MAC 0460 / 5832
Introduction to Machine Learning

07 – Is learning feasible ?

IME/USP (05/05/2021)

(RECAP) Computational view of ML: A meta-programming approach

We consider a **generic input-to-output mapping model** and adjust its parameters from available training data

 $\Longrightarrow$ generic model $w(0) \rightarrow w(1)$ $\Longrightarrow$ `output=eliptic?`

 $\Longrightarrow$ generic model $w(1) \rightarrow w(2)$ $\Longrightarrow$ `output=spiral ?`

 $\Longrightarrow$ generic model $w(2) \rightarrow w(3)$ $\Longrightarrow$ `output=spiral ?`

So far, we have been choosing the best data fitting parameters

We discussed some ML algorithms that work with linear models

- Perceptron
- Linear regression
- Logistic regression

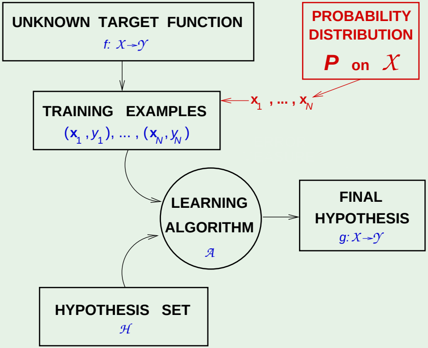
Training ML algorithms

Training dataset $D = \{(\mathbf{x}^{(n)}, y^{(n)}) : \ n = 1, \ldots, N\}$

Hypothesis space $\mathcal{H}$

Choose $g \in \mathcal{H}$ that minimizes some error measure J over D

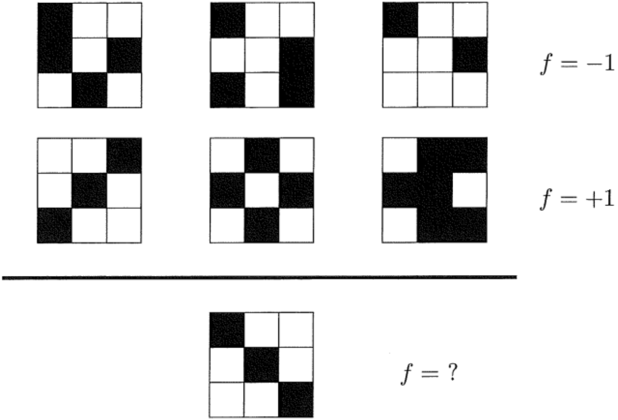
The learning diagram - where we left it



Learning or memorizing?

How does the chosen hypothesis g behave out of sample ?

Puzzle (out of sample behavior)



Fact: Target f is unknown

We are stuck:
There is no guaranteed way to choose g that matches f

Is there a way out?
Are we able to choose a hypothesis g that has *small error* ?

Our ultimate goal is to pick the optimal $g \in \mathcal{H}$; one with minimum

$$E_{out}(g) = \mathbb{E}\Big[\, Err\big(y, g(\mathbf{x})\big) \,\Big] \quad \text{(Expected error wrt } p(\mathbf{x}, y) \text{)}$$

We pick g based on E_{in}:

$$E_{in}(g) = \frac{1}{N} \sum_{i=1}^{N} Err\Big(y^{(i)}, g(\mathbf{x}^{(i)})\Big) \quad \text{(Empirical error)}$$

Important

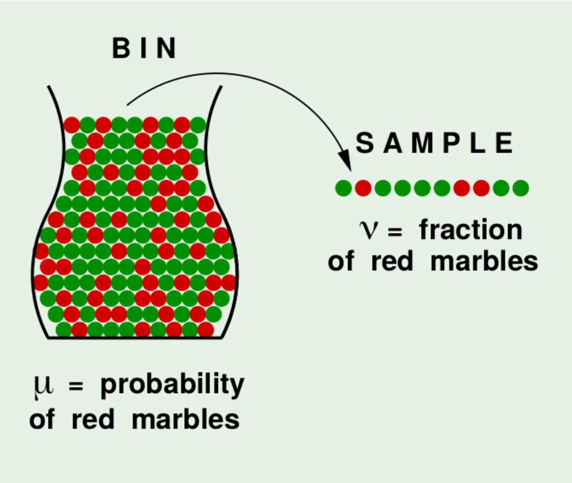
in-sample error: E_{in} (empirical error)

out-of-sample error: E_{out} (true error)

A QUESTION:

Does $E_{in}(g)$ say anything about $E_{out}(g)$?

A probabilistic view



A related experiment

- Consider a 'bin' with red and green marbles.

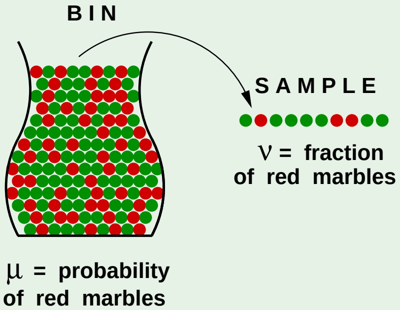
$$\mathbb{P}[\text{ picking a red marble }] = \mu$$

$$\mathbb{P}[\text{ picking a green marble }] = 1 - \mu$$

- The value of μ is <u>unknown</u> to us.

- We pick N marbles independently.

- The fraction of red marbles in sample $= \nu$



Does ν say anything about μ?

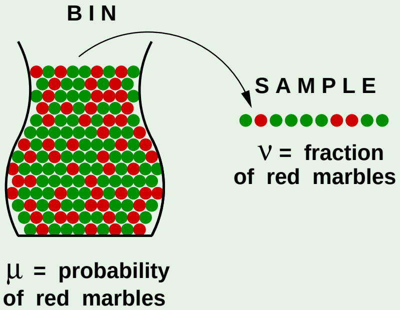
No!

Sample can be mostly green while bin is mostly red.

Yes!

Sample frequency ν is likely close to bin frequency μ.

possible versus **probable**



© AML Creator: Yaser Abu-Mostafa - LFD Lecture 2

ν is an estimate of μ

Is it good enough? Is $|\nu - \mu|$ small ??

Central Limit Theorem

Take samples of size N and compute the fraction of red marbles ν

Repeat this several times

The distribution of ν will be a normal distribution with mean μ

The larger N, the smaller the standard deviation of ν

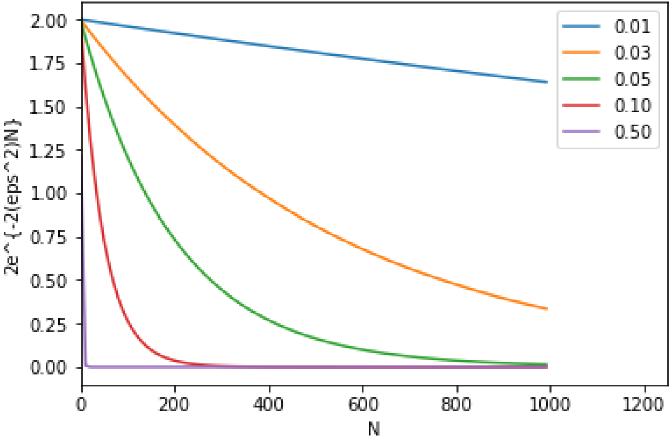
There are other "laws" that establish a relationship between μ (unknown parameter) and ν (its estimate)

Hoeffding inequality

$$P\Big(\ |\nu - \mu| > \epsilon\ \Big) \leq 2e^{-2\epsilon N^2}$$

Bound variation in function of N

$$P\Big(|\nu - \mu| > \epsilon\Big) \leq 2e^{-2\epsilon^2 N}$$



Each color represents a different value of ϵ

We will cast the $E_{in}(h)$ / $E_{out}(h)$ as the **Red** / **Green** marble problem

---

Conceptually, we can color every $\mathbf{x} \in \mathcal{X}$:

- $\mathbf{x}$ if $h(\mathbf{x}) = f(\mathbf{x})$ (green)
- $\mathbf{x}$ if $h(\mathbf{x}) \neq f(\mathbf{x})$ (red)

$E_{out}(h)$ is, then, the fraction of red colored instances in $\mathcal{X}$

$E_{in}(h)$ is, then, the fraction of red colored instances in D

$E_{out}(h)$ unknown parameter

$E_{in}(h)$ an estimate of $E_{out}(h)$

$\Longrightarrow \quad |E_{in}(h) - E_{out}(h)| > \epsilon$??

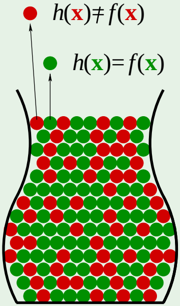
Connection to learning

Bin: The unknown is a number μ

Learning: The unknown is a function $f : \mathcal{X} \to \mathcal{Y}$

Each marble ● is a point $\mathbf{x} \in \mathcal{X}$

- ● : Hypothesis got it right $h(\mathbf{x}) = f(\mathbf{x})$
- ● : Hypothesis got it wrong $h(\mathbf{x}) \neq f(\mathbf{x})$



Notation for learning

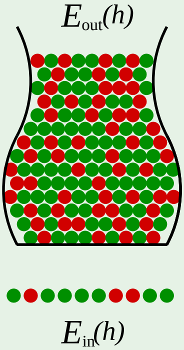
Both μ and ν depend on which hypothesis h

ν is '**in sample**' denoted by $E_{\text{in}}(h)$

μ is '**out of sample**' denoted by $E_{\text{out}}(h)$

The Hoeffding inequality becomes:

$$\mathbb{P}\left[\,|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\,\right] \;\leq\; 2e^{-2\epsilon^2 N}$$



Are we done?

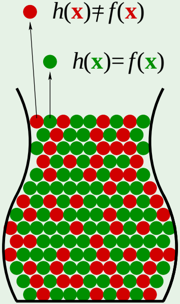
Not so fast! h is fixed.

For <u>this</u> h, ν generalizes to μ.

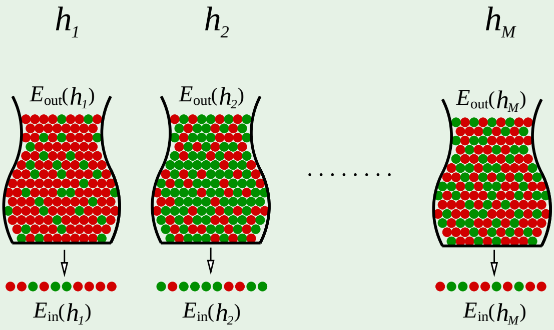
'**verification**' of h, not **learning**

No guarantee ν will be small.

We need to **choose** from multiple h's.

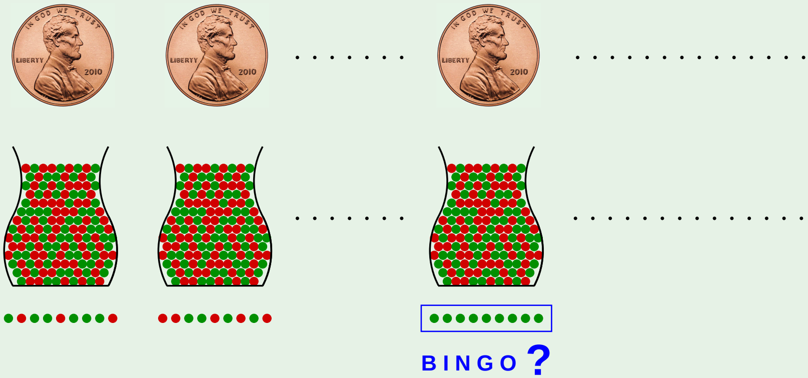


Notation with multiple bins



© Creator: Yaser Abu-Mostafa - LFD Lecture 2

From coins to learning



We can not simply apply Hoeffding to the chosen hypothesis g

WHY ? Because g is not a fixed hypothesis; it is a chosen one

Let us think about this using the coin flipping problem as an example

Suppose a fair coin ($\mu = 0.5$ is the probability of getting head)

Toss the coin $N = 10$ times and count the number of heads
X: number of heads

We expect $X \approx 5$

We compute an estimate of μ as $\nu = X/10$

What is the chance of $|\nu - \mu|$ be a large number??

One coin

If we flip a fair coin $N = 10$ times, what is the probability of 10 heads ?

$$P(X = 10) = (0.5)^{10} \approx 0.0001$$

Multiple coins

Repeat the above experiment for 1000 fair coins
What is the probability that at least one of the coins will yield 10 heads ?

$$\begin{aligned} & P(\text{at least one coin yields } X = 10) = \\ & = \; 1 - P(\text{no coin yields } \; X = 10) \approx 0.623 \end{aligned}$$

Hoeffding in the context of ML

We should consider the probability of some htpothesis h_m be such that $|E_{in}(h_m) - E_{out}(h_m)| > \epsilon$

If we have M hypothesis $h_1, h_2, \ldots, h_M$,
and we choose one, which we denote g,

$$\begin{aligned}
\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ \leq\ \mathbb{P}[&\quad\ |E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon \\
&\textbf{or } |E_{\text{in}}(h_2) - E_{\text{out}}(h_2)| > \epsilon \\
&\cdots \\
&\textbf{or } |E_{\text{in}}(h_M) - E_{\text{out}}(h_M)| > \epsilon\ ] \\
\leq\ & \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\right]
\end{aligned}$$

The final verdict

$$\begin{aligned}\mathbb{P}[\ |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon\ ] \ &\leq\ \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon\right] \\ &\leq\ \sum_{m=1}^{M} 2e^{-2\epsilon^2 N}\end{aligned}$$

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2{\color{red}M}e^{-2\epsilon^2 N}$$

Discussion

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

Consistent with our intuition:

- negative exponential $\rightarrow$ larger N implies smaller bound

Contrary to our intuition:

- number of hypothesis M $\rightarrow$ the larger the hypothesis space $\mathcal{H}$, the larger the bound

Discussion

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

QUESTION: Should we, then, choose a small hypothesis space ??