机器学习基石 · 笔记（第 3、4 周）

Machine Learning Foundations by Prof. Hsuan-Tien Lin

Haoming Wang

Spring 2020

这篇笔记是台湾大学林轩田教授的 Coursera 课程：機器學習基石上 (Machine Learning Foundations)—Mathematical Foundations 的课程笔记. 您可以点击这里获取更多笔记.

This is a note I took while studying the Coursera course taught by Prof. Hsuan-Tien Lin at National Taiwan University. You can click here for more notes.

目录

1 机器学习的种类 Types of Learning

1.1 由输出空间分类 Learning with Different Output Space $\mathcal{Y}$

回顾之前学习的感知器学习算法, 其输出空间 $\mathcal{Y} = \{+1, -1\}$, 这样的问题我们将其称为**二元分类问题** (binary classification). 除了 PLA 算法所处理的线性可分数据, 二元分类问题也包含非线性可分的数据.

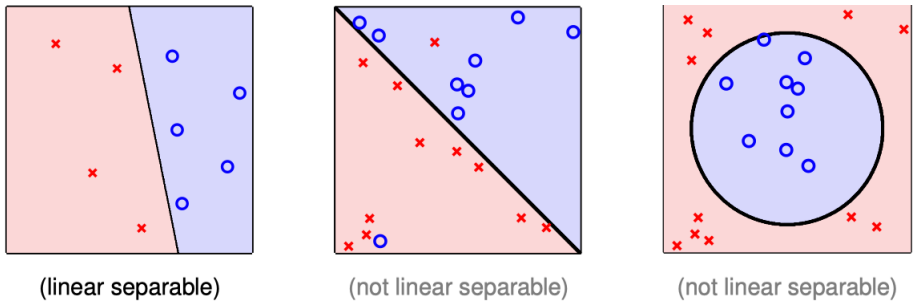

图 1: 二元分类

当 $\mathcal{Y} = \{1, 2, \cdot, k\}$ 时, 二元分类问题便扩展为**多元分类问题** (Multiclass Classification). 例如根据硬币的尺寸与重量, 将一堆硬币分为 1 美分, 5 美分, 10 美分, 25 美分四类.

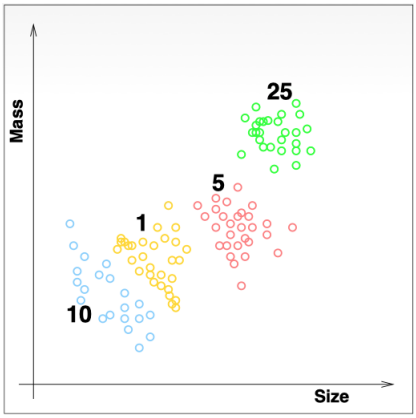

图 2: 多元分类问题

当 $\mathcal{Y} = \mathbb{R}$ 或为 $\mathbb{R}$ 的子区间时, 我们称之为**回归问题** (Regression) 或有界回归问题 (Bounded Regression). 例如预测股票价格或者天气气温.

考虑一个多分类的问题, 向机器输入一个单词, 由机器判断其词性. 该问题的输

入为 $\mathcal{X} = \{\text{word}\}$, 输出为 $\mathcal{Y} = \{\text{pronoun, noun, verb, adjective, adverb}, \cdots\}$. 现在扩展这个问题, 向机器输入一个句子, 并由机器判断句子中每个单词的词性, 即 $\mathcal{Y} = \{\text{pvn, pvp, nvn, pv}, \cdots\}$, 即输出是输入的某种结构, 我们称这样的问题为**结构学习** (Structured Learning).

1.2 由数据标签分类 Learning with Different Data Label y_n

对于输入的数据 $\mathcal{D}$, 若每个 $\mathbf{x}_n$ 都对应这一个标签 y_n, 则称这类机器学习问题为**监督学习** (Supervised Learning). 如上一节的感知器学习算法, 每个输入的点都存在一个标签-1 或 +1. 如图2, 该硬币分类问题中, 每个硬币都记上其所属标签, 这是监督学习.

如果每个 $\mathbf{x}_n$ 都没有标签 y_n, 这类问题称为**无监督学习** (Unsupervised Learning), 一种常见的无监督学习是聚类 (Clustering). 对应到硬币分类问题中, 如果不给每个硬币标签, 而仅以每个硬币的尺寸重量分布特征将其分类, 这便是聚类.

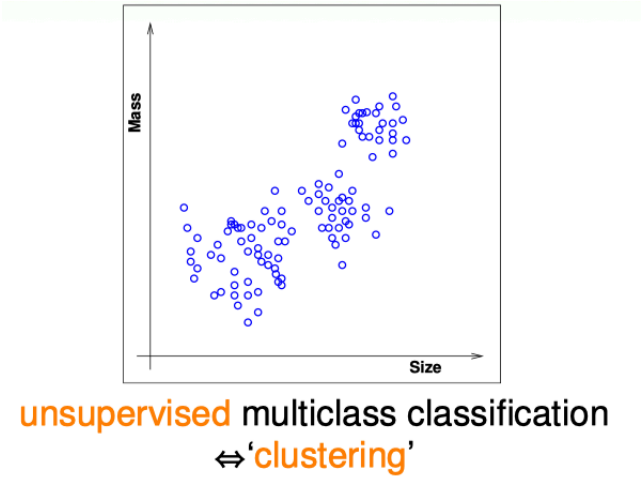

图 3: 无监督学习

从上图可看, 这一堆硬币大致可聚为三类. 无监督学习的应用除了聚类还有密度估计 (Density Estimation), 即输入一组数据, 找出数据最集中分布的位置. 如输入各地的交通信息, 找出易发生事故的路段; 异常检测 (Outlier Detection), 输入一组数据, 找出其中的异常值, 因为某些异常值发生频率不高, 数据有限, 因此我们只能用无监督的形式将其晒出. 如输入网络日志, 发出黑客入侵警报.

除了完全标记所有的输入数据和完全不标记, 我们还可以标记部分数据, 这样的问题称为**半监督学习** (Semi-supervised Learning)

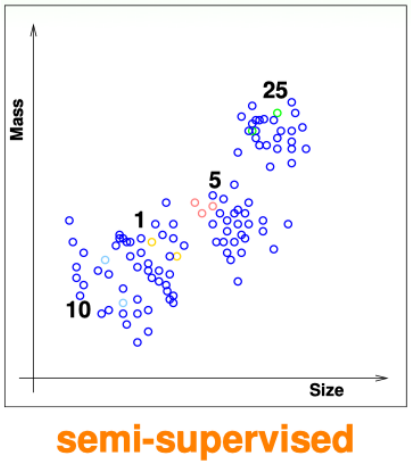

图 4: 半监督学习

为了建立一个树木识别系统，一家公司决定在互联网上收集 100 万张照片。然后，它要求 10 个公司成员查看 100 张照片，并记录每张照片是否包含一棵树。然后，这些图片和记录被输入一个学习算法来建立这个系统. 这便是一个半监督学习的场景.

除了以上三种, 当不能明确指定标签 y_n 时, 我们可以通过评价有偏标签 $\tilde{y}_n$ 使机器进行学习, 这种学习方式成为**强化学习** (Reinforcement Learning). 例如训练狗时, 当人发出指定"sit down"($\mathbf{x}_n$), 人无法向狗说明对应的坐下 (y_n) 是什么样的, 但若狗反应出坐下, 趴下, 握手等动作 ($\tilde{y}_n$) 时, 我们可以对其进行一些奖励, 从而促进其学习.

1.3 由学习方式分类 Learning with Different Protocol $f \Rightarrow (\mathbf{x}_n, y_n)$

一种最常见的学习方式是**批量学习** (Batch Learning), 即一次性输入所有数据. 除此之外还有一种学习方式是**线上学习** (Online Learning), 即先输入一个数据 $\mathbf{x}_n$, 得到预测结果 $g_t(\mathbf{x}_n)$, 然后通过 $(\mathbf{x}_n, y_n)$ 升为 g_t 为 g_{t+1}. 我们前面实现的 PLA 算法就是一种线上学习. 强化学习一般也是通过线上方式进行的.

除此之外还有一种学习方式: **主动学习** (Active Learning), 它可以通过策略性的向人提出问题从而以更少的标签来改进 hypothesis g. 例如在识别手写数字时, 主动学习可以让人类判断某些没有标签且机器认为置信度不高的数字图片, 从而提高学习的精度.

1.4 由输入空间分类 Learning with Different Input Space $\mathcal{X}$

在硬币分类问题中, 我们输入给机器的是**具体特征** (Concrete Features), 即硬币的尺寸和重量. 机器通过这些具体特征来识别硬币种类. 但是在另一些问题中, 我们只有数据

的**原始特征** (Raw Features).

例如在识别手写数字的问题中, 当我们想要识别数字 1 和数字 5 时, 我们可以人为的为两个数字创建两个具体特征: 对称性和密度 (数字 1 的对称性高, 密度低; 数字 5 的对称性低, 密度高).

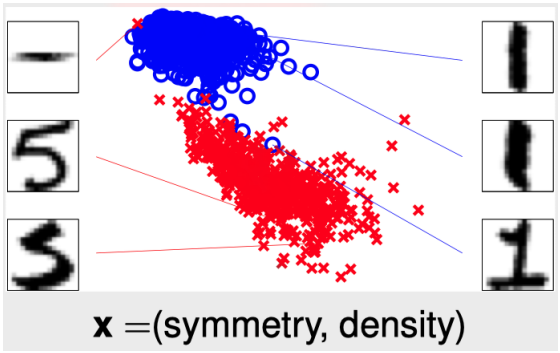

图 5: 数字 1 和数字 5 的特征

因为每一个数字的图片是一个 16×16 的像素矩阵, 所以另外一种方式是将每一个数字图片视为一个 256 为的特征向量, 这便是一种原始特征.

对于原始特征的数据, 我们一般将其抽象出一些具体特征, 这一过程可以交由机器, 也可以由人类完成. 对于前者我们可以称为**深度学习** (Deep Learning); 对于后者我们称之为**特征工程** (Feature Engineering).

需要注意, 具体特征和原始特征均具有物理含义 (上例中原式特征的物理含义是像素的位置). 但是现实中还有一些特征没有 (或几乎没有) 物理含义, 我们称为**抽象特征** (Abstract Features), 例如使用者的 id, 项目的 id 等. 这时我们要对抽象特征抽取出其具体特征, 如使用者的的偏好和项目的类别等.

2 机器学习可行性 Feasibility of Learning

2.1 机器学习不可行的例子 Learning is Impossible?

在很多时候是无法进行机器学习的, 如下给出六个样本, 其中第一行给予标签 $y_n = -1$, 第二行给予标签 $y_n = +1$, 问第三行的标签应该是多少.

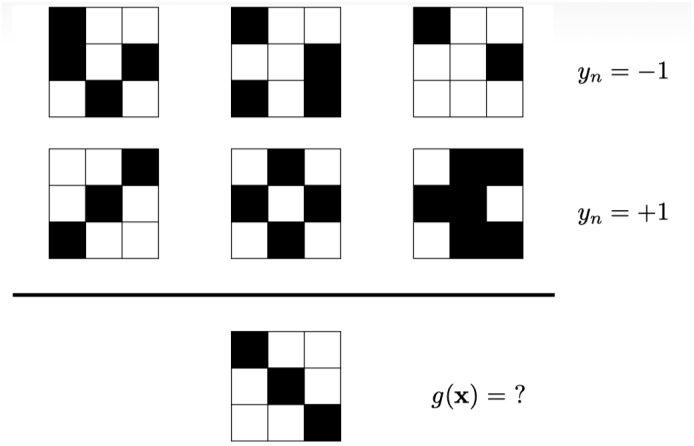

图 6: A Learning Puzzle

当我们令

$$f = \begin{cases} +1 & \text{图形是对称的;} \\ -1 & \text{图形是不对称的} \end{cases}$$

时, 第三行图形的标签为 +1. 但是当我们令

$$f = \begin{cases} +1 & \text{左上角是黑的;} \\ -1 & \text{左上角是白的} \end{cases}$$

时, 第三行图形的标签为-1.

这个例子是想说明, 在某些情况下我们是无法运用机器学习预测 $\mathcal{D}$ 以外的数据的. 例如某数列 a_n 满足: $a_1 = 1, a_2 = 2, a_3 = 3, a_4 = 8, a_5 = 26$, 问 $a_6 =?$. 事实上这个问题是无解的, 因为使用拉格朗日插值法, a_6 可以是任意值. 如令:

$$\begin{aligned} y1 &= -24 + 55x - \frac{523}{12}x^2 + \frac{391}{24}x^3 - \frac{35}{12}x^4 + \frac{5}{24}x^5 \\ y2 &= -36 + \frac{412}{5}x - \frac{793}{12}x^2 + \frac{595}{24}x^3 - \frac{53}{12}x^4 + \frac{37}{120}x^5 \\ y3 &= -40 + \frac{1373}{15}x - \frac{883}{12}x^2 + \frac{221}{8}x^3 - \frac{59}{12}x^4 + \frac{41}{120}x^5 \\ y4 &= -48 + \frac{549}{5}x - \frac{1063}{12}x^2 + \frac{799}{24}x^3 - \frac{71}{12}x^4 + \frac{49}{120}x^5 \end{aligned}$$

则 y_1, y_2, y_3, y_4 在 $x = 1, 2, 3, 4, 5$ 均满足 $y = 1, 2, 3, 8, 26$. 但是 $y_1(6) = 96, y_2(6) = 108, y_3(6) = 112, y_4(6) = 120$.

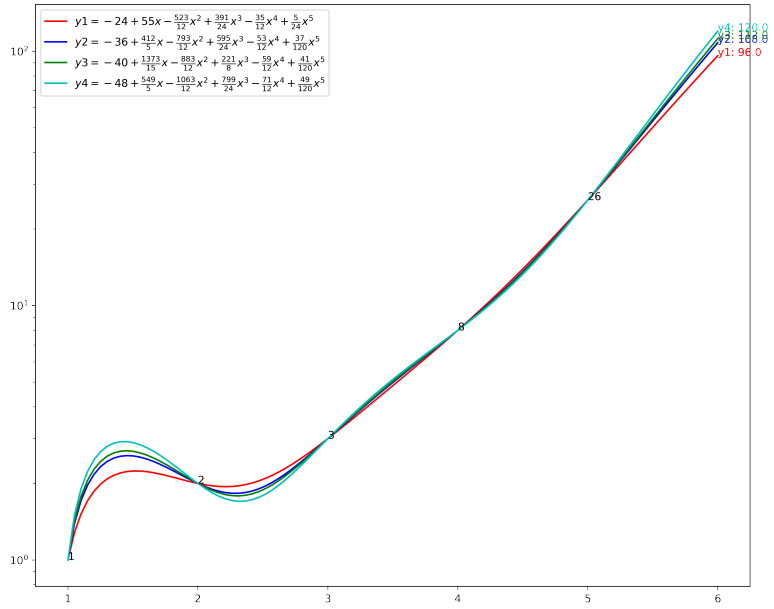

图 7: 四个多项式函数

实现代码如下:

```
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline

x = np.linspace(1, 6, 100)
y1 = -24 + 55*x - (523/12)*x**2 + (391/24)*x**3 - (35/12)*x**4 + (5/24)*x**5
y2 = -36 + (412/5)*x - (793/12)*x**2 + (595/24)*x**3 - (53/12)*x**4 + (37/120
                                                    )*x**5
y3 = -40 + (1373/15)*x - (883/12)*x**2 + (221/8)*x**3 - (59/12)*x**4 + (41/
                                                    120)*x**5
y4 = -48 + (549/5)*x - (1063/12)*x**2 + (799/24)*x**3 - (71/12)*x**4 + (49/
                                                    120)*x**5
plt.figure(figsize=(12,10))
plt.plot(x,y1, "r", label=r"$y1=-24+55x-\frac{523}{12}x^2+\frac{391}{24}x^3-\
                                                    frac{35}{12}x^4+\frac{5}{24}x^5$")
plt.plot(x,y2, "b", label=r"$y2=-36+\frac{412}{5}x-\frac{793}{12}x^2+\frac{
                                                    595}{24}x^3-\frac{53}{12}x^4+\frac{37}
                                                    {120}x^5$")
plt.plot(x,y3, "g", label=r"$y3=-40+\frac{1373}{15}x-\frac{883}{12}x^2+\frac{
                                                    221}{8}x^3-\frac{59}{12}x^4+\frac{41}{
```

```
                                        120}x^5$")
plt.plot(x,y4, "c", label=r"$y4=-48+\frac{549}{5}x-\frac{1063}{12}x^2+\frac{
                                        799}{24}x^3-\frac{71}{12}x^4+\frac{49}
                                        {120}x^5$")
plt.yscale("log")
for i in range(1, 6):
    plt.text(i ,np.round(y1[np.int((i-1) * 99/5)]), np.int(np.round(y1[np.int
                                            ((i-1) * 99/5)])))
plt.text(6 ,np.round(y1[99]), f"y1: {np.round(y1[99])}", c="r")
plt.text(6 ,np.round(y2[99]), f"y2: {np.round(y2[99])}", c="b")
plt.text(6 ,np.round(y3[99]), f"y3: {np.round(y3[99])}", c="g")
plt.text(6 ,np.round(y4[99]), f"y4: {np.round(y4[99])}", c="c")
plt.legend()
plt.savefig("/Users/wanghaoming/Documents/LaTeX_doc/Machine_Learning/
                                        unlearnexamp.png", bbox_inches='tight'
                                        , dpi=500)
```

2.2 霍夫丁不等式 Hoeffding's inequality

上一节我们指出, 某些情况下, 在数据集 $\mathcal{D}$ 以外推断位置的目标 f 是困难的; 那么在其他的情形下, 我们能否推断出一些未知的事物呢? 考虑一个装有橙色和绿色的桶, 假设桶中橙色球的比例为 μ, 绿色球的比例为 $1-\mu$, 其中 μ 未知. 现从桶中独立地取出 N 个球 (样本), 这些样本中橙色球的比例为 ν, 绿色球的比例为 $1-\nu$. 我们想知道 ν 是否能说明 μ 的某些性质.

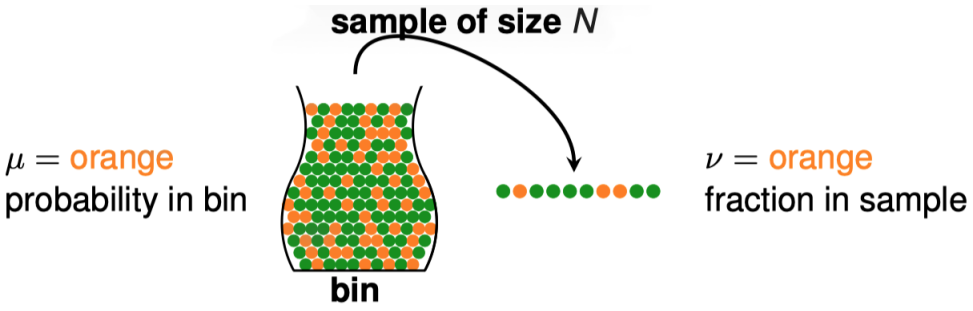

图 8: bin model

Lemma 2.1 (霍夫丁引理 (Hoeffding's Lemma)). *对于随机变量 X, 若 $\mathbb{P}(X \in [a,b]) = 1, \mathbb{E}[X] = 0$, 则对于 $\forall s > 0$ 有:*

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\frac{s^2(b-a)^2}{8}}.$$

Theorem 2.1 (霍夫丁不等式 (Hoeffding's Inequality))**.** 设有两两独立的一系列随机变量 $X_i, i = 1, 2, \cdots, n$, 设 X_i *a.e.* 有界, 即 $\mathbb{P}(X_i \in [a_i, b_i]) = 1$, 令

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

则对任意 $\epsilon > 0$

$$\mathbb{P}\left(|\bar{X} - \mathbb{E}\left[\bar{X}\right]| \geq \epsilon\right) \leq 2e^{-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

证明. 取 $s > 0, \epsilon > 0$, 令 $S_n = \sum_{i=1}^{n} X_i$, 由马尔科夫不等式有:

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq \epsilon\right) &= \mathbb{P}\left(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{s\epsilon}\right) \\
&\leq e^{-s\epsilon}\mathbb{E}\left[e^{s(S_n - \mathbb{E}[S_n])}\right] \\
&= e^{-s\epsilon}\mathbb{E}\left[e^{s\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])}\right] \\
&= e^{-s\epsilon}\prod_{i}^{n}\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right] \quad X_i \text{ 两两独立}
\end{aligned}$$

由霍夫丁引理有

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq \epsilon\right) &\leq e^{-s\epsilon}\prod_{i=1}^{n} e^{\frac{s^2(b_i - a_i)^2}{8}} \\
&= e^{-s\epsilon + \frac{1}{8}s^2\sum_{i=1}^{n}(b_i - a_i)^2}.
\end{aligned} \tag{1}$$

令 $s^* = -\frac{4\epsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$, 代入上式有

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

令 $S_n = \bar{X}n$, 则有

$$\mathbb{P}\left(\bar{X} - \mathbb{E}\left[\bar{X}\right] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

将 $\mathbb{E}\left[S_n\right] - S_n$ 代入式 (1) 有

$$\mathbb{P}\left(\mathbb{E}\left[\bar{X} - \bar{X}\right] \geq \epsilon\right) \leq e^{-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

因此

$$\mathbb{P}\left(|\bar{X} - \mathbb{E}\left[\bar{X}\right]| \geq \epsilon\right) \leq 2e^{-\frac{2\epsilon^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}.$$

□

回到上文的桶模型, 在 N 次抽样中, 令每次抽取的小球结果为 X_i, 其中

$$X_i = \begin{cases} 1, & \text{第 } i \text{ 次抽取的结果为橙色,} \\ 0, & \text{第 } i \text{ 次抽取的结果为绿色} \end{cases}$$

则有 $\nu = \frac{\sum_{i=1}^{N} X_i}{N}$, 以及

$$\begin{aligned} \mathbb{E}[\nu] &= \mathbb{E}\left[\frac{\sum_{i=1}^{N} X_i}{N}\right] = \frac{\sum_{i=1}^{N} \mathbb{E}[X_i]}{N} \\ &= \frac{\sum_{i=1}^{N}(1-\mu)\cdot 0 + \mu \cdot 1}{N} = \frac{N\mu}{N} = \mu. \end{aligned}$$

又因为 $\mathbb{P}(X_i \in [0,1]) = 1$, 所以有

$$\begin{aligned} \mathbb{P}(|\nu - \mu| \geq \epsilon) &= \mathbb{P}(|\nu - \mathbb{E}[\nu]| \geq \epsilon) \\ &\leq 2e^{-\frac{2\epsilon^2 N^2}{\sum_{i=1}^{N}(1-0)^2}} \\ &= 2e^{-2\epsilon^2 N}. \end{aligned}$$

因此我们有结论: 在一个大样本中 (N 很大), 样本 ν 可能在误差 ϵ 内接近于总体 μ.

换句话说, 命题 $\nu = \mu$, **概率近似正确** (大概-差不多-对 probably approximately correct (PAC)), i.e. 在概率 $1 - 2e^{-2\epsilon^2 N}$ 内 ν 与 μ 不超过 ϵ. 当我们有更大的样本数 N 或更小的误差 ϵ, 我们就能提高 $\mu \approx \nu$ 的概率.

2.3 概率近似正确学习框架 PAC Learning Framework

比较桶模型与机器学习的过程, 对于 hypothesis h 与样本 $\mathbf{x}_i \in \mathcal{X} (i = 1, 2, \cdots, N)$, 若 $h(\mathbf{x}_i) \neq f(\mathbf{x}_i)$, 则令该样本 $\mathbf{x}_i$ 为" 橙色球"; 若 $h(\mathbf{x}_i) = f(\mathbf{x}_i)$, 则令该样本 $\mathbf{x}_i$ 为" 绿色球". 依次在 $\mathcal{D} = (\mathcal{X} \times \mathcal{Y})$(注意 $f(\mathbf{x}_i) = y_i \in \mathcal{Y}$) 上检查 h. 则在样本 $\mathcal{D}$ 中 $h(\mathbf{x}_i) \neq f(\mathbf{x}_i)$ 的概率即为桶模型中的 ν, 在总体中 $h(\mathbf{x}) \neq f(\mathbf{x})$ 的概率即为桶模型中的 μ.

如果 $\mathbf{x}_i \sim i.i.d$ 且在 N 较大的条件下, 我们可以由霍夫丁不等式通过 $\mathbb{P}(h(\mathbf{x}_i) \neq f(\mathbf{x}_i))$ 去推断 $\mathbb{P}(h(\mathbf{x}) \neq f(\mathbf{x}))$.

换句话说, 对于任意给定的 hypothesis h 和样本空间 $\mathcal{D}$, 我们可以通过已知的样本内误差 (in-sample error):

$$E_{in}(h) = \frac{1}{N}\sum_{i=1}^{N} [\![h(\mathbf{x}_i) \neq f(\mathbf{x}_i)]\!]$$

去推断未知的样本外误差 (out-sample error):

$$E_{out}(h) = \mathop{\mathcal{E}}_{\mathbf{x}\sim P} [\![h(\mathbf{x}) \neq f(\mathbf{x})]\!]$$

当样本容量 N 很大时, 有

$$\mathbb{P}\left(|E_{in}(h) - E_{out}(h)| > \epsilon\right) \leq 2e^{-2\epsilon^2 N}.$$

正如桶模型, 对于给定的 hypothesis h, 样本空间 $\mathcal{D}$ (with 样本容量 N) 以及 ϵ, 我们可以说 $E_{in}(h) = E_{out}(h)$ 是概率近似正确的 Probably Approximately Correct (PAC).

需要注意的是, 我们目前讨论的都是对某个给定的 hypothesis h, 即对一个 h 进行评价. 对于任何一个给定的 h, 当样本容量足够大时, 且 $E_{in}(h)$ 很小时, 算法 $\mathcal{A}$ 选择 h 作为 g, 则可以推断出 $g = f$ PAC, 此时我们可以说这是一个好的学习过程. 但是一般而言 $E_{in}(h)$ 都不会很小, 这时如果选 h 为 g, 我们可推断出 $g \neq f$ PAC, 这自然就不是好的学习过程.

真正的学习是从 $\mathcal{H}$ 中选择最好的 h 作为 g, 而不是指定选择某个 h 为 g. 因此我们上面所讨论的范畴仅是 verification 的过程, 并未涉及到 Learnign. 整个 verification 的框架如下:

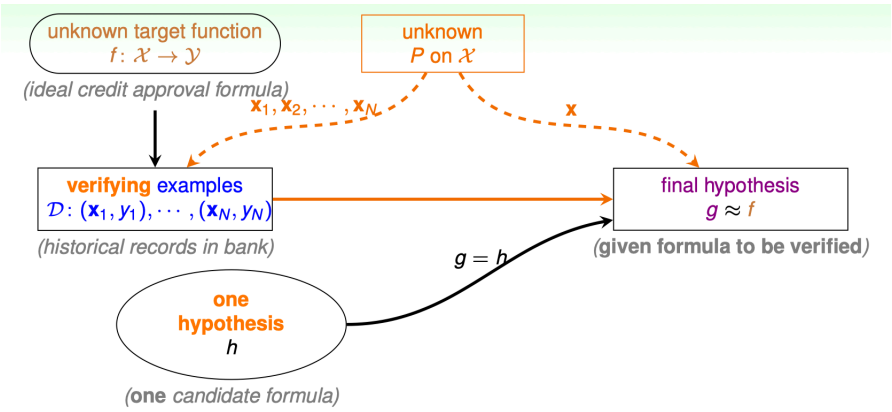

图 9: The 'Verification'Flow

服从某个分布 P 的 $i.i.d.r.v.$ $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N$ 以及其由 target function f 生产的标签 $\mathcal{Y} = \{y_i\}_N$ (真实值) 构成我们的样本空间 $\mathcal{D}$. $\mathcal{X} = \{\mathbf{x}_i\}_N$ 与某个 hypothesis h 的结果 $h(\mathbf{x}_i)$ 同样本标签中的真实值 $y_i = f(\mathbf{x}_i)$ 相检验, 从而计算出 $E_{in}(h)$, 通过霍夫丁不等式推断 $h = g = f$ PAC.

下面考虑多个 h 的情况. 如果我们有 $h_1, \cdots, h_M$ 共 M 个 hypothesis. 若 h_M 在其 $N(= 10)$ 个样本中全为绿色, 我们能否认定 h_M 是一个好的 hypothesis? 霍夫丁不等式告诉我们的是在大概率下 $E_{in}(h_M) \approx E_{out}(h_M)$, 但是观察下图我们会发现, 在这种情况下, 小概率事件发生了.

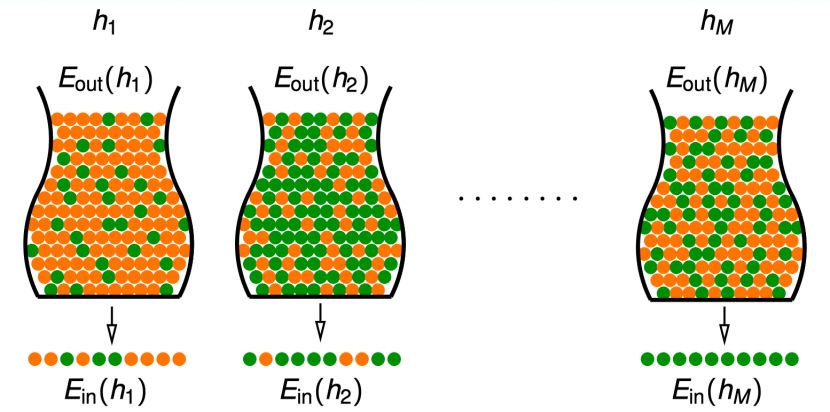

图 10: 多个 hypothesis

我们称这种情况下 h_M 选出的那 $N(=10)$ 个的样本 $\mathcal{D}'$ 为**坏样本** (BAD sample), 即 $|E_{in_{\mathcal{D}'}}(h) - E_{out}(h)| \geq \epsilon$.

考虑某一个 hypothesis h, 其有多个样本空间 D_i, 其坏样本的情况如下:

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	...	Hoeffding
h	BAD					BAD		$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h] $\leq$ ...

霍夫丁不等式告诉我们

$$\mathbb{P}_{\mathcal{D}}(\text{BAD } \mathcal{D}) = \sum_{\text{all possible } \mathcal{D}} \mathbb{P}(\mathcal{D}) \cdot [\![\mathcal{D} \text{ is BAD}]\!]$$

是小某一值的.

下面考虑多样本情况.

	$\mathcal{D}_1$	$\mathcal{D}_2$	...	$\mathcal{D}_{1126}$	...	$\mathcal{D}_{5678}$	Hoeffding
h_1	BAD					BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_1] $\leq$ ...
h_2		BAD					$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_2] $\leq$ ...
h_3	BAD	BAD				BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_3] $\leq$ ...
...							
h_M	BAD					BAD	$\mathbb{P}_{\mathcal{D}}$ [BAD $\mathcal{D}$ for h_M] $\leq$ ...
all	BAD	BAD				BAD	?

我们定义: 对于任意 $h_i \in \mathcal{H}$ 而言, 若 $\mathcal{D}_j$ 是 BAD 样本, 则称 $\mathcal{D}_j$ 对于 $\mathcal{H}$ 是一个 BAD 样本.

若 $\mathcal{D}'$ 对于 $\mathcal{H}$ 而言是一个 BAD 样本, 则算法 $\mathcal{A}$ 不能在 $\mathcal{D}'$ 上自由的在 $\mathcal{H}$ 中选择 h (如 $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_{5678}$). 即 $\mathcal{A}$ 只能在对于任意 hypothesis 都没有坏样本的样本上从 $\mathcal{H}$ 中选择 h, 如 $\mathcal{D}_{1126}$. 这样我们选择了 $\mathcal{H}$ 的坏样本的概率为:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}}(\text{BAD } \mathcal{D}) &= \mathbb{P}_{\mathcal{D}}\left(\bigcup_{i=1}^{M} \text{BAD } \mathcal{D} \text{ for } h_i\right) \\
&\leq \sum_{i=1}^{M} \mathbb{P}_{\mathcal{D}}\left(\text{BAD } \mathcal{D} \text{ for } h_i\right) \\
&\leq \sum_{i=1}^{M} 2e^{-2\epsilon^2 N} \\
&= 2Me^{-2\epsilon^2 N}.
\end{aligned}$$

这说明, 我们有 $1 - 2Me^{-2\epsilon^2 N}$ 的概率选择了对于任意 $h \in \mathcal{H}$ 都是好的样本. 即如果 $|\mathcal{H}| = M < \infty$, N 足够大, 则对于任何 $\mathcal{A}$ 选择的 g, 有 $E_{in}(g) = E_{out}(g)$ PAC. 如果 $\mathcal{A}$ 可以选择一个满足 $E_{in}(g) \approx 0$, 则我们可以说 $E_{out}(g) \approx 0$ PAC. 这说明学习是可行的.

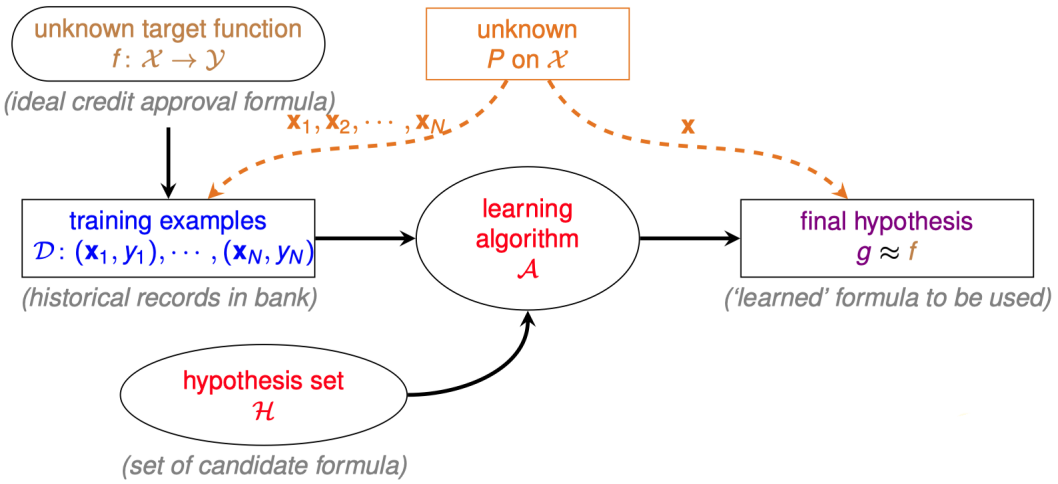

图 11: PAC 学习框架