Term $\ln|\mathcal{H}|$ measures the <u>complexity</u> of the hypothesis space.

================================

Complexity:

If we want to give each hyp. a name in $\mathcal{H}$, how many bits we need to do that?

The number of bits you need would be $log_2|\mathcal{H}|$, where $|\mathcal{H}|$ stands for the <u>number of hypothesis</u> in this class. So, here, we use $\ln|\mathcal{H}|$ roughly measures the complexity of $\mathcal{H}$.

================================

Recall three conditions we need to do machine learning:

1. enough data

2. fit the training set well

3. use simple classifier

Equation (2) gives us an accurate classifier. Equation (1) tells us we have enough data. Remember, Alg. A finds a h_A is consistent with training data. And $\ln|\mathcal{H}|$ bound the complexity. The more complex it is, the higher value would $\ln|\mathcal{H}|$ be. Then, m will also go up. It means with more complex hypothesis space, we need more data to get an accuracy prediction. So, <u>we need to limit the complexity</u>. Note, $|\mathcal{H}|$ is the cardinality, not absolute value.

Theorem 2

Assume given $m \geq \frac{1}{\varepsilon}(\ln |\mathcal{H}| + \ln \frac{1}{\delta})$ examples, with $prob. \geq 1 - \delta$,

$\forall h \in \mathcal{H}$:

if h is consistent, then $err_D(h) \leq \varepsilon$.

Note,

if $err_D(h) \leq \varepsilon$, we say h is ε-good

if $err_D(h) \leq \varepsilon$, we say h is ε-bad

Proof:

$Pr(\forall h \in \mathcal{H} : h \text{ consistent} \Rightarrow h \text{ is } \varepsilon\text{-good}) \geq 1 - \delta$

can be written as

$Pr(\exists h \in \mathcal{H} : h \text{ } cons. \text{ and } h \text{ is } \varepsilon\text{-bad}) \leq \delta$.

================================

Remember h is a random variable because it depends on training samples. But "e is ε-bad"

is not a RV.

================================

Let's go back to the proof.

Define

$$B = \{h \in \mathcal{H} : \text{ } h \text{ is } \varepsilon\text{-bad } \}$$

So we can write

$$Pr(\exists h \in \mathcal{H}: \ h \text{ cons.} \ \Rightarrow \ h \text{ is } \varepsilon\text{-bad }) \tag{3}$$

$$=Pr(\exists h \in B: h \text{ is cons. }) \tag{4}$$

$$=Pr\left(\bigcup_{h \in B} (h \text{ is cons. })\right) \tag{5}$$

$$=(h_1 \in B \text{ cons. }) \cup (h_2 \in B \text{ cons. }) \cdots (h_n \in B \text{ cons. }) \tag{6}$$

$$\leq \sum_{h \in B} Pr(h \text{ cons. }) \quad \text{recall, } Pr(a \cup b) \leq Pr(a) + Pr(b) \tag{7}$$

Then what is the $Pr(h \text{ cons. })$?

Because x_i are $i.i.d.$,

$$Pr(h \text{ cons. }) = Pr(h(x_1) = c(x_1) \cap \cdots \cap h(x_m) = c(x_m)) \tag{8}$$

$$= \prod_{i=1}^{m} Pr(h(x_i) = c(x_i)) \tag{9}$$

$$\leq (1 - \varepsilon)^m \tag{10}$$

Note, since $Pr(h(x_i) \neq c(x_i)) \leq \varepsilon$, so $Pr(h(x_i) = c(x_i)) \leq 1 - \varepsilon$.

Hence, from equation (7) and (10), we can write

$$Pr\left(\bigcup_{h \in B} (h \text{ cons. })\right)$$

$$\leq \sum_{h \in B} Pr(h \text{ cons. })$$

$$\leq |B| \, (1 - \varepsilon)^m \quad \text{Note, } |B| \leq |\mathcal{H}|$$

$$\leq |\mathcal{H}| \, e^{-\varepsilon m}$$

$$\leq \delta$$

How do get δ from $|\mathcal{H}|\, e^{-\varepsilon m}$?

Take log, we have $\ln |\mathcal{H}| - \varepsilon m$. Given $m \geq \frac{1}{\varepsilon}(\ln |\mathcal{H}|)$

$$em \geq \ln |\mathcal{H}| + \ln \frac{1}{\delta}$$

$$-\varepsilon m \leq -\ln |\mathcal{H}| - \ln \frac{1}{\delta} = -\ln |\mathcal{H}| + \ln \frac{1}{\delta}$$

$$\ln |\mathcal{H}| - \varepsilon m \leq \ln \delta$$

$$|\mathcal{H}|\, e^{-\varepsilon m} \leq \delta \quad \text{take exponential}$$

$$Q.E.D.$$

1.5 Infinite hypothesis space

Suppose we have four examples on the real line. h_1 and h_2 give us the same prediction.
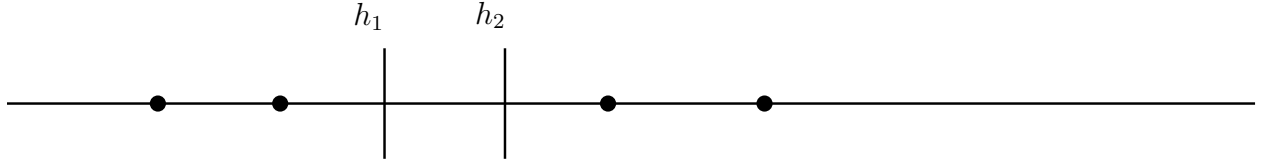

Figure 5: same prediction

The prediction would be different if we do this,

For each of the four instances, hypothesis's behavior can be either labeling instance as + or

-. So, for m instances, the max behavior is 2^m. Here we have $m + 1$, 5, hypothesis.

1.5.1 Growth Function

A set of examples $\mathcal{S} = \{x_1, \cdots, x_m\}$ contains m labeled instances. For particular hypothesis h,

$$\Pi_{\mathcal{H}}(\mathcal{S}) = \{\{h(x_1), \cdots, h(x_m)\} : h \in \mathcal{H}\} \tag{11}$$

where $\Pi_{\mathcal{H}}(\mathcal{S})$ is a set of hypothesis behaviors.

$$h_1 \quad\quad h_2 \quad\quad\quad h_3 \quad\quad\quad h_4 \quad\quad\quad h_5$$

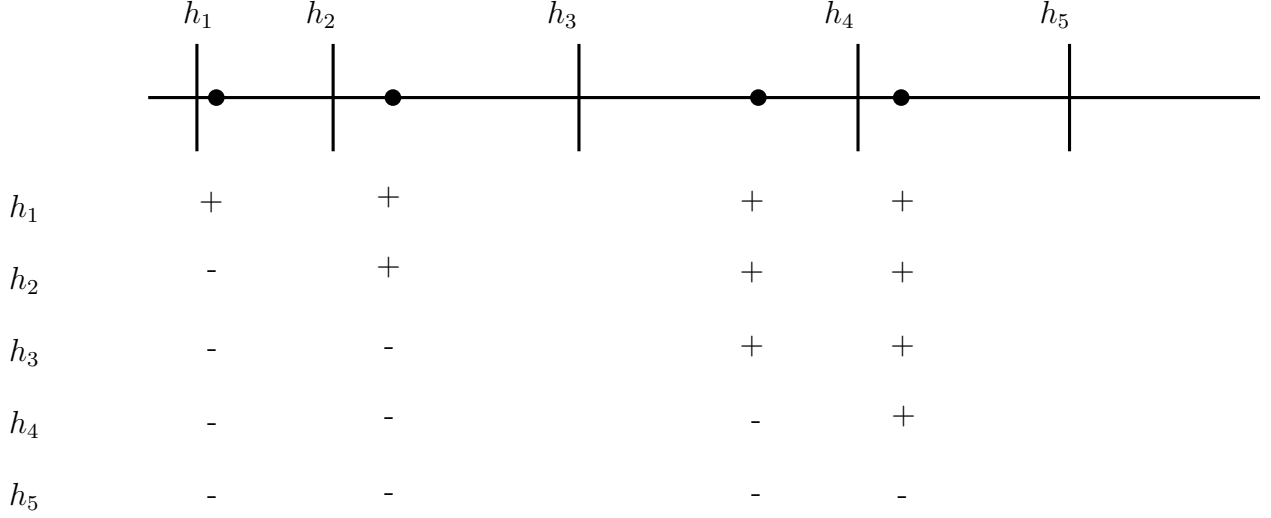|	h_1	h_2	h_3	h_4
h_1	+	+	+	+
h_2	-	+	+	+
h_3	-	-	+	+
h_4	-	-	-	+
h_5	-	-	-	-

Figure 6: different hyp.

We can look at how bad (maximum number) this set can be.

$$\Pi_H(m) = \max_{|\mathcal{S}|=m} |\Pi_{\mathcal{H}}(\mathcal{S})|$$

These are all cardinality, not absolute value. They measures the size.

The Growth function tells us the maximum number (the worst case) of labeling behavior a hypothesis space can make given a sample with size m.

In figure 6, we have $m+1$ hypothesis to make it an effective hypothesis space. And $\Pi_{\mathcal{H}}(m)$ is called the growth function.

Recall, when $|\mathcal{H}| < \infty$, we have

$$err_D(h) \leq \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{m}$$

Now, we relax this finite assumption, and we can use the growth function to measure the complexity of the hypothesis space.

Theorem 3

Given m training examples, with $prob. \geq 1 - \delta$, $\forall h \in \mathcal{H}$,

if h is consistent, then

$$err_D(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right) \tag{12}$$

If the growth function is polynomial (this is a nice case),

$$\Pi_{\mathcal{H}}(m) = O(m^d)$$

where d is a constant. We use big O here to hide constant terms. Hence the error becomes,

$$err_D(h) \leq \frac{d \ln m + \ln \frac{1}{\delta}}{m} \tag{13}$$

In the nice case,
$$\Pi_{\mathcal{H}}(m) = O(m^d),$$

it is the finite case.

In the worst case, the growth function realizes all possible behaviors,

$$\Pi_H(m) = 2^m,$$

it is an infinite case.

For any hypothesis space $\mathcal{H}$, it will be either the nice case, or the worst case. There's NO other cases!

In the nice case, learning is possible because we can solve the $err_D(h)$ according to the theorem.

In the worst case, learning is not possible.

NB:

In $\Pi_H(m) = O(m^d)$, d is called the VC-dimension, where VC is the short form of Vapnik-Chervnenkis.

Before we introduce VC-dimension, we need to know <u>shattering</u> first.

Shattering:

Sample $\mathcal{S}$ with size m is shattered by $\mathcal{H}$, if all behaviors are possible, $|\Pi_H(\mathcal{S})| = 2^m$.

Equivalently, we say sample $\mathcal{S}$ is shattered by $\mathcal{H}$, if $\mathcal{H}$ can dichotomize all elements in sample $\mathcal{S}$, i.e., $\Pi_H(\mathcal{S}) = 2^m$.
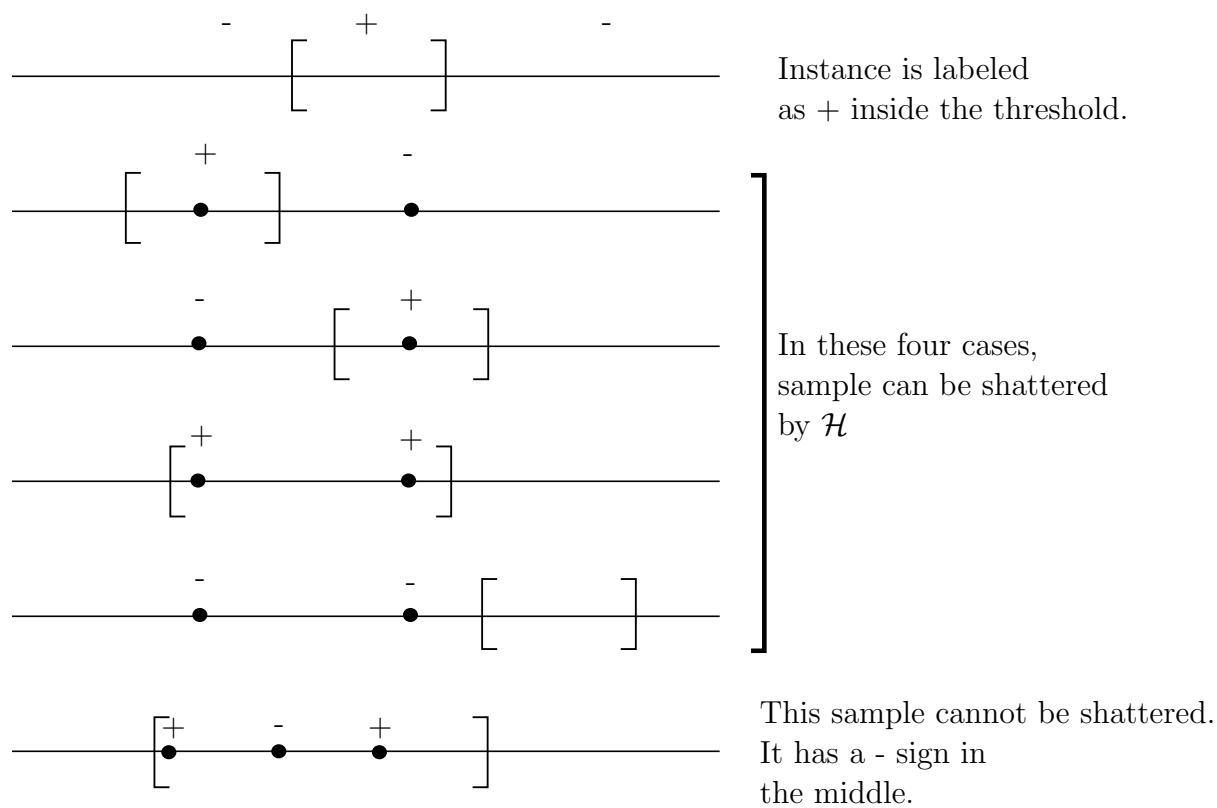

Here's an example

Figure 7: shattering example

In a three-point example, sample is not shattered because we <u>cannot</u> label them using one threshold. Hence, we can shatter two points, but we can never shatter three points.

Definition:

The VC-dimension for $\mathcal{H}$ is the maximum size of a sample, which can be shattered by $\mathcal{H}$.

$$\text{VC-dim } (\mathcal{H}) = \max\left\{m : \Pi_{\mathcal{H}}(m) = 2^m\right\}$$

In this case, VC-dim(intervals) = 2 because we can only shatter a two-point sample.

Extension:

Linear threshold function (LTF) in $\mathbb{R}^n$ looks like this. The VC-dim(LTF in $\mathbb{R}^n$) $= n+1$

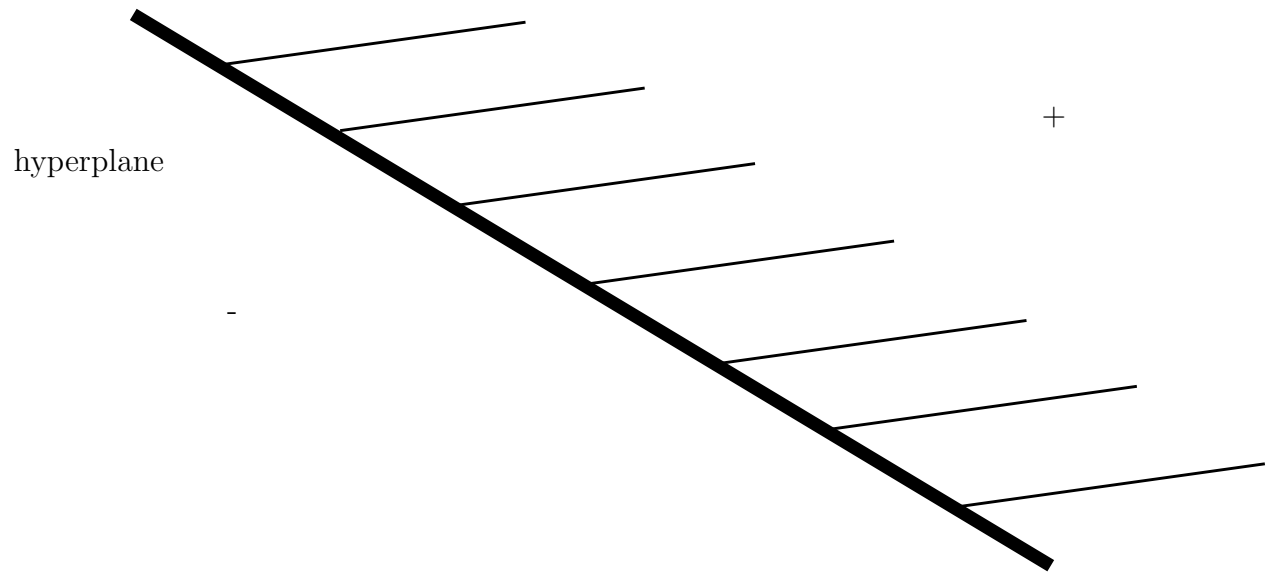

Figure 8: Linear threshold function

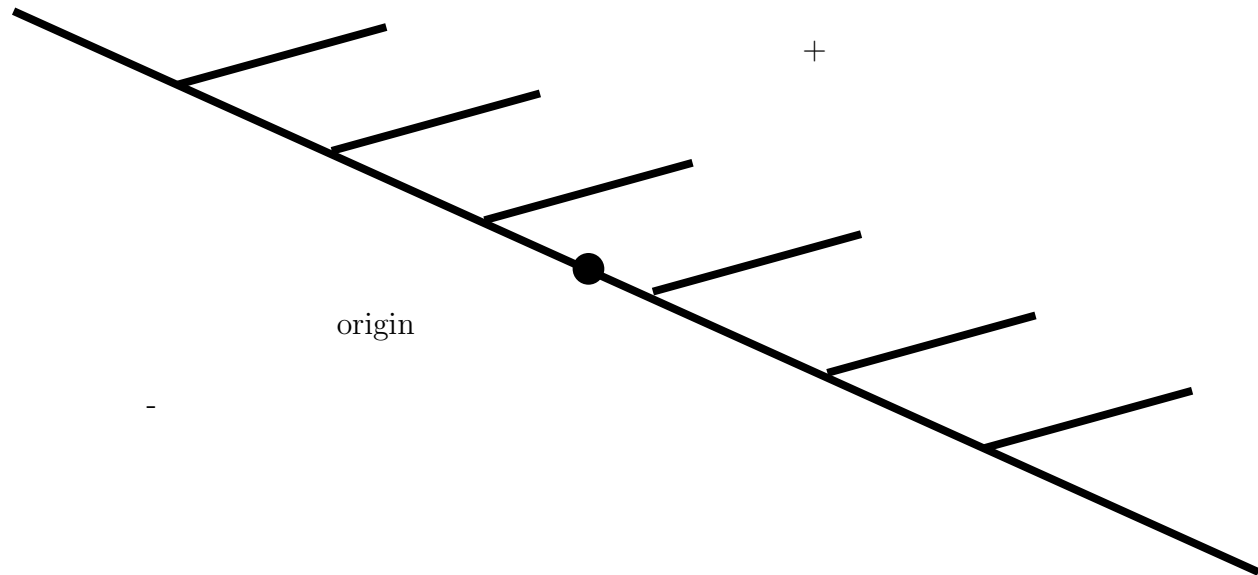VC-dim(LTF that go through the origin) $= n$

Figure 9: LTF go through the origin

For finite hypothesis space $\mathcal{H}$,

$$\text{VC-dim } (\mathcal{H}) \leq \ln |\mathcal{H}|$$

where $\ln |\mathcal{H}|$ measures the complexity of the hypothesis space.

Summary:

Given m training examples, with $prob. \geq 1 - \delta$, $\forall h \in \mathcal{H}$:

If h is consistent, then

$$err_D(h) \leq \frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{m}, \quad \text{if } |\mathcal{H}| < \infty \tag{14}$$

$$err_D(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}}(2m) + \ln \frac{1}{\delta}}{m}\right), \quad \text{for all } \mathcal{H} \text{ not only finite ones,} \tag{15}$$

1.5.2   Sauer's Lemma

Review on textbook page 277

Sauer's Lemma:

Given d = VC-dim $(\mathcal{H})$,

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d, \quad \text{if } m \geq d \geq 1 \tag{16}$$

where we can solve binomial term using

$$\binom{m}{i} = \frac{m!}{i!(m-i)!}$$

Recall, for any $\mathcal{H}$, either

$\Pi_H(m) = 2^m \quad \forall m$, this means $d = $ VC-dim is infinite. The worst case

or

$\Pi_{\mathcal{H}}(m) = O(m^d)$ for some constant d. This is the nice case. It says $d = $VC-dim is finite.

If we plug equation (16) into (15),

$$err_D(h) \leq O\left(\frac{\ln \Pi_{\mathcal{H}} + \ln \frac{1}{\delta}}{m}\right)$$

$$\leq O\left(\frac{\ln \left(\frac{em}{d}\right)^d + \ln \frac{1}{\delta}}{m}\right)$$

$$\leq O\left(\frac{d \ln \frac{m}{d} + d + \ln \frac{1}{\delta}}{m}\right)$$

So, now, the VC-dim $= d$ acts like a measure of complexity.

NB: the VC-dim also gives us a lower bound of how many examples we need for training.

================================

Note, VC-dim$(\mathcal{H}) = d$ means that there exists ($\exists$) a sample with size d which can be shattered by hypothesis space $\mathcal{H}$. Here, we say there exists ($\exists$). Hence, it does NOT mean that for all sample with size d can be shattered. One is enough.

================================

1.5.3 Empirical Risk Minimization Model (ERM)

There might be some randomness between today's weather and tomorrow's weather. So, instead of write y as a direct function of x, $y = f(x)$, we say instance x and label y are in a pair (x, y). And this pair is followed a distribution D, $(x, y) \sim D$.

Hence, the measure of hypothesis $h(\cdot)$ becomes

$$err_D(h) = Pr_{(x,y)\sim D}[h(x) \neq y]$$

Now the problem becomes this:

Given sample $(x_1, y_1), \cdots, (x_m, y_m)$ where $(x_i, y_i) \sim D$.

We want: $\min_{h \in \mathcal{H}} err_D(h)$.

This says we want to minimize the error, $err_D(h)$ over all hypothesis in $\mathcal{H}$.

21

Notice, $err_D(h)$ is the generalized error. Our **training error** from the machine would be

$$\widehat{err_D(h)} = \frac{1}{m}\sum_{i=1}^{m} 1\left\{h(x_i \neq y_i)\right\}$$

where $1\left\{h(x_i \neq y_i)\right\}$ is an indicator function. We can write it in ECON way:

$$\widehat{err_D(h)} = \frac{1}{m}\sum_{i=1}^{m} I, \quad I = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & o.w. \end{cases} \tag{17}$$

Clearly, $\frac{1}{m}\sum_i I$ is the sample mean of the indicator function. It says in what percent the outcome hypothesis is wrong. Remember $I = 1$ when the outcome $h(x_i)$ is NOT the same as true label y_i.

Hence, $\widehat{err_D(h)} < 1$.

And the hypothesis we get from the training machine would be

$$\widehat{h} = \arg\min_{h \in \mathcal{H}} err_D(h) \tag{18}$$

This method is called empirical risk minimization, or ERM for short. The error term $\widehat{err_D(h)}$ is called empirical risk.

So what we what to prove is the following theorem.

Theorem 4

For sufficiently large m, with $prob. \geq 1 - \delta$, $\forall h \in \mathcal{H}$:

$$\left|err_D(h) - \widehat{err_D(h)}\right| \leq \varepsilon, \quad \text{here is abs rather cardinality.}$$

It says that the training error is always close to the generalization error for all hypothesis in $\mathcal{H}$.

This is called <u>the uniform convergence result</u>. If the $h(\cdot)$ is consistent, then $\widehat{err_D(h)} = 0$, and $err_D(h)$ is a small number.

Recall,

$$err_D(h) = Pr(h(x) \neq y)$$

The indicator function appears in equation (17) can be written in this way

$$I = \begin{cases} 1 & \text{with } prob. = err_D(h) \\ 0 & o.w. \end{cases}$$

And the training error is the sample mean,

$$\widehat{err_D(h)} = \frac{1}{m} \sum_{i=1}^{m} I$$

So, we can prove how fast $\widehat{err_D(h)}$ converge to $err_D(h)$.

Example:

Given RVs, $Z_1, \cdots, Z_m$, $i.i.d.$, $Z_i \in \{0, 1\}$. Let $p = E(Z_i)$, expectation. We want to estimate p by $\hat{p} = \frac{1}{m} \sum_i Z_i$, which is the sample mean. And the <u>Hoeffing inequality</u> tells us (Hoeffding, 1963)

$$Pr\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} E(x_i) \geq \varepsilon \right) \leq exp(-2m\varepsilon^2) \tag{19}$$

$$Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} E(x_i)\right| \geq \varepsilon \right) \leq 2exp(-2m\varepsilon^2) \tag{20}$$

Hence, in our example, we have

$$Pr(\hat{p} - p \geq \varepsilon) \leq exp(-2m\varepsilon^2),$$

where m stands for sample size, $\varepsilon \in (0, 1)$.

So, we have the sample mean converge to the expectation,

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \to E(Z_i)$$

Or we can write

$$\frac{1}{m}\sum_i Z_i \to E(\frac{1}{m}\sum_i Z_i),$$

since $E(\frac{1}{m}\sum_i Z_i) = \frac{1}{m}\sum_i E(Z_i) = E(Z_i)$

Note,

$\frac{1}{m}\sum_i Z_i$ can be written as $f(Z_1, \cdots, Z_m)$,

$E(\frac{1}{m}\sum_i Z_i)$ can be written as $E(f(Z_1, \cdots, Z_m))$. Hence, we have

$$f(Z_1, \cdots, Z_m) \to E(f(Z_1, \cdots, Z_m))$$

1.5.4 McDiarmid's Inequality (McDiarmid, 1989)

Suppose:

1. $f(Z_1, \cdots, Z_m)$ is real-valued.

2. Changing in Z_i will change $f(\cdot)$ by <u>at most</u> c_i, i.e., $\forall Z_1, \cdots, Z_m, Z_i'$

$$|f(Z_1, \cdots, Z_i, \cdots, Z_m) - f(Z_1, \cdots, Z_i', \cdots, Z_m)| \leq c_i \tag{21}$$

Note, for each Z_i, we have a equation (21) and c_i. If all Z_i are changed, then we will have m number of c_i.

3. Instances $Z_1, \cdots, Z_m$ are independent, but NOT necessarily identical.

If ALL of the above conditions are hold, then we can prove this convergence result,

$$Pr\left(f(Z_1, \cdots, Z_m) - E(f(Z_1, \cdots, Z_m)) \geq \varepsilon\right) \leq exp\left(\frac{-2\varepsilon^2}{\sum\limits_{i=1}^{m} c_i^2}\right) \tag{22}$$

In our example,

$$f(Z_1, \cdots, Z_m) = \frac{1}{m} \sum_{i=1}^{m} Z_i, \quad Z_i \in \{0, 1\}$$

So, $c_i = \frac{1}{m}$ because Z_i is an indicator function. It can be either 0 or 1. If one of Z_i change from 0 to 1, $f(\cdot)$ increases $\frac{1}{m}$, vice versa.

1.5.5 Proof for Theorem 4

Recall, theorem 4 says that

For sufficiently large m, with $prob. \geq 1 - \delta$, $\forall h \in \mathcal{H}$:

$$\left| err_D(h) - \widehat{err_D(h)} \right| \leq \varepsilon$$

Let's prove an easier one first, saying $\exists h \in \mathcal{H}$, the upper inequality holds.

Here,

$$Z_i = I, \quad \text{where } I = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\ 0 & o.w. \end{cases}$$

$$p = E(Z_i) = err_D(h)$$

$$\widehat{p} = \widehat{err_D(h)} = \frac{1}{m} \sum_{i=1}^{m} I$$

From Hoeffding's inequality, for $|\mathcal{H}| < \infty$,

$$Pr\left(\exists h \in \mathcal{H} : \left|err_D(h) - \widehat{err_D(h)} \geq \varepsilon\right|\right) \leq 2\,|\mathcal{H}|\,exp\left(-2m\varepsilon^2\right) \tag{23}$$

where $|\mathcal{H}|$ is the cardinality of $\mathcal{H}$.

How do we derive equation (23):

$$LHS = Pr\left(\left(\left|err_D(h_1) - \widehat{err_D(h_1)}\right| > \varepsilon\right) \cup \cdots \cup \left(\left|err_D(h_{|\mathcal{H}|}) - \widehat{err_D(h_{|\mathcal{H}|})}\right| > \varepsilon\right)\right)$$

$$\leq \sum_{h \in \mathcal{H}} Pr\left(\left|err_D(h) - \widehat{err_D(h)}\right| > \varepsilon\right)$$

From Hoeffding's inequality,

$$Pr\left(\left|err_D(h) - \widehat{err_D(h)}\right| \geq \varepsilon\right) \leq 2exp(-2m\varepsilon^2)$$

Hence,

$$\sum_{h \in \mathcal{H}} Pr\left(\left|err_D(h) - \widehat{err_D(h)}\right| > \varepsilon\right) \leq 2\,|\mathcal{H}|\,exp(-2m\varepsilon^2)$$

Now, let δ equals to the RHS of equation (23), then we can solve for ε.

$$2\,|\mathcal{H}|\,exp(-2m\varepsilon^2) = \delta$$

$$exp(-2m\varepsilon^2) = \frac{\delta}{2\,|\mathcal{H}|}$$

$$-2m\varepsilon^2 = \ln\frac{\delta}{2} - \ln|\mathcal{H}|$$

$$\varepsilon^2 = \frac{\ln|\mathcal{H}| - \ln\frac{\delta}{2}}{2m}$$

$$\varepsilon = \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}$$

$$\varepsilon = O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m}}\right)$$

Recall, theorem 4, plug in the value of ε in to $\left|err_D(h) - \widehat{err_D(h)}\right| \le \varepsilon.$

$$\left|err_D(h) - \widehat{err_D(h)}\right| \le O\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m}}$$

Hence, we can write,

$$\widehat{err_D(h)} - O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{m}}\right) \le err_D(h) \le \widehat{err_D(h)} + O\left(\sqrt{\frac{|\ln\mathcal{H}| + \ln\frac{1}{\delta}}{m}}\right) \qquad (24)$$

Clearly, the generalization error, $err_D(h)$, is related to $\widehat{err_D(h)}, \ln|\mathcal{H}|, m$, and δ.

The thing is that we expect the $\widehat{err_D(h)}$ goes down as we increase the complexity of $|\mathcal{H}|$.

Higher complexity of $\mathcal{H}$ means that we add more restrictions to the model. And our model

would fit the training data better. Note, if we complexity if too high even though the machine

fit the training data better, we would have over fitting problem.

For $err_D(h)$, as $|\mathcal{H}|$ goes up, it decreases initially, because it is affected by the training error.

However, as $\mathcal{H}$ getting much more complex, training error becomes very small, and $O(\cdot)$

dominates the trend. Hence, $err_D(h)$ goes up. The graph is shown below.
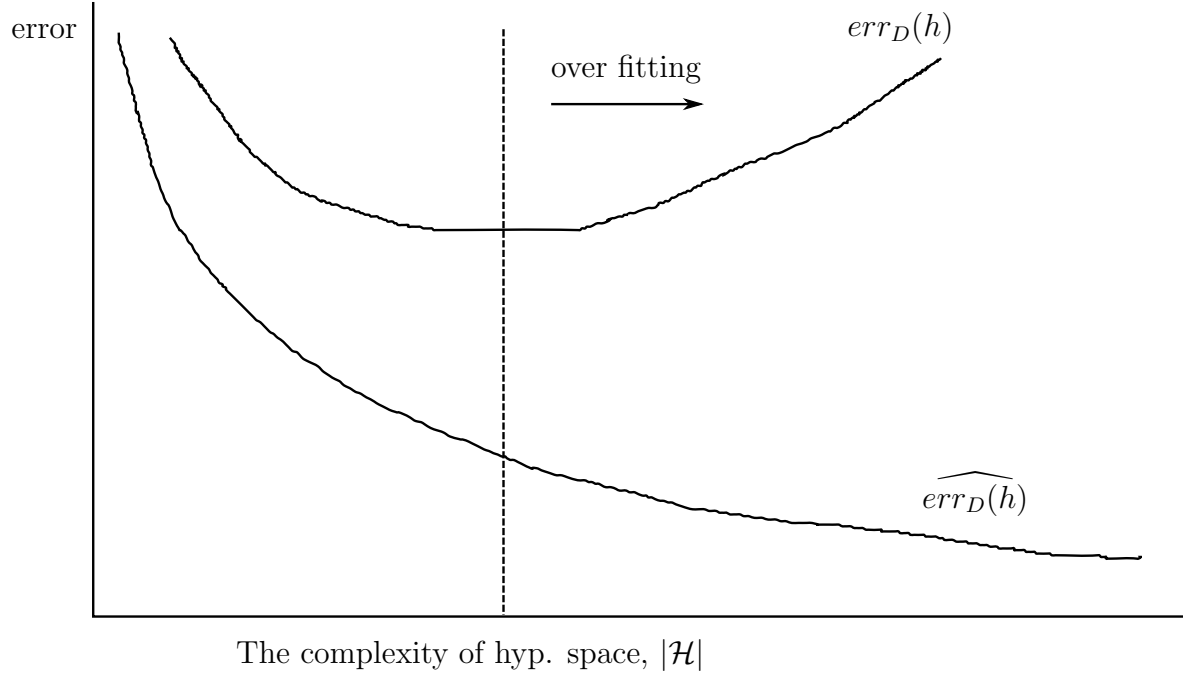

The complexity of hyp. space, $|\mathcal{H}|$

Figure 10: Complexity of hyp. space and error trend

1.5.6 Rademacher Complexity

Another to measure the complexity is using Rademacher complexity. Radermacher is a Germany mathematician.

Given sample space $\mathcal{S} = (x_1, y_1), \cdots, (x_m, y_m)$,

$y_i \in \{-1, +1\}$, y is label.

$h : X \longmapsto \{-1, +1\}$

$\widehat{err_D(h)} = \frac{1}{m} \sum\limits_{i=1}^{m} I$

Since indicator I can be either 0 or 1, we can rewrite it as

$$I = \frac{1 - y_i h(x_i)}{2}$$

Let me show you why it make sense. Remember, $h(x_i)$ is the prediction, $h_{x_i} \in \{-1, +1\}$.

If true label $y_i = 1$:

$$I = \frac{1 - h(x_i)}{2} \quad \begin{cases} \text{if } h(x_i) = 1, h(x_i) = y_i, I = \frac{1-1}{2} = 0 \\[2mm] \text{if } h(x_i) = -1, h(x_i) \neq y_i, I = \frac{1-(-1)}{2} = 1 \end{cases}$$

If true label $y_i = -1$:

$$I = \frac{1 + h(x_i)}{2} \quad \begin{cases} \text{if } h(x_i) = 1, h(x_i) \neq y_i, I = \frac{1+1}{2} = 1 \\[2mm] \text{if } h(x_i) = -1, h(x_i) = y_i, I = \frac{1-1}{2} = 0 \end{cases}$$

It says, $I = \begin{cases} 1 & \text{if } h(x_i) \neq y_i \\[2mm] 0 & o.w. \end{cases}$

It is exactly the same thing! Genius!

So we can rewrite the training error as the following,

$$\begin{aligned}
\widehat{err_D(h)} &= \frac{1}{m} \sum_{i=1}^{m} I \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1 - y_i h(x_i)}{2} \\
&= \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(1 - y_i h(x_i)) \\
&= \frac{1}{m} \sum_{i} \frac{1}{2} - \frac{1}{2m} \sum_{i} y_i h(x_i) \\
&= \frac{1}{2} - \frac{1}{2m} \sum_{i=1}^{m} y_i h(x_i)
\end{aligned}$$

Now, we are using the <u>average of</u> $y_i h(x_i)$ to measure how good a hypothesis h is. Since

$$\widehat{err_D(h)} = \frac{1}{2} - \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$$

Our optimization problem becomes this,

$$\arg\min \widehat{err_D(h)} \implies \arg\max_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} y_i h(x_i)$$

================================

Since in real world data the label y_i in (x_i, y_i) can be affected by random noise, y_i in (x_i, y_i) no longer be the true label. How do we deal with this problem?

Let σ_i be an assumed label (a RV),

$$\sigma_i = \begin{cases} 1 & \text{with } prob. = \frac{1}{2} \\ -1 & \text{with } prob. = \frac{1}{2} \end{cases}$$

Then we can rewrite the problem as,

$$\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i) \tag{25}$$

Here we use SUP rather MAX because $\mathcal{H}$ is <u>infinite</u>, it is possible that we cannot find the maximum.

Now we take expectation for equation (25) w.r.t. σ

$$R = E_\sigma\left(\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right) \tag{26}$$

where $R \in [0, 1]$. Here, we use R to measure the complexity of $\mathcal{H}$.

If $|\mathcal{H}| = 1$ then $R = 0$.

If $|\mathcal{H}| = 2^m$ and $\mathcal{S}$ can be shattered by $\mathcal{H}$ then $R = 1$.

So, R would be closed to 1 if $\mathcal{H}$ is very complicated.

===============================

To see why R = 0 if $|\mathcal{H}| = 1$,

If $|\mathcal{H}| = 1$ we no longer have SUP

$$R = E_\sigma \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right)$$

$$= \frac{1}{m} \sum_i E_\sigma \sigma_i h(x_i), \quad \sigma_i = \begin{cases} 1 \text{ with } prob. = \frac{1}{2} \\ -1 \text{ with } prob. = \frac{1}{2} \end{cases}$$

Hence,

$$E_\sigma \sigma_i h(x_i) = \frac{1}{2} \cdot (1) \cdot h(x_i) + \frac{1}{2} \cdot (-1) \cdot h(x_i)$$
$$= \frac{1}{2} h(x_i) - \frac{1}{2} h(x_i)$$
$$= 0$$

So, $R = 0$ ===============================

If $|\mathcal{H}| = 2^m$, and $\mathcal{S}$ can be shattered by $\mathcal{H}$, then there would be one hypothesis that make $h(x_i) = \sigma_i$. This says that one hypothesis makes $R = 1$ which is the maximum because all predictions are correct.

It makes sense because when we have <u>infinite</u> number of hypothesis, there would be <u>one</u> who can fit all values.

1.6　Empirical and expected RC

Note, from here, I use $\mathbb{E}$ for expectation, and E for the error of prediction.

Recall,

$$h : X \to \{-1, 1\}, h \in \mathcal{H}$$

$$\mathcal{S} = \{x_1, x_2, \cdots, x_m\}$$

$$\mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$$

We want to show:

$$\forall h \in \mathcal{H}, \quad \widehat{err}(h) \to err(h)$$

Now, let's have a new setting.

$$f : Z \to \mathbb{R}, \quad f \in F$$

$$Z = \{z_1, z_2, \cdots, z_m\}$$

$$\widehat{R}_Z(F) = \mathbb{E}_\sigma \left[\sup_{f \in F} \frac{1}{m} \sum_{i=i}^{m} \sigma_i f(z_i) \right]$$

where $\widehat{R}_Z(F)$ is called the empirical Rademacher complexity, F stands for the hypothesis space, Z stands for a sample set, and f stands for a hypothesis.

Now we want:

$\forall f \in F$:

$$\widehat{E}_Z(f) \to E(f)$$

where $\widehat{E}_Z(f) = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$, and $E(f) = E_{Z \sim D}(f(z))$.

Since $\widehat{R}_Z(F)$ is defined on a particular sample, but, normally, samples are chosen at random

from a distribution. So the expected value with respect to the random sample Z is called expected Rademacher Complexity (RC).

$$R_m(F) = \mathbb{E}_{Z \subseteq \mathbb{Z}:|Z|=m}\left[\widehat{R}_Z(F)\right]$$

Theorem 5

For function space maps sample to the real number between 0 and 1, $F : Z \to [0,1]$

$Z = \{z_1, \cdots, z_m\}$ $z_i \sim D, iid, \forall f \in F$ and with $prob. \geq 1 - \delta$ we have the following results:

$$E(f(z)) \leq \frac{1}{m}\sum_{i=1}^{m} f(z_i) + 2R_m(F) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right), \quad O(\cdot) = \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \qquad (27)$$

$$E(f(z)) \leq \frac{1}{m}\sum_{i=1}^{m} f(z_i) + 2\widehat{R}_Z(f) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right), \quad O(\cdot) = 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}} \qquad (28)$$

---

where $\widehat{E}_Z(f) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$, stands for empirical error and $E[f(z)]$ stands for the true error.

How to prove equation (27)?

The intuition is that we want to find the bound (the max difference) between the true error and the empirical error. Hence, let's define

$$\Phi(Z) = \sup_{f \in F}\left(E(f) - \widehat{E}_Z(f)\right)$$

Step 1

Now we claim that the relationship between the true supreme difference and the empirical supreme difference is the following,

with $prob. \geq 1 - \delta$

$$\Phi(Z) \leq \mathbb{E}_Z(\Phi(Z)) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$$

$\Phi(Z) :$ true supreme difference

$\mathbb{E}_Z(\Phi(Z)) :$ empirical supreme difference

Z is a RV, since the samples are drawn from distribution D

How do we know this?

Since $f(\cdot)$ will change at most $\frac{1}{m}$, and $E(f)$ is constant, so, the difference $E(f) - \widehat{E}_Z(f)$ will change at most $\frac{1}{m}$. This means that $\sup\left(E(f) - \widehat{E}_Z(f)\right)$ will change at most $\frac{1}{m}$. It says that

$$\Phi(Z) - \Phi(Z') \leq \frac{1}{m}$$
$$|\Phi(Z) - \Phi(Z')| \leq \frac{1}{m}$$

So from McDiarmid inequality, we can write

$$\Phi(Z) \leq \mathbb{E}_Z(\Phi(Z)) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right)$$

Step 2

Since the empirical error, $\widehat{E}_Z(f)$ is based on a particular sample Z, it is finite. But the true error, $E(f)$ is based on all possible samples from the distribution D. Hence, $E(f)$ if infinite. We need to work on this.

The idea is that we <u>replace</u> $E(f)$ by another empirical average from a second (imaginary) sample.

34

sample we have: $Z = \{z_1, \cdots, z_m\}$, called real sample

imaginary sample: $Z' = \{z'_1, \cdots, z'_m\}$, called ghost sample

where, $Z, Z' \sim D, iid.$

So the idea is that we replace $E(f)$ in $\Phi(Z) = \sup_{f \in F} = \left(E(f) - \widehat{E}_Z(f) \right)$ by the empirical average from Z'. Hence, we can write

$$\mathbb{E}_Z(\Phi(Z)) = \mathbb{E}_Z \left[\sup_{f \in F} \left(E(f) - \widehat{E}_Z(f) \right) \right] \tag{29}$$

$$\leq \mathbb{E}_{Z,Z'} \left[\sup_f \left(\widehat{E}_{Z'}(f) - \widehat{E}_Z(f) \right) \right] \tag{30}$$

We are going to prove this relationship!

Since z_i and z'_i are all drawn from D, iid. It will not change the result if we swap them between two samples. Hence we do a "flip-coin" here. If Head, we put z_i to Z' and z'_i to Z; If Tail, we do nothing. After this swap process we will receive two new samples T and T'

Before swap:

$$Z = \{z_1, z_2, \cdots, z_m\}$$
$$Z' = \{z'_1, z'_2, \cdots, z'_m\}$$

After swap:

$$T = \{z'_1, z_2, \cdots, z_m\}$$
$$T' = \{z_1, z'_2, \cdots, z'_m\}$$

Now we go back to our equation

$$\widehat{E}_{Z'}(f) - \widehat{E}_Z(f) = \frac{1}{m} \sum_{i=1}^{m} (f(z_i') - f(z_i))$$

Since we swap some of the examples(instances), the upper equation becomes this

$$\widehat{E}_{T'}(f) - \widehat{E}_T(f) = \frac{1}{m} \sum_{i=1}^{m} \begin{cases} f(z_i') - f(z_i) \text{ if no swap} \\ f(z_i) - f(z_i') \text{ if swap} \end{cases} \tag{31}$$

To simplify this equation, we introduce and indicator function σ_i where

$$\sigma_i = \begin{cases} 1 & \text{if no swap} \\ -1 & \text{if swap} \end{cases}$$

Note, σ_i here is not the same as the one we defined previously.

Hence we can rewrite equation(31),

$$\widehat{E}_{T'}(f) - \widehat{E}_T(f) = \frac{1}{m} \sum_{i=1}^{m} \sigma_i(f(z_i') - f(z_i)), \quad \sigma_i = \begin{cases} 1 & \text{if no swap} \\ -1 & \text{if swap} \end{cases}$$

Hence, we can now rewrite equation(29),

$$\mathbb{E}_Z(\Phi(Z)) = \mathbb{E}_Z \left[\sup_{f \in F} \left(E(f) - \widehat{E}_Z(f) \right) \right] \tag{32}$$

$$\leq \mathbb{E}_{Z,Z'} \left[\sup_f \left(\widehat{E}_{Z'}(f) - \widehat{E}_Z(f) \right) \right] \tag{33}$$

$$= \mathbb{E}_{Z,Z',\boldsymbol{\sigma}} \left[\sup_f \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i') - f(z_i) \right) \right] \tag{34}$$

$$= \mathbb{E}_{Z,Z',\boldsymbol{\sigma}} \left[\sup_f \left(\frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i') + \frac{1}{m} \sum_{i=1}^m (-\sigma_i) f(z_i) \right) \right] \tag{35}$$

$$\leq \mathbb{E}_{Z,Z',\boldsymbol{\sigma}} \left[\sup_f \frac{1}{m} \sum_{i=1}^m \sigma_i f(z_i') + \sup_f \frac{1}{m} \sum_{i=1}^m (-\sigma_i) f(z_i) \right] \tag{36}$$

$$= \mathbb{E}_{Z',\boldsymbol{\sigma}} \left(\sup_f \frac{1}{m} \sum_i \sigma_i f(z_i') \right) + \mathbb{E}_{Z,\boldsymbol{\sigma}} \left(\sup_f \frac{1}{m} \sum_i (-\sigma_i) f(z_i) \right) \tag{37}$$

$$= 2 R_m(F) \tag{38}$$

Note, empirical Rademacher Complexity is

$$\widehat{R}_Z(F) = \mathbb{E}_{\boldsymbol{\sigma}} \left(\sup_f \frac{1}{m} \sum_i \sigma_i f(z_i) \right)$$

And the expected RC is this

$$R_m(F) = \mathbb{E}_Z \left(\widehat{R}_Z(F) \right) = \mathbb{E}_{Z,\boldsymbol{\sigma}} \left(\sup_f \frac{1}{m} \sum_i \sigma_i f(z_i) \right)$$

Hence, the first part in equation(37) is definitely an expected RC, $R_m(F)$.

What about the second part? We know

$$\sigma_i = \begin{cases} 1 & \text{with } prob. = \frac{1}{2} \\ -1 & \text{with } prob. = \frac{1}{2}. \end{cases}$$

So, $-\sigma_i$ and σ are the same thing. The have the same distribution, and do not affect the

results. Hence, the second part is also a RC, $R_m(F)$.

And we have the result:

$$\mathbb{E}_Z(\Phi(Z)) \leq 2R_m(F) \tag{39}$$

Recall, $\Phi(Z) = \sup_f \left[E(f) - \widehat{E}_Z(f) \right]$

We have now proved

$$E(f(z)) \leq \mathbb{E}_Z(f(z)) + 2R_m(F) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right), \quad O(\cdot) = \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{40}$$

or we can write in this way,

$$E(f(z)) \leq \frac{1}{m} \sum_{i=1}^{m} f(z_i) + 2R_m(F) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}} \right) \tag{41}$$

Recall

$$err_D(h) \le \widehat{err}_D(h) + \cdots$$

$$\text{where, } err_D(h) = E(I), \text{ true error}$$

$$\widehat{err}_D(h) = \widehat{E}_Z(I), \text{ empirical error}$$

where $I = \begin{cases} 1 & \text{if } h(x_i) \ne y_i \\ 0 & \text{o.w.} \end{cases}$

Then we have the following theorem

Theorem 6

For hyp. space $\mathcal{H} : X \to \{-1, +1\}$, given sample $\mathcal{S} = \{x_1, x_2, \cdots, x_m\}$ where $x_i \sim D, iid, x_i \in X$,

$\forall h \in \mathcal{H}$, with $prob. \ge 1 - \delta$ we have:

$$err_D(h) \le \widehat{err}_D(h) + R_m(\mathcal{H}) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right), \quad O(\cdot) = \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{42}$$

$$err_D(h) \le \widehat{err}_D(h) + \widehat{R}_\mathcal{S}(\mathcal{H}) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{m}}\right), O(\cdot) = 3\sqrt{\frac{\ln \frac{2}{\delta}}{2m}} \tag{43}$$

Proof:

For $\mathcal{H}$, define a new sample $Z = X \times \{-1, +1\}$, then hyp. $h(\cdot)$ in $\mathcal{H}$ becomes

$$l_h(z) = l_h(x, y) = I = \begin{cases} 1 & \text{if } h(x_i) \ne y_i \\ 0 & \text{o.w.} \end{cases}$$

Hence, we transfer $\mathcal{H}$ of $\{-1, 1\}$ to a function space $\mathscr{L}_\mathcal{H} = \{l_h : h \in \mathcal{H}\}, \quad \mathscr{L}_\mathcal{H} \in [0, 1]$

And from the definition of RC, we can write

$$\widehat{R}_Z(\mathscr{L}_{\mathcal{H}}) = \mathbb{E}_{\boldsymbol{\sigma}} \left(\sup_{l_h \in \mathscr{L}_{\mathcal{H}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i l_h(x_i, y_i) \right), \text{ where } \sigma_i = \begin{cases} 1 & \text{with } prob. = \frac{1}{2} \\ -1 & \text{with } prob. = \frac{1}{2} \end{cases}$$

$$= \mathbb{E}_{\boldsymbol{\sigma}} \left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \frac{1 - y_i h(x_i)}{2} \right)$$

$$= \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}} \left(\frac{1}{m} \sum_i \sigma_i + \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i (-y_i \sigma_i h(x_i)) \right)$$

$$= 0 + \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}} \left(\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i (-y_i \sigma_i h(x_i)) \right)$$

$$= \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}} \left(\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_i (\sigma_i h(x_i)) \right)$$

$$= \frac{1}{2} \widehat{R}_S(\mathcal{H})$$

Remember,

$$\sigma_i = \begin{cases} 1 & \text{with } prob. = \frac{1}{2} \\ -1 & \text{with } prob. = \frac{1}{2} \end{cases}$$

So, $\mathbb{E}(\sigma_i) = 0$, and then $\frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}} \left(\frac{1}{m} \sum_i \sigma_i \right) = 0$

Also, y_i stands for the true label, and

$$y_i = \begin{cases} 1 \\ -1 \end{cases}$$

$-y_i \sigma_i$ and σ_i have the same distribution. So, we can replace $-y_i \sigma_i$ with σ_i.

They are exactly the same thing!!

So far, we have

$$\widehat{R}_Z(\mathscr{L}_{\mathcal{H}}) = \frac{1}{2} \widehat{R}_S(\mathcal{H})$$

take expectation we obtain the expected Rademacher Complexity,

$$R_m(\mathscr{L}_\mathcal{H}) = \frac{1}{2} R_m(\mathcal{H}) \tag{44}$$

Recall from the theorem,

$$E(f(x_i)) \leq \widehat{E}_\mathcal{S}(f(x_i)) + 2R_m(F) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right)$$

In our case,

$$E(f(x_i)) \text{ stands for } err_D(h) \tag{45}$$

$$\widehat{E}_\mathcal{S}(f(x_i)) \text{ stands for } \widehat{err}_D(h) \tag{46}$$

$$F \text{ stands for } \mathscr{L}_\mathcal{H} \tag{47}$$

Hence we can write this,

$$2R_m(F) = 2R_m(\mathscr{L}_\mathcal{H}) = 2 \times \frac{1}{2} R_m(\mathcal{H}) = R_m(\mathcal{H})$$

So, we have proved the result:

$$err_D(h) \leq \widehat{err}_D(h) + R_m(\mathcal{H}) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right), \text{ where } O(\cdot) = \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

or

$$err_D(h) \leq \widehat{err}_D(h) + \widehat{R}_\mathcal{S}(\mathcal{H}) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right), \text{ where } O(\cdot) = 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}$$

So, how do we compute $\widehat{R}_\mathcal{S}(\mathcal{H})$?

1. If $|\mathcal{H}| < \infty$:

$$\widehat{R}_\mathcal{S}(\mathcal{H}) \leq \sqrt{\frac{2\ln|\mathcal{H}|}{m}}$$

2. If $|\mathcal{H}| = \infty$:

Recall the definition of empirical RC:

$$\widehat{R}_\mathcal{S} = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^m \sigma_i h(x_i) \right]$$

Theorem 7

According to Mohri et al., (2012), based on the following three previously proved result,

$$R_m(\mathcal{H}) \leq \sqrt{\frac{2 \ln \Pi_\mathcal{H}(m)}{m}}$$

$$\Pi_\mathcal{H}(m) \leq \left(\frac{em}{d} \right)^d$$

$$E(h) \leq \widehat{E}(h) + R_m(\mathcal{H}) + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$$

we have the following result:

$$E(h) \leq \widehat{E}(h) + \sqrt{\frac{2d \ln \frac{em}{m}}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{48}$$