计算学习理论基础

南开大学网络空间安全学院 计算机学院

主要问题

- 约束归纳学习的普遍规则
- 什么样的问题是可学习的
- 何时可以信任学习算法的结果

主要目标

- 样本复杂度(sample complexity)
  - 学习算法成功（或较高概率）收敛到假设需要多少训练样本?
- 计算复杂度(computational complexity)
  - 学习算法成功（或较高概率）收敛到假设需要多少计算量?
- 错误边界(mistake bound)
  - 学习算法成功（或较高概率）收敛到假设之前对训练样本错误分类的次数是多少?

计算学习理论

- 成功学习的概率
- 训练样本的数目
- 假设空间的复杂度
- 近似目标概念的准确度
- 表达训练样本的方式

概念学习原型

- 已知
  - 实例空间 X：使用属性描述的实例集合
  - 概念空间 C：所有可能目标概念的集合
  - 假设空间 H：所有可能假设的集合
  - 训练样本集合 S：满足目标概念 $c \in C$ 的正例和反例集合

$$\{\langle x_1, c(x_1)\rangle, \langle x_2, c(x_2)\rangle, \ldots, \langle x_N, c(x_N)\rangle\}$$

- 求解
  - 问题一：对于所有 $x \in \mathcal{S}$，存在 $h \in \mathcal{H}$，满足 $h(x) = c(x)$。
  - 问题二：对于所有 $x \in \mathcal{X}$，存在 $h \in \mathcal{H}$，满足 $h(x) = c(x)$。

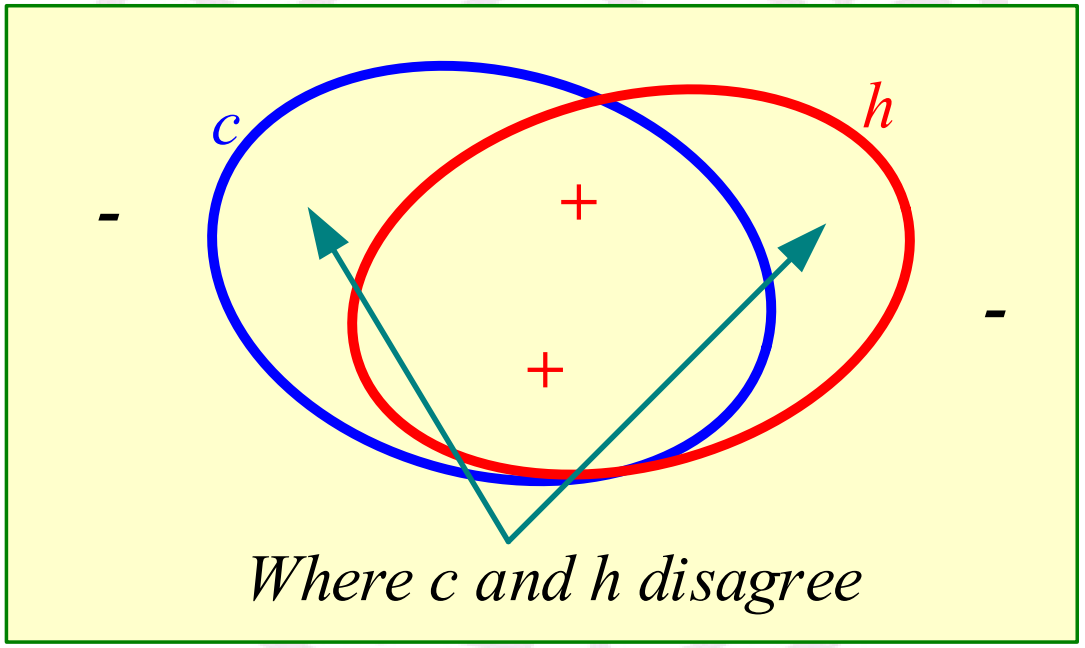
概念学习原型

- 学习过程
  - 训练样本集合 S 是随机均匀地抽取于实例空间 X 的，力图做到使二者的分布一致。
  - 通过学习 $h \in \mathcal{H}$ 来估计概念 c 。
  - 通过对独立抽取于实例空间 X 中实例 $x \in \mathcal{X}$ 的决策性能来评估假设 h 。
- 决策性能：使用错误率来度量
  - 假设 h 对目标概念 c 的样本错误率：在整个训练样本集合中，被假设错误决策的训练样本数目占训练样本总数的比例。
  - 假设 h 对目标概念 c 的真实错误率：实际样本被假设错误决策的样本数目占样本总数的比例。

真实错误率

- 定义：假设 h 对目标概念 c 和分布 $\mathscr{D}$ 的真实错误率为假设 h 对根据概率分布 $\mathscr{D}$ 独立抽取实例的错误决策概率。即：

$$\text{error}_{\mathscr{D}}(h) = \Pr_{x \in \mathscr{D}}\left[h(x) \neq c(x)\right]$$



学习合取式

- 问题：学习合取式

$$h(x_1, x_2, \ldots, x_{100}) = (x_2 = v_1) \wedge (x_3 = v_2) \wedge (x_5 = v_4) \wedge (x_{100} = v_5)$$

- 获取训练样本的方法
  - 学习算法给出样本属性向教师询问决策结果：学习算法需要进行100次询问
  - 教师提出训练样本：学习算法需要5个训练样本
  - 随机数据源提供样本，教师提供对应决策结果：所有文字组合作为候选，消除在正例中为假的那些文字。

学习合取式

- 假设：z 是一个文字(literal)，$p(z)$ 是正例中 z 为假的概率
  - 其中正例是按照分布 $\mathcal{D}$ 随机采样的
- 错误率
  - 如果　z 在目标概念中
    则　　$p(z)=0$
    否则　$p(z)$ 是随机采样正例中 z 为假的概率
    故，有

$$\text{error}(h) \le \sum_{z \in h} p(z)$$

- 结论：假设 h 只会对正例决策错误

学习合取式

- 劣质文字：在正例中存在有效的出现概率，但没有出现在训练样本正例中的文字。
- 定义：如果 $p(z) > \varepsilon/n$，则称文字 z 是劣质的，其中 n 为属性数目。
- 结论：如果不存在劣质文字，则 $\text{error}(h) < \varepsilon$。

学习合取式

- 令：z 是一个劣质文字
  - z 存在一个样本中的概率为

$$\begin{aligned}\Pr(z \text{ survives one example}) &= 1-\Pr(z \text{ is eliminated by one example}) \\ &= 1-p(z) \\ &< 1-\varepsilon/n\end{aligned}$$

  - z 存在 m 个独立样本中的概率为

$$\begin{aligned}\Pr(z \text{ suevives } m \text{ independent examples}) &= \left[1-p(z)\right]^m \\ &< (1-\varepsilon/n)^m\end{aligned}$$

- 因为最多存在 n 个劣质文字，所以某些劣质文字残存在 m 个样本正例中概率的边界为：

$$n(1-\varepsilon/n)^m$$

学习合取式

- 期望：选择足够大的 m，使得 z 残存在 m 个样本中的概率小于 δ，则

$$\Pr\left(z \text{ survives } m \text{ independent examples}\right) = n\left(1-\varepsilon/n\right)^m < \delta$$

  - 利用公式 $1-x<e^{-x}$，得到 $ne^{-m\cdot\varepsilon/n}<\delta$。
  - 保证发生错误概率小于 δ 的样本数目 m 为：

$$m > \frac{n}{\varepsilon}\left[\ln\left(n\right)+\ln\left(1/\delta\right)\right]$$

- 例
  - $\because \delta=0.1, \varepsilon=0.1, n=100, \therefore m=6907$
  - $\because \delta=0.1, \varepsilon=0.1, n=10, \therefore m=460$
  - $\because \delta=0.01, \varepsilon=0.1, n=10, \therefore m=690$

学习的期望

- 不要期望学习算法能够完全地学习某个概念
  - 一般存在着多个概念与有效训练样本相一致
  - 未见到的样本可能属于任何类别
  - 接受不在训练样本集合中且罕见样本被错误决策的事实
- 不要期望能够学习到对目标概念的最佳近似
  - 训练样本集合中的某些样本不具有代表性
- 对学习算法的实际期望是以较高的概率学习到对目标概念的最佳近似。

可近似正确学习(PCA)

- 定义：可近似正确学习(PCA)
  - 已知两个较小的参数 ε 和 δ，学习算法以至少 $1-\delta$ 的概率学习到决策错误率最多为 ε 的假设。
  - 原因：没有对概率分布进行任何前提假设

- 已知：在实例空间 X（包含 n 个文字的实例）上定义的概念 C，学习算法 L 使用假设空间 H 学习假设 h。
- 概念 C 可近似正确学习(PAC) 的条件
  - 对于所有 $f \in \mathcal{C}$
  - 对于所有在 X 上的分布 $\mathscr{D}$ 和 $0<\varepsilon,\delta<1$
  - 根据分布 $\mathscr{D}$ 独立采样包含 m 个样本的集合，学习算法以至少 $1-\delta$ 的概率学习到错误率最大为 ε 的假设 $h \in \mathcal{H}$。

可近似正确学习

- 如果学习算法 L 学习假设的时间是 $1/\delta, 1/\varepsilon, n, \text{size}(C)$ 的多项式，则称概念 C 是高效可学习的。
- 对 **PCA** 的限制
  - 多项式样本复杂度（信息论约束）：是否在训练样本中存在着足够信息来区分近似 f 的假设 h？
  - 多项式时间复杂度（算法复杂性）：是否存在着有效算法能够处理训练样本数据和生成较好的假设 h？
- 为了实现 PCA，在假设 $\mathcal{H} \supseteq \mathcal{C}$ 的前提下，对 $f \in \mathcal{C}$，必须存在具有任意小错误决策的假设 $h \in \mathcal{H}$。

一致学习算法

- 最差情形：对于任何概率分布，对于任何存在概念 $\mathcal{C}$ 中的目标概念 f，必须满足正确性。
- 一般流程
  - 已知 m 个实例的训练样本集合 $\mathcal{D}$
  - 寻找某些 $h \in \mathcal{H}$ 与所有 m 个样本相一致
- 学习合取式
  - 使用消除算法寻找与训练样本相一致的假设 h
  - 实验表明：如果拥有足够多的训练样本，则 h 接近 f。

奥坎姆剃刀

- 结论：如果存在着与 m 个样本相一致的假设 $h \in \mathcal{H}$，则假设错误率 $\text{error}(h) > \varepsilon$ 的概率小于 $|\mathcal{H}|(1-\varepsilon)^m$。
- 证明：
  - 令 h 是一个劣质假设
  - h 与 f 的一个样本相一致的概率为

$$\Pr_{x \in \mathcal{D}}\left[f(x) = h(x)\right] < 1 - \varepsilon$$

  - 因为 m 个样本的抽取是彼此独立的，所以 h 与 f 的 m 个样本相一致的概率小于 $(1-\varepsilon)^m$。
  - $\mathcal{H}$ 中的某些假设与 m 个样本相一致的概率小于

$$|\mathcal{H}|(1-\varepsilon)^m$$

  - 期望

$$|\mathcal{H}|(1-\varepsilon)^m < \delta$$

$$\ln(|\mathcal{H}|) + m\ln(1-\varepsilon) < \ln(\delta)$$

奥坎姆剃刀

利用公式 $\ln(1-\varepsilon)<-\varepsilon$，有

$$m>\frac{1}{\varepsilon}\left[\ln(|\mathcal{H}|)+\ln\left(\frac{1}{\delta}\right)\right]$$

- 上式称为奥坎姆剃刀，即：偏爱较小的假设空间。

例

- 问题：学习布尔属性的合取式
- 求解：因为每个属性有三种取值，即：不出现、正变量和反变量，所以假设空间的尺度为3^n，有

$$m > \frac{1}{\varepsilon}\left[\ln\left(3^n\right)+\ln\left(\frac{1}{\delta}\right)\right]=\frac{1}{\varepsilon}\left[n\ln 3+\ln\left(\frac{1}{\delta}\right)\right]$$

  - 如果期望保证学习算法对10个布尔变量的合取式以95%的机会学到正确率不小于90%的假设，则

$$m > \frac{1}{1-0.90}\left[10\ln 3+\ln\left(\frac{1}{1-0.95}\right)\right]=140$$

  - 学习特征为100个布尔变量时，$m=1130$，与 n 的关系是线性的。
  - 当$\delta=1\%$ 时，$m=1145$，与 δ 成对数关系。
  - 注：上述结论对所有一致学习算法均成立。

合取范式(CNF)

- 问题：学习 $f \in k-\mathrm{CNF}$（k 元合取范式）的奥坎姆算法

$$f=\bigwedge_{i=1}^{m}\left(l_{i_1}\vee l_{i_2}\vee\cdots\vee l_{i_k}\right)$$

- 求解：
  - 抽取包含 m 个样本的样本集合 $\mathcal{D}$
  - 寻找一个假设与 $\mathcal{D}$ 中的所有样本相一致
  - 确定样本复杂度

$$f=C_1\wedge C_2\wedge\ldots\wedge C_m, C_i=l_1\vee l_2\vee\cdots\vee l_k$$

$$\ln\left(\left|k-\mathrm{CNF}\right|\right)=O\left(n^k\right)$$

- 根据样本复杂度，结果假设 h 保证是一个 PAC 假设。

合取范式

- 问题：如何寻找一个一致性假设
- 求解：
  - 定义一个新特征集合

$$y_i = l_{i_1} \vee l_{i_2} \vee \cdots \vee l_{i_k}, j = 1,2,\ldots,n^k$$

  - 在新特征集合上使用学习单调合取式算法。

例

- 问题： $n=4,\ k=2,$ 单调k-CNF。

$$y_1 = x_1 \vee x_2 \quad y_2 = x_1 \vee x_3 \quad y_3 = x_1 \vee x_4$$
$$y_4 = x_2 \vee x_3 \quad y_5 = x_2 \vee x_4 \quad y_6 = x_3 \vee x_4$$

- 求解：
  - 原始样本集合：$\{(0000,1),(1010,1),(1110,1),(1111,1)\}$
  - 新样本集合：$\{(000000,1),(111101,1),(111111,1),(111111,1)\}$
  - 求解.....

无偏学习

- 问题：包含 n 个布尔特征所有函数的假设空间
- 求解：
  - n 个布尔特征组成的实例空间大小为 $|\mathcal{X}|=2^n$。
  - 无偏概念 C 对应于 X 的幂集，大小为 $|\mathcal{C}|=2^{|\mathcal{X}|}$。
  - 为了学习无偏概念，学习算法必须使用无偏假设空间，即

$$\because \mathcal{H}=\mathcal{C},\because |\mathcal{H}|=2^{2^n}$$

  - 学习无偏概念的样本复杂度为

$$m>\frac{1}{\varepsilon}\left[2^n\ln 2+\ln\left(\frac{1}{\delta}\right)\right]\cdots\cdots\cdots O\left(2^n\right)$$

无偏学习

- *k*-CNF：任意数目子句的合取范式，其中每个析取子句最多包含 k 个文字。

$$f = C_1 \wedge C_2 \wedge \ldots \wedge C_m; C_i = l_{i_1} \vee l_{i_2} \vee \ldots \vee l_{i_k}$$

$$\because |\mathcal{H}| = 2^{(2n)^k}, \therefore \ln(|k-\text{CNF}|) \to O(n^k)$$

- *k-clause*-CNF：最多 k 个析取子句的合取范式

$$f = C_1 \wedge C_2 \wedge \ldots \wedge C_k; C_i = l_1 \vee l_2 \vee \ldots \vee l_m$$

$$\because |\mathcal{H}| = 3^{kn}, \therefore \ln(|k-\text{clause-CNF}|) \to O(kn)$$

无偏学习

- *k*-DNF：任意数目合取项的析取式，其中每个合取项最多包含 k 个文字。

$$f = T_1 \vee T_2 \vee \ldots \vee T_m; T_i = l_1 \wedge l_2 \wedge \ldots \wedge l_k$$

- *k-term*-DNF：最多 k 个合取项的析取式。

$$f = T_1 \vee T_2 \vee \ldots \vee T_k; T_i = l_1 \wedge l_2 \wedge \ldots \wedge l_m$$

其中 $|\mathcal{H}| = 3^{km}$ 为过高估计，因为可能存在 $T_i = T_j$ 或者 T_i 比 T_j 更一般的情形。

无偏学习

- 学习前面四种范式的样本复杂度均是多项式阶的。
- *k-term*-DNF
  - 确定是否存在与训练样本相一致的 *k-term*-DNF 是一个 NP 难的问题。
  - 根据计算复杂度，*k-term*-DNF 类型的学习问题不是一个 PAC 可学习的。
- *k*-CNF
  - 已经提出 *k*-CNF 学习算法
  - *k*-CNF是 *k-term*-DNF 的超集，即每个 *k-term*-DNF 均可以写成 *k*-CNF 的形式

$$T_1 \vee T_2 \vee T_3 = \prod_{x \in T_1, y \in T_2, z \in T_3} \{x \vee y \vee z\}$$

$$(a \wedge b \wedge c) \vee (b \wedge d \wedge e) = \prod \{a \vee b; a \vee d; a \vee e; b; b \vee d; b \vee e; c \vee b; c \vee d; c \vee e\}$$

- 无偏概念 $\mathcal{C} = k\text{-}term\text{-DNF}$ 可以使用 $\mathcal{H} = k\text{-CNF}$ 做为假设空间来学习。

不可知学习

- 问题：使用 $\mathcal{H}$ 中的假设来学习概念 f，但是 $f \notin \mathcal{H}$。
- 目标：寻找具有最小训练决策错误的假设

$$\text{error}_{\text{train}}(h) = \frac{1}{m}\left|\{x \in \text{training-samples}: f(x) \neq h(x)\}\right|$$

- 期望：保证具有最小训练决策错误的假设对未知样本具有较好的正确性

$$\text{error}_{\mathcal{D}}(h) = \Pr_{x \in \mathcal{D}}\left[f(x) \neq h(x)\right]$$

- 霍夫丁(Hoeffding)边界说明了某些事件的真实概率与在 m 次独立实验中得到的观测频度之间的差异。

$$\Pr\left[p > \hat{p} + \varepsilon\right] \leq e^{-2\varepsilon^2 m}$$

不可知学习

- 对于$h \in \mathcal{H}$，有

$$\Pr\left[\text{error}_{\mathcal{D}}(h) > \text{error}_{\text{train}}(h) + \varepsilon\right] \leq e^{-2\varepsilon^2 m}$$

- 对于产生训练样本和测试样本的任何分布$\mathcal{D}$，当所有的$h \in \mathcal{H}$，训练样本集合的大小为m，存在

$$\text{error}_{\mathcal{D}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln(|\mathcal{H}|) + \ln\left(\frac{1}{\delta}\right)}{2m}}$$

不可知学习

- 如果，一个不可知学习算法
  - 不保证 f 是否在假设空间 $\mathcal{H}$ 中，且
  - 在包含至少 m 个训练样本的集合上，返回最小训练决策错误的一个假设 h 。
- 那么，算法保证
  - 真实错误率不超过训练错误率加上 ε 值的概率至少为 $1-\delta$。其中：

$$m > \frac{1}{2\varepsilon^2}\left[\ln\left(|\mathcal{H}|\right) + \ln\left(\frac{1}{\delta}\right)\right]$$

无限假设空间

- 前面的分析限制在有限的假设空间中
- 某些无限假设空间具有更好的表现能力
- 对于无限假设空间的表现能力需要使用假设空间大小之外的其它量来度量
- VC维就是这样一个度量量
- 与假设空间大小$|\mathcal{H}|$相似，$\mathrm{VC}(\mathcal{H})$也用于表示样本复杂度的边界。

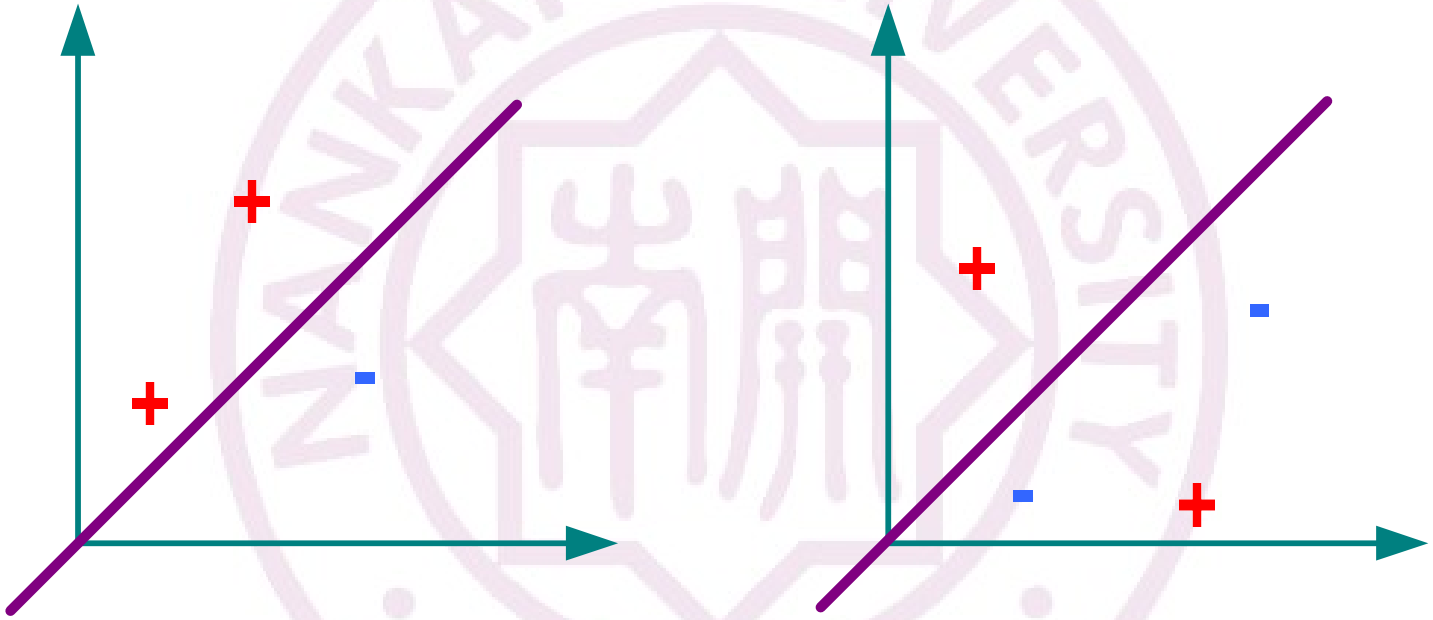
基本概念

- 二分(dichotomy)：集合 S 的二分将集合 S 分成两个互不相交的子集合。
- 打散(shatter)：如果对实例集合 S 中实例的每次正例和反例的划分，均在函数集合 H 中存在某个函数与此划分相一致，那么，我们称实例集合 S 被函数集合 H 打散。
  - 实数轴上存在左界的区间 $[0,a), a>0$，可以打散单点点集，而不能打散两点点集。

  - 实数轴上的区间 $[a,b], a>b$，可以打散两点点集，而不能打散三点点集。

基本概念

- 平面上的半空间可以打散单点点集、两点点集和三点点集，而不能打散四点点集。



VC维

- 无偏假设空间 $\mathcal{H}$ 打散整个实例空间 $\mathcal{X}$，即它能够归纳实例集合每种可能的划分。
- 能被打散的实例空间 $\mathcal{X}$ 的子集合越大，假设空间 $\mathcal{H}$ 就有越好的表现能力，即稍微有偏。
- 在实例空间 $\mathcal{X}$ 上假设空间 $\mathcal{H}$ 的 VC维是能够被 $\mathcal{H}$ 打散的实例空间 $\mathcal{X}$ 中最大有限子集的大小。
  - 如果　存在着大小为 d 的子集被打散
    则　　$\mathrm{VC}(\mathcal{H}) = d$
  - 如果　不存在着大小为 d 的子集合被打散
    则　　$\mathrm{VC}(\mathcal{H}) < d$
  - 对于前面三种打散的情形，有

$$\because d < 2, \therefore \mathrm{VC}(\text{half intervals}) = 1$$

$$\because d < 3, \therefore \mathrm{VC}(\text{intervals}) = 2$$

$$\because d < 4, \therefore \mathrm{VC}(\text{half-space in the plane}) = 3$$

使用VC维的样本复杂度

- 无限假设空间的奥坎姆算法
  - 已知包含 m 个样本的集合 S
  - 寻找与所有 m 个样本相一致的假设
  - 如果

$$m > \frac{1}{\varepsilon}\left[8 \cdot \mathrm{VC}(\mathcal{H}) \cdot \ln\frac{13}{\varepsilon} + 4 \cdot \ln\frac{2}{\delta}\right]$$

则，假设 h 决策错误率小于 ε 的概率至少为 $1-\delta$。

样本复杂度的下界

- 在一般情形下，对 PAC 学习也存在着所需最小训练样本数目的下界。
- 如果考虑任意概念集合 $\mathcal{C}$，有 $\mathrm{VC}(\mathcal{C})>2$，任意学习算法 L 和足够小的正实数 ε 与 δ，那么，存在着一种分布 $\mathscr{D}$ 和 $\mathcal{C}$ 中的目标函数，使得：
  - 如果学习算法 L 观测的样本数目小于

$$m=\max\left\{\frac{1}{\varepsilon}\ln\frac{1}{\delta},\frac{\mathrm{VC}(\mathcal{C})-1}{32\varepsilon}\right\}$$

则学习算法 L 输出错误率大于 ε 假设的概率至少为 δ。

总结

- 可近似正确学习(PAC)框架为理论分析学习算法提供了较好的模型
- 理论框架便于在各种前提下对学习算法复杂性的具体分析
- 理论结果阐明了一些重要问题，如表达式的重要性、样本复杂度和计算复杂度等
- 学习理论对实际学习系统的影响将越来越明显