

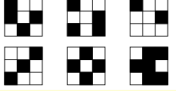
Lecture 4: Feasibility of Learning——学习的可行性

4.1 Learning is impossible? 学习的窘境

First Puzzle: 第一个谜题

Two Controversial Answers

whatever you say about $g(\mathbf{x})$,



$y_0 = -1$
 $y_0 = +1$

$g(\mathbf{x}) = ?$

truth $f(\mathbf{x}) = +1$ because ...

- symmetry $\Leftrightarrow +1$
- (black or white count = 3) or (black count = 4 and middle-top black) $\Leftrightarrow +1$

truth $f(\mathbf{x}) = -1$ because ...

- left-top black $\Leftrightarrow -1$
- middle column contains at most 1 black and right-top white $\Leftrightarrow -1$

all valid reasons, your **adversarial teacher** can always call you 'didn't learn'. :-(

不同的人，不同的答案，但在样本上的正确率均为100%，哪个是正确的好像说不清楚——Learning is impossible.

Second Puzzle: 第二个谜题

No Free Lunch

\mathbf{x}	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	0	0	0	0	0	0	0	0	0	0
0 0 1	1	1	1	1	1	1	1	1	1	1
0 1 0	1	1	1	1	1	1	1	1	1	1
0 1 1	0	0	0	0	0	0	0	0	0	0
1 0 0	1	1	1	1	1	1	1	1	1	1
1 0 1	?	?	0	0	0	0	1	1	1	1
1 1 0	?	?	0	0	1	1	0	0	1	1
1 1 1	?	?	0	1	0	1	1	0	0	1

• $g \approx f$ inside \mathcal{D} : sure!

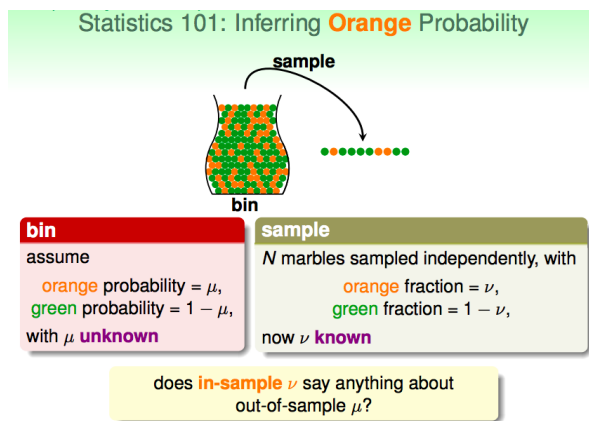
• $g \approx f$ outside \mathcal{D} : **No!** (but that's really what we want!)

learning from \mathcal{D} (to infer something outside \mathcal{D}) is doomed if **any 'unknown' f can happen**. :-(

理论上共有 2^8 个hypothesis，但在样本上表现全对的hypothesis只有 $2^3 = 8$ 种，而这8种中只有1种是 f ，其余7种虽然在样本上的正确率为100%，但在样本外会有偏差，甚至全错。这个时候用类似PLA的算法挑出在样本上表现全对的hypothesis好像没有什么用——Learning is impossible.

4.2 Probability to the Rescue: 运用概率去解决

试想一个情景：有一个罐子(bin)，里面有橙色(orange)的球(marble)和绿色(green)的球，我们并不知道其中orange球的占比(portion or probability)，那么我们能不能去infer呢？好像是可以的，我们可以抽样：



对于罐子(bin), 我们假设:

- Orange probability = μ
- Green probability = $1 - \mu$
- 但 μ 未知

对于样本(sample), 我们假设 N 个marbles都是independently抽取(i.i.d)的:

- Orange fraction = ν
- Green fraction = $1 - \nu$
- ν 已知!

所以我们自然而然地想, 可不可以通过 ν 得到一些关于 μ 的信息?

- No! 因为可能抽出来的全是green而罐子里全是orange! → 这种情况是possible的, 但可能性不大;
- Yes! 当 N 很大的时候, ν 很可能接近 μ ! → 这种情况是probable的, 可能性很大;
- 用概率的视角来思考。

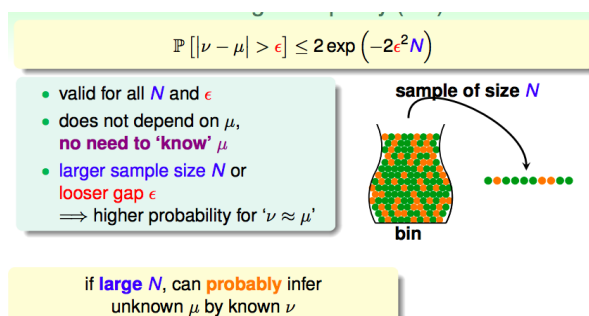
那么 ν 和 μ 之间的关系到底是什么呢? **Hoeffding's Inequality**!

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

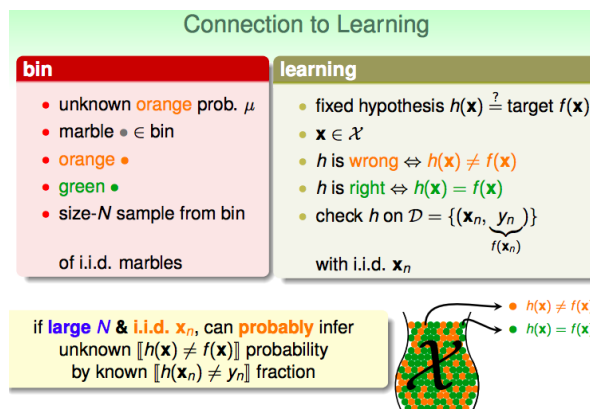
$$P[|\nu - \mu| > \epsilon] \leq 2 \cdot e^{-2\epsilon^2 N}$$

该不等式说明: N 足够大时, ν 和 μ 相似的概率很高—— N 非常大, 则不等式右侧 e 项接近于0, 那么不等式左侧事件发生的概率也就接近于0, 即 ν 和 μ 相差很多的概率几乎为0, 即 ν 和 μ 相近的概率很大。

" $\mu = \nu$ " is "probably approximately correct" —— **PAC** ——那么就可以用已知的 ν 来推断未知的 μ ——但是要注意条件, N 需要比较大才可以。



4.3 Connection to Learning: 与学习联系起来



对于罐子：

- $Marble \in bin$
- orange球
- green球
- orange的概率 μ 未知
- Size-n sample
- i.i.d. marbles

对于学习（fixed hypothesis）：

- $\vec{x} \in X$
- $h(\vec{x}) \neq f(\vec{x})$
- $h(\vec{x}) = f(\vec{x})$
- $h(\vec{x}) \neq f(\vec{x})$ 在总体中的概率未知—— E_{out} 未知
- $D = \{(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)\}$ —— n 个样本
- i.i.d. \vec{x}_n

注意： i.i.d. marbles——marble有一个概率分布，每个marble都有一定被选取的概率；所以，也有 Probability distribution P on X ，但是：

- No assumption about P
- P can be anything

也就是说， x_1, x_2, \dots, x_n 是根据 P 从input space X 中generated出来的。

对比bin和learning可以得到：对于fixed hypothesis h ，if large N & i.i.d. \vec{x}_n ，可以probably推断未知的 $\llbracket h(\vec{x}) \neq f(\vec{x}) \rrbracket$ 的prob. BY 已知的 $\llbracket h(\vec{x}_n) \neq y_n \rrbracket$ 的frac.。

即：对固定的 h ，可以probably推断未知的**Out-of-sample error**：

$$E_{out}(h) = \mathbb{E}_{\vec{x} \sim P} [h(\vec{x}) \neq f(\vec{x})]$$

by 已知的**In-sample error**：

$$E_{in}(h) = \frac{1}{N} \sum_{n=1}^N [h(\vec{x}_n) \neq y_n]$$

其中， $E_{in}(h)$ 与 $E_{out}(h)$ 满足：

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

The Formal Guarantee

for any fixed h , in 'big' data (N large),
 $\boxed{\text{in-sample error } E_{in}(h)}$ is probably close to
 $\boxed{\text{out-of-sample error } E_{out}(h)}$ (within ϵ)

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

same as the 'bin' analogy ...

- valid for all N and ϵ
- does not depend on $E_{out}(h)$, **no need to 'know' $E_{out}(h)$**
 $\rightarrow f$ and P can stay unknown
- ' $E_{in}(h) = E_{out}(h)$ ' is **probably approximately correct (PAC)**

if ' $E_{in}(h) \approx E_{out}(h)$ ' and ' $E_{in}(h)$ small'
 $\Rightarrow E_{out}(h)$ small $\Rightarrow h \approx f$ with respect to P

但上述内容，并不是learning，而是verification！因为 h 是fixed的。也就是说，我们拿到了一个candidate h ，霍夫汀不等式向我们保证了，这个 h 的 E_{in} 和 E_{out} 是PAC程度上相等的——但霍夫汀不等式的应用条件是 **h is fixed**！如果 h 不是固定的，那么 E_{in} 和 E_{out} 是PAC程度上相等这个结论就不一定成立了。

h is not fixed，这才是现实中的learning，因为现实中并不是给你一个 h 让你去验证，而是让你从一堆 h 中挑出一个最好的 h 作为 g ，即在 H 中运用算法 A 选出 g 来。因此，real learning实质上是一个multiple bins的问题，而不是single bin的问题，需要区别一下。

Verification of One h

for any fixed h , when data large enough,
 $E_{in}(h) \approx E_{out}(h)$
Can we claim 'good learning' ($g \approx f$)?

Yes! if $E_{in}(h)$ small for the fixed h and A pick the h as g $\Rightarrow 'g = f'$ PAC	No! if A forced to pick THE h as g $\Rightarrow E_{in}(h)$ almost always not small $\Rightarrow 'g \neq f'$ PAC!
---	--

real learning:
 A shall **make choices** $\in \mathcal{H}$ (like PLA)
 rather than **being forced to pick one h** . $\therefore \{$

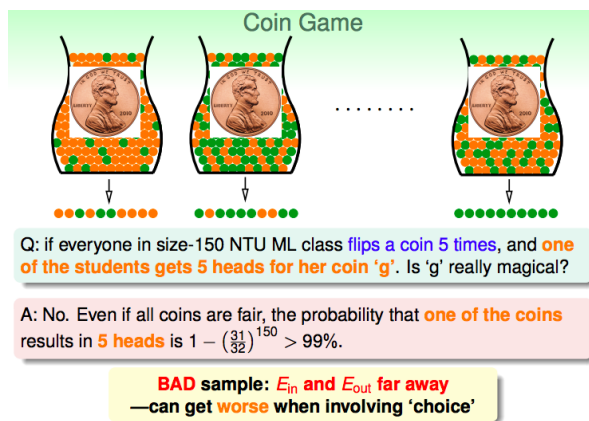
4.4 Connection to Real Learning: 与真实的学习相联系

Real Learning：下面两个问题是等价的

- 【希望拿到全是(绝大多数)是绿球的bin】有 M 个bins，某个bin抽样的球全是green，是否说明这个bin中全是green？
- 【希望拿到 E_{out} 是(接近)0的hypothesis】有 M 个hypothesis，某个 h 对样本的预测全对，是否这个 h 就很好？

好像都说明不了，因为霍夫汀不等式只是对单个fixed的hypothesis成立，对multiple hypothesis就不能成立了，所以在multiple的情况下即使拿到的hypothesis g 的 E_{in} 很小，但并不能通过霍夫汀不等式得到 E_{out} 也很小的推论，因为并不能保证 $E_{out} \approx E_{in}$ 。

再考虑如下问题：



- If you toss a fair coin 10 times, what is the prob. That will get 10 heads? $\approx 0.1\%$
- If you toss 1000 fair coins 10 times each, what is the prob. that some coin will get 10 heads? $\approx 63\%$

通过上述问题可以知道：

- 对于某一个hypothesis h ，通过Hoeffding不等式可以知道，在任给一组sample的情况下，有很高的可能性 $E_{out}(h)$ 和 $E_{in}(h)$ 非常接近；
- 但对于一组hypotheses \mathcal{H} ，通过掷硬币问题类比可知，在任给一组sample的情况下，至少出现某个hypothesis其 E_{out} 和 E_{in} 相差很大的可能性是很高的——简单想， h 有很多，那么对于任何一组sample，总是很可能有一个hypothesis能够瞎猫碰见死耗子使得in sample error为0，但是out of sample error很差；
- 此时，对于某个演算法 A ，就有可能选到比较差的hypothesis：例如，当hypotheses有很多很多时，最小经验误差演算法就很容易踩到雷(overfit)。

对于任意的某个演算法，我们将其回传的结果记为 g ；将“ $|E_{in}(g) - E_{out}(g)| > \epsilon$ ”记作event A，将“ $|E_{in}(h_1) - E_{out}(h_1)| > \epsilon$ or $|E_{in}(h_2) - E_{out}(h_2)| > \epsilon$ or ... or $|E_{in}(h_M) - E_{out}(h_M)| > \epsilon$ ”记作event B。易知，event A属于event B，故：

$$Probability(A) \leq Probability(B)$$

即：

$$\begin{aligned} P[|E_{in}(g) - E_{out}(g)| > \epsilon] &\leq P[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon \text{ or } |E_{in}(h_2) - E_{out}(h_2)| > \epsilon \text{ or } \dots \text{ or } |E_{in}(h_M) - E_{out}(h_M)| > \epsilon] \\ &\stackrel{\text{union bound}}{\leq} P[|E_{in}(h_1) - E_{out}(h_1)| > \epsilon] + P[|E_{in}(h_2) - E_{out}(h_2)| > \epsilon] + \dots + P[|E_{in}(h_M) - E_{out}(h_M)| > \epsilon] \\ &= 2Me^{-2\epsilon^2 N} \end{aligned}$$

上式相当于multiple h 情景下的霍夫汀不等式，通过该不等式可知：

- 如果 $|H| = M$ ，即假说集是有限(finite)的，在 N large enough的条件下，由任何演算法 A 选出的 g ，probably有 $E_{in} = E_{out}$ 。

The 'Statistical' Learning Flow

if $|\mathcal{H}| = M$ finite, N large enough,

for whatever g picked by \mathcal{A} , $E_{\text{out}}(g) \approx E_{\text{in}}(g)$

if \mathcal{A} finds one g with $E_{\text{in}}(g) \approx 0$,

PAC guarantee for $E_{\text{out}}(g) \approx 0 \Rightarrow$ **learning possible :-)**

