

Big Data Analytics

The learning problem

Souhaib Ben Taieb

March 2, 2021

University of Mons

Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

A learning puzzle

A Learning puzzle



$$f = -1$$



$$f = +1$$



$$f = ?$$

Is it -1 or +1?

Is learning feasible?

The target function is **unknown**. How could a limited data set reveal enough information to pin down the entire target function?

- More than one function fits the 6 training examples.
 - If the true f is +1 when the pattern is symmetric, then the solution is +1
 - If the true f is -1 when the top left square of the pattern is white, then the solution is -1
- We know the values of f on all the points in the training data \mathcal{D} . But since f is an unknown function, f remains unknown outside of \mathcal{D} .
- The whole purpose of learning f is to be able to predict the value of f on new points.
- Is learning feasible? Yes, in a **probabilistic sense**.

Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

- A related experiment
- Connection to learning
- Connection to real learning
- The solution

Bin and marbles

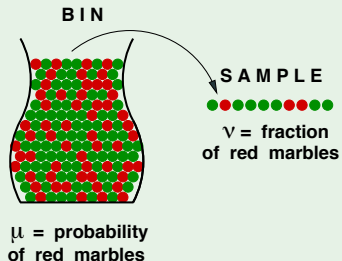
A related experiment

- 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 unknown to us.
- We pick N marbles independently.
- The fraction of red marbles in sample = ν



Bin and marbles

Does ν say anything about μ ?

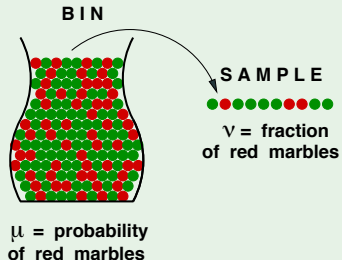
No!

Sample can be mostly green while bin is mostly red.

Yes!

Sample frequency ν is likely close to bin frequency μ .

possible versus probable



Hoeffding's inequality

What does ν say about μ ?

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

Formally,

$$\mathbb{P} [|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

This is called **Hoeffding's Inequality**.

In other words, the statement " $\mu = \nu$ " is P.A.C.

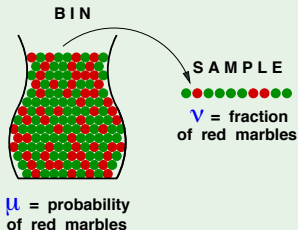
P.A.C = Probably Approximately Correct

We want the probability of the “bad event” (ν far from μ) to be small.

Hoeffding's inequality

$$\mathbb{P} [|\nu - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- Valid for all N and ϵ
- Bound does not depend on μ
- Tradeoff: N , ϵ , and the bound.
- $\nu \approx \mu \implies \mu \approx \nu \quad \text{☺}$



Two rules of probability

Let $\mathcal{B}_1, \mathcal{B}_2$ be any two events. If $\mathcal{B}_1 \implies \mathcal{B}_2$ (i.e. event \mathcal{B}_1 implies event \mathcal{B}_2 or equivalently, $\mathcal{B}_1 \subseteq \mathcal{B}_2$), then

$$\mathbb{P}(\mathcal{B}_1) \leq \mathbb{P}(\mathcal{B}_2).$$

Let $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_M$ be any M events, then

$$\mathbb{P}(\mathcal{B}_1 \text{ or } \mathcal{B}_2 \text{ or } \dots \text{ or } \mathcal{B}_M) \leq \mathbb{P}(\mathcal{B}_1) + \mathbb{P}(\mathcal{B}_2) + \dots + \mathbb{P}(\mathcal{B}_M).$$

The second rule is known as the *union bound* or *Boole's inequality*.

Exercise I

If the probability of red marbles is $\mu = 0.9$, what is the probability that a sample of $N = 10$ marbles will have a fraction of red marbles $\nu \leq 0.1$? [Hint: use a binomial distribution]

Exercise I

If the probability of red marbles is $\mu = 0.9$, what is the probability that a sample of $N = 10$ marbles will have a fraction of red marbles $\nu \leq 0.1$? [Hint: use a binomial distribution]

Solution. If X is the number of red marbles among $N = 10$ marbles, we have that

$$\nu \leq 0.1 \iff X \leq 1,$$

Then, we can write

$$\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = 9.2 \times 10^{-9},$$

where

$$\mathbb{P}(X = x) = \binom{N}{x} \mu^x (1 - \mu)^{N-x}, \quad x = 0, 1, \dots, N,$$

since $X \sim \text{Binomial}(N, \mu)$.

Exercise II

If the probability of red marbles is $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have a fraction of red marbles $\nu \leq 0.1$ and compare the answer to the previous exercise. [Hint: Use one of the previous rule]

Exercise II

If the probability of red marbles is $\mu = 0.9$, use the Hoeffding Inequality to bound the probability that a sample of 10 marbles will have a fraction of red marbles $\nu \leq 0.1$ and compare the answer to the previous exercise. [Hint: Use one of the previous rule]

Solution. If $\mu = 0.9$ and $\nu \leq 0.1$, it implies that $|\nu - \mu| \geq 0.8$ (or $|\nu - \mu| > 0.8^-$, the largest number less than 0.8). We can write

$$\begin{aligned}\mathbb{P}(\nu \leq 0.1) &\leq \mathbb{P}(|\nu - \mu| \geq 0.8) \\ &= \mathbb{P}(|\nu - \mu| > 0.8^-) \\ &\leq 2e^{-2 \times (0.8^-)^2 \times 10} = 5.52 \times 10^{-6}\end{aligned}$$

Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

From marbles to learning

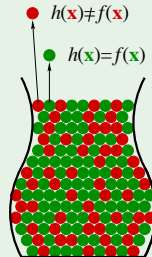
Connection to learning

Bin: The unknown is a number μ

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

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

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



Learning diagram updated

Back to the learning diagram

The bin analogy:

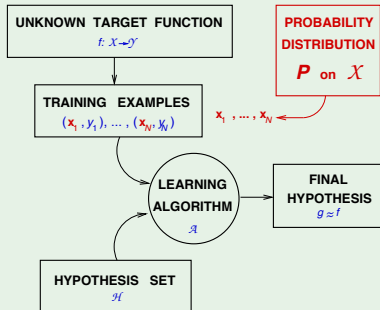


Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

Verification vs learning

Are we done?

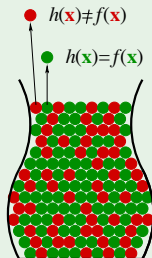
Not so fast! h is fixed.

For this h , ν generalizes to μ .

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

No guarantee ν will be small.

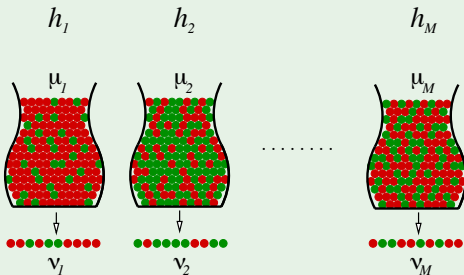
We need to **choose** from multiple h 's.



Multiple bins

Multiple bins

Generalizing the bin model to more than one hypothesis:



Notation for learning

Notation for learning

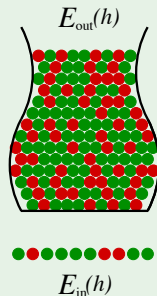
Both μ and ν depend on which hypothesis h

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

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

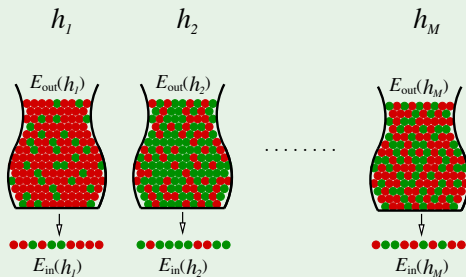
The Hoeffding inequality becomes:

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



Notation with multiple bins

Notation with multiple bins



Hoefdding does not apply to multiple bins - coin analogy

Hoefdding does not apply to multiple bins!

- If you toss **one fair coin** 10 times, what is the probability that you will get 10 heads?
- If you toss **1000 fair coins** 10 times each, what is the probability that some coin will get 10 heads?

Hoefdding does not apply to multiple bins - coin analogy

If you toss **one fair coin** 10 times, what is the probability that you will get 10 heads?

$$\begin{aligned} &P(10 \text{ heads in } 10 \text{ tosses}) \\ &= P(\text{first toss is head and } \dots \text{ and tenth toss is head}) \\ &= P(\text{first toss is head}) \times \dots \times P(\text{tenth toss is head}) \\ &= [P(\text{a toss is head})]^{10} \\ &= 1/2^{10} \\ &\approx 1/1000 \\ &= 0.1\% \end{aligned}$$

Hoefdding does not apply to multiple bins - coin analogy

If you toss **1000 fair coins** 10 times each, what is the probability that some coin will get 10 heads?

$$\begin{aligned} &P(\text{ no heads in 10 tosses for one coin}) \\ &= 1 - P(10 \text{ heads in 10 tosses for one coin}) \\ &= 1 - 1/1000 \end{aligned}$$

$$\begin{aligned} &P(\text{ no heads in 10 tosses for 1000 coins}) \\ &= (1 - 1/1000)^{1000} \\ &\approx 0.37 \end{aligned}$$

$$P(10 \text{ heads in 10 tosses for at least one coin}) = 1 - 0.37 \approx 0.63$$

Hoefdding does not apply to multiple bins - coin analogy

Coin analogy

Question: If you toss a fair coin 10 times, what is the probability that you will get 10 heads?

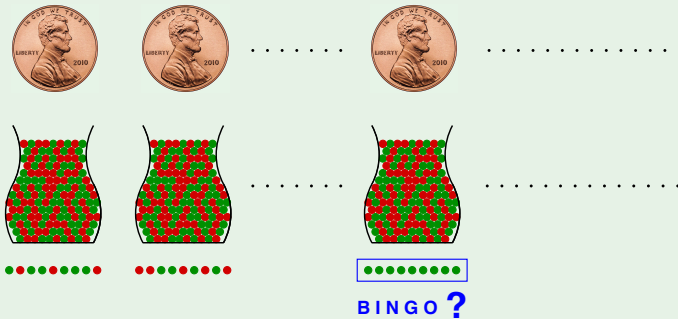
Answer: $\approx 0.1\%$

Question: If you toss 1000 fair coins 10 times each, what is the probability that some coin will get 10 heads?

Answer: $\approx 63\%$

From coins to learning

From coins to learning



From coins to learning

The Hoeffding inequality applies to each bin individually. The inequality states that

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

where

1. the hypothesis h is fixed before the data is generated,
2. the probability is with respect to random data sets \mathcal{D} .

The assumption “ h is fixed before the data set is generated” is critical to the validity of the bound.

From coins to learning

In learning, we consider an entire hypothesis set, say $\mathcal{H} = \{h_1, h_2, \dots, h_M\}$ (with a finite number of hypotheses), instead of just one hypothesis h . Then, the learning algorithm picks the final hypothesis $g \in \mathcal{H}$ based on \mathcal{D} .

The statement we would like to make is **not**

$$\mathbb{P}[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \epsilon] \text{ is small for any fixed } h_m \in \mathcal{H},$$

where $m = 1, 2, \dots, M$, but **rather**

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \text{ is small for the final hypothesis } g.$$

A simple solution

$$|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon$$

\implies

$$|E_{\text{in}}(h_1) - E_{\text{out}}(h_1)| > \epsilon$$

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

...

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

A simple solution

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

The final verdict

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

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

Generalization error

- The **out-of-sample error** E_{out} measures how well our training on \mathcal{D} has generalized to unseen data points. E_{out} is based on the performance over the entire input space \mathcal{X} .
- The **in-sample error** E_{in} is based on the training data points.
- The **generalization error** is the discrepancy between E_{in} and E_{out} . Generalization error is also used as another name for E_{out} (but not here).
- The Hoeffding inequality provides a way to *characterize the generalization error* with a probabilistic bound.

Generalization bound

The Hoeffding inequality states that

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

This can be rephrased as follows. Pick a tolerance level δ , for example $\delta = 0.01$, and assert with probability at least $1 - \delta$ that

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

This is called a *generalization bound* since it bounds E_{out} in terms of E_{in} .

Generalization bound

We can rewrite the Hoeffding inequality as follows

$$\begin{aligned}\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] &\leq 2Me^{-2\epsilon^2 N} \\ \implies 1 - \mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon] &\leq 2Me^{-2\epsilon^2 N} \\ \implies \mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon] &\geq 1 - 2Me^{-2\epsilon^2 N}\end{aligned}$$

In other words, with probability at least $1 - 2Me^{-2\epsilon^2 N}$, we have

$$|E_{\text{in}}(g) - E_{\text{out}}(g)| \leq \epsilon,$$

or, equivalently,

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

Generalization bound

If $\delta = 2Me^{-2\epsilon^2 N}$, then we have $\epsilon = \sqrt{\frac{1}{2N} \ln\left(\frac{2M}{\delta}\right)}$.

Since $E_{\text{out}}(g) \leq E_{\text{in}}(g) + \epsilon$, we can say that, with probability at least $1 - \delta$,

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

In other words, the hypothesis g that we choose will continue to do well out of sample.

We also want to be sure that there is no other hypothesis $h \in \mathcal{H}$ where $E_{\text{out}}(h)$ is significantly better than $E_{\text{out}}(g)$. The other direction of the bound (i.e. $E_{\text{out}}(g) \geq E_{\text{in}}(g) - \epsilon$) assures us that it is unlikely that any other hypothesis in \mathcal{H} was unlucky on the training set but is actually much better than the g we have chosen.

The VC inequality

- The error bound depends on M , the size of the hypothesis set \mathcal{H}
- If \mathcal{H} is an infinite set, the bound goes to infinity and becomes useless
- M can be replaced with something finite (the effective number of hypotheses), so that the bound is meaningful.

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8} \epsilon^2 N}$$

The Vapnik-Chervonenkis Inequality

Table of contents

Is learning feasible?

A probabilistic perspective of learning

From marbles to learning

Verification vs learning

Feasibility of learning

- If we insist on a deterministic answer, i.e. \mathcal{D} tells us something certain about f outside of \mathcal{D} , then the answer is no.
- If we accept a probabilistic answer, i.e. \mathcal{D} tells us something likely about f outside of \mathcal{D} , then the answer is yes.

What we know so far

Learning is feasible. It is likely that

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

Is this learning? We need $g \approx f$, which means

$$E_{\text{out}}(g) \approx 0.$$

The two questions of learning

$E_{\text{out}}(g) \approx 0$ is achieved through

$$E_{\text{out}}(g) \approx E_{\text{in}}(g) \text{ and } E_{\text{in}}(g) \approx 0$$

Learning can be reduced to two questions:

1. Can we make sure that $E_{\text{out}}(g)$ is close enough to $E_{\text{in}}(g)$?
2. Can we make $E_{\text{in}}(g)$ small enough?
 - The Hoeffding Inequality addresses the **first question** only.
 - We answer the **second question** after running the learning algorithm on the the training data.

The two questions of learning

The complexity of \mathcal{H} .

Question 1: According to the Hoeffding Inequality, a larger M increases the risk that $E_{\text{in}}(g)$ will be a poor estimate of $E_{\text{out}}(g)$
 \implies we need to control M (a measure of the complexity of \mathcal{H}).

Question 2: We stand a better chance if \mathcal{H} is more complex \implies
a more complex \mathcal{H} gives us more flexibility in finding some g that fits the data well.

The two questions of learning

The complexity of f .

Question 1: If we fix the hypothesis set and the number of training examples, the inequality provides the same bound \implies The complexity of f does not affect how well $E_{\text{in}}(g)$ approximates $E_{\text{out}}(g)$.

Question 2: The data from a complex f are harder to fit than the data from a simple f (large $E_{\text{in}}(g)$). We can increase the complexity of \mathcal{H} , but then $E_{\text{out}}(g)$ will not be as close to $E_{\text{in}}(g)$.

What the theory will achieve

Characterizing the feasibility of learning for
infinite M

Characterizing the tradeoff:

Model complexity \uparrow	E_{in}	\downarrow
Model complexity \uparrow	$E_{\text{out}} - E_{\text{in}}$	\uparrow

