VC 维学习笔记

李雨佳，统计系，15420170155241

2019 年 3 月 8 日

[摘要]：本文通过整理与 VC 维相关的重要概念，对分（Dichotomy）、增长函数（Growth Function）、打散（Shatter）、Break Point，对 VC 维的概念作出理解。

[关键词]：对分；增长函数；打散；Break Point；VC 维

VC 维（Vapnik-Chervonenkis Dimension）是机器学习中很基础同时也很重要的概念，于 1971 年由 Vapnik 和 Chervonenkis[1]提出。VC 维用于衡量假设空间的容量。这个容量反映了函数集的复杂程度，并通过评估集合内函数的摇摆程度来衡量函数集的表现力、丰富性和灵活性[2]。VC 维反映了函数集的学习性能，一般而言，VC 维越大，学习机器的学习能力越强，但学习机器也越复杂。要学习 VC 维，首先需要了解 Dichotomy、Growth Function、Shattering、Break Point 的概念。

1 相关概念理解

1.1 可学习的条件——基于 Hoeffding 不等式

根据 Hoeffding 不等式，我们有：

[1] Vapnik V N , Chervonenkis A Y . On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities[J]. Theory of Probability and Its Applications, 1971, 17(2):264-280.

[2] Sewell M.,VC Dimension[EB/OL].http://www.svms.org/vc-dimension/vc-dimension.pdf,2008.

$$P_r[|E_{in}(h) - E_{out}(h)| > \epsilon] \leq 2\exp(-2\epsilon^2 N) \qquad (1.1)$$

假定假设空间 H 中有 M 个假设，$h_1, h_2, ...h_m$, 则有：

$$P_r[E(h_1) > \epsilon \cup E(h_2) > \epsilon ... \cup E(h_M) > \epsilon]$$

$$\leq P_r[E(h_1) > \epsilon] + P_r[E(h_2) > \epsilon] + ... + P_r[E(h_M) > \epsilon]$$

$$\leq 2M\exp(-2\epsilon^2 N)$$

其中 $E(h_i) = |E_{in}(h_i) - E_{out}(h_i)|$。根据上式可知：

$$\forall g \in H, P_r[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 2M\exp(-2\epsilon^2 N) \qquad (1.2)$$

式（1.2）的含义即在假设空间 H 中，对于任意一个假设 g,$E_{in}(g)$和$E_{out}(g)$的差距大于一个任意小的值 ϵ 的概率的上界是 $2M\exp(-2\epsilon^2 N)$。当 M 较小时，有式（1.2）可知，$E_{out}(g)$ 比较接近 $E_{in}(g)$，但由于可选的 g 的数目 M 较小，我们无法实现 $E_{in}(g)$ 尽可能的小。当 M 较大时，由于可选的 g 的数目 M 较大，$E_{in}(g)$ 能够实现尽可能的小，但 $E_{out}(g) \approx E_{in}(g)$ 难以实现。因此要满足可学习的条件，需要合理选取假设数 M。在一个假设空间中，其假设数常常是很大甚至是无限的，这样就无法实现约束的意义，因此需要引入有效假设数概念。

1.2 有效假设数

通俗的理解，假设一个二分问题，共有 N 个数据，那么对于整个训练集，最多的分类可能性有 2^N 个。如果不限制模型的种类，则不论 M 取多大，假设空间 H 中的有效模型只有 2^N 个，大于这个数量的假设空间中，必然存在两个模型的分类结果是完全一样的。在现实中，假设空间里的 h_i 之间并不是完全独立，即可以归类的。在限制了模型的种类情况下，不论 M 取值多大，假设空间中有效的模型应该是小于 2^N

用二维假设空间来理解，当有 3 个不同的数据点时，假设空间 H 种的无数条直线可以分为 8 类，即 2^3，如图 1.1 所示。但在 4 个数据点时，H 中最多有 14 类而不是 16 类。

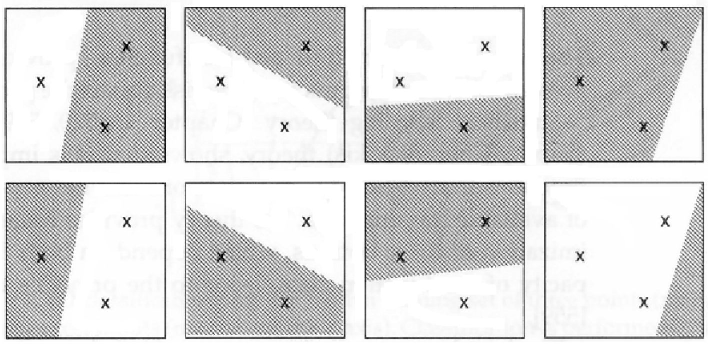

图 1.1

因此，可以定义在特定的样本集上，有效的假设数 M_{eff} 满足：

$$\forall g \in H, P_r[|E_{in}(g) - E_{out}(g)| > \epsilon] \le 2M_{eff}\exp(-2\epsilon^2 N) \qquad (1.3)$$

1.3 对分和增长函数

对分的定义：对于假设空间 $H=\{h:\chi\rightarrow\{+1，-1\}\}$，称 $h(X_1,X_2,...,X_N)=(h(X_1),h(X_2),...,(h(X_N))\in\{+1，-1\}^N$ 为一个对分，一个对分表示样本的一种标记结果，$H(X_1,X_2,...,X_N)$ 表示假设空间 H 在训练集 D 上的所有对分。

通俗的理解，在平面里用一条直线对 2 个点进行二元分类，输出结果可能为 {1,-1}，{-1,1}，{1,1}，{-1,-1}，这样每个输出向量就称为一个对分。对分与有效假设数的关系为 $M_{eff}=H$作用于 D“最多”能产生的不同对分数。由此可见，“H作用于 D“最多”能产生多少不同的对分”是取决于 具体的数据集 D 的，为了摆脱对 D 的依赖，需要引入增长函数。

增长函数的定义：假设空间 H 的增长函数 $m_H(N)=\max\limits_{X_1,X_2,...,X_N\in\chi}|H(X_1,X_2,...,X_N)|$，即增长函数表示假设空间 H 对任意 N 个样本所能赋予标记的最大可能结果数，其上界为 2^N。因此增长函数的引入将 H 的势 M 极大的缩小至 2^N，它代表的是真正能得出不同结果的、有意义的假设的最大数量。$m_H(N)$ 越大，H 的表示能力越强。因此增长函数反映了假设空间的表示能力和复杂度。

2 VC 维

2.1VC 界（Bound）理解

2.1.1 打散

打散的概念：当假设空间 H 作用于大小为 N 的样本集 D 时，产生的对分数量等于 2^N，即 $m_H(N)=2^N$ 时，就称 D 被 H 打散了。意思是，N 个点的所有可能情形都被 H 产生了。相对应的就是不打散，有些情况下，增长函数达不到对应的 2^N 值，如在二维实平面上的线性划分情况中，存在 4 个点时，$m_H(N)=14\neq 2^4$，如图 2.1 所示。

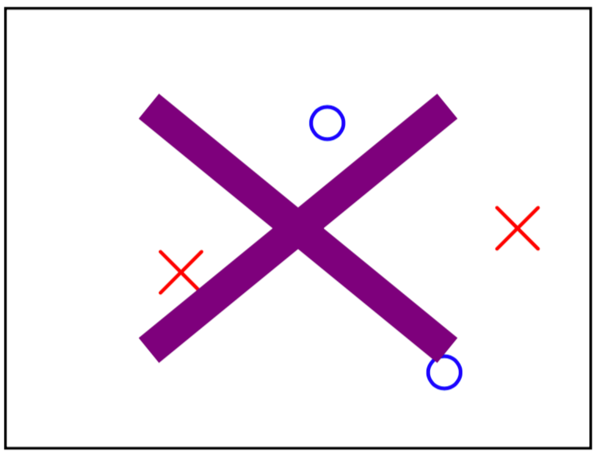

图 2.1

2.1.2BreakPoint

虽然增长函数把假设数从无穷缩小到 2^N，但是这个量级仍然太大，因此就有了 Break Point 的概念。

它的定义是：对于假设空间 H 的增长函数 $m_H(N)$，从 $N=1$ 出发逐渐增大，当增大到 k 时，出现 $m_H(N)<2^N$ 的情形，就称 k 是 H 的 BreakPoint。通俗的理解就是：对于任何大小为 N（$N\geq k$）的数据集，H 无法打碎它。

运用 Sauer's Lemma 可以证明 BreakPoint 存在且为 k 的假设空间的

增长函数上界为 $B(N,k)$，它满足：

$$m_H(N) \le B(N,k) \le \sum_{i=0}^{k-1} \binom{N}{i} \le N^{k-1}$$

最后一个不等式仅在 $N \ge 2$ 且 $k \ge 2$ 时成立。

Break Point 表明增长函数在 k 这个点开始变缓，它的意义是将式（1.3）的上界缩小了，使得学习可行了。

2.1.2VC 界

以上的关系可以理解为：对分的数量上界是增长函数，增长函数的上界是 $B(N,k)$；如果 Break Point 存在，则增长函数是多项式的，它远小于 2^N。也就是说，H 作用于样本量为 N 的样本集 D，其方程数量看似无穷，但真正有效的数量是有限的，这个数量为 $m_H(N)$。

因此现在的问题是，是否能用 $m_H(N)$ 替换掉式（1.3）中的 M。可以证明不能直接替换，因为 H 中每一个 h 作用于 D 都能算出一个 E_{in}, 一共有 $m_H(N)$ 个 E_{in}，但是 E_{out} 的可能取值是无限的，所以就有 VC 界，即正确的替换是：

$$\forall g \in H, P_r[|E_{in}(g) - E_{out}(g)| > \epsilon] \le 4m_H(2N)\exp(-\frac{1}{8}\epsilon^2 N) \qquad (2.1)$$

VC 界的意义就是：（1）如果假设空间 H 存在有限的 Break Point k，即 $m_H(2N)$ 会被最高幂次为 k-1 的多项式约束住，那么随着 N 的逐渐增大，$\exp(-\frac{1}{8}\epsilon^2 N)$ 的下降速度比 $m_H(2N)$ 的增长速度快得多。（2）当 N 足够大时，对于 H 中的任意 g，E_{in} 都接近于 $E_{out}(g)$, 即学习是可行的。

2.2VC 维理解

VC 维的定义：假设空间 H 的 VC 维是能被 H 打散的最大数据集的大小，即:

$$VC(H) = max\{N : m_H(N) = 2^N\} \qquad (2.2)$$

对于一个假设空间 H，如果存在 m 个数据样本能够被假设空间 H 中的函数按所有可能的 2^h 种形式分开，则称假设空间 H 能够把 m 个数据样本打散。假设空间的 VC 维就是能打散的最大数据样本数目。根据式（2.2）有

VC(H) = k − 1，结合式（2.1）和 $m_H(N) \leq N^{k-1}$ 可得：

$$\forall g \in H, P_r[|E_{in}(g) - E_{out}(g)| > \epsilon] \leq 4(2N)^{VC(H)} \exp(-\frac{1}{8}\epsilon^2 N) \qquad (2.3)$$

假设空间 H 的 VC 维是有限的，且 N 足够大时，可保证式（2.3）上界趋于 0，即从 H 中选出的任意假设 g 都满足 $E_{out}(g) \approx E_{in}(g)$。

VC 维的大小与数据集、学习算法、目标函数都无关，只与假设空间有关。VC 维越大，能够学习到的模型越复杂，也就是说它反映了函数集的学习能力。但是它同时也类似一个惩罚项，VC 维越大，$E_{in}(g)$ 与 $E_{out}(g)$ 越远，泛化能力越差。

3 小结

总的来看，VC 维是机器学习中一个很基础的概念，它给机器学习可学性提供了理论支撑。VC 维可以帮助研究者选择风险更低的模型。

4 参考文献

[1] Vapnik V N , Chervonenkis A Y . On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities[J]. Theory of Probability and Its Applications, 1971, 17(2):264-280.

[2] Sewell M.,VC Dimension[EB/OL].http://www.svms.org/vc-dimension/vc-dimension.pdf,2008.

[3] Abumostafa Y S, Magdonismail M, Lin H T. Learning from Data: A Short Course[J]. Amlbook, 2012.

[4] 周志华. 机器学习 [M]. 清华大学出版社, 2016.