Introduction to Machine Learning

Vapnik–Chervonenkis Dimension

Andres Mendez-Vazquez

July 22, 2018

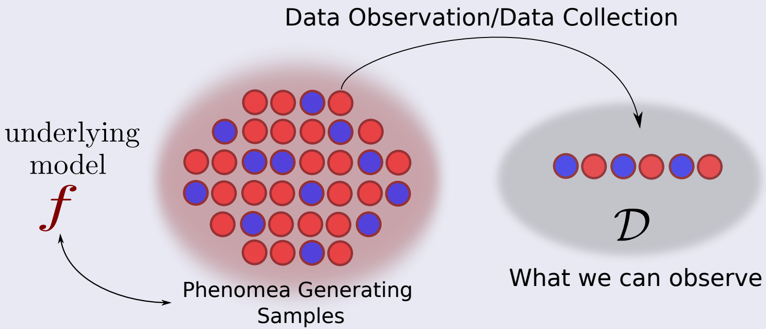
Outline

Outline

Until Now

We have been learning

- A Lot of functions to approximate the models f of a given data set $\mathcal{D}$.



The Question

But Never asked ourselves if

- **Are we able to really learn f from $\mathcal{D}$?**

Example

Consider the following data set $\mathcal{D}$

- Consider a Boolean target function over a three-dimensional input space $\mathcal{X} = \{0, 1\}^3$

With a data set $\mathcal{D}$

n	x_n	y_n
1	000	0
2	001	1
3	010	1
4	011	0
5	100	1

Example

Consider the following data set $\mathcal{D}$

- Consider a Boolean target function over a three-dimensional input space $\mathcal{X} = \{0, 1\}^3$

With a data set $\mathcal{D}$

n	$\boldsymbol{x}_n$	y_n
1	000	0
2	001	1
3	010	1
4	011	0
5	100	1

We have the following

We have the space of input has 2^3 possibilities

- Therefore, we have 2^{2^3} possible functions for f

Learning outside the data $\mathcal{D}$, basically we want a g that generalize outside $\mathcal{D}$

n	x_n	y_n	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
1	000	0	0	0	0	0	0	0	0	0	0
2	001	1	1	1	1	1	1	1	1	1	1
3	010	1	1	1	1	1	1	1	1	1	1
4	011	0	0	0	0	0	0	0	0	0	0
5	100	1	1	1	1	1	1	1	1	1	1
6	101		?	0	0	0	0	1	1	1	1
7	110		?	0	0	1	1	0	0	1	1
7	110		?	0	1	0	1	0	1	0	1

We have the following

We have the space of input has 2^3 possibilities

- Therefore, we have 2^{2^3} possible functions for f

Learning outside the data $\mathcal{D}$, basically we want a g that generalize outside $\mathcal{D}$

n	$\boldsymbol{x}_n$	y_n	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
1	000	0	**0**	0	0	0	0	0	0	0	0
2	001	1	**1**	1	1	1	1	1	1	1	1
3	010	1	**1**	1	1	1	1	1	1	1	1
4	011	0	**0**	0	0	0	0	0	0	0	0
5	100	1	**1**	1	1	1	1	1	1	1	1
6	101		?	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**
7	110		?	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**
7	110		?	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**

Outline

Here is the Dilemma!!!

Each of the $f_1, f_2, ..., f_8$

- It is a possible real f, the true f.
- Any of them is a possible good f

Therefore

- The quality of the learning will be determined by how close our prediction is to the true value.

Here is the Dilemma!!!

Each of the $f_1, f_2, ..., f_8$

- It is a possible real f, the true f.
- Any of them is a possible good f

Therefore

- The quality of the learning will be determined by how close our prediction is to the true value.

Therefore, we have

In order to select a g, we need to have an hypothesis $\mathcal{H}$

- To be able to select such g by our training procedure.

Further, any of the $f_1, f_2, ..., f_k$ is a good choice for f

- Therefore, it does not matter how near we are to the bits in $\mathcal{D}$

Our problem, we want to generalize to the data outside $\mathcal{D}$

- However, it does not make any difference if our Hypothesis is correct or incorrect in $\mathcal{D}$

Therefore, we have

In order to select a g, we need to have an hypothesis $\mathcal{H}$

- To be able to select such g by our training procedure.

Further, any of the $f_1, f_2, ..., f_8$ is a good choice for f

- Therefore, it does not matter how near we are to the bits in $\mathcal{D}$

Our problem, we want to generalize to the data outside $\mathcal{D}$

- However, it does not make any difference if our Hypothesis is correct or incorrect in $\mathcal{D}$

Therefore, we have

In order to select a g, we need to have an hypothesis $\mathcal{H}$

- To be able to select such g by our training procedure.

Further, any of the $f_1, f_2, ..., f_8$ is a good choice for f

- Therefore, it does not matter how near we are to the bits in $\mathcal{D}$

Our problem, we want to generalize to the data outside $\mathcal{D}$

- However, it does not make any difference if our Hypothesis is correct or incorrect in $\mathcal{D}$

We want to Generalize

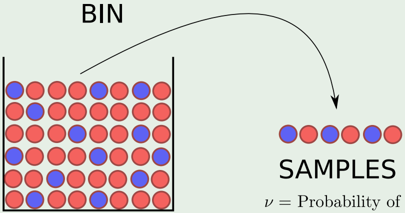
But, If we want to use only a deterministic approach to $\mathcal{H}$

- Our Attempts to use $\mathcal{H}$ to learn g is a waste of time!!!

Outline

Consider a “bin” with red and green marbles

Going back to our example



ν = Probability of Red Marbles in sample

Real μ = Probability of Red Marbles

Therefore

We have the "Real Probabilities"

- $P\left[\text{Pick a Red marble}\right] = \mu$
- $P\left[\text{Pick a Blue marble}\right] = 1 - \mu$

However, the value of μ is not know

- Thus, we sample the space for N samples in an independent way.

Here, the fraction of real marbles is equal to ν

- Question: Can ν can be used to know about μ?

Therefore

We have the "Real Probabilities"

- $P\,[\text{Pick a Red marble}] = \mu$
- $P\,[\text{Pick a Blue marble}] = 1 - \mu$

However, the value of μ is not know

- Thus, we sample the space for N samples in an independent way.

Here, the fraction of real marbles is equal to ν

- Question: Can ν can be used to know about μ?

Therefore

We have the "Real Probabilities"

- $P\,[\text{Pick a Red marble}] = \mu$
- $P\,[\text{Pick a Blue marble}] = 1 - \mu$

However, the value of μ is not know

- Thus, we sample the space for N samples in an independent way.

Here, the fraction of real marbles is equal to ν

- Question: Can ν can be used to know about μ?

Two Answers... Possible vs. Probable

No!!! Because we can see only the samples

- For example, Sample an be mostly blue while bin is mostly red.

Yes!!!

- Sample frequency ν is likely close to bin frequency μ.

Two Answers... Possible vs. Probable

No!!! Because we can see only the samples

- For example, Sample an be mostly blue while bin is mostly red.

Yes!!!

- Sample frequency ν is likely close to bin frequency μ.

What does ν say about μ?

We have the following hypothesis

- In a big sample (large N), ν is probably close to μ (within ϵ).

How?

- Hoeffding's Inequality .

What does ν say about μ?

We have the following hypothesis

- In a big sample (large N), ν is probably close to μ (within ϵ).

How?

- Hoeffding's Inequality .

Outline

We have the following theorem

Theorem (Hoeffding's inequality)

- Let $Z_1, ..., Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, where $-\infty < a \leq b < \infty$. Then

$$P\left(\frac{1}{N}\sum_{i=1}^{N}(Z_i - E[Z_i]) \geq t\right) \leq \exp^{-\frac{2Nt^2}{(b-a)^2}}$$

and

$$P\left(\frac{1}{N}\sum_{i=1}^{N}(Z_i - E[Z_i]) \leq -t\right) \leq \exp^{-\frac{2Nt^2}{(b-a)^2}}$$

for all $t \geq 0$.

Therefore

Assume that the Z_i are the random variables from the N samples

- Then, we have that values for $Z_i \in \{0, 1\}$ therefore we have that...

First inequality, for any $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \geq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Second inequality, for $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \leq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Therefore

Assume that the Z_i are the random variables from the N samples

- Then, we have that values for $Z_i \in \{0, 1\}$ therefore we have that...

First inequality, for any $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \geq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Second inequality, for $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \leq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Therefore

Assume that the Z_i are the random variables from the N samples

- Then, we have that values for $Z_i \in \{0, 1\}$ therefore we have that...

First inequality, for any $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \geq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Second inequality, for $\epsilon > 0$ and N

$$P\left[\left(\frac{1}{N}\sum_{i=1}^{N} Z_i\right) - \mu \leq \epsilon\right] \leq \exp^{-2N\epsilon^2}$$

Here

We can use the fact that

$$\nu = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

Putting all together, we have

$$P\left(\nu - \mu \geq \epsilon \text{ or } \nu - \mu \leq \epsilon\right) \leq P\left(\nu - \mu \geq \epsilon\right) + P\left(\nu - \mu \leq \epsilon\right)$$

Finally

$$P\left(|\nu - \mu| \geq \epsilon\right) \leq 2\exp^{-2N\epsilon^2}$$

Here

We can use the fact that

$$\nu = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

Putting all together, we have

$$P\left(\nu - \mu \geq \epsilon \text{ or } \nu - \mu \leq \epsilon\right) \leq P\left(\nu - \mu \geq \epsilon\right) + P\left(\nu - \mu \leq \epsilon\right)$$

Finally

$$P\left(|\nu - \mu| \geq \epsilon\right) \leq 2\exp^{-2N\epsilon^2}$$

Here

We can use the fact that

$$\nu = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

Putting all together, we have

$$P\left(\nu - \mu \geq \epsilon \text{ or } \nu - \mu \leq \epsilon\right) \leq P\left(\nu - \mu \geq \epsilon\right) + P\left(\nu - \mu \leq \epsilon\right)$$

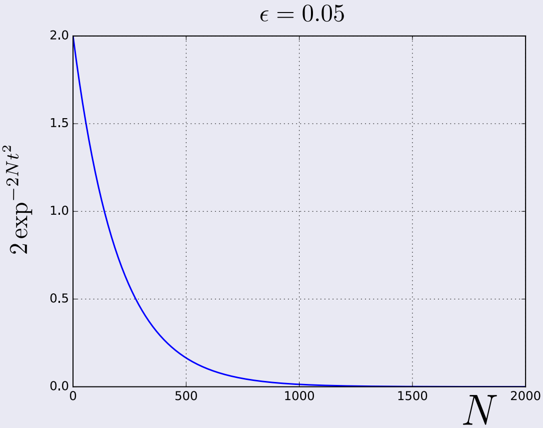
Finally

$$P\left(|\nu - \mu| \geq \epsilon\right) \leq 2\exp^{-2N\epsilon^2}$$

Therefore

We have the following

- If ϵ is small enough and as long as N is large



Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

  - Here we assume that each ball in the bin is a sample $x \in \mathcal{X}$

- $h(x) = f(x)$ we color the sample blue.
- $h(x) \neq f(x)$ we color the sample red.

Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

Learning

- We want to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which is unknown!!!
    - Here we assume that each ball in the bin is a sample $x \in \mathcal{X}$.

- $h(x) = f(x)$ we color the sample blue.
- $h(x) \neq f(x)$ we color the sample red.

Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

Learning

- We want to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which is unknown!!!
  - Here we assume that each ball in the bin is a sample $\boldsymbol{x} \in \mathcal{X}$.

Thus, it is necessary to select an hypothesis

Basically, we want to have an hypothesis h:

- $h(\boldsymbol{x}) = f(\boldsymbol{x})$ we color the sample blue.
- $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$ we color the sample red.

Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

Learning

- We want to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which is unknown!!!
  - Here we assume that each ball in the bin is a sample $\boldsymbol{x} \in \mathcal{X}$.

Thus, it is necessary to select an hypothesis

Basically, we want to have an hypothesis h:

- $h(\boldsymbol{x}) = f(\boldsymbol{x})$ we color the sample blue.
- $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$ we color the sample red.

Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

Learning

- We want to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which is unknown!!!
  - Here we assume that each ball in the bin is a sample $\boldsymbol{x} \in \mathcal{X}$.

Thus, it is necessary to select an hypothesis

Basically, we want to have an hypothesis h:

- $h(\boldsymbol{x}) = f(\boldsymbol{x})$ we color the sample **blue**.
- $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$ we color the sample **red**.

Making Possible

Possible to estimate $\nu \approx \mu$

- How do we connect with Learning?

Learning

- We want to find a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ which is unknown!!!
  - Here we assume that each ball in the bin is a sample $\boldsymbol{x} \in \mathcal{X}$.

Thus, it is necessary to select an hypothesis

Basically, we want to have an hypothesis h:

- $h(\boldsymbol{x}) = f(\boldsymbol{x})$ we color the sample **blue**.
- $h(\boldsymbol{x}) \neq f(\boldsymbol{x})$ we color the sample **red**.

Here a Small Remark

Here, we are not talking about classes

- When talking about blue and red balls, but if we are able to identify the correct label:

$$\widehat{y}_h = h(\boldsymbol{x}) = f(\boldsymbol{x}) = y$$

or

$$\widehat{y}_h = h(\boldsymbol{x}) \neq f(\boldsymbol{x}) = y$$

Still, the use of blue and red balls allows

- to see our Learning Problem as a Bernoulli distribution

Here a Small Remark

Here, we are not talking about classes

- When talking about blue and red balls, but if we are able to identify the correct label:

$$\widehat{y}_h = h(\boldsymbol{x}) = f(\boldsymbol{x}) = y$$

or

$$\widehat{y}_h = h(\boldsymbol{x}) \neq f(\boldsymbol{x}) = y$$

Still, the use of blue and red balls allows

- to see our Learning Problem as a Bernoulli distribution

Swiss mathematician Jacob Bernoulli

Definition

- The Bernoulli distribution is a discrete distribution having two possible outcomes $X = 0$ or $X = 1$.

With the following probabilities

$$P(X|p) = \begin{cases} 1-p & \text{if } X = 0 \\ p & \text{if } X = 1 \end{cases}$$

Also expressed as

$$P(X = k|p) = (p)^k (1-p)^{1-k}$$

Swiss mathematician Jacob Bernoulli

Definition

- The Bernoulli distribution is a discrete distribution having two possible outcomes $X = 0$ or $X = 1$.

With the following probabilities

$$P\left(X|p\right) = \begin{cases} 1 - p & \text{if } X = 0 \\ p & \text{if } X = 1 \end{cases}$$

Also expressed as

$$P\left(X = k|p\right) = \left(p\right)^k \left(1 - p\right)^{1-k}$$

Swiss mathematician Jacob Bernoulli

Definition

- The Bernoulli distribution is a discrete distribution having two possible outcomes $X = 0$ or $X = 1$.

With the following probabilities

$$P(X|p) = \begin{cases} 1-p & \text{if } X = 0 \\ p & \text{if } X = 1 \end{cases}$$

Also expressed as

$$P(X = k|p) = (p)^k (1-p)^{1-k}$$

Outline

Thus

We define E_{in} (in-sample error)

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} I\left(h\left(\boldsymbol{x}_n\right) \neq f\left(\boldsymbol{x}_n\right)\right)$$

- We have made explicit the dependency of E_{in} on the particular h that we are considering.

Now E_{out} (out-of-sample error)

$$E_{out}(h) = P\left(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) = \mu$$

Where

- The probability is based on the distribution P over $\mathcal{X}$ which is used to sample the data points $\boldsymbol{x}$.

Thus

We define E_{in} (in-sample error)

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)\right)$$

- We have made explicit the dependency of E_{in} on the particular h that we are considering.

Now E_{out} (out-of-sample error)

$$E_{out}(h) = P\left(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) = \mu$$

Where

- The probability is based on the distribution P over $\mathcal{X}$ which is used to sample the data points $\boldsymbol{x}$.

Thus

We define E_{in} (in-sample error)

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h\left(\boldsymbol{x}_n\right) \neq f\left(\boldsymbol{x}_n\right)\right)$$

- We have made explicit the dependency of E_{in} on the particular h that we are considering.

Now E_{out} (out-of-sample error)

$$E_{out}(h) = P\left(h\left(\boldsymbol{x}\right) \neq f\left(\boldsymbol{x}\right)\right) = \mu$$

Where

- The probability is based on the distribution P over $\mathcal{X}$ which is used to sample the data points $\boldsymbol{x}$.

Outline

Generalization Error

Definition (Generalization Error/out-of-sample error)

Given a **hypothesis/proposed** model $h \in \mathcal{H}$, a target **concept/real** model $f \in \mathcal{F}$, and an underlying distribution $\mathcal{D}$, the generalization error or risk of h is defined by

$$R(h) = P_{\boldsymbol{x} \sim \mathcal{D}}\left(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) = E_{\boldsymbol{x} \sim \mathcal{D}}\left[I_{h(\boldsymbol{x}) \neq f(\boldsymbol{x})}\right]$$

[a]

where I_ω is the indicator function of the event ω.

[a] This comes the fact that $1 * P(A) + 0 * P\left(\overline{A}\right) = E\left[I_A\right]$

Empirical Error

Definition (Empirical Error/in-sample error**)**

Given a **hypothesis/proposed** model $h \in \mathcal{H}$, a target **concept/real** model $f \in \mathcal{F}$, a sample $\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, the empirical error or empirical risk of h is defined by:

$$\widehat{R} = \frac{1}{N} \sum_{i=1}^{N} I_{h(\boldsymbol{x}_i) \neq f(\boldsymbol{x}_i)}$$

Outline

Basically

We have

$$P\left(\left|E_{in}\left(h\right)-E_{out}\left(h\right)\right|\geq\epsilon\right)\leq 2\exp^{-2Nt^{2}}$$

Now, we need to consider an entire set of hypothesis, $\mathcal{H}$

$$\mathcal{H}=\{h_1,h_2,\dots,h_M\}$$

Basically

We have

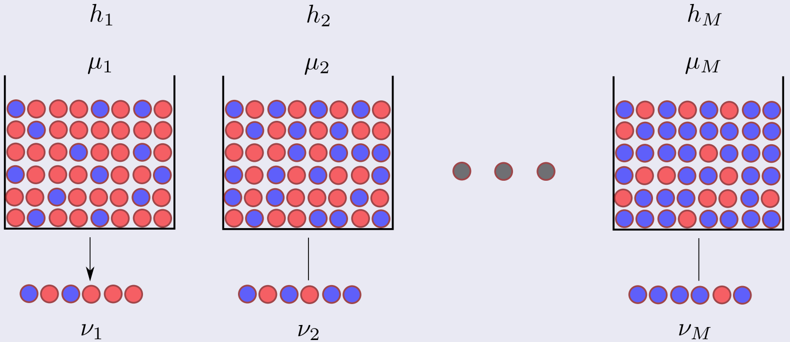
$$P\left(\left|E_{in}\left(h\right)-E_{out}\left(h\right)\right|\geq\epsilon\right)\leq2\exp^{-2Nt^{2}}$$

Now, we need to consider an entire set of hypothesis, $\mathcal{H}$

$$\mathcal{H}=\left\{h_{1},h_{2},...,h_{M}\right\}$$

Therefore

Each Hypothesis is a scenario in the bin space



Remark

The Hoeffding Inequality still applies to each bin individually

- Now, we need to consider all the bins simultaneously.

Here, we have the following situation

- h is fixed before the data set is generated!!!

If you are allowed to change h after you generate the data set

- The Hoeffding Inequality no longer holds

Remark

The Hoeffding Inequality still applies to each bin individually

- Now, we need to consider all the bins simultaneously.

Here, we have the following situation

- h is fixed before the data set is generated!!!

If you are allowed to change h after you generate the data set

- The Hoeffding Inequality no longer holds

Remark

The Hoeffding Inequality still applies to each bin individually

- Now, we need to consider all the bins simultaneously.

Here, we have the following situation

- h is fixed before the data set is generated!!!

If you are allowed to change h after you generate the data set

- The Hoeffding Inequality no longer holds

Outline

Therefore

With multiple hypotheses in $\mathcal{H}$

- The Learning Algorithm chooses the final hypothesis g based on $\mathcal{D}$ after generating the data.

The statement we would like to make is not

$P\left(\left|E_{in}\left(h_m\right) - E_{out}\left(h_m\right)\right| \geq \epsilon\right)$ is small.

We would rather

$P\left(\left|E_{in}\left(g\right) - E_{out}\left(g\right)\right| \geq \epsilon\right)$ is small for the final hypothesis g.

Therefore

With multiple hypotheses in $\mathcal{H}$

- The Learning Algorithm chooses the final hypothesis g based on $\mathcal{D}$ after generating the data.

The statement we would like to make is not

$$P\left(\left|E_{in}\left(h_m\right) - E_{out}\left(h_m\right)\right| \geq \epsilon\right) \text{ is small.}$$

We would rather

$$P\left(\left|E_{in}\left(g\right) - E_{out}\left(g\right)\right| \geq \epsilon\right) \text{ is small for the final hypothesis } g.$$

Therefore

With multiple hypotheses in $\mathcal{H}$

- The Learning Algorithm chooses the final hypothesis g based on $\mathcal{D}$ after generating the data.

The statement we would like to make is not

$$P\left(\left|E_{in}\left(h_m\right) - E_{out}\left(h_m\right)\right| \geq \epsilon\right) \text{ is small.}$$

We would rather

$$P\left(\left|E_{in}\left(g\right) - E_{out}\left(g\right)\right| \geq \epsilon\right) \text{ is small for the final hypothesis } g.$$

Therefore

Something Notable

- The hypothesis g is not fixed ahead of time before generating the data

Thus we need to bound

$$P\left(|E_{in}(g) - E_{out}(g)| \geq \epsilon\right)$$

- Which it does not depend on which g the algorithm picks.

Therefore

Something Notable

- The hypothesis g is not fixed ahead of time before generating the data

Thus we need to bound

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)$$

- Which it does not depend on which g the algorithm picks.

We have two rules

First one

$$\text{if } A_1 \Longrightarrow A_2, \text{ then } P(A_1) \leq P(A_2)$$

If you have any set of events $A_1, A_2, ..., A_M$

$$P(A_1 \cup A_2 \cup \cdots \cup A_M) \leq \sum_{m=1}^{M} P(A_m)$$

We have two rules

First one

$$\text{if } A_1 \implies A_2, \text{ then } P(A_1) \leq P(A_2)$$

If you have any set of events $A_1, A_2, ..., A_M$

$$P(A_1 \cup A_2 \cup \cdots \cup A_M) \leq \sum_{m=1}^{M} P(A_m)$$

Therefore

Now assuming independence between hypothesis

$$
\begin{aligned}
|E_{in}(g) - E_{out}(g)| \geq \epsilon \Longrightarrow & |E_{in}(h_1) - E_{out}(h_1)| \geq \epsilon \\
& \text{or } |E_{in}(h_2) - E_{out}(h_2)| \geq \epsilon \\
& \cdots \\
& \text{or } |E_{in}(h_M) - E_{out}(h_M)| \geq \epsilon
\end{aligned}
$$

Thus

We have

$$
\begin{aligned}
P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq & P\left[\left|E_{in}\left(h_1\right)-E_{out}\left(h_1\right)\right|\geq\epsilon\right.\\
& \text{or }\left|E_{in}\left(h_2\right)-E_{out}\left(h_2\right)\right|\geq\epsilon\\
& \cdots\\
& \left.\text{or }\left|E_{in}\left(h_M\right)-E_{out}\left(h_M\right)\right|\geq\epsilon\right]
\end{aligned}
$$

Then

We have

$$P\left(|E_{in}(g) - E_{out}(g)| \geq \epsilon\right) \leq \sum_{m=1}^{M} \left[|E_{in}(h_m) - E_{out}(h_m)| \geq \epsilon\right]$$

Thus

$$P\left(|E_{in}(g) - E_{out}(g)| \geq \epsilon\right) \leq 2M \exp^{-2N\epsilon^2}$$

Then

We have

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq\sum_{m=1}^{M}\left[\left|E_{in}\left(h_m\right)-E_{out}\left(h_m\right)\right|\geq\epsilon\right]$$

Thus

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq 2M\exp^{-2N\epsilon^2}$$

Outline

We have

Something Notable

- We have introduced two apparently conflicting arguments about the feasibility of learning.

- The other say it is possible!!!

We have

Something Notable

- We have introduced two apparently conflicting arguments about the feasibility of learning.

We have two possibilities

- One argument says that we cannot learn anything outside of $\mathcal{D}$.
- The other say it is possible!!!

We have

Something Notable

- We have introduced two apparently conflicting arguments about the feasibility of learning.

We have two possibilities

- One argument says that we cannot learn anything outside of $\mathcal{D}$.
- The other say it is possible!!!

Here, we introduce the probabilistic answer

- This will solve our conundrum!!!

We have

Something Notable

- We have introduced two apparently conflicting arguments about the feasibility of learning.

We have two possibilities

- One argument says that we cannot learn anything outside of $\mathcal{D}$.
- The other say it is possible!!!

Here, we introduce the probabilistic answer

- This will solve our conundrum!!!

Then

The Deterministic Answer

- Do we have something to say about f outside of $\mathcal{D}$? The answer is NO.

The Probabilistic Answer

- Is $\mathcal{D}$ telling us something likely about f outside of $\mathcal{D}$? The answer is YES

The reason why

- We approach our Learning from a Probabilistic point of view!!!

Then

The Deterministic Answer

- Do we have something to say about f outside of $\mathcal{D}$? The answer is NO.

The Probabilistic Answer

- Is $\mathcal{D}$ telling us something likely about f outside of $\mathcal{D}$? The answer is YES

The reason why

- We approach our Learning from a Probabilistic point of view!!!

Then

The Deterministic Answer

- Do we have something to say about f outside of $\mathcal{D}$? The answer is NO.

The Probabilistic Answer

- Is $\mathcal{D}$ telling us something likely about f outside of $\mathcal{D}$? The answer is YES

The reason why

- We approach our Learning from a Probabilistic point of view!!!

Outline

For example

We could have hypothesis based in hyperplanes

- Linear regression output:

$$h(\boldsymbol{x}) = \sum_{i=1}^{d} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$$

Therefore

$$E_{in}(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \left(h(\boldsymbol{x}_n) - y_n\right)^2$$

For example

We could have hypothesis based in hyperplanes

- Linear regression output:

$$h(\boldsymbol{x}) = \sum_{i=1}^{d} w_i x_i = \boldsymbol{w}^T \boldsymbol{x}$$

Therefore

$$E_{in}(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} \left(h(\boldsymbol{x}_n) - y_n\right)^2$$

Clearly, we have used loss functions

Mostly to give meaning $h \approx f$

- By Error Measures $E(h, f)$

By using pointwise definitions

$$e(h(x), f(x))$$

Examples

- Squared Error $e(h(x), f(x)) = [h(x) - f(x)]^2$
- Binary Error $e(h(x), f(x)) = I[h(x) \neq f(x)]$

Clearly, we have used loss functions

Mostly to give meaning $h \approx f$

- By Error Measures $E(h, f)$

By using pointwise definitions

$$e(h(\boldsymbol{x}), f(\boldsymbol{x}))$$

Examples

- Squared Error $e(h(\boldsymbol{x}), f(\boldsymbol{x})) = [h(\boldsymbol{x}) - f(\boldsymbol{x})]^2$
- Binary Error $e(h(\boldsymbol{x}), f(\boldsymbol{x})) = I[h(\boldsymbol{x}) \neq f(\boldsymbol{x})]$

Clearly, we have used loss functions

Mostly to give meaning $h \approx f$

- By Error Measures $E(h, f)$

By using pointwise definitions

$$e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)$$

Examples

- Squared Error $e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right) = \left[h\left(\boldsymbol{x}\right) - f\left(\boldsymbol{x}\right)\right]^2$
- Binary Error $e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right) = I\left[h\left(\boldsymbol{x}\right) \neq f\left(\boldsymbol{x}\right)\right]$

Outline

Therefore, we have

The Overall Error

$$E\left(h, f\right) = \text{Average of pointwise errors } e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)$$

In-Sample Error

$$E_{in}\left(h\right) = \frac{1}{N}\sum_{i=1}^{N} e\left(h\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_i\right)\right)$$

Out-of-sample error

$$E_{in}\left(h\right) = E_{\mathcal{X}}\left[e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)\right]$$

Therefore, we have

The Overall Error

$$E\left(h, f\right) = \text{Average of pointwise errors } e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)$$

In-Sample Error

$$E_{in}\left(h\right) = \frac{1}{N}\sum_{i=1}^{N} e\left(h\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_i\right)\right)$$

Out-of-sample error

$$E_{in}\left(h\right) = E_{\mathcal{X}}\left[e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)\right]$$

Therefore, we have

The Overall Error

$$E\left(h, f\right) = \text{Average of pointwise errors } e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)$$

In-Sample Error

$$E_{in}\left(h\right) = \frac{1}{N}\sum_{i=1}^{N} e\left(h\left(\boldsymbol{x}_i\right), f\left(\boldsymbol{x}_i\right)\right)$$

Out-of-sample error

$$E_{in}\left(h\right) = E_{\mathcal{X}}\left[e\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)\right]$$

We have the following Process

Assuming $P(y|\boldsymbol{x})$ instead of $y = f(\boldsymbol{x})$

- Then a data point $(\boldsymbol{x}, y)$ is now generated by the joint distribution $P(\boldsymbol{x}, y) = P(\boldsymbol{x}) P(y|\boldsymbol{x})$

Therefore

- Noisy target is a deterministic target plus added noise.

$$f(\boldsymbol{x}) \approx E[y|\boldsymbol{x}] + (y - f(\boldsymbol{x}))$$

We have the following Process

Assuming $P(y|\boldsymbol{x})$ instead of $y = f(\boldsymbol{x})$

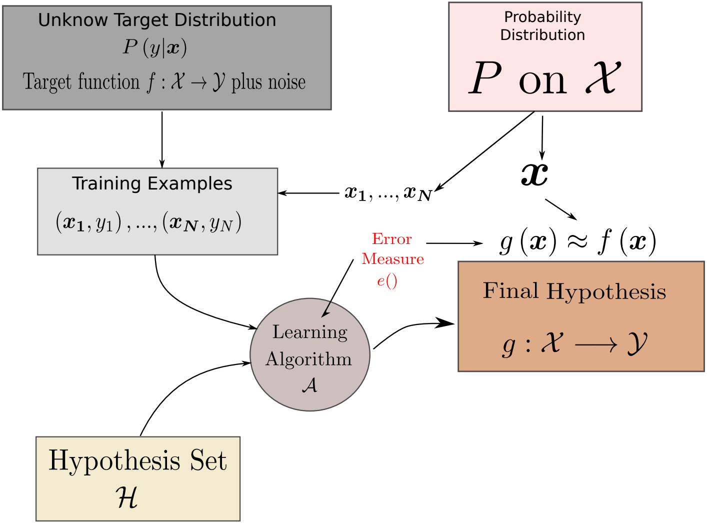
- Then a data point $(\boldsymbol{x}, y)$ is now generated by the joint distribution $P(\boldsymbol{x}, y) = P(\boldsymbol{x}) P(y|\boldsymbol{x})$

Therefore

- Noisy target is a deterministic target plus added noise.

$$f(\boldsymbol{x}) \approx E[y|\boldsymbol{x}] + (y - f(\boldsymbol{x}))$$

Finally, we have as Learning Process



Therefore

Distinction between $P(y|\boldsymbol{x})$ and $P(\boldsymbol{x})$

- Both convey probabilistic aspects of $\boldsymbol{x}$ and y.

Therefore

1. The Target distribution $P(y|\boldsymbol{x})$ is what we are trying to learn.
2. The Input distribution $P(\boldsymbol{x})$ quantifies relative importance of $\boldsymbol{x}$.

Finally

- Merging $P(\boldsymbol{x}, y) = P(y|\boldsymbol{x}) P(\boldsymbol{x})$ mixes the two concepts

Therefore

Distinction between $P(y|\boldsymbol{x})$ and $P(\boldsymbol{x})$

- Both convey probabilistic aspects of $\boldsymbol{x}$ and y.

Therefore

1. The Target distribution $P(y|\boldsymbol{x})$ is what we are trying to learn.
2. The Input distribution $P(\boldsymbol{x})$ quantifies relative importance of $\boldsymbol{x}$.

Finally

- Merging $P(\boldsymbol{x}, y) = P(y|\boldsymbol{x})P(\boldsymbol{x})$ mixes the two concepts

Therefore

Distinction between $P(y|\boldsymbol{x})$ and $P(\boldsymbol{x})$

- Both convey probabilistic aspects of $\boldsymbol{x}$ and y.

Therefore

1. The Target distribution $P(y|\boldsymbol{x})$ is what we are trying to learn.
2. The Input distribution $P(\boldsymbol{x})$ quantifies relative importance of $\boldsymbol{x}$.

Finally

- Merging $P(\boldsymbol{x}, y) = P(y|\boldsymbol{x}) P(\boldsymbol{x})$ mixes the two concepts

Therefore

Learning is feasible because It is likely that

$$E_{out}(g) \approx E_{in}(g)$$

Therefore, we need $g \approx f$

$$E_{out}(g) = P(g(x) \neq f(x)) \approx 0$$

How do we achieve this?

$$E_{out}(g) \approx E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N} I(g(x_n) \neq f(x_n))$$

Therefore

Learning is feasible because It is likely that

$$E_{out}(g) \approx E_{in}(g)$$

Therefore, we need $g \approx f$

$$E_{out}(g) = P\left(g(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) \approx 0$$

How do we achieve this?

$$E_{out}(g) \approx E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N} I\left(g(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)\right)$$

Therefore

Learning is feasible because It is likely that

$$E_{out}(g) \approx E_{in}(g)$$

Therefore, we need $g \approx f$

$$E_{out}(g) = P\left(g(\boldsymbol{x}) \neq f(\boldsymbol{x})\right) \approx 0$$

How do we achieve this?

$$E_{out}(g) \approx E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N} I\left(g(\boldsymbol{x}_n) \neq f(\boldsymbol{x}_n)\right)$$

Then

We make at the same time

$$E_{in}(g) \approx 0$$

- To Make the Error in our selected hypothesis g with respect to the real function f

Learning splits in two questions

1. Can we make $E_{out}(g)$ is close enough $E_{in}(g)$?
2. Can we make $E_{in}(g)$ small enough?

Then

We make at the same time

$$E_{in}(g) \approx 0$$

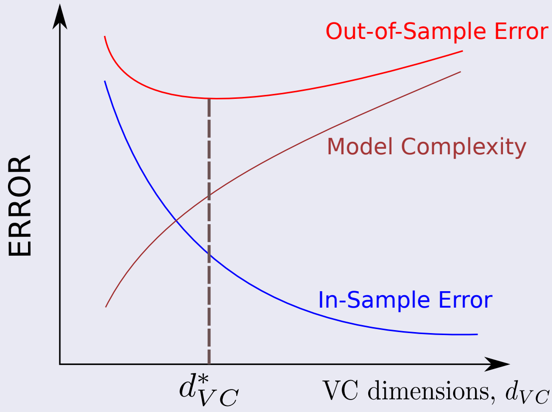
- To Make the Error in our selected hypothesis g with respect to the real function f

Learning splits in two questions

1. Can we make $E_{out}(g)$ is close enough $E_{in}(g)$?
2. Can we make $E_{in}(g)$ small enough?

Therefore, we have

Nice Connection with Bias-Variance Trade-off



Outline

We have that

The out-of-sample error

$$E_{out}(h) = P(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$$

It Measures how well our training on D

- It has generalized to data that we have not seen before.

Remark

- E_{out} is based on the performance over the entire input space $\mathcal{X}$.

We have that

The out-of-sample error

$$E_{out}(h) = P\left(h(\boldsymbol{x}) \neq f(\boldsymbol{x})\right)$$

It Measures how well our training on $\mathcal{D}$

- It has generalized to data that we have not seen before.

Remark

- E_{out} is based on the performance over the entire input space $\mathcal{X}$.

We have that

The out-of-sample error

$$E_{out}(h) = P(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))$$

It Measures how well our training on $\mathcal{D}$

- It has generalized to data that we have not seen before.

Remark

- E_{out} is based on the performance over the entire input space $\mathcal{X}$.

Testing Data Set

Intuitively

- we want to estimate the value of E_{out} using a sample of data points.

Something Notable

- These points must be 'fresh' test points that have not been used for training.

Basically

- Out Testing Set.

Testing Data Set

Intuitively

- we want to estimate the value of E_{out} using a sample of data points.

Something Notable

- These points must be '**fresh**' test points that have not been used for training.

Basically

- Out Testing Set.

Testing Data Set

Intuitively

- we want to estimate the value of E_{out} using a sample of data points.

Something Notable

- These points must be '**fresh**' test points that have not been used for training.

Basically

- Out Testing Set.

Outline

Thus

It is possible to define

- The **generalization error** as the discrepancy between E_{in} and E_{out}

Therefore

- The Hoeffding Inequality is a way to characterize the generalization error with a **probabilistic bound**

$$P\left(|E_{in}(g) - E_{out}(g)| \geq \epsilon\right) \leq 2M \exp^{-2N\epsilon^2}$$

  - For any $\epsilon > 0$.

Thus

It is possible to define

- The **generalization error** as the discrepancy between E_{in} and E_{out}

Therefore

- The Hoeffding Inequality is a way to characterize the generalization error with a **probabilistic bound**

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq 2M\exp^{-2N\epsilon^{2}}$$

  - For any $\epsilon > 0$.

Outline

Reinterpreting This

Assume a Tolerance Level δ, for example $\delta = 0.0005$

- It is possible to say that with probability $1 - \delta$:

$$E_{out}\left(g\right) < E_{in}\left(g\right) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

Proof

We have the complement Hoeffding Probability using the absolute value

$$P\left(\left|E_{out}\left(g\right)-E_{in}\left(g\right)\right|<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

Therefore, we have

$$P\left(-\epsilon<E_{out}\left(g\right)-E_{in}\left(g\right)<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

This imply

$$E_{out}\left(g\right)<E_{in}\left(g\right)+\epsilon$$

Proof

We have the complement Hoeffding Probability using the absolute value

$$P\left(\left|E_{out}\left(g\right)-E_{in}\left(g\right)\right|<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

Therefore, we have

$$P\left(-\epsilon<E_{out}\left(g\right)-E_{in}\left(g\right)<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

This imply

$$E_{out}\left(g\right)<E_{in}\left(g\right)+\epsilon$$

Proof

We have the complement Hoeffding Probability using the absolute value

$$P\left(\left|E_{out}\left(g\right)-E_{in}\left(g\right)\right|<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

Therefore, we have

$$P\left(-\epsilon<E_{out}\left(g\right)-E_{in}\left(g\right)<\epsilon\right)\leq 1-2M\exp^{-2N\epsilon^{2}}$$

This imply

$$E_{out}\left(g\right)<E_{in}\left(g\right)+\epsilon$$

Therefore

We simply use

$$\delta = 2M \exp^{-2N\epsilon^2}$$

Then

$$\ln 1 - \ln \frac{\delta}{2M} = 2N\epsilon^2$$

Therefore

$$\epsilon = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

Therefore

We simply use

$$\delta = 2M \exp^{-2N\epsilon^2}$$

Then

$$\ln 1 - \ln \frac{\delta}{2M} = 2N\epsilon^2$$

Therefore

$$\epsilon = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

Therefore

We simply use

$$\delta = 2M \exp^{-2N\epsilon^2}$$

Then

$$\ln 1 - \ln \frac{\delta}{2M} = 2N\epsilon^2$$

Therefore

$$\epsilon = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

Generalization Bound

This inequality is know as a generalization Bound

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

Outline

We have

The following inequality also holds

$$-\epsilon < E_{out}(g) - E_{in}(g) \Rightarrow E_{out}(g) > E_{in}(g) - \epsilon$$

Thus

- Not only we want our hypothesis g to do well int the out samples, $E_{out}(g) < E_{in}(g) + \epsilon$

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!
  - Given any hypothesis with higher

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I(h(x_n) \neq f(x_n))$$

than g.

We have

The following inequality also holds

$$-\epsilon < E_{out}(g) - E_{in}(g) \Rightarrow E_{out}(g) > E_{in}(g) - \epsilon$$

Thus

- Not only we want our hypothesis g to do well int the out samples, $E_{out}(g) < E_{in}(g) + \epsilon$

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!
  - Given any hypothesis with higher

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h(x_n) \neq f(x_n)\right)$$

than g.

We have

The following inequality also holds

$$-\epsilon < E_{out}(g) - E_{in}(g) \Rightarrow E_{out}(g) > E_{in}(g) - \epsilon$$

Thus

- Not only we want our hypothesis g to do well int the out samples, $E_{out}(g) < E_{in}(g) + \epsilon$

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!
    - Given any hypothesis with higher

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h\left(\boldsymbol{x}_n\right) \neq f\left(\boldsymbol{x}_n\right)\right)$$

than g.

Therefore

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!

Given any hypothesis h with higher than g

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^{N} I\left(h\left(x_n\right) \neq f\left(x_n\right)\right)$$

It will have a higher $E_{out}(h)$ given

$$E_{out}(h) > E_{in}(h) - \epsilon$$

Therefore

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!

Given any hypothesis h with higher than g

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h\left(\boldsymbol{x}_n\right) \neq f\left(\boldsymbol{x}_n\right)\right)$$

It will have a higher $E_{out}(h)$ given

$$E_{out}(h) > E_{in}(h) - \epsilon$$

Therefore

But, we want to know how well we did with our $\mathcal{H}$

- Thus, $E_{out}(g) > E_{in}(g) - \epsilon$ assures that it is not possible to do better!!!

Given any hypothesis h with higher than g

$$E_{in}(h) = \frac{1}{N}\sum_{n=1}^{N} I\left(h\left(\boldsymbol{x}_n\right) \neq f\left(\boldsymbol{x}_n\right)\right)$$

It will have a higher $E_{out}(h)$ given

$$E_{out}(h) > E_{in}(h) - \epsilon$$

Outline

The Infiniteness of $\mathcal{H}$

A Problem with the Error Bound given its dependency on M

$$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

What happens when M becomes infinity

- The number of hypothesis in $\mathcal{H}$ becomes infinity.

Thus, the bound becomes infinity

- Problem, almost all interesting learning models have infinite $\mathcal{H}$....
  - For Example... in our linear Regression... $f(x) = w^T x$

The Infiniteness of $\mathcal{H}$

A Problem with the Error Bound given its dependency on M

$$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

What happens when M becomes infinity

- The number of hypothesis in $\mathcal{H}$ becomes infinity.

Thus, the bound becomes infinity

- Problem, almost all interesting learning models have infinite $\mathcal{H}$....
  - For Example... in our linear Regression... $f(x) = w^T x$

The Infiniteness of $\mathcal{H}$

A Problem with the Error Bound given its dependency on M

$$\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

What happens when M becomes infinity

- The number of hypothesis in $\mathcal{H}$ becomes infinity.

Thus, the bound becomes infinity

- Problem, almost all interesting learning models have infinite $\mathcal{H}$....
  - For Example... in our linear Regression... $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$

Therefore, we need to replace M

We need to find a finite substitute with finite range values

- For this, we notice that

$$|E_{in}(h_1) - E_{out}(h_1)| \geq \epsilon \text{ or } |E_{in}(h_2) - E_{out}(h_2)| \geq \epsilon \cdots$$

$$\text{or } |E_{in}(h_M) - E_{out}(h_M)| \geq \epsilon$$

We have

This guarantee $|E_{in}(g) - E_{out}(g)| \geq \epsilon$

- Thus, we can take a look at the events $\mathcal{B}_m$ events for which you have $|E_{in}(h_m) - E_{out}(h_m)| \geq \epsilon$

Then

$$P\left[\mathcal{B}_1 \quad \text{or } \mathcal{B}_2 \quad \cdots \quad \text{or } \mathcal{B}_M\right] \leq \sum_{m=1}^{M} P\left[\mathcal{B}_m\right]$$

We have

This guarantee $|E_{in}(g) - E_{out}(g)| \geq \epsilon$

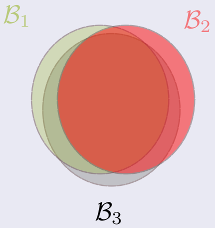
- Thus, we can take a look at the events $\mathcal{B}_m$ events for which you have $|E_{in}(h_m) - E_{out}(h_m)| \geq \epsilon$

Then

$$P\left[\ \mathcal{B}_1 \quad \text{or } \mathcal{B}_2 \quad \cdots \quad \text{or } \mathcal{B}_M\ \right] \leq \sum_{m=1}^{M} P\left[\mathcal{B}_m\right]$$

Now, we have the following

Example



We have a gross overestimate

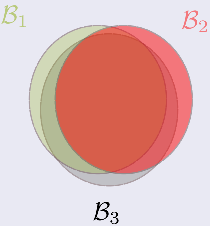
- Basically, if h_i and h_j are quite similar the two events

$$\left|E_{in}(h_i) - E_{out}(h_i)\right| \geq \epsilon \text{ and } \left|E_{in}(h_j) - E_{out}(h_j)\right| \geq \epsilon$$

are likely to coincide!!!

Now, we have the following

Example



We have a gross overestimate

- Basically, if h_i and h_j are quite similar the two events

$$|E_{in}(h_i) - E_{out}(h_i)| \geq \epsilon \text{ and } |E_{in}(h_j) - E_{out}(h_j)| \geq \epsilon$$

are likely to coincide!!!

We have

Something Notable

- In a typical learning model, many hypotheses are indeed very similar.

The mathematical theory of generalization hinges on this observation

- We only need to account for the overlapping on different hypothesis to substitute M.

We have

Something Notable

- In a typical learning model, many hypotheses are indeed very similar.

The mathematical theory of generalization hinges on this observation

- We only need to account for the overlapping on different hypothesis to substitute M.

Outline

Consider

A finite data set

$$\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$$

And we consider a set of hypothesis $h \in \mathcal{H}$ such that $h : \mathcal{X} \to \{-1, +1\}$

- We get a N-tuple, when applied to $\mathcal{X}$, $h(x_1), h(x_2), ..., h(x_N)$ of ± 1.

Such N-tuple is called a Dichotomy

- Given that it splits $x_1, x_2, ..., x_N$ into two groups...

Consider

A finite data set

$$\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$$

And we consider a set of hypothesis $h \in \mathcal{H}$ such that $h : \mathcal{X} \to \{-1, +1\}$

- We get a N-tuple, when applied to $\mathcal{X}$, $h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), ..., h(\boldsymbol{x}_N)$ of ± 1.

Such N-tuple is called a Dichotomy

- Given that it splits $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ into two groups...

Consider

A finite data set

$$\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$$

And we consider a set of hypothesis $h \in \mathcal{H}$ such that $h : \mathcal{X} \to \{-1, +1\}$

- We get a N-tuple, when applied to $\mathcal{X}$, $h(\boldsymbol{x}_1), h(\boldsymbol{x}_2), ..., h(\boldsymbol{x}_N)$ of ± 1.

Such N-tuple is called a Dichotomy

- Given that it splits $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$ into two groups...

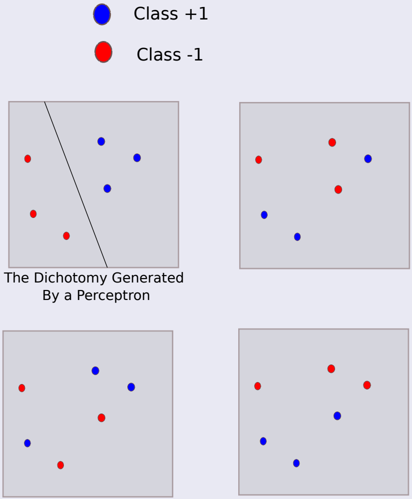
Dichotomy

Definition

- Given a hypothesis set $\mathcal{H}$, a **dichotomy** of a set $\mathcal{X}$ is **one of the possible ways** of labeling the points of $\mathcal{X}$ using a hypothesis in $\mathcal{H}$.

Examples of Dichotomies

Here the first Dichotomy can be generated by a perceptron



The Dichotomy Generated By a Perceptron

Something Important

Each $h \in \mathcal{H}$ generates a dichotomy on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

- However, two different h's may generate the same dichotomy if they generate the same pattern

Remark

Definition

- Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by

$$\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \left\{\left(h\left[\boldsymbol{x}_1\right], h\left[\boldsymbol{x}_2\right], ..., h\left[\boldsymbol{x}_N\right]\right) | h \in \mathcal{H}\right\}$$

Therefore

- We can see $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ as a set of hypothesis by using the geometry of the points.

Thus

- A large $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ means $\mathcal{H}$ is more diverse.

Remark

Definition

- Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by

$$\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \left\{\left(h\left[\boldsymbol{x}_1\right], h\left[\boldsymbol{x}_2\right], ..., h\left[\boldsymbol{x}_N\right]\right) | h \in \mathcal{H}\right\}$$

Therefore

- We can see $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ as a set of hypothesis by using the geometry of the points.

Thus

- A large $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ means $\mathcal{H}$ is more diverse.

Remark

Definition

- Let $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by

$$\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \left\{\left(h\left[\boldsymbol{x}_1\right], h\left[\boldsymbol{x}_2\right], ..., h\left[\boldsymbol{x}_N\right]\right) | h \in \mathcal{H}\right\}$$

Therefore

- We can see $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ as a set of hypothesis by using the geometry of the points.

Thus

- A large $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right)$ means $\mathcal{H}$ is more diverse.

Growth function, Our Replacement of M

Definition

- The growth function is defined for a hypothesis set $\mathcal{H}$ by

$$m_{\mathcal{H}}\left(N\right) = \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathcal{X}} \#\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\right)$$

  - where $\#$ denotes the cardinality (number of elements) of a set.

Therefore

- $m_{\mathcal{H}}\left(N\right)$ is the **maximum number of dichotomies** that be generated by $\mathcal{H}$ on any N points.
  - We remove dependency on the entire $\mathcal{X}$

Growth function, Our Replacement of M

Definition

- The growth function is defined for a hypothesis set $\mathcal{H}$ by

$$m_{\mathcal{H}}\left(N\right) = \max_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \mathcal{X}} \#\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\right)$$

  - where $\#$ denotes the cardinality (number of elements) of a set.

Therefore

- $m_{\mathcal{H}}\left(N\right)$ is the **maximum number of dichotomies** that be generated by $\mathcal{H}$ on any N points.
  - **We remove dependency on the entire $\mathcal{X}$**

Therefore

We have that

- M and $m_{\mathcal{H}}(N)$ is a measure of the of the number of hypothesis in $\mathcal{H}$

However, we avoid considering all of $\mathcal{X}$

- Now we only consider N points instead of the entire $\mathcal{X}$.

Therefore

We have that

- M and $m_{\mathcal{H}}(N)$ is a measure of the of the number of hypothesis in $\mathcal{H}$

However, we avoid considering all of $\mathcal{X}$

- Now we only consider N points instead of the entire $\mathcal{X}$.

Upper Bound for $m_{\mathcal{H}}(N)$

First, we know that

$$\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N) \subseteq \{-1, +1\}^N$$

Hence, we have the value of $m_{\mathcal{H}}(N)$ is at most $\#\{-1, +1\}^N$

$$m_{\mathcal{H}}(N) \leq 2^N$$

Upper Bound for $m_{\mathcal{H}}(N)$

First, we know that

$$\mathcal{H}(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N) \subseteq \{-1, +1\}^N$$

Hence, we have the value of $m_{\mathcal{H}}(N)$ is at most $\#\{-1, +1\}^N$

$$m_{\mathcal{H}}(N) \leq 2^N$$

Outline

Therefore

If $\mathcal{H}$ is capable of generating all possible dichotomies on $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

- Then,
  - $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \{-1, +1\}^N$ and $\#\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = 2^N$

We can say that

- $\mathcal{H}$ can shatter $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

Meaning

- $\mathcal{H}$ is as diverse as can be on this particular sample.

Therefore

If $\mathcal{H}$ is capable of generating all possible dichotomies on $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

- Then,
  - $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \{-1, +1\}^N$ and $\#\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = 2^N$

We can say that

- $\mathcal{H}$ can shatter $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

Meaning

- $\mathcal{H}$ is as diverse as can be on this particular sample.

Therefore

If $\mathcal{H}$ is capable of generating all possible dichotomies on $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

- Then,
  - $\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = \{-1, +1\}^N$ and $\#\mathcal{H}\left(\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right) = 2^N$

We can say that

- $\mathcal{H}$ can shatter $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N$

Meaning

- $\mathcal{H}$ is as diverse as can be on this particular sample.

Shattering

Definition

- A set $\mathcal{X}$ of $N \geq 1$ points is said to be shattered by a hypothesis set $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of $\mathcal{X}$, that is when

$$m_{\mathcal{H}}(N) = 2^N$$

Outline

Example

Positive Rays

- Imagine a input space on $\mathbb{R}$, with $\mathcal{H}$ consisting of all hypotheses $h : \mathbb{R} \rightarrow \{-1, +1\}$ of the form

$$h\left(x\right) = sign\left(x - a\right)$$

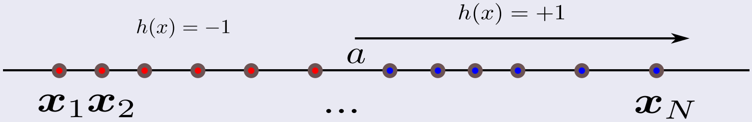
Example

Positive Rays

- Imagine a input space on $\mathbb{R}$, with $\mathcal{H}$ consisting of all hypotheses $h : \mathbb{R} \to \{-1, +1\}$ of the form

$$h\left(x\right) = sign\left(x - a\right)$$

Example



Thus, we have that

As we change a, we get $N+1$ different dichotomies

$$m_{\mathcal{H}}(N) = N+1$$

Now, we have the case of positive intervals

- $\mathcal{H}$ consists of all hypotheses in one dimension that return $+1$ within some interval and -1 otherwise.

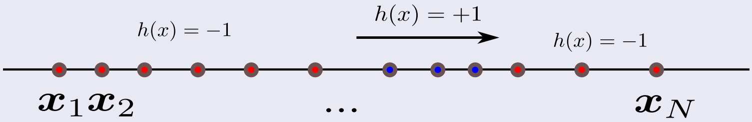
Thus, we have that

As we change a, we get $N+1$ different dichotomies

$$m_{\mathcal{H}}(N) = N + 1$$

Now, we have the case of positive intervals

- $\mathcal{H}$ consists of all hypotheses in one dimension that return $+1$ within some interval and -1 otherwise.



Therefore

We have

- The line is again split by the points into $N+1$ regions.

Furthermore

- The dichotomy we get is decided by which two regions contain the end values of the interval

Therefore, we have the number of possible dichotomies

$$\binom{N+1}{2}$$

Therefore

We have

- The line is again split by the points into $N+1$ regions.

Furthermore

- The dichotomy we get is decided by which two regions contain the end values of the interval

Therefore, we have the number of possible dichotomies

$$\binom{N+1}{2}$$

Therefore

We have

- The line is again split by the points into $N+1$ regions.

Furthermore

- The dichotomy we get is decided by which two regions contain the end values of the interval

Therefore, we have the number of possible dichotomies

$$\binom{N+1}{2}$$

Additionally

If the two points fall in the same region, the $\mathcal{H} = -1$

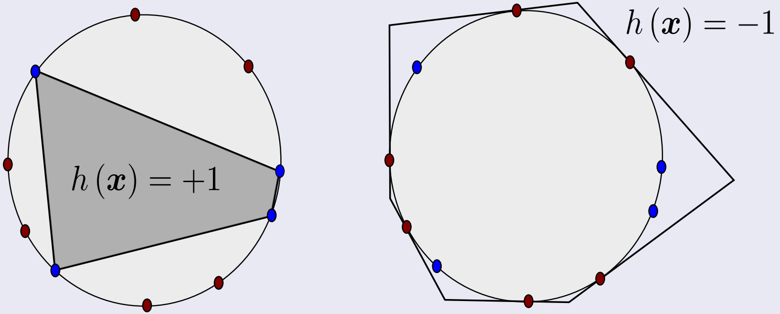
- Then

$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

Finally

In the case of a Convex Set in $\mathbb{R}^2$

- $\mathcal{H}$ consists of all hypothesis in two dimensions that are positive inside some convex set and negative elsewhere.



Therefore

We have the following

$$m_{\mathcal{H}}(N) = 2^N$$

By using the "Radon's theorem"

Outline

Remember

We have that

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq 2M\exp^{-2N\epsilon^{2}}$$

What if $m_{\mathcal{H}}(N)$ replaces M

- If $m_{\mathcal{H}}(N)$ is polynomial, we have an excellent case!!!

Therefore, we need to prove that

- $m_{\mathcal{H}}(N)$ is polynomial

Remember

We have that

$$P\left(\left|E_{in}\left(g\right) - E_{out}\left(g\right)\right| \geq \epsilon\right) \leq 2M \exp^{-2N\epsilon^2}$$

What if $m_{\mathcal{H}}\left(N\right)$ replaces M

- If $m_{\mathcal{H}}\left(N\right)$ is polynomial, we have an excellent case!!!

Therefore, we need to prove that

- $m_{\mathcal{H}}\left(N\right)$ is polynomial

Remember

We have that

$$P\left(\left|E_{in}\left(g\right)-E_{out}\left(g\right)\right|\geq\epsilon\right)\leq 2M\exp^{-2N\epsilon^2}$$

What if $m_{\mathcal{H}}\left(N\right)$ replaces M

- If $m_{\mathcal{H}}\left(N\right)$ is polynomial, we have an excellent case!!!

Therefore, we need to prove that

- $m_{\mathcal{H}}\left(N\right)$ is polynomial

Outline

Break Point

Definition

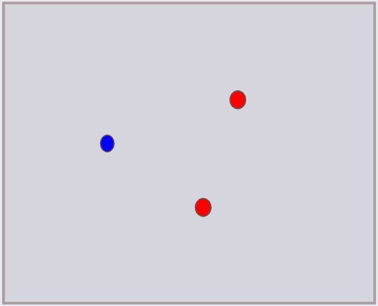
- If **no data set of size** k can be shattered by $\mathcal{H}$, then k is said to be a break point for $\mathcal{H}$:

$$\underline{m_{\mathcal{H}}(k) < 2^k}$$

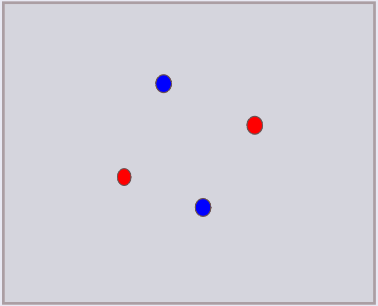
Example

For the Perceptron, we have $k = 4$

Shatter



Non-Shatter



Important

Something Notable

- In general, it is easier to find a break point for $\mathcal{H}$ than to compute the full growth function for that $\mathcal{H}$.

Using this concept

We are ready to define the concept of Vapnik–Chervonenkis (VC) dimension.

Important

Something Notable

- In general, it is easier to find a break point for $\mathcal{H}$ than to compute the full growth function for that $\mathcal{H}$.

Using this concept

We are ready to define the concept of Vapnik–Chervonenkis (VC) dimension.

Outline

VC-Dimension

Definition

- The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$ (Those points need to be in "General Position"):

$$VC_{dim}\left(\mathcal{H}\right) = \max\left\{k|m_{\mathcal{H}}\left(k\right) = 2^k\right\}$$

  - A set containing k points, for arbitrary k, is in **general linear position** if and only if no $(k-1)$ −dimensional flat contains them all

Important Remarks

Remark 1

- if $VC_{dim}(\mathcal{H}) = d$, there exists a set of size d that can be fully shattered.

Remark2

- This does not imply that all sets of size d or less are fully shattered
    - This is typically the case!!!

Important Remarks

Remark 1

- if $VC_{dim}(\mathcal{H}) = d$, there exists a set of size d that can be fully shattered.

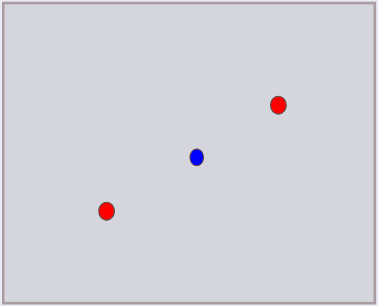
Remark2

- This does not imply that all sets of size d or less are fully shattered
  - This is typically the case!!!

Why? **General Linear Position**

For example in the Perceptron

No General Position



Now, we define $B(N, k)$

Definition

- $B(N, k)$ is the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies.

Something Notable

- The definition of $B(N, k)$ assumes a break point k!!!

Now, we define $B(N, k)$

Definition

- $B(N, k)$ is the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies.

Something Notable

- The definition of $B(N, k)$ assumes a break point k!!!

Further

Since $B(N,k)$ is a maximum

- It is an upper bound for $m_{\mathcal{H}}(N)$ under a break point k.

$$m_{\mathcal{H}}(N) \leq B(N,k) \text{ if } k \text{ is a break point for } \mathcal{H}.$$

Then

- We need to find a Bound for $B(N,k)$ to prove that $m_{\mathcal{H}}(k)$ is polynomial.

Further

Since $B(N,k)$ is a maximum

- It is an upper bound for $m_{\mathcal{H}}(N)$ under a break point k.

$$m_{\mathcal{H}}(N) \leq B(N,k) \text{ if } k \text{ is a break point for } \mathcal{H}.$$

Then

- We need to find a Bound for $B(N,k)$ to prove that $m_{\mathcal{H}}(k)$ is polynomial.

Therefore

Thus, we start with two boundary conditions $k = 1$ and $N = 1$

$$B(N,1) = 1$$
$$B(1,k) = 2 \ k > 1$$

Why?

Something Notable

- $B(N, 1) = 1$ for all N since **if no subset of size 1 can be shattered**
  - Then only one dichotomy can be allowed.
  - Because a second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered.

  - Because the constraint is vacuously true and we have 2 possible dichotomies $+1$ and -1.

Why?

Something Notable

- $B(N,1)=1$ for all N since **if no subset of size 1 can be shattered**
  - Then only one dichotomy can be allowed.
  - Because a second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered.

Second

- $B(1,k)=2$ for $k>1$ since there do not even exist subsets of size k.
  - Because the constraint is vacuously true and we have 2 possible dichotomies $+1$ and -1.

Why?

Something Notable

- $B(N, 1) = 1$ for all N since **if no subset of size 1 can be shattered**
  - Then only one dichotomy can be allowed.
  - Because a second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered.

Second

- $B(1, k) = 2$ for $k > 1$ since there do not even exist subsets of size k.
  - Because the constraint is vacuously true and we have 2 possible dichotomies $+1$ and -1.

Why?

Something Notable

- $B(N, 1) = 1$ for all N since **if no subset of size 1 can be shattered**
  - Then only one dichotomy can be allowed.
  - Because a second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered.

Second

- $B(1, k) = 2$ for $k > 1$ since there do not even exist subsets of size k.
  - Because the constraint is vacuously true and we have 2 possible dichotomies $+1$ and -1.

Why?

Something Notable

- $B(N, 1) = 1$ for all N since **if no subset of size 1 can be shattered**
  - Then only one dichotomy can be allowed.
  - Because a second different dichotomy must differ on at least one point and then that subset of size 1 would be shattered.

Second

- $B(1, k) = 2$ for $k > 1$ since there do not even exist subsets of size k.
  - Because the constraint is vacuously true and we have 2 possible dichotomies $+1$ and -1.

Outline

$B(N,k)$ Dichotomies, $N \geq 2$ and $k \geq 2$

	# of rows	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_{N-1}$	$\boldsymbol{x}_N$
S_1	α	+1	+1	$\cdots$	+1	+1
		-1	+1	$\cdots$	+1	-1
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
		+1	-1	$\cdots$	-1	-1
		-1	+1	$\cdots$	-1	+1
S_2: S_2^+	β	+1	-1	$\cdots$	+1	+1
		-1	-1	$\cdots$	+1	+1
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
		+1	-1	$\cdots$	+1	+1
		-1	+1	$\cdots$	-1	+1
S_2: S_2^-	β	+1	-1	$\cdots$	+1	-1
		-1	-1	$\cdots$	+1	-1
		$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
		+1	-1	$\cdots$	+1	-1
		-1	+1	$\cdots$	-1	-1

What is this partition mean

First, Consider the dichotomies on $\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_{N-1}$

- Some appear once (Either $+1$ or -1 at x_N), but only ONCE!!!
- We collect them in S_1

What is this partition mean

First, Consider the dichotomies on $\boldsymbol{x}_1 \boldsymbol{x}_2 \cdots \boldsymbol{x}_{N-1}$

- Some appear once (Either $+1$ or -1 at x_N), but only ONCE!!!
- We collect them in S_1

The Remaining Dichotomies appear Twice

- Once with $+1$ and once with -1 in the x_N column.

What is this partition mean

First, Consider the dichotomies on $\boldsymbol{x}_1\boldsymbol{x}_2\cdots\boldsymbol{x}_{N-1}$

- Some appear once (Either $+1$ or -1 at x_N), but only ONCE!!!
- We collect them in S_1

The Remaining Dichotomies appear Twice

- Once with $+1$ and once with -1 in the $\boldsymbol{x}_N$ column.

Therefore, we collect them in three sets

The ones with only one Dichotomy

- We use the set S_1

The other in two different sets

- S_2^+ the ones with $x_N = +1$.
- S_2^- the ones with $x_N = -1$.

Therefore, we collect them in three sets

The ones with only one Dichotomy

- We use the set S_1

The other in two different sets

- S_2^+ the ones with $x_N = +1$.
- S_2^- the ones with $x_N = -1$.

Therefore

We have the following

$$B(N,k) = \alpha + 2\beta$$

The total number of different dichotomies on the first $N-1$ points

- They are $\alpha + \beta$.

Additionally, no subset of k of these first $N-1$ points can be shattered

- Since no k-subset of all N points can be shattered:

$$\alpha + \beta \leq B(N-1,k)$$

By definition of B.

Therefore

We have the following

$$B(N,k) = \alpha + 2\beta$$

The total number of different dichotomies on the first $N-1$ points

- They are $\alpha + \beta$.

Additionally, no subset of k of these first $N-1$ points can be shattered

- Since no k-subset of all N points can be shattered:

$$\alpha + \beta \leq B(N-1,k)$$

By definition of B.

Therefore

We have the following

$$B(N,k) = \alpha + 2\beta$$

The total number of different dichotomies on the first $N-1$ points

- They are $\alpha + \beta$.

Additionally, no subset of k of these first $N-1$ points can be shattered

- Since no k-subset of all N points can be shattered:

$$\alpha + \beta \leq B(N-1,k)$$

By definition of B.

Then

Further, no subset of size $k-1$ of the first $N-1$ points can be shattered by the dichotomies in S_2^+

- If there existed such a subset, then taking the corresponding set of dichotomies in S_2^- and $\boldsymbol{x}_N$
  - You finish with a subset of size k that can be shattered a contradiction given the definition of $B(N,k)$.

Then

Further, no subset of size $k-1$ of the first $N-1$ points can be shattered by the dichotomies in S_2^+

- If there existed such a subset, then taking the corresponding set of dichotomies in S_2^- and $\boldsymbol{x}_N$
  - You finish with a subset of size k that can be shattered a contradiction given the definition of $B(N,k)$.

Therefore

$$\beta \leq B(N-1,k-1)$$

Then

Further, no subset of size $k-1$ of the first $N-1$ points can be shattered by the dichotomies in S_2^+

- If there existed such a subset, then taking the corresponding set of dichotomies in S_2^- and $\boldsymbol{x}_N$
  - You finish with a subset of size k that can be shattered a contradiction given the definition of $B(N,k)$.

Therefore

$$\beta \leq B(N-1, k-1)$$

Then, we have

$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

Then

Further, no subset of size $k-1$ of the first $N-1$ points can be shattered by the dichotomies in S_2^+

- If there existed such a subset, then taking the corresponding set of dichotomies in S_2^- and $\boldsymbol{x}_N$
  - You finish with a subset of size k that can be shattered a contradiction given the definition of $B(N,k)$.

Therefore

$$\beta \leq B(N-1,k-1)$$

Then, we have

$$B(N,k) \leq B(N-1,k) + B(N-1,k-1)$$

Outline

Connecting the Growth Function with the VC_{dim}

Sauer's Lemma

- For all $k \in \mathbb{N}$, the following inequality holds:

$$B\left(N,k\right) \leq \sum_{i=0}^{k-1} \left(\begin{array}{c} N \\ i \end{array} \right)$$

Proof

Proof

- For $k = 1$

$$B\left(N,1\right) \leq B(N-1,1) + B\left(N-1,0\right) = 1 + 0 = \begin{pmatrix} N \\ 0 \end{pmatrix}$$

Then, by induction

- We assume that the statement is true for $N \leq N_0$ and all k.

Proof

Proof

- For $k = 1$

$$B(N,1) \leq B(N-1,1) + B(N-1,0) = 1 + 0 = \binom{N}{0}$$

Then, by induction

- We assume that the statement is true for $N \leq N_0$ and all k.

Now

We need to prove this for $N = N_0 + 1$ and all k

- Observation: This is true for $k = 1$ given

$$B(N, 1) = 1$$

Now, consider $k \geq 2$

$$B(N_0, k) + B(N_0, k-1)$$

Therefore

$$B(N_0 + 1, k) \leq \sum_{i=0}^{k-1} \binom{N_0}{i} + \sum_{i=0}^{k-2} \binom{N_0}{i}$$

Now

We need to prove this for $N = N_0 + 1$ and all k

- Observation: This is true for $k = 1$ given

$$B\left(N, 1\right) = 1$$

Now, consider $k \geq 2$

$$B(N_0, k) + B\left(N_0, k-1\right)$$

Therefore

$$B\left(N_0 + 1, k\right) \leq \sum_{i=0}^{k-1} \binom{N_0}{i} + \sum_{i=0}^{k-2} \binom{N_0}{i}$$

Now

We need to prove this for $N = N_0 + 1$ and all k

- Observation: This is true for $k = 1$ given

$$B(N, 1) = 1$$

Now, consider $k \geq 2$

$$B(N_0, k) + B(N_0, k-1)$$

Therefore

$$B(N_0 + 1, k) \leq \sum_{i=0}^{k-1} \binom{N_0}{i} + \sum_{i=0}^{k-2} \binom{N_0}{i}$$

Therefore

We have the following

$$
\begin{aligned}
B\left(N_0+1,k\right) &\leq 1+\sum_{i=1}^{k-1}\binom{N_0}{i}+\sum_{i=1}^{k-1}\binom{N_0}{i-1}\\
&= 1+\sum_{i=1}^{k-1}\left[\binom{N_0}{i}+\binom{N_0}{i-1}\right]\\
&= 1+\sum_{i=1}^{k-1}\binom{N_0+1}{i}=\sum_{i=0}^{k-1}\binom{N_0+1}{i}
\end{aligned}
$$

- Because $\binom{N_0}{i}+\binom{N_0}{i-1}=\binom{N_0+1}{i}$

Therefore

We have the following

$$
\begin{aligned}
B\left(N_0+1,k\right) &\leq 1+\sum_{i=1}^{k-1}\binom{N_0}{i}+\sum_{i=1}^{k-1}\binom{N_0}{i-1}\\
&= 1+\sum_{i=1}^{k-1}\left[\binom{N_0}{i}+\binom{N_0}{i-1}\right]\\
&= 1+\sum_{i=1}^{k-1}\binom{N_0+1}{i}=\sum_{i=0}^{k-1}\binom{N_0+1}{i}
\end{aligned}
$$

- Because $\binom{N_0}{i}+\binom{N_0}{i-1}=\binom{N_0+1}{i}$

Therefore

We have the following

$$
\begin{aligned}
B\left(N_0+1,k\right) &\leq 1+\sum_{i=1}^{k-1}\binom{N_0}{i}+\sum_{i=1}^{k-1}\binom{N_0}{i-1}\\
&=1+\sum_{i=1}^{k-1}\left[\binom{N_0}{i}+\binom{N_0}{i-1}\right]\\
&=1+\sum_{i=1}^{k-1}\binom{N_0+1}{i}=\sum_{i=0}^{k-1}\binom{N_0+1}{i}
\end{aligned}
$$

- Because $\binom{N_0}{i}+\binom{N_0}{i-1}=\binom{N_0+1}{i}$

Therefore

We have the following

$$B\left(N_0+1,k\right) \leq 1+\sum_{i=1}^{k-1}\binom{N_0}{i}+\sum_{i=1}^{k-1}\binom{N_0}{i-1}$$
$$=1+\sum_{i=1}^{k-1}\left[\binom{N_0}{i}+\binom{N_0}{i-1}\right]$$
$$=1+\sum_{i=1}^{k-1}\binom{N_0+1}{i}=\sum_{i=0}^{k-1}\binom{N_0+1}{i}$$

- Because $\binom{N_0}{i}+\binom{N_0}{i-1}=\binom{N_0+1}{i}$

Now

We have in conclusion for all k

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Therefore

$$m_{\mathcal{H}}(N) \leq B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Now

We have in conclusion for all k

$$B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Therefore

$$m_{\mathcal{H}}(N) \leq B(N,k) \leq \sum_{i=0}^{k-1} \binom{N}{i}$$

Then

Theorem

- If $m_{\mathcal{H}}(k) < 2^k$ for some value k, then

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \begin{pmatrix} N \\ i \end{pmatrix}$$

Finally

Corollary

- Let $\mathcal{H}$ be a hypothesis set with $VC_{dim}(\mathcal{H}) = k$. Then, for all $N \geq k$

$$m_{\mathcal{H}}(N) \leq \left(\frac{eN}{k}\right)^{k-1} = O\left(N^k\right)$$

We have

Proof

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k} \binom{N}{i}$$

We have

Proof

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \sum_{i=0}^{k} \binom{N}{i} \\
&\leq \sum_{i=0}^{k} \binom{N}{i} \left[\frac{N}{k}\right]^{k-i}
\end{aligned}
$$

We have

Proof

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \sum_{i=0}^{k} \binom{N}{i} \\
&\leq \sum_{i=0}^{k} \binom{N}{i} \left[\frac{N}{k}\right]^{k-i} \\
&\leq \sum_{i=0}^{N} \binom{N}{i} \left[\frac{N}{k}\right]^{k-i}
\end{aligned}
$$

We have

Proof

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \sum_{i=0}^{k} \binom{N}{i} \\
&\leq \sum_{i=0}^{k} \binom{N}{i} \left[\frac{N}{k}\right]^{k-i} \\
&\leq \sum_{i=0}^{N} \binom{N}{i} \left[\frac{N}{k}\right]^{k-i} \\
&\left[\frac{N}{k}\right]^{k} \sum_{i=0}^{N} \binom{N}{i} \left[\frac{k}{N}\right]^{i}
\end{aligned}
$$

Therefore

We have

$$m_{\mathcal{H}}(N) \leq \left[\frac{N}{k}\right]^k \sum_{i=0}^{N} \binom{N}{i} \left[\frac{k}{N}\right]^i$$

Therefore

We have

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \left[\frac{N}{k}\right]^k \sum_{i=0}^{N} \binom{N}{i} \left[\frac{k}{N}\right]^i \\
&= \left[\frac{N}{k}\right]^k \left[1 + \frac{k}{N}\right]^N
\end{aligned}
$$

Given that $(1 - x) = e^{-x}$

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \left[\frac{N}{k}\right]^k e^{\frac{k}{N}} \\
&\leq \left[\frac{N}{k}\right]^{k-1} e^{k-1} = \left[\frac{e}{k}\right]^k N^k = O\left(N^k\right)
\end{aligned}
$$

Therefore

We have

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \left[\frac{N}{k}\right]^k \sum_{i=0}^{N} \binom{N}{i} \left[\frac{k}{N}\right]^i \\
&= \left[\frac{N}{k}\right]^k \left[1+\frac{k}{N}\right]^N
\end{aligned}
$$

Given that $(1-x)=e^{-x}$

$$
m_{\mathcal{H}}(N) \leq \left[\frac{N}{k}\right]^k e^{\frac{k}{N}}
$$

$$
\leq \left[\frac{N}{k}\right]^{k-1} e^{k-1} = \left[\frac{e}{k}\right]^k N^k = O\left(N^k\right)
$$

Therefore

We have

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \left[\frac{N}{k}\right]^{k} \sum_{i=0}^{N} \left(\begin{array}{c} N \\ i \end{array} \right) \left[\frac{k}{N}\right]^{i} \\
&= \left[\frac{N}{k}\right]^{k} \left[1 + \frac{k}{N}\right]^{N}
\end{aligned}
$$

Given that $(1-x) = e^{-x}$

$$
\begin{aligned}
m_{\mathcal{H}}(N) &\leq \left[\frac{N}{k}\right]^{k} e^{\frac{k}{N}} \\
&\leq \left[\frac{N}{k}\right]^{k-1} e^{k-1} = \left[\frac{e}{k}\right]^{k} N^{k} = O\left(N^{k}\right)
\end{aligned}
$$

Therefore

We have that

- $m_{\mathcal{H}}(N)$ is bounded by N^{k-1} i.e. if $m_{\mathcal{H}}(k) < 2^k$ we have that $m_{\mathcal{H}}(N)$ is polynomial
- We are not depending on the number of hypothesis!!!!

Therefore

We have that

- $m_{\mathcal{H}}(N)$ is bounded by N^{k-1} i.e. if $m_{\mathcal{H}}(k) < 2^k$ we have that $m_{\mathcal{H}}(N)$ is polynomial
- We are not depending on the number of hypothesis!!!!

Therefore

We have that

- $m_{\mathcal{H}}(N)$ is bounded by N^{k-1} i.e. if $m_{\mathcal{H}}(k) < 2^k$ we have that $m_{\mathcal{H}}(N)$ is polynomial
- We are not depending on the number of hypothesis!!!!

Outline

Remark about $m_{\mathcal{H}}(k)$

We have bounded the number of effective hypothesis

- **Yes!!! we can have M hypotheses but the number of dichotomies generated by them is bounded by $m_{\mathcal{H}}(k)$**

VC-Dimension Again

Definition

- The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$ (Those points need to be in "General Position"):

$$VC_{dim}\left(\mathcal{H}\right) = \max\left\{k | m_{\mathcal{H}}\left(k\right) = 2^k\right\}$$

Something Notable

- If $m_{\mathcal{H}}\left(N\right) = 2^N$ for all N, $VC_{dim}\left(\mathcal{H}\right) = \infty$

VC-Dimension Again

Definition

- The VC-dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be fully shattered by $\mathcal{H}$ (Those points need to be in "General Position"):

$$VC_{dim}\left(\mathcal{H}\right) = \max\left\{k|m_{\mathcal{H}}\left(k\right) = 2^k\right\}$$

Something Notable

- If $m_{\mathcal{H}}\left(N\right) = 2^N$ for all N, $VC_{dim}\left(\mathcal{H}\right) = \infty$

Remember

We have the following

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$$

We instead of using M, we use $m_{\mathcal{H}}(N)$

- We can use our growth function as the effective way to bound

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{1}{2N}\ln\frac{2m_{\mathcal{H}}(N)}{\delta}}$$

Remember

We have the following

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

We instead of using M, we use $m_{\mathcal{H}}(N)$

- We can use our growth function as the effective way to bound

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{1}{2N} \ln \frac{2m_{\mathcal{H}}(N)}{\delta}}$$

VC Generalized Bound

Theorem (VC Generalized Bound)

- For any tolerance $\delta > 0$ and $\mathcal{H}$ be a hypothesis set with $VC_{dim}(\mathcal{H}) = k$.,

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{2k}{N} \ln \frac{eN}{k}} + \sqrt{\frac{1}{2N} \ln \frac{1}{\delta}}$$

  - with probability $\geq 1 - \delta$

Something Notable

This Bound only fails when $VC_{dim}(\mathcal{H}) = \infty$!!!

VC Generalized Bound

Theorem (VC Generalized Bound)

- For any tolerance $\delta > 0$ and $\mathcal{H}$ be a hypothesis set with $VC_{dim}(\mathcal{H}) = k.$,

$$E_{in}(g) < E_{out}(g) + \sqrt{\frac{2k}{N}\ln\frac{eN}{k}} + \sqrt{\frac{1}{2N}\ln\frac{1}{\delta}}$$

  - with probability $\geq 1 - \delta$

Something Notable

This Bound only fails when $VC_{dim}(\mathcal{H}) = \infty$!!!

Proof

Although we will not talk about it

- We will remark the that is possible to use the Rademacher complexity
  - To manage the number of overlapping hypothesis (Which can be infinite)

We will stop here, but

- But I will encourage to look at more about the proof...

Proof

Although we will not talk about it

- We will remark the that is possible to use the Rademacher complexity
    - To manage the number of overlapping hypothesis (Which can be infinite)

We will stop here, but

- But I will encourage to look at more about the proof...

About the Proof

For More, take a look at

- "A Probabilistic Theory of Pattern Recognition" by Luc Devroye et al.
- "Foundations of Machine Learning" by Mehryar Mohori et al.

This is the equivalent to use Measure Theory to understand the innards of Probability

- We are professionals, we must understand!!!

About the Proof

For More, take a look at

- "A Probabilistic Theory of Pattern Recognition" by Luc Devroye et al.
- "Foundations of Machine Learning" by Mehryar Mohori et al.

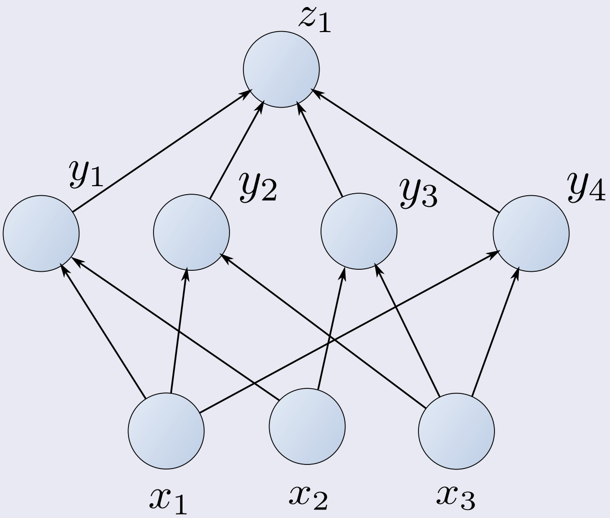
This is the equivalent to use Measure Theory to understand the innards of Probability

- We are professionals, we must understand!!!

Outline

As you remember from previous classes

We have architectures like



G-composition of $\mathcal{H}$

Let G be a layered directed acyclic graph

Where directed edges go from one layer l to the next layer $l + 1$.

- Intermediate Nodes with in-degree r
- Single Output node with out-degree 0

G-composition of $\mathcal{H}$

Let G be a layered directed acyclic graph

Where directed edges go from one layer l to the next layer $l + 1$.

Now, we have a set of hypothesis $\mathcal{H}$

- NInput Nodes with in-degree 0
- Intermediate Nodes with in-degree r
- Single Output node with out-degree 0

G-composition of $\mathcal{H}$

Let G be a layered directed acyclic graph

Where directed edges go from one layer l to the next layer $l+1$.

Now, we have a set of hypothesis $\mathcal{H}$

- NInput Nodes with in-degree 0
- Intermediate Nodes with in-degree r
- Single Output node with out-degree 0

$\mathcal{H}$ our hypothesis over the space Euclidean space $\mathbb{R}^r$

- Basically each node represent the hypothesis $c_i : \mathbb{R}^r \rightarrow \{-1, 1\}$ by mean of tanh.

G-composition of $\mathcal{H}$

Let G be a layered directed acyclic graph

Where directed edges go from one layer l to the next layer $l+1$.

Now, we have a set of hypothesis $\mathcal{H}$

- NInput Nodes with in-degree 0
- Intermediate Nodes with in-degree r
- Single Output node with out-degree 0

$\mathcal{H}$ our hypothesis over the space Euclidean space $\mathbb{R}^r$

- Basically each node represent the hypothesis $c_i : \mathbb{R}^r \to \{-1, 1\}$ by mean of tanh.

G-composition of $\mathcal{H}$

Let G be a layered directed acyclic graph

Where directed edges go from one layer l to the next layer $l+1$.

Now, we have a set of hypothesis $\mathcal{H}$

- NInput Nodes with in-degree 0
- Intermediate Nodes with in-degree r
- Single Output node with out-degree 0

$\mathcal{H}$ our hypothesis over the space Euclidean space $\mathbb{R}^r$

- Basically each node represent the hypothesis $c_i : \mathbb{R}^r \rightarrow \{-1, 1\}$ by mean of $\tanh$.

Therefore

We have that

- The Neural concept represent an hypothesis from $\mathbb{R}^N$ to $\{-1, 1\}$

Therefore the entire hypothesis is a composition of concepts.

- This is called a G-composition of $\mathcal{H}$.

Therefore

We have that

- The Neural concept represent an hypothesis from $\mathbb{R}^N$ to $\{-1, 1\}$

Therefore the entire hypothesis is a composition of concepts

- This is called a G-composition of $\mathcal{H}$.

We have the following theorem

Theorem (Kearns and Vazirani, 1994)

- Let G be a layered directed acyclic graph with N input nodes and $r \geq 2$ internal nodes each of indegree r.
- Let $\mathcal{H}$ hypothesis set over $\mathbb{R}^r$ of $VC_{dim}(\mathcal{H}) = d$, and let G-composition of $\mathcal{H}$. then

$$VC_{dim}(\mathcal{H}_G) \leq 2ds \log_2(es)$$

We have the following theorem

Theorem (Kearns and Vazirani, 1994)

- Let G be a layered directed acyclic graph with N input nodes and $r \geq 2$ internal nodes each of indegree r.
- Let $\mathcal{H}$ hypothesis set over $\mathbb{R}^r$ of $VC_{dim}(\mathcal{H}) = d$, and let G-composition of $\mathcal{H}$. then

$$VC_{dim}(\mathcal{H}_G) \leq 2ds \log_2 (es)$$