Fundmentals of Machine Learning

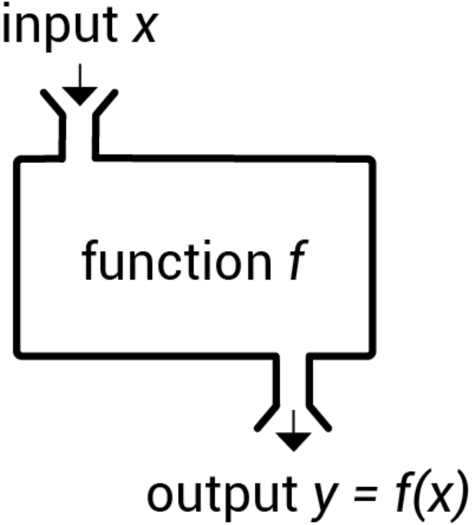
Master Degree in Computer Science - IAS Curriculum

Probabilistic Learning - I

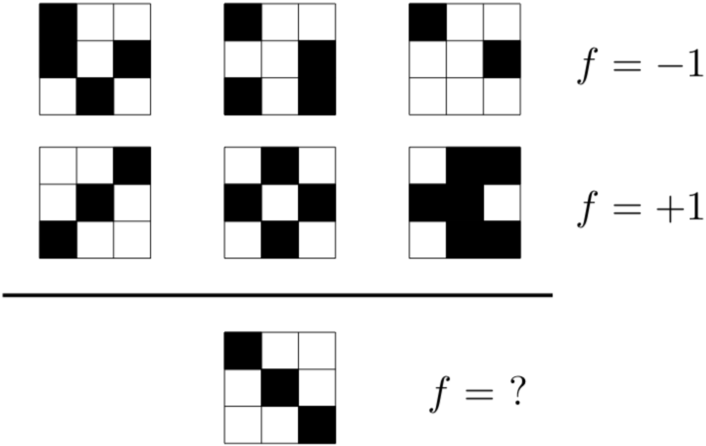
Marco Piangerelli
marco.piangerelli@unicam.it

13 December 2022 - 09 January 2023

What are we learning?



What are we learning?



Learning VS Machine Learning

Learning

“ Learning is about acquiring skills $\rightarrow$ using experience from a set of observations”

Learning VS Machine Learning

Learning

" Learning is about acquiring skills → using experience from a set of observations"

Machine Learning

" Machine Learning is about acquiring skills → using experience derived from data "

Learning is about " acquiring **skills**"

What do mean with "skill"?

- predict energy consumption
- recognizing objects
- ...
- uncovering an hidden process
- improving a performance measure (e.g accuracy, recall, f1-score ...)

Learning VS Machine Learning

Definition [Mitchell (1997)]

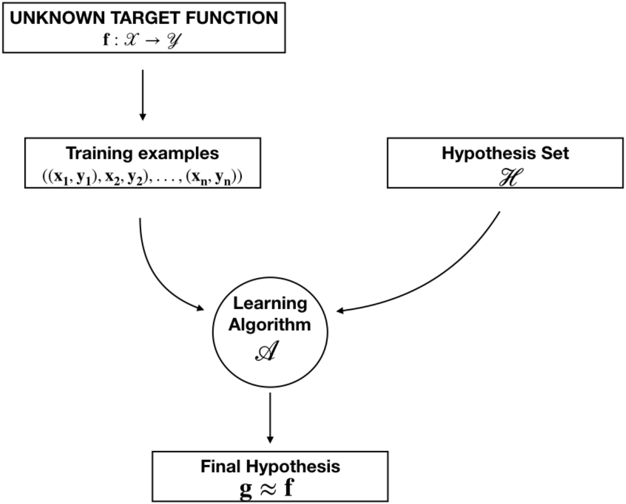
" A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

Notation

$\mathbf{x}$ the input $\mathbf{x} \in \mathcal{X}$. Often a column vector $\mathbf{x} \in \mathbb{R}^d$ or $\mathbf{x} \in \{1\} \times \mathbb{R}^d$. x is used if input is scalar. $\mathbf{y}$ the output $\mathbf{y} \in \mathcal{Y}$.
$\mathcal{X}$ input space whose elements are $\mathbf{x} \in \mathcal{X}$, $\mathcal{Y}$ output space whose elements are $\mathbf{y} \in \mathcal{Y}$
Data, $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)...(\mathbf{x}_n, y_n)\}$
Unknown function to be learned $f : \mathcal{X} \rightarrow \mathcal{Y}$
Approximation of the **Unknown** function $g : \mathcal{X} \rightarrow \mathcal{Y}$
$\mathcal{A}$ learning algorithm, $\mathcal{H}$ set of candidates formulas for g

Notation

$\mathbf{x}$ the input $\mathbf{x} \in \mathcal{X}$. Often a column vector $\mathbf{x} \in \mathbb{R}^d$ or $\mathbf{x} \in \{1\} \times \mathbb{R}^d$. x is used if input is scalar. $\mathbf{y}$ the output $\mathbf{y} \in \mathcal{Y}$.
$\mathcal{X}$ input space whose elements are $\mathbf{x} \in \mathcal{X}$, $\mathcal{Y}$ output space whose elements are $\mathbf{y} \in \mathcal{Y}$
Data, $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)...(\mathbf{x}_n, y_n)\}$
Unknown function to be learned $f : \mathcal{X} \rightarrow \mathcal{Y}$
Approximation of the **Unknown** function $g : \mathcal{X} \rightarrow \mathcal{Y}$
$\mathcal{A}$ learning algorithm, $\mathcal{H}$ set of candidates formulas for g


UNKNOWN TARGET FUNCTION
$f : \mathscr{X} \to \mathscr{Y}$
Training examples
$((x_1, y_1), x_2, y_2), \ldots, (x_n, y_n))$
Hypothesis Set
$\mathscr{H}$
Learning Algorithm
$\mathscr{A}$
Final Hypothesis
$g \approx f$

A daily example...

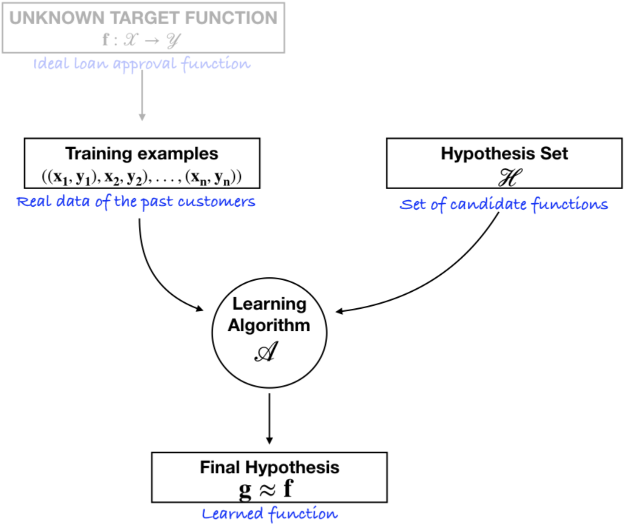
Let suppose we need a bank loan. We go to the bank explaining why we need money and then we ask a certain amount.

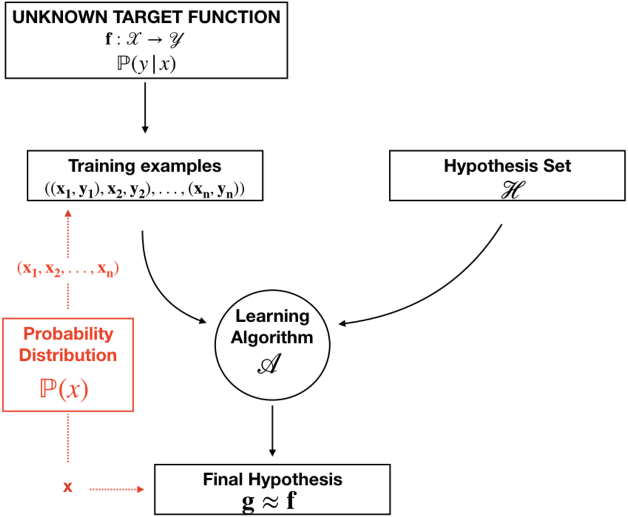
Do we get those money?

A daily example...

Now, let suppose that a lot of people need a bank loan and the bank want to set up an automatic procedure for approving or rejecting the applications

What does the bank do?(hint: Remember that the bank has a lot of data)


UNKNOWN TARGET FUNCTION
$f : \mathcal{X} \to \mathcal{Y}$
Ideal loan approval function
Training examples
$((x_1, y_1), x_2, y_2), \ldots, (x_n, y_n))$
Real data of the past customers
Hypothesis Set
$\mathcal{H}$
Set of candidate functions
Learning Algorithm
$\mathcal{A}$
Final Hypothesis
$g \approx f$
Learned function


UNKNOWN TARGET FUNCTION
$\mathbf{f} : \mathcal{X} \to \mathcal{Y}$
$\mathbb{P}(y \mid x)$
Training examples
$((\mathbf{x}_1, \mathbf{y}_1), \mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$
Hypothesis Set
$\mathcal{H}$
$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
Probability Distribution
$\mathbb{P}(x)$
Learning Algorithm
$\mathcal{A}$
x
Final Hypothesis
$\mathbf{g} \approx \mathbf{f}$

A “simple” model

$\mathcal{X}$ is the set of data, $\mathbf{x}$, namely the information about the clients that requested a bank loan
$\mathcal{Y}$ is the binary set $\{-1, 1\}$ (yes or no)

A “simple” model

$\mathcal{X}$ is the set of data, $\mathbf{x}$, namely the information about the clients that requested a bank loan
$\mathcal{Y}$ is the binary set $\{-1, 1\}$ (yes or no)
A simple model could be a “thresholded” model:

- $\sum_{i=1}^{k} w_i x_i > threshold \rightarrow +1 \rightarrow$ YES
- $\sum_{i=1}^{k} w_i x_i < threshold \rightarrow$ -1 $\rightarrow$ NO

A "simple" model

$\mathcal{X}$ is the set of data, $\mathbf{x}$, namely the information about the clients that requested a bank loan

$\mathcal{Y}$ is the binary set $\{-1, 1\}$ (yes or no)

A simple model could be a "thresholded" model:

- $\sum_{i=1}^{k} w_i x_i > threshold \rightarrow +1 \rightarrow$ YES
- $\sum_{i=1}^{k} w_i x_i < threshold \rightarrow$ -1 $\rightarrow$ NO

In a more compact way we can write:

- $h(\mathbf{x}) = sign((\sum_{i=1}^{k} w_i x_i) + threshold)$

A “simple” model

$$h(\mathbf{x}) = sign((\sum_{i=1}^{k} w_i x_i) + threshold)$$

A “simple” model

$$h(\mathbf{x}) = sign((\sum_{i=1}^{k} w_i x_i) + threshold)$$
$$h(\mathbf{x}) = \gamma(\mathbf{w}^T \mathbf{x})$$
$$h(\mathbf{x}) = \gamma(\mathbf{w}^T \phi(\mathbf{x}))$$

$$\gamma(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$

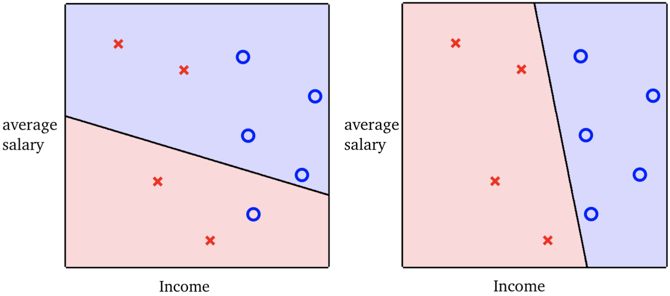
A "simple" model

$$h(\mathbf{x}) = sign((\sum_{i=1}^{k} w_i x_i) + threshold)$$
$$h(\mathbf{x}) = \gamma(\mathbf{w}^T \mathbf{x})$$
$$h(\mathbf{x}) = \gamma(\mathbf{w}^T \phi(\mathbf{x}))$$

$$\gamma(a) = \begin{cases} +1 & a \geq 0 \\ -1 & a < 0 \end{cases}$$

$$\gamma(\mathbf{w}^T \phi(\mathbf{x})) = y(\mathbf{x}) = \{-1, +1\}$$

The Perceptron (Rosenblatt 1958)

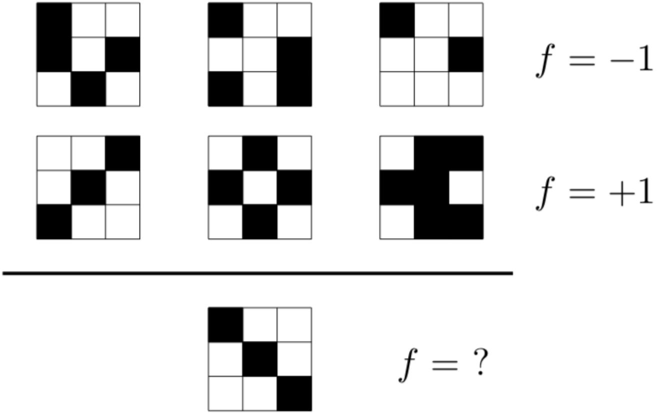


The role of f and g

In ML we are interested in learning f but ....

The role of f and g

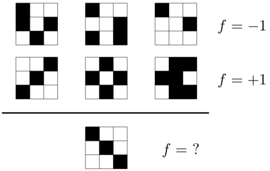
In ML we are interested in learning f but ....f is unknown



We know the value of f for each sample but how can we generalize and say that f is able to predict something that it has never seen before?

We know the value of f for each sample but how can we generalize and say that f is able to predict something that it has never seen before? Can $\mathcal{D}$ tell us anything outside of $\mathcal{D}$?

Let's see an example....



- An easy visual learning problem just got very messy.

  For *every* f that fits the data and is "+1" on the new point, there is one that is "−1".

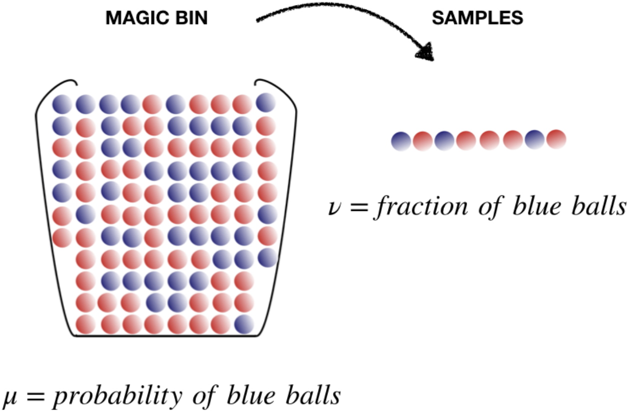
  Since f is *unknown*, it can take on any value outside the data, no matter how large the data.

- This is called **No Free Lunch (NFL)**.

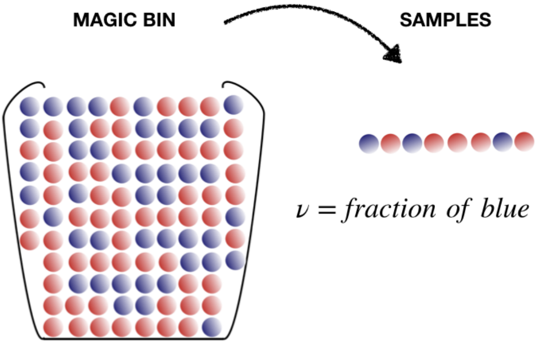
  You cannot know anything *for sure* about f outside the data without making assumptions.

- **What now!**

  Is there *any hope* to know *anything* about f outside the data set *without* making assumptions about f?

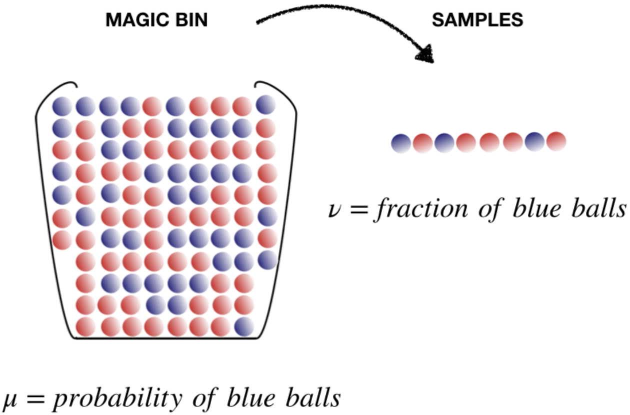


The marbles are indefinitely many and μ is **Unknown**.

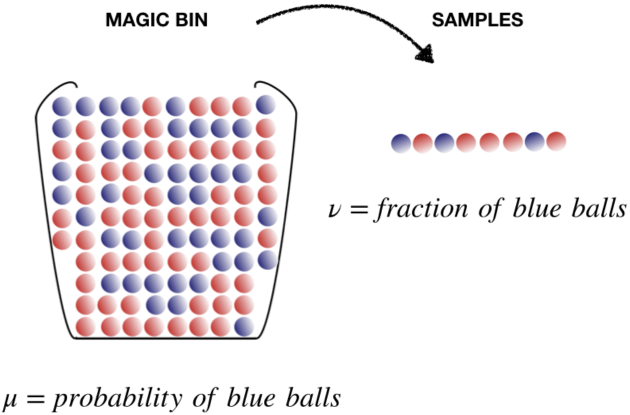


$\nu = \textit{fraction of blue balls}$

$\mu = \textit{probability of blue balls}$



We pick N marbles. one marble at time, independently from the previous one and check the color of the marble.



We pick N marbles. one marble at time, independently from the previous one and check the color of the marble. Can we use ν for saying something about μ?

The Law of large numbers

If $x1, x2, \ldots, x_m$are m i.i.d. samples of a random variable $\mathbb{X}$ distributed over $\mathbb{P}$, then for a small positive non-zero value ϵ:

$$\lim_{m\to\infty} \mathbb{P}\left[\left|\mathbb{E}[X]_{X\sim P} - \frac{1}{m}\sum_{i=1}^{m} x_i\right| > \epsilon\right] = 0$$

Hoeffding's Inequality

$\mathbb{P}[\cdot] \leq x$, for some conditions

$\mathbb{P}[\bar{\cdot}] \geq 1 - x$, for some conditions

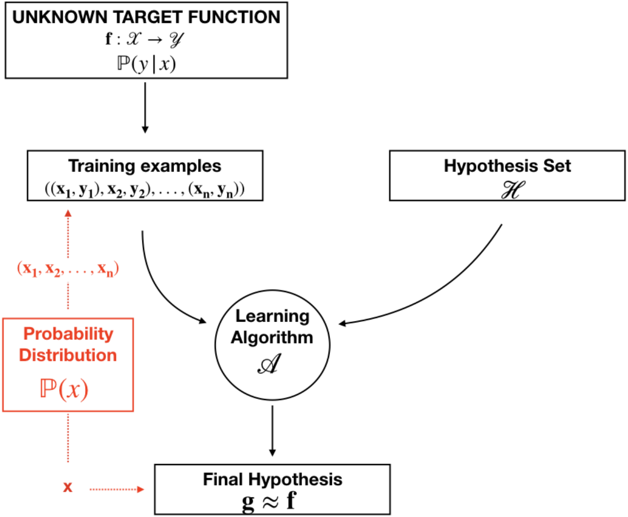
Hoeffding's Inequality

$\mathbb{P}[\cdot] \leq x$, for some conditions

$\mathbb{P}[\bar{\cdot}] \geq 1 - x$, for some conditions

$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$, for any $\epsilon \geq 0$

$\mathbb{P}[|\nu - \mu| \leq \epsilon] \geq 1 - 2e^{-2\epsilon^2 N}$, for any $\epsilon \geq 0$


UNKNOWN TARGET FUNCTION
$\mathbf{f}: \mathscr{X} \to \mathscr{Y}$
$\mathbb{P}(y \mid x)$
Training examples
$((\mathbf{x}_1, \mathbf{y}_1), \mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$
Hypothesis Set
$\mathscr{H}$
$(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
Probability Distribution
$\mathbb{P}(x)$
Learning Algorithm
$\mathscr{A}$
x
Final Hypothesis
$\mathbf{g} \approx \mathbf{f}$

Choose an Hypthesis $h \in \mathcal{H}$ and and compare it to f in each point $x \in \mathcal{X}$ and if $h(x) = f(x)$ color marble blue otherwise it is red; but since f is unknown the color is unknown too; but...

Choose an Hypthesis $h \in \mathcal{H}$ and and compare it to f in each point $x \in \mathcal{X}$ and if $h(x) = f(x)$ color marble blue otherwise it is red; but since f is unknown the color is unknown too; but...

The training samples play the role of the samples form the bin.

$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \cdots, \mathbf{x_N}$ are picked *independently* according to $\mathbf{P}$ we will get a random sample of blue marbles (μ) and a random sample of red ones $(1 - \mu)$.

Choose an Hypthesis $h \in \mathcal{H}$ and and compare it to f in each point $x \in \mathcal{X}$ and if $h(x) = f(x)$ color marble blue otherwise it is red; but since f is unknown the color is unknown too; but...
The training samples play the role of the samples form the bin.
$\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \cdots, \mathbf{x_N}$ are picked *independently* according to **P** we will get a random sample of blue marbles (μ) and a random sample of red ones ($1 - \mu$). Now we see the color....so we know $f(\mathbf{x}_n)$ and we can compare it with our h. In this case ν depends on h....(Why??)

The role of h - Verification

How can we compare the two situations?

- take any single hypothesis h$\in \mathcal{H}$
- compare it to f on each point $\mathbf{x} \in \mathcal{X}$
- if h($\mathbf{x}$) $=$ f($\mathbf{x}$) $\rightarrow$ color $\mathbf{x}$ red, otherwise color $\mathbf{x}$ blue
- since f is unknown we do not know which color $\mathbf{x}$ has
- we pick $\mathbf{x}$ at random accordingly to some probaility distribution P $\rightarrow$ $\mathbf{x}$ will be blue with some probability ,μ, and red with $1 - \mu$
- the training examples play the role of the sample from the bin $\rightarrow$ we know μ and ν
- ν is based on the particular hypothesis h

In learning we need many hypothesis to choose from....in this case we are just verifying, non learning....

Introducing the Error (Risk)

- In-sample Error

$$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} l(h(x_i), f(x_i))$$

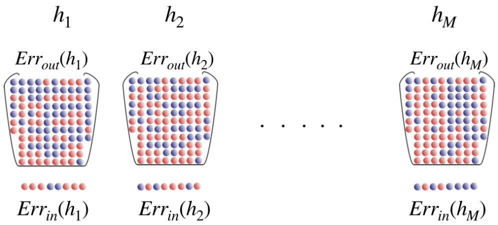
- Out-of-sample Error

$$E_{out}(h) = \mathbb{E}_X[l(h(x), f(x))]$$

The role of h

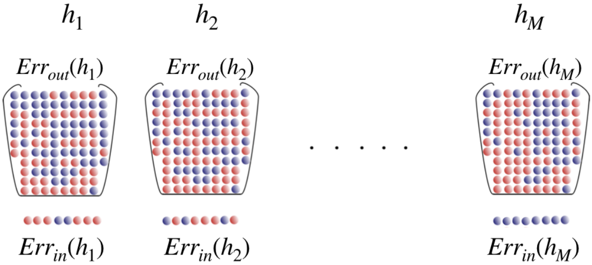
Hoeffding's Inequality revised

$$\mathbb{P}[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2e^{-2\epsilon^2 N}, \text{ for any } \epsilon \geq 0$$



Almost done....

Now we have a problem $\rightarrow$ The Hoeffding's Inequality DOES NOT apply to multiple bins



Pick the hypothesis with minimum E_{in}; will E_{out} be small?

Basic probability notions

Implications

If $A \Rightarrow B$ $(A \subseteq B)$ *then* $\mathbb{P}[A] \leq \mathbb{P}[B]$

Union Bound

If $A \Rightarrow B$ $(A \subseteq B)$ *then* $\mathbb{P}[A \text{ or } B] = \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B]$

In general $\qquad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$

Almost done....

$$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq \mathbb{P}[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \textit{ or}$$

$$|E_{in}(h_1) - E_{out}(h_2)| > \epsilon \textit{ or}$$

$$\textit{or}....$$

$$|E_{in}(h_M) - E_{out}(h_M)| > \epsilon]$$

$$\leq \sum_{m=1}^{M} 2e^{-2\epsilon^2 N}, \text{ for any } \epsilon \geq 0$$

Almost done....

$\mathbb{P}[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$, for any $\epsilon \geq 0$

M can be see as the “complexity” of the model

Is learning feasible?

- No, in a deterministic perspective
- Yes, in probabilistic perspective
  - only assumption we make is : the samples in $\mathcal{D}$ are to be generate independently
  - if $g \approx f \;\Rightarrow\; E_{out}(g) = 0$, but f in unknown
    The only information we get from the probabilistic analysis, i.e. Hoeffding Inequality, is $E_{in}(g) \approx Err_{out}(g)$
  - we control $E_{in}(g)$

Is learning feasible?

Finally, the answer to the question is....

Is learning feasible?

Finally, the answer to the question is....
YES., in PROBABILISTIC WAY

Is learning feasible?

Finally, the answer to the question is....
YES., in PROBABILISTIC WAY
but, HOW?

Is learning feasible?

Finally, the answer to the question is....
YES., in PROBABILISTIC WAY
but, HOW? $\rightarrow E_{out}(g) \approx 0$

Is learning feasible?

Finally, the answer to the question is....
YES., in PROBABILISTIC WAY
but, HOW? $\rightarrow E_{out}(g) \approx 0$

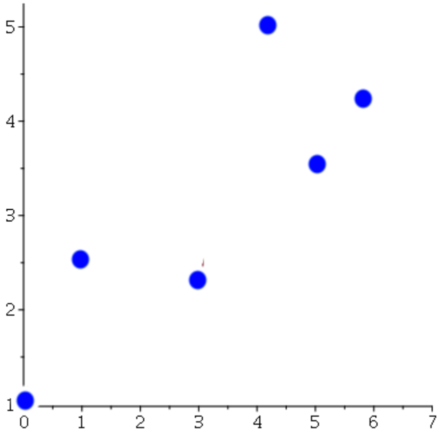
1. make sure that $E_{in}(g) \approx E_{out}(g)$
2. $Err_{in}(g) \approx 0$

Is learning feasible?

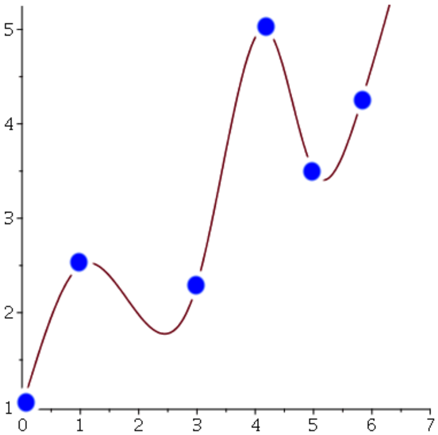
Finally, the answer to the question is....
YES., in PROBABILISTIC WAY
but, HOW? $\rightarrow E_{out}(g) \approx 0$

1. make sure that $E_{in}(g) \approx E_{out}(g) \rightarrow$ Hoeffdind's Inequality
2. $Err_{in}(g) \approx 0$

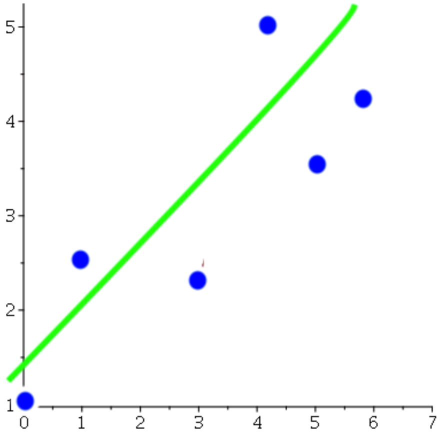
Learning is not memorizing



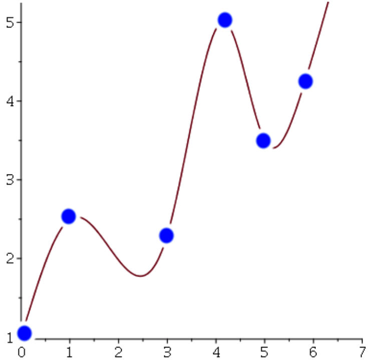
Learning is not memorizing (er the effect of M)

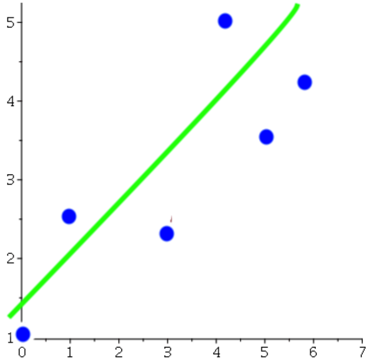


Learning is not memorizing



Learning is not memorizing





Memorizing VS Learning

Generalization Bound

$|E_{in}(g) - E_{out}(g)| = Generalization\ Error < \epsilon$

Generalization Bound

$|E_{in}(g) - E_{out}(g)| = Generalization\ Error < \epsilon$

Theorem

With probability at least $1 - \delta$

$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}} \leftarrow Generalization\ Error$

This Inequality is known as the *Generalization Bound*

Generalization Bound

$|E_{in}(g) - E_{out}(g)| = Generalization\ Error < \epsilon$

Theorem

With probability at least $1 - \delta$

$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}} \leftarrow Generalization\ Error$

This Inequality is known as the *Generalization Bound*

Proof

Let $M = |\mathcal{H}|$

Let $\delta = 2|\mathcal{H}|e^{-2\epsilon^2 N}$.

Then, $\mathbb{P}[|E_{in}(g) - E_{out}(g)| \leq \epsilon] \geq 1 - \delta$

In words, with probability at least $1 - \delta$, $|E_{in}(g) - E_{out}(g)| < \epsilon$.

Hence $E_{out}(g) \leq E_{in}(g) + \epsilon$

From the definition of δ, solving for ϵ :

$\epsilon = \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}}$

Generalization Bound

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

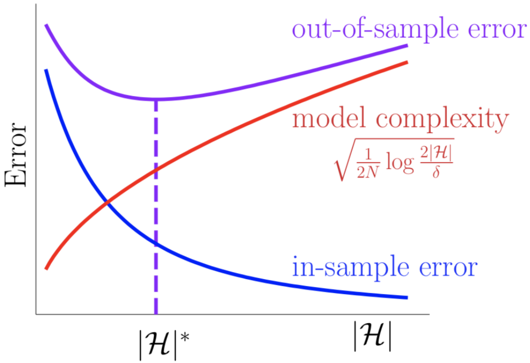
- $E_{out}(g) \leq E_{in}(g) + \epsilon$
- $E_{out}(g) \geq E_{in}(g) - \epsilon$

Generalization Bound

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

- $E_{out}(g) \leq E_{in}(g) + \epsilon \Rightarrow$ the hypothesis g continues to perform well out of samples
- $E_{out}(g) \geq E_{in}(g) - \epsilon \Rightarrow$ there is no other hypothesis $h \in \mathcal{H}$ whose $Err_{out}(h)$ is not significantly better than $Err_{out}(g)$

Almost done....


out-of-sample error
model complexity
$\sqrt{\frac{1}{2N}\log\frac{2|\mathcal{H}|}{\delta}}$
in-sample error
Error
$|\mathcal{H}|^*$
$|\mathcal{H}|$

The dependance on $\mathcal{H}$

With probability at least $1-\delta$
$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}}$

1
2

The dependance on $\mathcal{H}$

With probability at least $1 - \delta$

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

1. $N \gg ln|\mathcal{H}|, then\ E_{out}(g) \approx E_{in}(g)$

The dependance on $\mathcal{H}$

With probability at least $1 - \delta$

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

1. $N \gg ln|\mathcal{H}|, then\ E_{out}(g) \approx E_{in}(g)$
2. $|\mathcal{H}| \to +\infty$, then $E_{out}(g) \leq +\infty$

The dependance on $\mathcal{H}$

The second condition does not make sense and unfortunately almost all learning models have infinite $M = \mathcal{H}$

We need to replace M with "something" that is finite, M goes to $+\infty$

Infinite number of $\mathcal{H}$

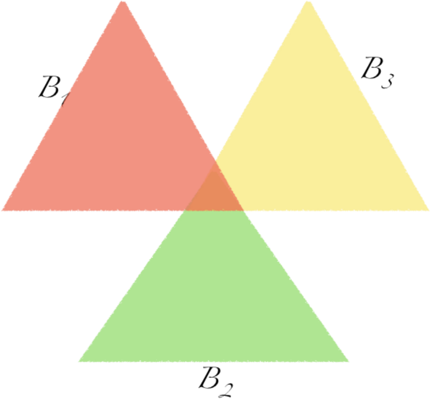
$$|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ or}$$
$$|E_{in}(h_1) - E_{out}(h_2)| > \epsilon \text{ or}$$
$$or....$$
$$|E_{in}(h_M) - E_{out}(h_M)| > \epsilon]$$

Infinite number of $\mathcal{H}$

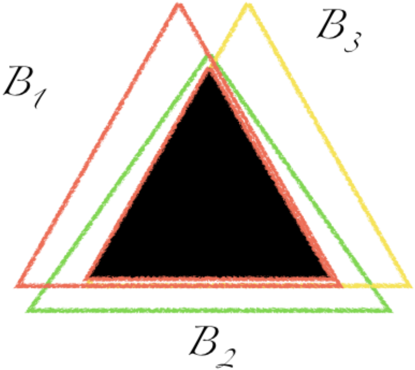
$$|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ or}$$
$$|E_{in}(h_1) - E_{out}(h_2)| > \epsilon \text{ or}$$
$$or....$$
$$|E_{in}(h_M) - E_{out}(h_M)| > \epsilon]$$

USING THE UNION BOUND WE ARE OVER-ESTIMATING THE PROBABILITY OF THE EVENT $|E_{in}(g) - E_{out}(g)| > \epsilon$

Infinite number of $\mathcal{H}$



Infinite number of $\mathcal{H}$



Infinite number of $\mathcal{H}$

The Union Bound states that the total area covered by $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ is smaller than the sum of the individual areas

It is true $\rightarrow$ but is a strong assumption when the areas overlap heavily

Infinite number of $\mathcal{H}$

The Union Bound states that the total area covered by $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ is smaller than the sum of the individual areas

It is true $\rightarrow$ but is a strong assumption when the areas overlap heavily
Overlapping events $\rightarrow \mathcal{B}_1 \sim \mathcal{B}_2 \sim \mathcal{B}_3$

Infinite number of $\mathcal{H}$

The Union Bound states that the total area covered by $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ is smaller than the sum of the individual areas

It is true $\rightarrow$ but is a strong assumption when the areas overlap heavily
Overlapping events $\rightarrow \mathcal{B}_1 \sim \mathcal{B}_2 \sim \mathcal{B}_3$

Overlapping events
$\rightarrow |Err_{in}(h_1) - Err_{out}(h_1)| > \epsilon$ *coincides to* $|Err_{in}(h_2) - Err_{out}(h_3)| > \epsilon$ *coincides to* $|Err_{in}(h_3) - Err_{out}(h_3)| > \epsilon$

$$\rightarrow h_1 \sim h_2 \sim h_3$$

From $|\mathcal{H}|$ to $m_{|\mathcal{H}|}(N)$

Hoeffding's Inequality revised

$$\mathbb{P}[|Err_{in}(h) - Err_{out}(h)| > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N}, \text{ for any } \epsilon \geq 0$$

The Hoeffding's Inequality DOES NOT apply to multiple bins

for $|\mathcal{H}| \rightarrow \infty$ the generalization bound $Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}}$ does not make any sense

From $|\mathcal{H}|$ to $m_{|\mathcal{H}|}(N)$

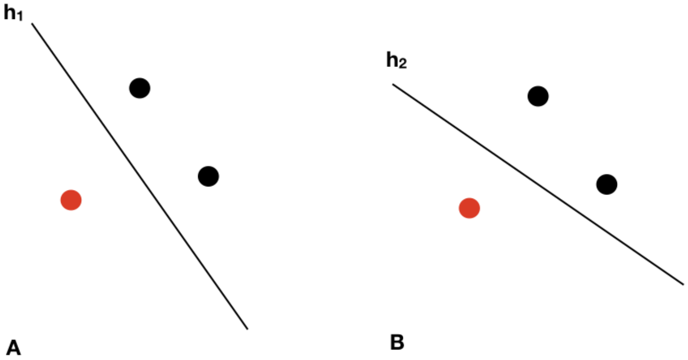
We NEED to substitute $|\mathcal{H}|$ with another quantity that does not go to ∞

From $|\mathcal{H}|$ to $m_{|\mathcal{H}|}(N)$

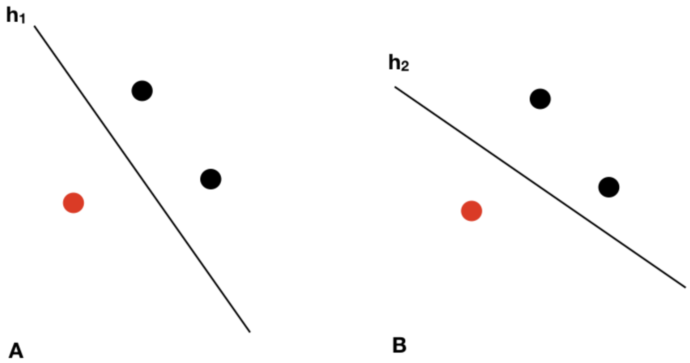
We NEED to substitute $|\mathcal{H}|$ with another quantity that does not go to ∞
We call this quantity "The growth function" $\rightarrow$ It is a combinatorial

quantity that captures HOW different the hypothesis are and HOW much they overlap.

Dichotomies



Dichotomies



Between h_1 and h_2 we can found "infinite" straight -lines (hypothesis) that can split the plane into 2 sub- planes

Dichotomies

- A hypothesis $h : \mathcal{X} \to -1, +1$
- a dichotomy $h : x_1, x_2, ..., x_N \to -1, +1$, a Dichotomy is an Hypothesis that is defined only on finite subset of the input space
- number of hypothesis $|\mathcal{H}|$ can be infinite
- number of dichotomies $|\mathcal{H}(x_1, x_2, ..., x_N)|$

Dichotomies

For defining the growth function we take into consideration a problem of Binary Classification

$$h \in \mathcal{H}, h : (\mathbf{x}_1 ... \mathbf{x}_N) \rightarrow \{-1, +1\}$$

The hypothesis h splits the samples into two groups : those who are classified as -1 and those who are classified as $+1$

Dichotomies

For defining the growth function we take into consideration a problem of Binary Classification

$$h \in \mathcal{H}, h : (\mathbf{x}_1 ... \mathbf{x}_N) \rightarrow \{-1, +1\}$$

The hypothesis h splits the samples into two groups : those who are classified as -1 and those who are classified as $+1$

That is called a *dichotomy*

Dichotomies

Definition

Let $\mathbf{x}_1 ... \mathbf{x}_N \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by

$$\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N) = \{(h(\mathbf{x}_1), ..., h(\mathbf{x}_N)|h \in \mathcal{H}\}$$

One can think about $\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)$ as an $\mathcal{H}$ based only on that training set.
A larger $\mathcal{H}(\mathbf{x}_1, ..., \mathbf{x}_N)$ means $\mathcal{H}$ is more "diverse", i.e. it generates more dichotomies on $\mathbf{x}_1, ..., \mathbf{x}_N)$.
How many dichotomies? at most 2^N
Why?

Growth Function

Definition

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

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\ldots,\mathbf{x}_N \in \mathcal{H}} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)|$$

Where $|\cdot|$ denotes the cardinality of the set.

In words it means that $m_{\mathcal{H}}(N)$ is the maximum number of dichotomies that can be generated by $\mathcal{H}$ on any N points.

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

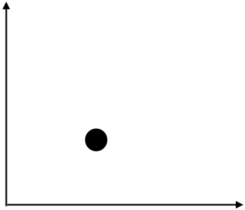
Dichotomies

To compute $m_{\mathcal{H}}(N)$, we need to:

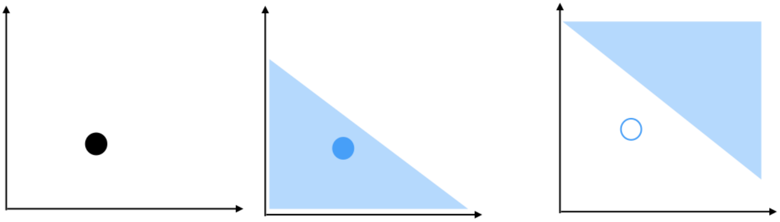
- consider the number of possible choices of N points from $\mathcal{X}$
- pick the one that gives us the most dichotomies

If $\mathcal{H}$ is capable to generate all the possible dichotomies for that number of points we say that $\mathcal{H}$ can *shatter* $\mathbf{x}_1, ..., \mathbf{x}_N$

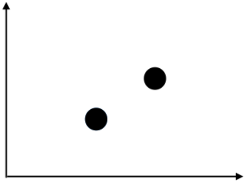
Dichotomies (N = 1)



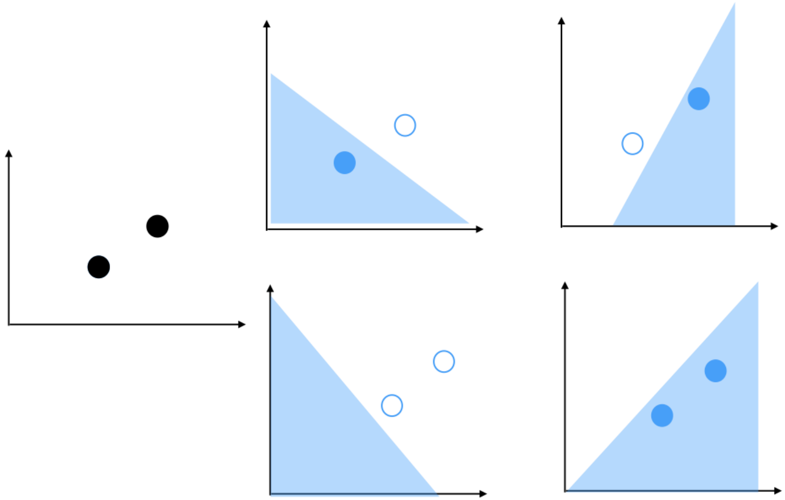
Dichotomies (N = 1) $\rightarrow m_{\mathcal{H}}(1) = 2$



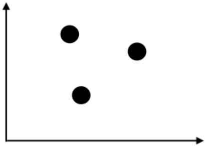
Dichotomies (N = 2)

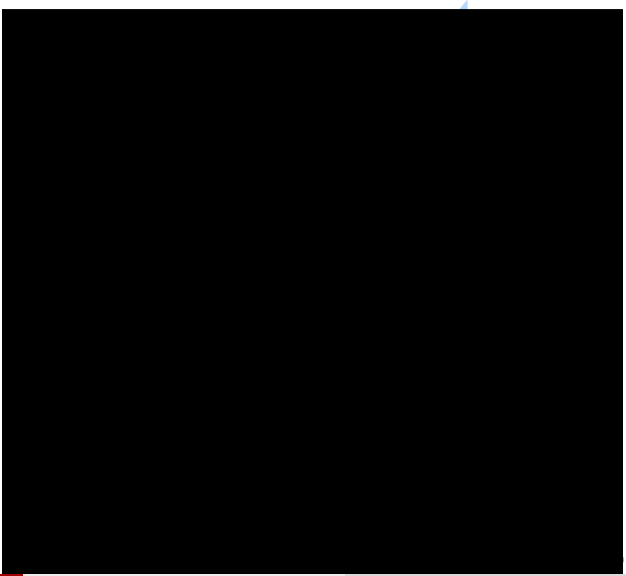


Dichotomies (N = 2) $\rightarrow m_{\mathcal{H}}(2) = 4$

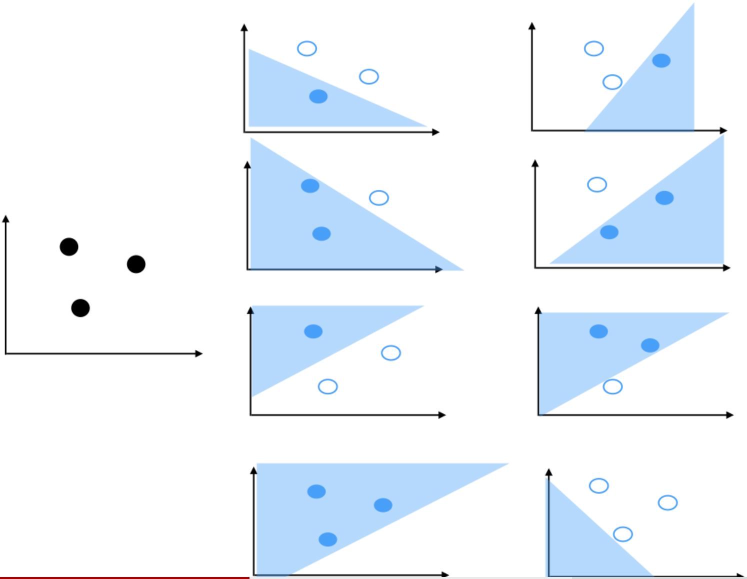


Dichotomies (N = 3)





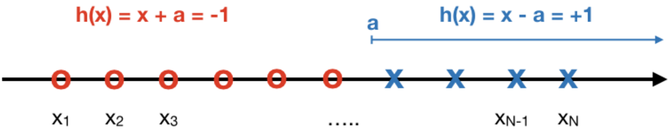
Dichotomies (N = 3) $\rightarrow m_{\mathcal{H}}(3) = 8$



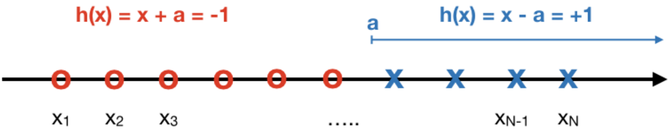
Dichotomies (N = 4)

Dichotomies (N = 4) $\rightarrow m_{\mathcal{H}}(3) = 14$

Example 1: Positive Rays

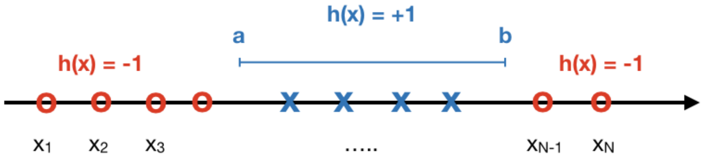


Example 1: Positive Rays

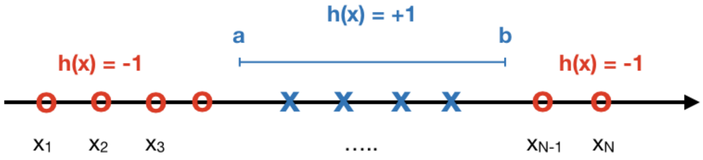


$$m_{\mathcal{H}}(N) = N + 1$$

Example 2: Intervals

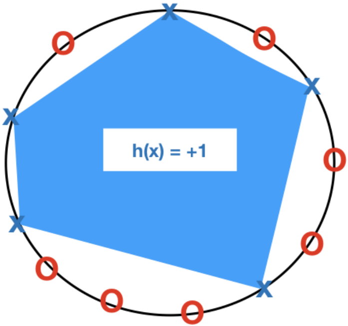


Example 2: Intervals



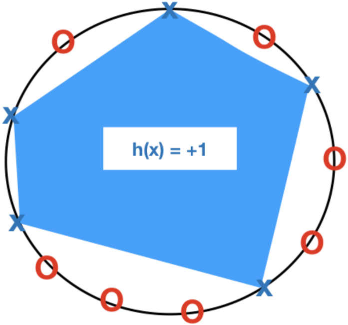
$$m_{\mathcal{H}}(N) = \binom{N+1}{2} + 1 = \frac{(N+1)!}{(N+1-2)!2!} + 1 = \frac{1}{2}N^2 + \frac{1}{2}N + 1$$

Example 3: Convex sets



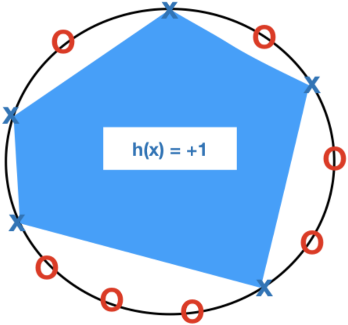
A convex set is a region where for any two points picked within a region, the entirety of the line segment connecting them lies within the region.

Example 3: Convex sets



A convex set is a region where for any two points picked within a region, the entirety of the line segment connecting them lies within the region.

Example 3: Convex sets



A convex set is a region where for any two points picked within a region, the entirety of the line segment connecting them lies within the region.

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

Dichotomies sets

- Positive Rays $m_{\mathcal{H}} = N + 1$
- Positive Intervals $m_{\mathcal{H}} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$
- Convex sets $m_{\mathcal{H}} = 2^N$

Dichotomies sets

- Positive Rays $m_{\mathcal{H}} = N + 1$
- Positive Intervals $m_{\mathcal{H}} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$
- Convex sets $m_{\mathcal{H}} = 2^N$

The number of dichotomies increase if the complexity of the model increse

Dichotomies sets

- Positive Rays $m_{\mathcal{H}} = N + 1$
- Positive Intervals $m_{\mathcal{H}} = \frac{1}{2}N^2 + \frac{1}{2}N + 1$
- Convex sets $m_{\mathcal{H}} = 2^N$

The number of dichotomies increase if the complexity of the model increse
The fact that the more complex h is, the bigger is the number of dichotomies is good

Can $m_{\mathcal{H}}(N)$ help us?

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

The 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 $\mathcal{H}$

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\dots,\mathbf{x}_N \in \mathcal{H}} |\mathcal{H}(\mathbf{x}_1, \dots, \mathbf{x}_N)|$$

By extension, this means that a bigger data set cannot be shattered either. In other words, given a hypothesis set, a break point is the point at which we fail to achieve all possible dichotomies.

The break point is important for computing a bound of the growth function. The most important fact about the growth function is that if the condition $m_{\mathcal{H}}(N) = 2^N$ breaks for any point, we can bound $m_{\mathcal{H}}(N)$ for all values of N by a simple polynomial based on the break point. For the bound, being Polynomial is crucial.

The 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 $\mathcal{H}$

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1,\ldots,\mathbf{x}_N \in \mathcal{H}} |\mathcal{H}(\mathbf{x}_1,\ldots,\mathbf{x}_N)|$$

By extension, this means that a bigger data set cannot be shattered either. In other words, given a hypothesis set, a break point is the point at which we fail to achieve all possible dichotomies.

The break point is important for computing a bound of the growth function. The most important fact about the growth function is that if the condition $m_{\mathcal{H}}(N) = 2^N$ breaks for any point, we can bound $m_{\mathcal{H}}(N)$ for all values of N by a simple polynomial based on the break point. For the bound, being Polynomial is crucial.

The break point- Example

- Positive Rays $m_{\mathcal{H}} = N + 1,\ k = 2$
- Positive Intervals $m_{\mathcal{H}} = \frac{1}{2}N^2 + \frac{1}{2}N + 1,\ k = 2$
- Convex sets $m_{\mathcal{H}} = 2^N,\ k = \infty$

Review

- Hoeffding's Inequality $\mathbb{P}\left[|E_{in}(g) - E_{out}(g)| > \epsilon\right] \leq 2Me^{-2\epsilon^2 N}$
- The Growth Function for a hypothesis set $\mathcal{H}$ is the maximum number of dichotomies (patterns) we can get on N data points.
  - $m_{\mathcal{H}}(N) = N + 1$ positive rays
  - $m_{\mathcal{H}}(N) = \frac{1}{2}N^2 + \frac{1}{2}N + 1$ positive interval
  - $m_{\mathcal{H}}(N) = 2^N$ convex sets
- The break point for a hypothesis set $\mathcal{H}$ is the value of N for which we fail to get all possible dichotomies

Bounding $m_{\mathcal{H}}(N)$

- Define a combinatorial quantity $B(N, k)$

$B(N, k)$

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

- Assuming that k is a break point for $\mathcal{H}$, $m_{\mathcal{H}}(N) \leq B(N, k)$

Bounding $m_{\mathcal{H}}(N)$

Sauer's Lemma

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

Proof ....

- The growth function $m_{\mathcal{H}}(N)$ is either 2^N or polynomial, nothing different
- For a given hyphotesis set $\mathcal{H}$, the break point k is fixed, and does not grow with N

Theorem

Theorem

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

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

for all N. The right hand side is polynomial in N of degree $k-1$

The Vapnik - Chervonenkis Dimension

The Vapnik - Chervonenkis Dimension

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$ or simply d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{VC} = \infty$

In simple words d_{VC} is the most points $\mathcal{H}$ can shatter.

The Vapnik - Chervonenkis Dimension

The Vapnik - Chervonenkis Dimension

The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{VC}(\mathcal{H})$ or simply d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{VC} = \infty$

In simple words d_{VC} is the most points $\mathcal{H}$ can shatter.

If d_{VC} is the VC dimension of $\mathcal{H}$, then $k = d_{VC} + 1$ is a break point for $m_{\mathcal{H}}(N)$ since $m_{\mathcal{H}}(N)$ can not be equal to 2^N for any $N > d_{VC}$ by definition. It is easy to see that no smaller break point exists since $\mathcal{H}$ can shatter d_{VC} points, hence it can also shatter any subset of these points.

d_{VC} + bounding the growth function

Since $k = d_{VC} + 1$ we can write

Theorem

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

for all N. The right hand side is polynomial in N of degree d_{VC} By induction it is possible to prove that :

$$m_{\mathcal{H}}(N) \leq N^{d_{VC}} + 1$$

From $|\mathcal{H}|$ to $m_{\mathcal{H}}(N)$

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{1}{2N} ln \frac{2|\mathcal{H}|}{\delta}}$$

$$\downarrow$$

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

VC generalization bound

Theorem

For any tolerance $\delta > 0$

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

with probability $\geq 1 - \delta$

The *VC* generalization bound holds for any binary target function f, any hypothesis set $\mathcal{H}$, any learning algorithm $\mathcal{A}$ and any input probability distribution P.

The VC generalization bound is the most important mathematical result in the theory of learning. It establishes the feasibility of learning with infinite hypothesis sets.

Putting it together

- For a hypothesis set $\mathcal{H}$, the existence of a finite d_{VC} means that the learning is feasible (i.e. generalization is possible)
  Finite d_{VC} means the existence of a polynomial bound for the growth function
- The value of d_{VC} tells us the resources needed to achieve e desired performance
- The larger d_{VC}, the more complex the hypothesis set $\mathcal{H}$
- Infinite d_{VC} means no break point for $\mathcal{H}$ because it shatters every set op points $\rightarrow$ good for fitting, bad for generalization

Interpreting the VC dimension

- What does the d_{VC} mean ?
- How to use d_{VC} in practice ?

Interpreting the VC dimension

- What does the d_{VC} mean ? $\rightarrow$ degrees of freedom
- How to use d_{VC} in practice ?

Interpreting the VC dimension

- What does the d_{VC} mean ? $\rightarrow$ degrees of freedom
- How to use d_{VC} in practice ?$\rightarrow$ number of data points needed

Interpreting the VC dimension

- The VC dimension is a measure of the “effective” number of parameters, or “ degrees of freedom” that enable the model to express a diverse set of hypothesis

Interpreting the VC dimension - Sample Complexity

How many training examples N are needed?

- the error tolerance ϵ indicates the allowed generalization error
- the confidence parameter δ indicates how often ϵ is violated
- how. much N grows w.r.t. the decreasing of ϵ and δ tells us how many data are needed for a good generalization

Fixed $\delta > 0$, we want the generalization error to be at most ϵ

$$\sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon$$

$$\downarrow$$

Interpreting the VC dimension - Sample Complexity

How many training examples N are needed?

- the error tolerance ϵ indicates the allowed generalization error
- the confidence parameter δ indicates how often ϵ is violated
- how. much N grows w.r.t. the decreasing of ϵ and δ tells us how many data are needed for a good generalization

Fixed $\delta > 0$, we want the generalization error to be at most ϵ

$$\sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon$$

$$\downarrow$$

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4m_{\mathcal{H}}(2N)}{\delta})$$

for having a generalization error at most of ϵ with $\mathbb{P}$ at least of $1 - \delta$

Interpreting the VC dimension - Sample Complexity

If we replace $m_{\mathcal{H}}(2N)$ with its polynomial upper bound, based on the d_{VC}
Fixed $\delta > 0$,

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{VC}}+1)}{\delta})$$

for having a generalization error at most of ϵ with $\mathbb{P}$ at least of $1 - \delta$

Interpreting the VC dimension - Sample Complexity

If we replace $m_{\mathcal{H}}(2N)$ with its polynomial upper bound, based on the d_{VC}
Fixed $\delta > 0$,

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{VC}}+1)}{\delta})$$

for having a generalization error at most of ϵ with $\mathbb{P}$ at least of $1-\delta$

Example

$\epsilon = 0.1$, $\delta = 0.1$

How many data do we need ?

Interpreting the VC dimension - Sample Complexity

If we replace $m_{\mathcal{H}}(2N)$ with its polynomial upper bound, based on the d_{VC}
Fixed $\delta > 0$,

$$N \geq \frac{8}{\epsilon^2} ln(\frac{4((2N)^{d_{VC}}+1)}{\delta})$$

for having a generalization error at most of ϵ with $\mathbb{P}$ at least of $1 - \delta$

Example

$\epsilon = 0.1$, $\delta = 0.1$

How many data do we need ?

Rule of thumb $\rightarrow N \geq 10 * d_{VC}$

Interpreting the VC dimension - Model Complexity

In most practical situation, however the number N is fixed ($\mathcal{D}$ is fixed)
In these cases the most important question "What performance can we expect with N"?

With probability $\mathbb{P}$ at least of $1-\delta$ we can say that :

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$$

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Interpreting the VC dimension - Model Complexity

In most practical situation, however the number N is fixed ($\mathcal{D}$ is fixed)
In these cases the most important question "What performance can we expect with N"?

With probability $\mathbb{P}$ at least of $1-\delta$ we can say that :

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$$

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Example

$N = 100$, $\delta = 0.1$, $d_{VC} = 1$

What is the error ?

Interpreting the VC dimension - Model Complexity

In most practical situation, however the number N is fixed ($\mathcal{D}$ is fixed)
In these cases the most important question "What performance can we expect with N"?

With probability $\mathbb{P}$ at least of $1-\delta$ we can say that :

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$$

$$Err_{out}(g) \leq Err_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_{\mathcal{H}}(2N)}{\delta}}$$

Example

$N = 100$, $\delta = 0.1$, $d_{VC} = 1$

What is the error ?

$$Err_{out}(g) \leq Err_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

Interpreting the VC dimension - Model Complexity

$$Err_{out}(g) \leq Err_{in}(g) + \Omega(N, \mathcal{H}, \delta)$$

- $\Omega(N, \mathcal{H}, \delta)$ is a "penalty" for the model complexity, more complex the model (larger d_{VC}), the worse the bound
- if δ decreases to much, the complexity increases
- if N increases, the complexity gets better