VC 学习笔记–Hoeffing 不等式

赵得涛　　15220172202792

2019 年 10 月 25 日

1　背景

VC 维的概念最早是在1971年，V. Vapnik and A. Chervonenkis在论文“On the uniform convergence of relative frequencies of events to their probabilities”中提出的。是机器学习中重要的概念。在机器学习中，我们不知道真实模型，我们只能用近似模型估计真实模型。我们的目标是 $E_{in} \approx 0$ 和 $E_{in} \approx E_{out}$，通常为了使E_{in}最小，人们会将使用比较复杂的模型(VC很大)，甚至过度拟合。然而这样只是满足的条件一却使条件二偏离很大。VC 维反映了函数集的学习性能，VC 维越大，学习机器的学习能力越强，但学习机器也越复杂。

2　Hoeffing 不等式

2.1　定义

Hoeffding不等式是关于一组随机变量均值的概率不等式。如果X1,X2,...,Xn为一组独立同分布的参数为p的伯努利分布随机变量，n为随机变量的个数,定义这组随机变量的均值为：

$$\overline{X} = \frac{X_1 + X_2 + X_3 + ... + X_n}{n}$$

对于任意的$\delta > 0$,Hoeffding不等式可以表示为:

$$P(\left|\overline{X} - E(\overline{X})\right| \geq \delta) \leq exp(-2\delta^2 n)$$

如果知道$X_i \in [a_i, b_i]$ 时，Horffing的应用更加广泛：

$$P(\overline{X} - E(\overline{X}) \geq \delta) \leq exp - \frac{2n^2\delta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

$$P(\left|\overline{X} - E(\overline{X})\right| \geq \delta) \leq 2exp - \frac{2n^2\delta^2}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

从不等式的证明可以看出当n趋于无限大时，我们可以用$\overline{X}$有效推断$E(\overline{X})$

2.2 Learning 中的应用

假设一个学习算法的实际假设是$f(x)$(一般是未知的)，$h(x)$是采样之后的假设，y_n 为对应样本的目标值：对于任意固定的h:

$$E_{out}(h) = \epsilon[h(x) \neq f(x)]$$

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

根据Hoeffing不等式，我们可以得出：

$$P(\|E_{in} - E_{out}\| > \epsilon) \leq 2exp(-2\epsilon^2 N)$$

对任意的$\epsilon > 0$ 成立，N 表示样本大小，当N越大，说明采样假设对于真实假设的推断是越好的。Hoeffding不等式只能保证$E_{in}(h)$和$E_{out}(h)$会有一个较大的几率说明他俩是相似的，但是无法保证在非常多得假设下(h(x))下，$E_{in}(h)$的值时比较小，那如果最终$E_{in}(h)$的值比较大，那么就背离了我们学习的两个目标之一。

假定假设空间H中有M个假设$h_1, h_2, h_3, ..., h_M$,从输入空间以分布P抽取样本数量为N的训练集D1，D2，D3 ... D5678 ... (可以是无穷个),如下图：

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

图 1: 矩阵表示

BAD表示当前的假设下对应的样本E_{in} 很大。对于任何一个h，任何一个训练集，发生BAD的概率最大为 $2exp(-2\epsilon^2 N)$, 对于M个假设，我们可以使用bound的方式来求总体的BAD几率：

$$P(BAD|D) = P_D[BADforh_1orBADforh_2or...orBADforh_M]$$

$$\leq P_D[BADforh_1] + P_D[BADforh_2] + ... + P_D[BADforh_M]$$

$$\leq 2exp(-2\epsilon^2 N) + 2exp(-2\epsilon^2 N) + ... + 2exp(-2\epsilon^2 N)$$

$$= 2Mexp(-2\epsilon^2 N)$$

2.3 关于M无限大的问题

拿最简单线性模型来说，假设空间 $H = \{\beta_0 + \beta_1 X\}$,我们可以有无数个$\beta_0$和$\beta_1$ 估计，这样是否意味着概率会变成无穷？首先开始几个PLA(线性可分)的例子：

当平面上只有一个线性可分的点时，有两条线(一虚一实)，如下图: 当

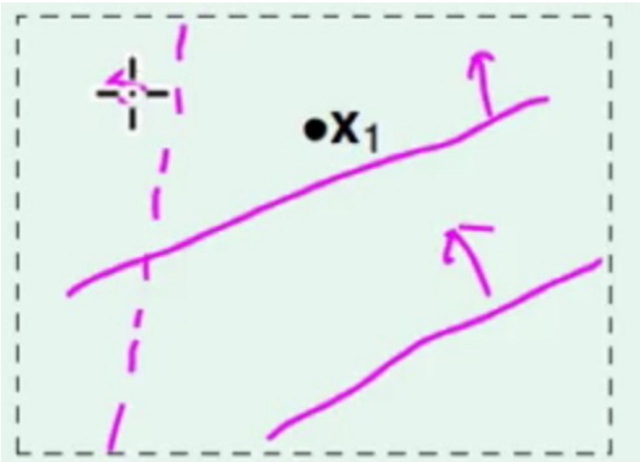

图 2: 1个线性可分得点

平面上只有一个线性可分的点时，有四条线，如下图： 在二维上，平面上

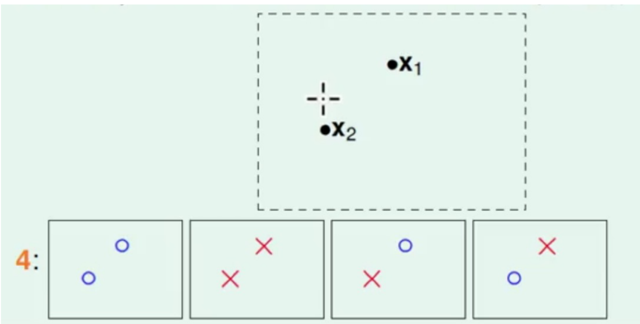

图 3: 2个线性可分得点

划分这些点的种类数总是小于2^n,现在假设使用$M_H(N)$来表示N个点的时候可能出现的最大分类线种数dichotomies:

$$M_H(N) = max \, \|H(X_1, X_2, ..., X_N)\|$$

那么这个$M_H(N)$的上界就是2^N，这个函数也叫做成长函数(Growth Function).如果$M_H(N)$能直接取代M,同时$M_H(N)$为多项式的话，N越大，那个

这个upbound会越来越小，或者接近于0，如果$M_H(N)$ 为指数的话，则不是很妙：

$$P(\|E_{in} - E_{out}\| > \epsilon) \leq 2M_H(N)exp(-2\epsilon^2 N)$$

2.4 Break point

我们希望假设空间H的增长函数越小越好，或者至少不要增长的太快，随着m的增大，一定会出现一个m使假设空间无法shatter。这种不满足的情况说明增长函数从这个点开始变缓了，所以我们把第一个不满足shatter的m值称为break point。其准确定义如下： 例如，在二维平面中，有1个点时2条

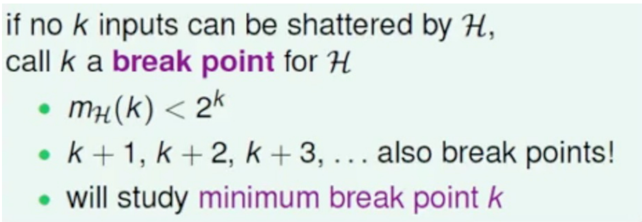

图 4: Break Point的定义

线，2个点4条线，3个点8条线，但是四个点就只有14条线而不是16条线，因此4就是二维平面的 break point. Break Point 表明增长函数在 k 这个点开始变缓,将 $M_H(N)$的上界变小或者有界，使得学习可行[1]。

3 总结

VC 维涉及的内容和概念比较广泛，本人只是介绍了Hoeffing不等式在机器学习中应用，讨论了假设函数趋于无限多的情况，引出解决该问题的一个概念 break point.

4 参考文献

[1] 《台大机器学习基石》 Hoeffding不等式
[2] https://www.zhihu.com/question/38607822/answer/149407083
[3] https://www.jianshu.com/p/97a4d3c35aa8

[1]李雨佳-VC学习笔记