统计学习理论调研报告 可能近似正确学习

张超
计研094 2009210934

摘要

本文对可能近似正确学习理论的一些基本内容结合实例做了初步的介绍，并将这些内容与统计学习理论做了比较，从而更深入的认识了可能正确学习学习理论的价值、地位以及它存在的问题。

1 关于统计学习机器的一些问题

在应用某种已知的学习机器解决机器学习的问题时，学习应使用的数据量、学习机器收敛的速度、学习过程中机器出现的错误、学习得到的函数与要学习的目标函数间的差异以及由此带来的错误率等在进行学习之前往往是未知的，通常需要依靠经验判断或者反复试验来确定。这样的学习过程是不可靠的。为了解决这个问题，在理论上得到上述问题的预期，“一些不完整的学习计算理论已经开始出现”[3]。“可能近似正确学习”就是这样的理论之一。

“可能近似正确学习”是用来解决学习机器的样本复杂度和计算复杂度问题的，

(1)样本复杂度，是指学习机器以高概率学习到某个“成功的”函数所需要的训练数据的数目。

(2)计算复杂度，是指学习机器以高概率学习到某个“成功的”函数所需要的花费的计算量。

对这两个问题的讨论将贯穿文章始终。

另外，问题中所谓“学习到某个‘成功的’函数”中的“成功”根据学习的不同要求可以有多种解释。可以认为学习机器恰好准确学习到目标函数才是成功的，但这样在实际中难以做到；也可以把问题弱化，认为算法输出的函数与目标函数在多数情况下结果一致即可。“可能近似正确学习”属于后者，这也是它名字的由来。同时，为了保证推导出的学习机器具有一般性，推导总是在对最差情况分析的基础上进行的。

本文将介绍“可能近似正确学习”理论中的一些基本内容，并将其与统计学习理论中的内容进行比较。讨论的范围将主要限制在布尔值目标函数和无噪声训练数据上。

2 可能近似正确学习

2.1 函数估计模型

基本符号

记X为所有实例的集合，X中的实例记为x。C为要学习的目标函数的集合，称为概念集，C中的概念记为c，$c : X \to \{0, 1\}$。若$c(x) = 1$则x是正例，反之x是反例。

假定每个实例按照非时变的分布$\mathscr{D}$独立的从X中随机产生，$\mathscr{D}$对学习机器未知。D为训练数据集。记H为学习机器可能学习到的所有函数的集合，称为假设集，其中的假设为h。

学习问题的一个例子

考虑以下例子。《应用随机过程》这门课通过是否交齐所有作业、是否通过期中考试、是否通过期末考试、是否上交读书报告四项指标来评判学生是否通过，指标取值为1表示学生满足该指标，取值为0表示不满足指标。假如在期末评分的过程中，任课教师中途出差，助教忘记了确切的及格标准，想从已经评判过的学生的记录中找出可能的标准，从而完成剩下的工作。

从已经评判过的学生的记录中找出评判标准相当于学习机器从训练集中学习目标函数的情况。在这里，由四种指标不同取值的各种组合，可以把所有学生分为$2 \times 2 \times 2 \times 2 = 16$种，$|X| = 16$；考虑到每种指标在目标函数中可能有真、假、不出现3种情况，所有假设共有$3 \times 3 \times 3 \times 3 = 81$种，$|H| = 81$。假如及格的标准是交齐所有作业、通过期中考试，且或通过期末考试或上交读书报告，定义$A = \{(1, 0, 1, 1), (1, 1, 1, 0), (1, 1, 1, 1)\}$则概念集$C = \{(x, 1) : \forall x \in A\} \bigcup \{(x, 0) : \forall x \in A^C\}$。

函数估计模型

样本学习的一般模型可以表述为图1的函数估计模型，产生器按照分布$\mathscr{D}$从X上抽取样例x，通过训练器得到x的真实分类$c(x)$；通过学习到假设h的学习机器得到的分类为$h(x)$。

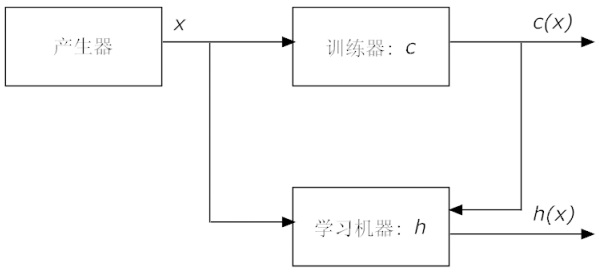

图 1: 样本学习的函数估计模型

真实错误率

由函数估计模型，把按照$\mathscr{D}$抽取的实例被h错分类

的概率，定义为假设h关于概念c和分布$\mathscr{D}$的真实错误率。记为，

$$\mathrm{error}_{\mathscr{D}}(h) \equiv P_{\mathscr{D}}(c \Delta h)$$

其中，Δ为对称差，$A \Delta B = (A \cup B) \setminus (A \cap B)$，$\mathrm{error}_{\mathscr{D}}(h)$即为随机抽取的实例落入$c$和$h$不一致的区间的概率（图2）。

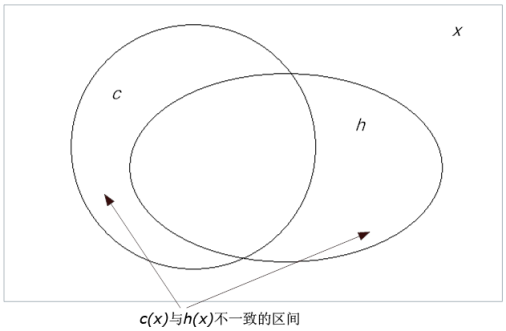

图 2: 关于概念c假设h的真实错误率

与统计学习理论的联系

按照统计学习理论中符号的定义，当定义损失泛函为

$$Q(z,\alpha) = \begin{cases} 1 & \text{if } c(x) \neq h(x) \\ 0 & \text{if } c(x) = h(x). \end{cases},$$

时，期望风险泛函[7]

$$R(\alpha) = \int Q(z,\alpha)\, dF(z) = \int_{c(x)\neq h(x)} dF(x,y)$$

其中，x在X上的分布$F(x)$即为$\mathscr{D}$，布尔值概念为确定性函数，输出条件概率$F(c(x)|x)$为1，所以

$$\begin{aligned} R(\alpha) &= \int_{c(x)\neq h(x)} dF(x,y) = \int_{c(x)\neq h(x)} d\mathscr{D} \\ &= \underset{x\in\mathscr{D}}{P}(c(x)\Delta h(x)) = \mathrm{error}_{\mathscr{D}}(h) \end{aligned}$$

于是可以看出，期望风险泛函在这种意义下就是真实错误率。

2.2 可能近似正确学习

PAC学习

如前面提到过的，由于在真实的学习问题中，通常不可能给训练器提供实例集中每个元素对应的训练样本；而随机抽取的、不完全的实例可能对学习机器产生误导。比如前面提到的判成绩的例子中，如果所有已经判为通过的学生中，没有通过上交读书报告方式来获得及格的例子，那么据此学习到的假设中就可能不会包括这种方式，从而带来误判。于是，不要求学习机器输出零错误率假设，而只要求它的错误率被限定在给定常数ε的范围内，ε可以任意的小；另外，也不要求学习机器对所有随机抽取的训练样本都可以成功输出这样的假设，而只要求失败的概率小于给定常数δ即可，δ也可以取任意小。套用例子，即助教以$1-\delta$的概率可以找到正确率为$1-\varepsilon$的替代准则。

于是，将这样学习机器可能学习到一个近似正确的假设称为可能近似正确学习（Probably Approximately Correct Learning），简称为PAC学习。下面给出可PAC学习的定义。

可PAC学习

对一个定义在长度为n的实例集X上的概念集C，以及假设空间为H的学习机器L，若对$\forall c \in C$、任何分布$\mathscr{D}$、任何$\varepsilon, 0 < \varepsilon < \frac{1}{2}$，以及任何$\delta, 0 < \delta < \frac{1}{2}$，学习机器$L$至少以概率$1-\delta$输出一个假设$h \in H$，使得$\mathrm{error}_{\mathscr{D}}(h) \leqslant \varepsilon$，且所用时间为$\frac{1}{\varepsilon}$、$\frac{1}{\delta}$、$n$和$\mathrm{size}(c)$的多项式函数，则称$C$是被$L$可PAC学习的。

这里，L必须满足能够以任意高的概率$1-\delta$输出错误率可以任意低（为ε）的假设，且这个学习过程是高效的，学习过程是多项式函数。$\mathrm{size}(c)$为假定对C采用某种描述时，c的编码的长度，例如本文中c为布尔值函数，$\mathrm{size}(c)$即为描述c的布尔特征的数量。

同时，由于样本增加也会导致学习时间变长，所以空间复杂度与时间复杂度紧密相关。即可PAC学习也隐含要求了C中的每个概念都可从多项式数目的样本被L学习到。于是，就可以用样本复杂度和时间复杂度来判断一个算法是否是可PAC学习的。

与统计学习理论的联系

统计学习理论中定义有不为零的期望风险，可能近似正确学习允许有小于任意值的真实错误率，这是两种理论相似的地方。不同之处在对待期望风险的角度上。对于统计学习理论，处理思路是在函数集上分类讨论，找到三类风险泛函对应的期望风险的上界；然后通过把期望风险最小化来构造学习机器；PAC学习中的期望风险的上界是一个任取的定值，在PAC学习中，往往是给定期望风险上界对已知的学习算法做分析。所以，两者的差别归根于在推导的出发点不同以及使用理论的方法不同。

可PAC学习中除包含真实错误率外，还允许学习机器学习失败；而统计学习理论在ERM学习机器推广能力的部分中，在推导概念性上界的时候引入了对损失函数集中所有函数同时成立的概率$1-\eta$，与$1-\delta$完全对应。可PAC学习的判定对算法的样本复杂度、计算复杂度做了限制，这两点在统计学习理论中有限样本下学习过程的收敛速度中有相关的内容，但统计学习理论中同时给出了推导出上界以及构造性上界。

3 有限假设空间的样本复杂度

由上面的讨论可以看出，可PAC学习很大程度上是由训练样本的数目决定的。随着问题规模的增长而导致所需训练样本数目的增长也称作该学习问题的样本复杂度[3]。在许多实际的学习问题中，往往由于可用训练数据的不足导致学习效果不理想，所以以下主要讨论样本复杂度[3]。

一致学习器

称一个假设在样本集D上是一致的，当且仅当$h(x) = c(x)$，即

$$\mathrm{Consistent}(h, D) = (\forall < x, c(x) > \in D), h(x) = c(x)$$

本节先对讨论的学习机器再加一条限制，只讨论所谓的一致学习器，而更一般的非一致学习器的情况在下

节。当一个学习机器可以在任何可能的时候都能够输出在训练集上一致的假设时，称这个学习机器为一致学习器。

假设在训练集上一致，即学习机器的训练错误率为零。类似真实错误率，定义训练错误率为

$$\text{error}_D(h) \equiv P_D(c\Delta h)$$

即学习到的假设在训练集上的错误率。

下面引入所谓变形空间的概念。

变形空间

$VS_{H,D} \equiv \{h \in H|\text{Consistent}(h,D)\}$称为变形空间。

即变形空间是H中所有能正确分类训练样例的假设构成的集合。

仍旧对于前述判成绩的例子，当给出训练集为$D = \{((1,0,1,1),1),((1,0,0,1),0)\}$时，假设空间$H$缩减为变形空间，$VS_{H,D} = \{((?,?,1,?),1) :? = 0\text{或}1\} \bigcup \{((?,?,0,?),0) :? = 0\text{或}1\}$。

无论X、H、D如何，每个一致学习器都输出属于一个变形空间的假设。这是因为变形空间中包含H中所有一致假设。于是，确定某个一致学习器所需的样例数量，就为确定对应变形空间中没有不可接受的假设所需要的样本的数量。

$\varepsilon-$详尽

$VS_{H,D}$对给定的c和$\mathscr{D}$被称作是$\varepsilon-$详尽的，当且仅当此时$VS_{H,D}$中的每个假设h都有小于ε的错误率，即

$$\forall h \in VS_{H,D}, \quad \text{error}_{\mathscr{D}}(h) < \varepsilon$$

下面变形空间的$\varepsilon-$详尽化给出了对任何c和D，得到$VS_{H,D}$为$\varepsilon-$详尽的概率的方法。

$\varepsilon-$详尽化

设H有限，D是目标概念c的m个$(m \geqslant 1)$独立随机抽取的样例，对任意$0 \leqslant \varepsilon \leqslant 1$，$VS_{H,D}$不是$\varepsilon-$详尽的概率小于等于

$$|H|e^{-\varepsilon m}$$

证明：

设$h_1, h_2, \cdots, h_k$是H中关于c真实错误率大于ε的所有假设，当且仅当k个假设中恰好有一个与D中所有m个样本一致时$VS_{H,D}$不能够$\varepsilon-$详尽化。其概率为，

$$C_k^1(1-\varepsilon)^m$$

又由于$k \leqslant |H|$，所以

$$k(1-\varepsilon)^m \leqslant |H|(1-\varepsilon)^m \leqslant |H|e^{-\varepsilon m}$$

又由PAC学习的定义，未能$\varepsilon-$详尽化即可能输出不满足真实错误率任意小的假设，这种失败率也应该可以任意的小，于是

$$|H|e^{-\varepsilon m} \leqslant \delta$$

可以解出m，

$$m \geqslant \frac{1}{\varepsilon}(\ln|H| + \ln(\frac{1}{\delta}))$$

由可PAC学习的定义易判断，有限假设空间中的一致学习器是可PAC学习的。

$|H|$的脆弱性

仍旧对于判成绩的例子，有$n = 4, |H| = 3^4$，给定$\varepsilon = 0.1, \delta = 0.05$，于是有，

$$m \geqslant \frac{1}{\varepsilon}(n\ln 3 + \ln(\frac{1}{\delta})) = 10 \times (4\ln 3 + \ln 20) \approx 74$$

$|X| = 16 < 74$，即算得的上界已经远超过实例集中实例的数目。可以看出，这是一个非常松的上界，这是由于使用$|H|$放缩时甚至把所有不一致假设都引入所致。

与统计学习理论的联系

对于有限假设空间上的一致学习器，如果用统计学习理论中结构风险的观点来看[7]，即

$$R_{\text{emp}}(\alpha) + \Phi(\frac{l}{h})$$

中，置信度$\Phi(\frac{l}{h})$固定，且由一致性保证经验风险$R_{\text{emp}} = 0$的学习机器。所谓变形空间，就是使经验风险最小时可能不唯一的解的集合。

又$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \Phi(\frac{l}{h}) = 0 + \Phi(\frac{l}{h})$时，$\text{error}_{\mathscr{D}}(h) < \varepsilon$，由前面的讨论已知$R(\alpha) = \text{error}_{\mathscr{D}}(h)$，所以上界$\varepsilon$与$R(\alpha)$有一定内在关联。

从另外角度来看，在统计学习理论中，在构造ERM学习机器的与分布无关的构造性界时有若[7]：$0 \leqslant Q(z,\alpha) \leqslant 1$时

$$\mathscr{E} = 2\frac{\ln N - \ln\eta}{l}$$

且，不等式

$$R(\alpha) \geqslant R_{\text{emp}}(\alpha) + \frac{B-A}{2}\sqrt{\mathscr{E}}$$

至少以$1-\eta$的概率对所有$Q(z,\alpha), \alpha \in \Lambda$成立。此时，$R_{\text{emp}}(\alpha) = 0, B = 1, A = 0$，于是得到，

$$2R^2(\alpha) \geqslant \frac{1}{l}\left(\ln N + \ln\frac{1}{\eta}\right)$$

而

$$\varepsilon \geqslant \frac{1}{m}(\ln|H| + \ln(\frac{1}{\delta}))$$

由此，可以看出虽然两种理论推导这组公式的目的不同，思路有别，但内在却有很深的联系，这是因为他们的出发点期望风险和真实错误率间有着深厚的关系。这是同一件事物被从不同角度定义的结果，殊途同归。

3.1 不可知学习和不一致假设

不可知学习器是指H中不包含概念c，于是并不总是存在一致学习器时的零训练错误率的情况。

于是下面考虑不一致假设，$\text{error}_D(h)$是训练错误率，h_{best}是H中有最小训练错误率的假设。于是有$\text{error}_{\mathscr{D}}(h_{\text{best}}) \leqslant \varepsilon + \text{error}_D(h_{\text{best}})$，即上节的一致学习器是这个问题的在$\text{error}_{\mathscr{D}}(h_{\text{best}})$时的特殊化。

使用Hoeffding边界[6]可以证得[3]，

$$P\left(\text{error}_{\mathscr{D}}(h)) > \text{error}_D(h) + \varepsilon\right) \leqslant e^{-2m\varepsilon^2}$$

可以看出边界与训练错误率有关。这样，为了保证h_{best}具有至少上面的边界，则所有$C^1_{|H|}$个假设中至少有一个具有上面的边界，即

$$P\left((\exists h \in H)(\text{error}_{\mathscr{D}}(h)) > \text{error}_D(h) + \varepsilon)\right) \leqslant |H| e^{-2m\varepsilon^2}$$

并将这个概率记为δ，则

$$m \geqslant \frac{1}{2\varepsilon^2}(\ln|H| + \ln\frac{1}{\delta})$$

3.2 可学习性

上面分析了有限假设空间下一般布尔值函数所需训练样本的下界，下面来列举几种有限假设空间中布尔值函数的PAC可学习性。

布尔文字的合取

布尔文字的合取是可PAC学习的，对n项合取，有$|H| = 3^n$，并有样本复杂度为，

$$m \geqslant \frac{1}{\varepsilon}(n\ln 3 + \ln(\frac{1}{\delta}))$$

无偏学习器

无偏的往往是无用的，概念集无偏的学习器也不例外。若X中的实例都有n个布尔值特征，于是$|X| = 2^n$；无偏的概念集为$|X|$的幂集，大小为$2^{|X|} = 2^{2^n}$。

于是，

$$m \geqslant \frac{1}{\varepsilon}(2^n\ln 2 + \ln\frac{1}{\delta})$$

样本复杂度是n的指数级，所以不是可PAC学习的。

k项DNF和$k-$CNF

由[1]，k项DNF有多项式的样本复杂度和非多项式的计算复杂度，所以不是可PAC学习的；但k项DNF包含于$k-$CNF，$k-$CNF却是可PAC学习的。

4 无限假设空间样本复杂度

4.1 VC维

示性函数集的VC维

由于X中任何实例x都是布尔文字，所以它们都可以表示成示性函数的形式。

一个示性函数集的VC维是能够被集合中的函数以所有肯能的2^d种方式分成两类的向量$z_1, \cdots, z_d$的最大数目d。如果对任意的n，总存在一个n个向量的集合可以被该函数集大打散，则函数集的VC维是无穷大[7]。

并且有

$$VC(H) = d = \log_2 2^d \leq \log_2 |H|$$

样本复杂度

所以利用示性函数的VC维在无限假设空间时代替$|H|$。经过推导，最后可以得到[1]，

$$m \geqslant \frac{1}{\varepsilon}(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\varepsilon})$$

这样，就能够解决布尔值函数的一般问题了。

与统计学习理论的联系

VC维即来自统计学习理论。

5 一般条件下的可能近似学习

文章第一部分末尾，我们将问题限制在了布尔值目标函数和无噪声训练数据上，对PAC学习这两条限制都可以取消。

(1)实值目标函数[1], [4]。

(2)在有噪声的数据中学习[5],[2]，但只有对有限几种噪声情况下的结论。

6 一些思考

PAC学习的一些优点，

(1)PAC学习提供了相对简单易行理论框架，在分析样本复杂度的同时还能够分析计算复杂度；

(2) PAC学习与变形空间、VC维等经典内容都有紧密的联系；

(3) PAC学习基于最坏情况分析，对学习机器具有一般性。

存在的一些问题，

(1) PAC学习基于最坏情况分析，以至在很多实际场合都不实用；

(2) PAC学习并不能很好的适用于训练数据有噪声的情况，因而降低了实用性；

(3) PAC学习事实上并不是一套新的学习理论，除去计算复杂度部分，统计学习理论的结构给出了它所有的内容[8],[1]。

7 总结

文章通过理论和实例介绍了布尔值目标函数、无噪声数据下可能近似学习理论在有限假设空间一致学习机器、不可知学习器以及无限假设空间中的样本复杂度等一些基本内容，并穿插了与统计学习理论的简单比较。

参考文献

[1] Blumer, A., Ehrenfeucht, A., Haussler, D., &Warmuth, M. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 34(4) (October), 929-965. 1989.

[2] Kearns, M. J., & Vazirani, U. V. *An introduction to computational learning theory.* Cambridge, MA: MIT Press, 1994.

[3] Mitchell. T. *Machine Learning.* McGraw Hill, 1997.

[4] Natarajan, B. K. *Machine Learning: A theoretical approach.* San Mateo, CA: Morgan Kaufmann, 1991.

[5] Laird, P. *Learning from good and bad data.* Dordrecht: Kluwer Academic Publishers, 1988.

[6] Vidyasagar, M. *A theory of learning and generalization : with applications to neural networks and control systems.* London ; New York : Springer, c1997

[7] V. Vapnik. *The Nature of Statistical Learning Theory.* 1995, Springer.

[8] V. Vapnik. *Statistical Learning Theory.* 1998, John Wiley and Sons.