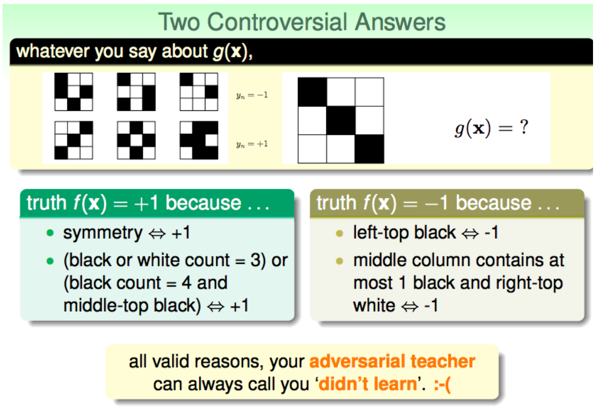
Lecture 4: Feasibility of Learning——学习的可行性

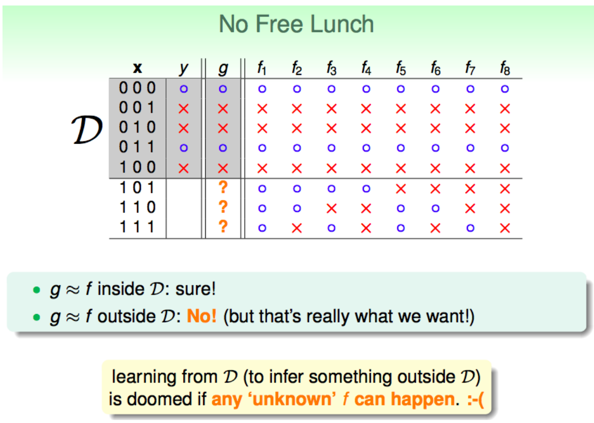
4.1 Learning is impossible? 学习的窘境

First Puzzle: 第一个谜题



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。

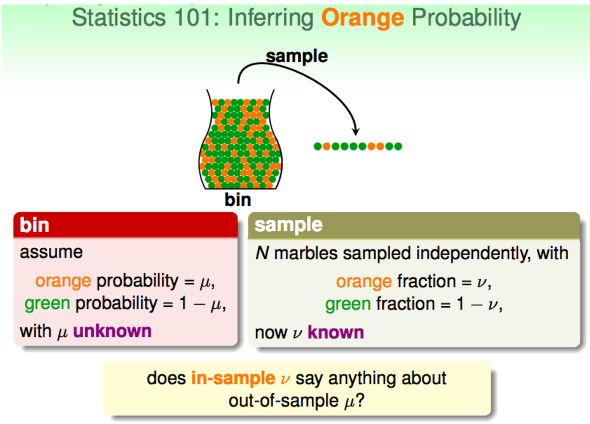
Second Puzzle: 第二个谜题



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 $= \mu$
- 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 $= \nu$
- 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！ $\rightarrow$ 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！ $\rightarrow$ 这种情况是probable的，可能性很大；
- 用概率的视角来思考。

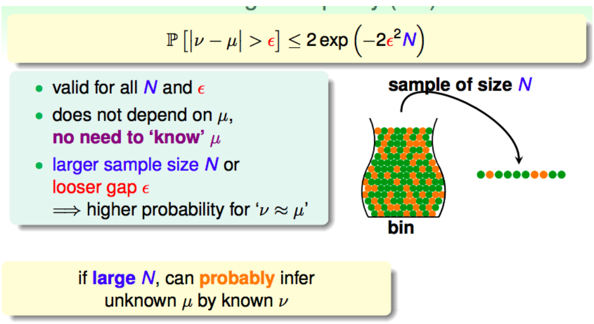
那么ν和μ之间的关系到底是什么呢？**Hoeffding's Inequality!**

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

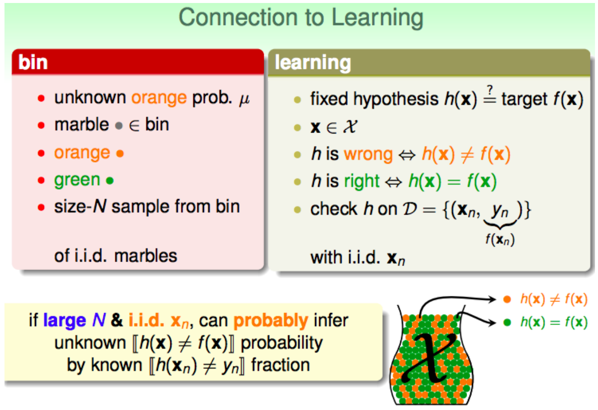
$P[|\nu - \mu| > \epsilon] \leq 2 \cdot e^{-2\epsilon^2 N}$

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
- $D = \{(\vec{x}_1, y_1), \cdots, (\vec{x}_n, y_n)\}$——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，$x_1, x_2, \cdots, x_n$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$[h(\vec{x}) \neq f(\vec{x})]$的prob. BY 已知的$[h(\vec{x}_n) \neq y_n]$的frac.。

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

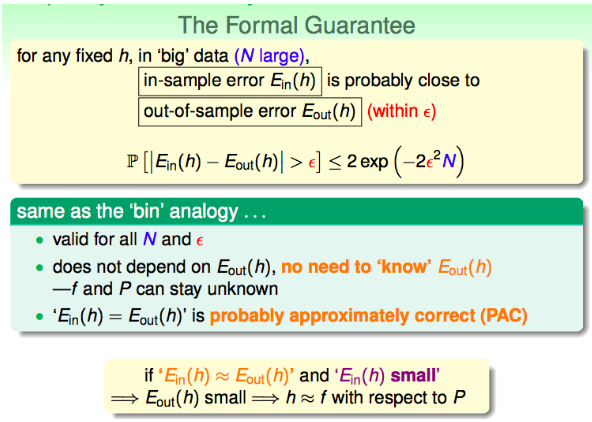
$$E_{out}(h) = \underset{\vec{x} \sim P}{\varepsilon}[h(\vec{x}) \neq f(\vec{x})]$$

by 已知的**In-sample error**：

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

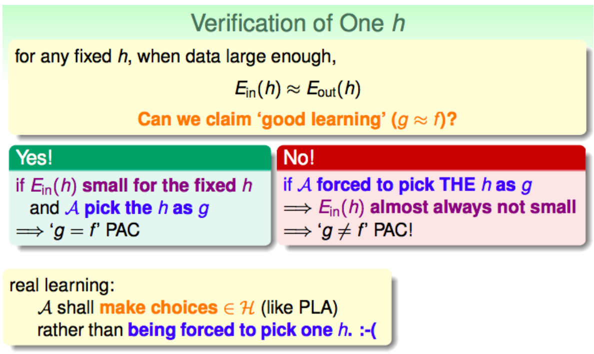
其中，$E_{in}(h)$与$E_{out}(h)$满足：

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



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。



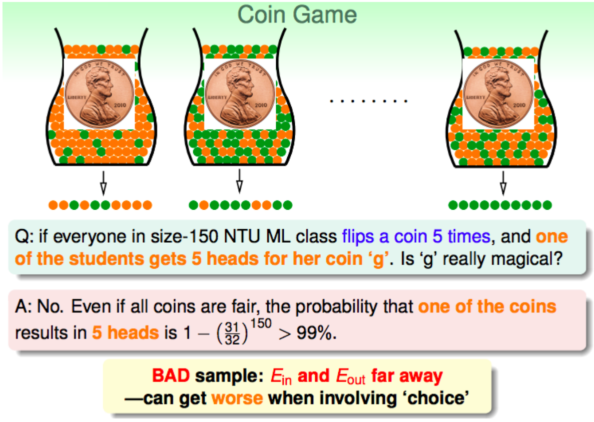
4.4 Connection to Real Learning: 与真实的学习相联系

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$\mathcal{A}$，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:

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

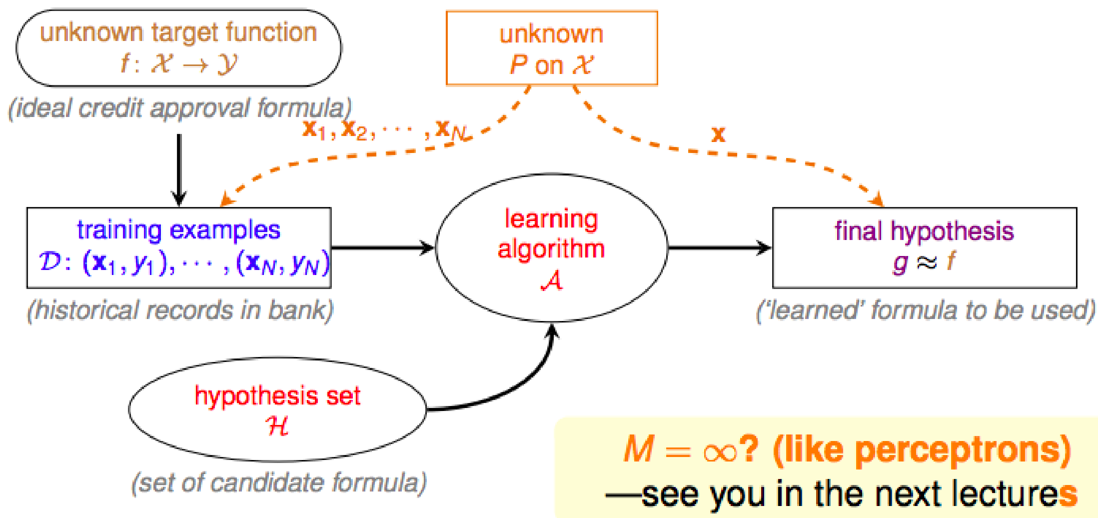
上式相当于multiple h情景下的霍夫汀不等式，通过该不等式可知:

- 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 $E_{\text{out}}(g) \approx 0 \Longrightarrow$ **learning possible :-)**



$M = \infty$? (like perceptrons)
—see you in the next lectures