四、Feasibility of Learning 机器学习的可能性

4.1 Learning is Impossible　学习可能是做不到的

在训练样本集（in-sample）中，可以求得一个最佳的假设g，该假设最大可能的接近目标函数f，但是在训练样本集之外的其他样本（out-of-sample）中，假设g和目标函数f可能差别很远。

4.2 Probability to the Rescue 可能的补救方式

通过上一小节，我们得到一个结论,机器学习无法求得近似目标函数f的假设函数g。

回忆在以前学过的知识中，有无遇到过类似的问题：通过少量的已知样本推论整个样本集的情况。

是否想到一个曾经学过的知识，其实就是概率统计中的知识。

通过一个例子来复习下该知识。有一个罐子，这个罐子里盛放着橙色和绿色两种颜色的小球，我们如何在不查遍所有小球的情况下，得知罐子中橙子小球所占的比例呢？抽取样本，如图4-1所示。

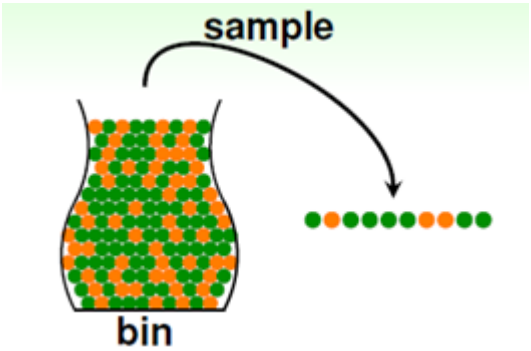

图4-1 抽取样本

假设罐子中橙色小球的概率为μ，不难得出绿色小球的概率为$1-\mu$，其中μ为未知值；

而通过抽样查出的橙色小球比例为ν，绿色小球的比例为$1-\nu$，是从抽样数据中计算出的，因此为已知值。

如何通过已知样本，求得未知的样本？

可以想象到，在很大的几率上接近的结果。因为在罐子里的小球均匀搅拌过后，抽出小球中的橙色小球比例很有可能接近整个罐子中橙色小球的比例，不难想象在抽出的小球数量等于罐中小球数量时，两者完全一致。

这其中不了解的是，到底有多大的可能性两者接近？此处使用数学的方式给予答案，如公式4-1所示。

$$P[|\nu-\mu|>\varepsilon]\leq 2\exp(-2\varepsilon_2 N)$$

该公式称之为霍夫丁不等式（Hoeffding's Inequality），其中P为概率符号，$|\nu-\mu|$表示ν与μ的接近程度，ε为此程度的下界，N表示样本数量，其中不等式左边表示ν与μ之间相差大于某值时的概率。从该不等式不难得出，随着样本量的增大，ν与μ相差较大的概率就不断变小。两者相差越多，即ε越大，该概率越低，就意味着ν与μ相等的结论大概近似正确（probably approximately correct PAC）。

同时可以得出当N足够大时，能够从已知的 ν推导出未知的μ 。

4.3 Connection to Learning 联系到机器学习上

上一节得出的结论可以扩展到其他应用场景，其中包括机器学习。

为方便理解，做一个对比表，如表4-1所示。

表4-1 机器学习与统计中的对比

罐子小球	机器学习
未知的橙色小球比例μ	某一确定的假设在整个X输入空间中，输入向量x满足条件$h(x)=f(x)$的占整个输入空间的比例
抽取的小球∈整个罐子中的小球	训练输入样本集x ∈整个数据集X
橙色小球	假设h作用于此输入向量x与给定的输出不相等⇔ $h(x)\neq f(x)$
绿色小球	假设h作用于此输入向量x与给定的输出相等⇔ $h(x)=f(x)$
小球样本是从罐子中独立随机抽取的	输入样本x是从整个数据集D中独立随机选择的

更通俗一点的解释上表表达的内容：训练输入样本集类比随机抽取的小球样本；此样本集中，先确定一个假设函数h，满足条件$h(x)\neq f(x)$的输入向量x占整个样本的比例类比于橙色小球在随机抽取小球样本的比例ν，写成公式的形式可以入公式4-2所示；因此使用上一节中的PAC（（可能近似正确的理论），在整个输入空间中这个固定的假设函数h同目标函数f不相等的输入量占整个输入空间数量的概率 μ（ μ的取值如公式4-3所示）与上述随机样本中两个函数不相等的样本数占抽样数的比例ν 相同，这一结论也是大概近似正确的。

$$\nu=\frac{1}{N}\sum_{i=1}^{N}[[h(x_i)\neq f(x_i)]] \qquad x_i\in X \qquad \text{公式}4\text{-}2$$

$$\mu=E_{x\sim P}[[h(x)\neq f(x)]] \qquad \text{公式}4\text{-}3$$

其中N为随机独立抽样的样本数，X为整个输入空间， [[ ]]满足条件为1否则为0，E为取期望值。

对1.4节的机器学习流程图进行扩展，得到如图4-2所示。

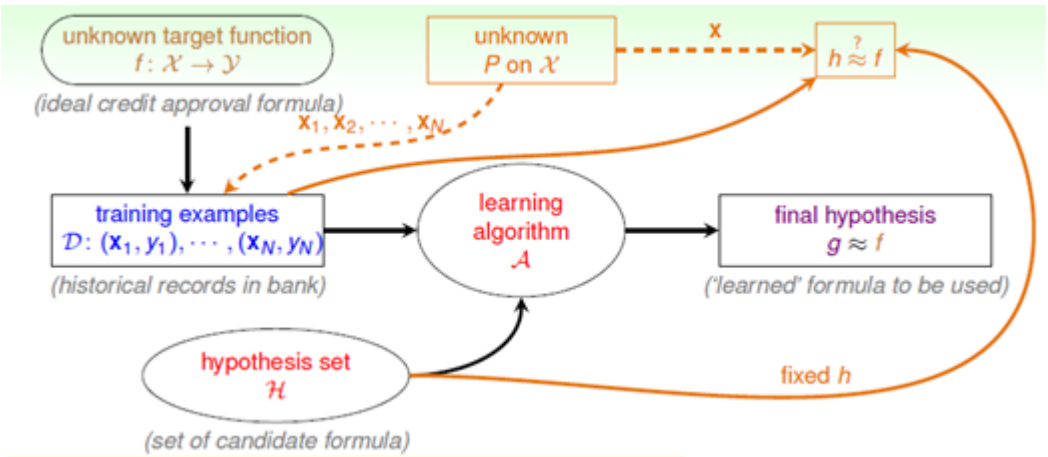

图4-2 引入统计学知识的机器学习流程图

其中虚线表示未知概率P对随机抽样以及概率μ 的影响，实线表示已经随机抽出的训练样本及某一确定的假设对比例 ν的影响。

得出的结论如下：对任意已确定的假设函数h，都可以通过已知的$E_{in} = \frac{1}{N}\sum_{i=1}^{N}[[h(x_i) \neq f(x_i)]]$求出未知的$E_{out} = E_{x\sim P}[[h(x) \neq f(x)]]$。

以后我们将使用和E_{in} 和E_{out}这种专业的符号，分别表示在某一确定的假设函数h中，随机抽样得到的样本错误率和整个输入空间的错误率，同样可以使用霍夫丁不等式对以上得到的结论做出相应的数学表达，如公式4-4所示。

$$P[|E_{in} - E_{out}| > \varepsilon] \leq 2\exp(-2\varepsilon^2 N) \qquad \text{公式 4-4}$$

但是，我们想得到的不是给定一个已确定的假设函数h，通过样本的错误比例来推断出在整个输入空间上的错误概率，而是在整个输入空间上同目标函数f最接近的假设函数h。

那如何实现最接近呢？说白了错误率最低。只需在上述结论上再加一个条件，即错误比例 E_{in}很小即可。总结下，在结论$E_{in}(h) \approx E_{out}(h)$基础之上，加上 $E_{in}(h)$很小，可以推出$E_{out}(h)$ 也很小，即在整个输入空间中h≈f。

上面说了那么多，可能很多人已经糊涂了，因为这并不是一个学习问题，而是一个固定假设函数h，判断该假设函数是否满足上述性质，这准确的讲是一种确认（Verification），确实如此，这种形式不能称为学习，如图4-3所示。

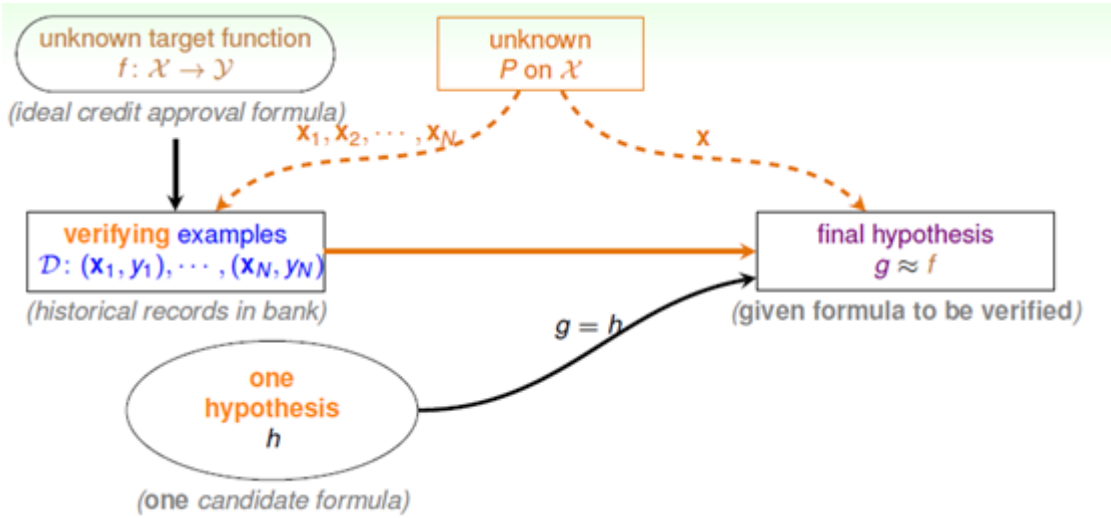

图4-3 确认流程图

4.4 Connection to Real Learning 联系到真正的学习上

首先我们要再次确认下我们上一小节确定的概念，要寻找的是一个使得E_{in}很小的假设函数h，这样就可以使得h和目标函数f在整个输入空间中也很接近。 继续以丢硬币为例，形象的观察这种学习方法有无问题，如图4-4所示。

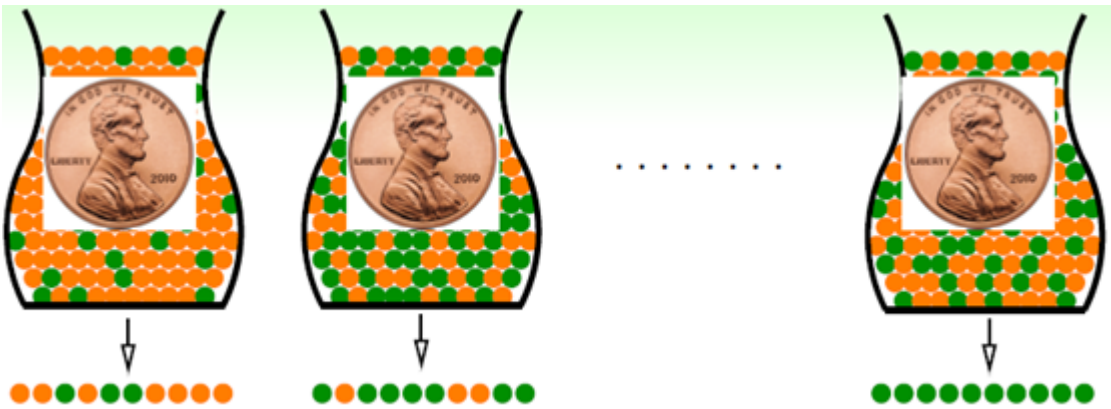

图4-4 丢硬币的例子

假设有150个人同时丢五次硬币，统计其中有一个人丢出五次全部正面向上的概率是多少，不难得出一个人丢出五次正面向上的概率为$\frac{1}{32}$，则150人中有一人丢出全正面向上的概率为$1-(\frac{31}{32})^{150}>99\%$。

这其中抛出正面类比于绿色小球的概率也就是$1-E_{in}$。当然从选择的角度肯定要选择犯错最小的，即正面尽可能多的情况，此例中不难发现存在全部都为正面的概率是非常大的，此处应注意，选择全为正面的或者说E_{in}为0并不正确（因为想得到的结果是$\frac{1}{32}$，而不是99%）这一结论与真实的情况或者说E_{out}差的太远（我们不仅仅要满足E_{in}很小条件，同时还要使得E_{in}与E_{out}不能有太大差距）。因此这种不好的样本的存在得到了很糟糕的结果。

上面介绍了坏的样例（bad sample），把本来很高的E_{out}，通过一个使得E_{in}的坏抽样样本进行了错误的估计。

到底是什么造成了这种错误，要深入了解。我们还需要介绍坏的数据（bad data）的概念。（这里写一下自己的理解，坏的样本bad sample∈坏的数据bad data）

坏的数据就是使得E_{in}与E_{out}相差很大时，抽样到的N个输入样本（我的理解不是这N个输入样本都不好，可能只是有几个不好的样本，导致该次抽样的数据产生不好的结果，但此次抽样的数据集被统一叫做坏的数据），根据霍夫丁不等式这种情况很少出现，但是并不代表没有，特别是当进行假设函数的选择时，它的影响会被放大，以下进行一个具体的说明

如果通过算法找出的g满足$E_{in}(g)\approx 0$，则通过PAC的规则可以保证$E_{out}(g)\approx 0$。