VC-dimension

方佳艳

1. 概述

对于大数据问题，通常采用随机采样的方法（random sampling），例如，机器学习算法所使用的各种各样的训练集。面对海量数据，我们需要用样本性质来估计整体数据的性质。如何确定样本大小来更真实反映出原始数据的特性，估计的误差和随机采样的样本大小之间满足何种关系，VC-dimension 提供了背后的理论依据。值得注意的是，这些使用的训练集并不等同于真实世界中的数据。例如，做图片/人脸识别，真实世界中的图像有无穷多个，但机器学习算法的训练集时有限的。因此，需要证明所使用的训练集的数量是够用的，或者是由于所选的数据集所导致的误差在多大的概率范围之内是可以接受的。VC-dimension 为以上这些问题提供了理论保证，因而是整个机器学习理论的基础。

2. 几个相关定义

2.1. Range Space

定义一个二元组 $\sum=(X,R)$ 是一个 range space。其中，X 是一个有限/无限集合，称为 Ground Set，$R=\{r|r是X的一个子集\}$（$r是X的一个range$）。

例如，X：二维平面上的所有点；r：二维平面上的圆/三角形/矩形

2.2. 投影

给定一个 $\sum=(X,R)$，$Y\subseteq X$，定义 $P_R(Y)=\{r\cap Y|r\in R\}$ 是 $Y在R上的投影$。

2.3. Shattered（打散）

如果 Y 是有限的，并且 $|P_R(Y)|=2^{|Y|}$，则称 Y 能被 R “打散”（shattered）。

2.4. 举例

①r：平面上的圆；Y：R^2 上的 3 个点

Y 有 2^3 个子集，易知，$R=\{平面上的圆\}$ 可以 shatter 平面上的 3 个点。

②r：平面上的圆；Y：R^2 上的 4 个点，则 Y 不能被 R shatter

思考：当 r 是平面上的矩形时，情况如何？

3. VC-dimension 理论

3.1. 定义

给定一个 $\sum=(X,R)$，$VC(\sum):=$ 最大的能被打散的 X 的子集大小，即 maximum

non-break point。

3.2. 举例

$$VC((R^2,\text{圆}))=3;\quad VC((R^2,\text{多边形}))=+\infty;\quad VC((R^d,\text{球}))=d+1$$

3.3. 针对机器学习算法的 VC-dimension [1]

记 H 为一个机器学习算法的假设空间（Hypothesis Set），$d_{VC}(H)$ 称为该假设空间的 VC-维。因此，VC-dimension 可以看作是假设空间的性质。不同的假设空间，它可以 shatter 掉的点数 N 的能力是不同的。从 VC-dimension 的角度而言，某个假设空间可以 shatter 掉某 N 个点（这 N 个点从某个 probability distribution 中产生），不一定是所有的 N 个点，但是的确存在一种 N 个点的摆放情形，使得这种情形可以产生的二分数量（dichotomy）为 2^N。但是，一旦超过这个假设空间的 VC-dimension 的时候，就不存在任何一种点的分布情形（即，不存在任何一个 probability distribution 产生的数据集）产生最大完全的二分数量了。

d_{VC} = ‘minimum k’ —1，其中，k 为 non-break point，即，最小的不能被打散的点数。

当 break point 不存在时，VC-dimension $\to\infty$

$N\le d_{VC}\Rightarrow H$ can shatter some N inputs.

$N>d_{VC}\Rightarrow N$ is a breake point for H.

给定假设空间 H 和数据集 $D=\{x_1,x_2,\cdots,x_m\}$，H 中的每个假设 h 都能对 D 中的数据赋予标记，标记结果可表示为

$$h\big|_D=\{(h(x_1),h(x_2),\cdots,h(x_m))\}$$

随着 m 的增大，H 中所有假设对 D 中的数据所能赋予标记的可能结果数也会增大。

定理 3.3.1. [2] 对所有 $m\in\mathrm{N},0<\varepsilon<1$和任意$h\in H,$有

$$P(\left|E(h)-\hat{E}(h)\right|>\varepsilon)\le 4\Pi_H(2m)\exp(-\frac{m\varepsilon^2}{8})$$

假设空间 H 中不同的假设对于 D 中数据赋予标记的结果可能相同，也可能不同；尽管 H 可能包含无穷多个假设，但其对 D 中数据赋予标记的可能结果数是有限的：对 m 个数据，最多有 2^m 个可能结果。对二分类问题来说，H 中的假设对 D 中数据赋予标记的每种可能结果称为对 D 的一种“对分”。若假设空间 H 能实现数据集 D 上的所有对分，即 $\Pi\big|_H(m)=2^m$，则称数据集 D 能被假设空间 H “打散”。

现在可以正式定义假设空间的 VC-dimension：

定义 3.3.2. 假设空间 H 的 VC-dimension 是能被 H 打散的最大数据集的大小，即

$$VC(H)=\max\{m:\Pi\big|_H(m)=2^m\}$$

$VC(H)=d$ 表明存在大小为 d 的数据集能被假设空间 H 打散。注意：这并不意味着所

有大小为d的数据集都能被假设空间 H 打散。因为，VC-dimension 的定义与数据分布D无关。因此，在数据分布未知时，仍能计算出假设空间 H 的 VC-dimension。

通常这样来计算 H 的 VC-dimension：若存在大小为 d 的数据集能被 H 打散，但不存在任何大小为 d+1 的数据集能被 H 打散，则 H 的 VC-dimension 是 d。

VC-dimension 和增长函数有着密切联系，以下引理给出了二者之间的定量关系：

引理 3.3.3. 若假设空间 H 的 VC-dimension 是 d，则对任意$m \in \mathrm{N}$有

$$\Pi_H(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

证明： 由数学归纳法证明。当$m=1, d=0$或$d=1$时，定理成立。假设定理对$(m-1,d-1)$和$(m-1,d)$成立。令$D=\{x_1,x_2,\cdots,x_m\}, D^{'}=\{x_1,x_2,\cdots,x_{m-1}\}$，

$$H\big|_D = \{(h(x_1),h(x_2),\cdots,h(x_m))\big|h \in H\}$$

$$H\big|_{D^{'}} = \{(h(x_1),h(x_2),\cdots,h(x_{m-1}))\big|h \in H\}$$

任何假设$h \in H$对x_m的分类结果或为+1，或为-1，因此任何出现在$H\big|_{D^{'}}$中的串都会在$H\big|_D$中出现一次或两次。令$H_{D^{'}|D}$表示在$H\big|_D$中出现两次的$H\big|_{D^{'}}$中串组成的集合，即

$$H_{D^{'}|D} = \{(y_1,y_2,\cdots,y_{m-1}) \in H_{|D^{'}} \big| \exists h, h^{'} \in H,$$
$$(h(x_i)=h^{'}(x_i)=y_i) \wedge (h(x_m) \ne h^{'}(x_m)), 1 \le i \le m-1\}$$

考虑到$H_{D^{'}|D}$中的串在$H_{|D}$中出现了两次，但在$H_{|D^{'}}$中仅出现了一次，有

$$\left|H_{|D}\right| = \left|H_{|D^{'}}\right| + \left|H_{D^{'}|D}\right|$$

$D^{'}$的大小为$m-1$，由假设可得：

$$\left|H_{|D^{'}}\right| \le \Pi_H(m-1) \le \sum_{i=0}^{d} \binom{m-1}{d}$$

令Q表示能被$H_{D^{'}|D}$打散的集合，由$H_{D^{'}|D}$定义可知$Q \cup \{x_m\}$必能被$H_{|D}$打散。由于 H 的 VC 维为 d，因此$H_{D^{'}|D}$的 VC 维最大为 d-1，于是有

$$\left|H_{D^{'}|D}\right| \le \Pi_H(m-1) \le \sum_{i=0}^{d-1} \binom{m-1}{i}$$

由上述式子可得：

$$\begin{aligned}\left|H_{|D}\right| &\le \sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{d-1}\binom{m-1}{i} \\ &= \sum_{i=0}^{d}\left(\binom{m-1}{i}+\binom{m-1}{i-1}\right) \\ &= \sum_{i=0}^{d}\binom{m}{i}\end{aligned}$$

由集合 D 的任意性，引理得证。

从引理 3.3.3.可计算出增长函数的上界：

推论 3.3.4. 若假设空间 H 的 VC 维为 d，则对任意整数 $m \ge d$ 有

$$\Pi_H(m) \le \left(\frac{e\cdot m}{d}\right)^d$$

证明：

$$\begin{aligned}\Pi_H(m) &\le \sum_{i=0}^{d}\binom{m}{i} \\ &\le \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} \\ &\le \left(\frac{m}{d}\right)^d\sum_{i=0}^{d}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \\ &\le \left(\frac{m}{d}\right)^d\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \\ &\le \left(\frac{m}{d}\right)^d\left(1+\frac{d}{m}\right)^{m} \\ &\le \left(\frac{e\cdot m}{d}\right)^d\end{aligned}$$

根据推论 3.3.4.和定理 3.3.1.可得基于 VC 维的泛化误差界：

定理 3.3.5. 若假设空间 H 的 VC 维为 d，则对任意 $m > d, 0 < \delta < 1$ 和 $h \in H,$ 有

$$P\left(E(h)-\hat{E}(h) \le \sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta}}{m}}\right) \ge 1-\delta$$

证明： 令 $4\Pi_H(2m)\exp\left(-\frac{m\varepsilon^2}{8}\right) \le 4\left(\frac{2em}{d}\right)^d\exp\left(-\frac{m\varepsilon^2}{8}\right) = \delta$，解得

$$\varepsilon = \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}$$

代入定理 3.3.1.，于是定理 3.3.5.

由定理 3.3.5.可知，泛化误差界只与样例数目 m 有关，收敛速率为$O(1/\sqrt{m})$，与数据分布和具体的数据集无关。因此，基于 VC 维的泛化误差界是分布无关（distribution-free）、数据独立（data-independent）的。

3.4. 重要结论

如果$VC((X,R)) = d$，令

$$R^{\cup K} = \{R\text{中}k\text{个}range\text{的并}\}$$
$$R^{\cap K} = \{R\text{中}k\text{个}range\text{的交}\}$$

$\Sigma_1 = (X, R^{\cup K}), \Sigma_2 = (X, R^{\cap K})$,则有：

$$dk \le VC(\Sigma_1), VC(\Sigma_2) \le dk\log k$$

4. 两个与随机采样相关的定理

4.1. $\varepsilon - net(0 < \varepsilon < 1)$

$Q \subseteq X$ 是$\sum = (X,R)$的$\varepsilon - net$，如果$\forall r \in R$，

$$\frac{X \cap r}{X} > \varepsilon \Rightarrow Q \cap r \neq \varnothing$$

4.2. $\varepsilon - sample(0 < \varepsilon < 1)$

$Q \subseteq X$ 是$\sum = (X,R)$的$\varepsilon - sample$，如果$\forall r \in R$，

$$\left|\frac{X \cap r}{X} - \frac{Q \cap r}{Q}\right| < \varepsilon$$

4.3. $\varepsilon - net$和$\varepsilon - sample$的比较

假如Q是$\varepsilon - sample$：

$$\frac{|X\cap r|}{|X|}>\varepsilon \Rightarrow \frac{|Q\cap R|}{|Q|}>0$$
$$\Rightarrow Q\cap r\neq\varnothing$$
$$\Rightarrow Q是\varepsilon-net$$

因此，$\varepsilon-sample$比$\varepsilon-net$的定义更强。

4.4. 两个定理

定理 4.4.1.

假设$\sum=(X,R)$的$VC-\dim=d$,Q是X的均匀采样，$\varepsilon,\delta\in(0,1)$

$$|Q|=\Theta\left(\frac{1}{\varepsilon^2}\left(d\log\frac{d}{\varepsilon}+\log\frac{1}{\delta}\right)\right),$$
$$\approx\tilde{\Theta}\left(\frac{1}{\varepsilon^2}d\right),(与|X|无关)$$

则Q是$\varepsilon-sample$的概率$\geq 1-\delta$

定理 4.4.2.

假设$\sum=(X,R)$的$VC-\dim=d$,Q是X的均匀采样，$\varepsilon,\delta\in(0,1)$

$$|Q|\geq\max\left\{\frac{4}{\varepsilon}\log\frac{2}{\delta},\frac{8d}{\varepsilon}\log\frac{8d}{\varepsilon}\right\}$$
$$\approx\Theta\left(\frac{d}{\varepsilon}\right)$$

则Q是$\varepsilon-net$的概率$\geq 1-\delta$

5. $\varepsilon-sample$ 和 $\varepsilon-net$ 的应用

$\varepsilon-sample$ 的应用：range counting

$\varepsilon-net$ 的应用：SVM（超平面定义了一种 range）

6. Probably Approximately Correct Learning (PAC Learning) [3]

定理 6.1. 任何 VC 维有限的假设空间 H 都是（不可知）PAC 可学习的。

证明：假设 $\Im$ 为满足经验风险最小化原则的算法，h 为学习算法 $\Im$ 输出的假设。令 g 表示 H 中具有最小泛化误差的假设，即

$$E(g)=\min_{h\in H}E(h)$$

令

$$\delta'=\frac{\delta}{2},$$

$$\sqrt{\frac{\left(\ln 2/\delta^{'}\right)}{2m}}=\frac{\varepsilon}{2},$$

由于

$$\hat{E}(g)-\frac{\varepsilon}{2}\le E(g)\le\hat{E}(g)+\frac{\varepsilon}{2}$$

至少以$1-\delta/2$的概率成立。令

$$\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta^{'}}}{m}}=\frac{\varepsilon}{2},$$

则由定理 3.3.5.可知

$$P\left(E(h)-\hat{E}(h)\le\frac{\varepsilon}{2}\right)\ge 1-\frac{\delta}{2},$$

从而可知

$$\begin{aligned}E(h)-E(g)&\le\hat{E}(h)+\frac{\varepsilon}{2}-\left(\hat{E}(g)-\frac{\varepsilon}{2}\right)\\&=\hat{E}(h)-\hat{E}(g)+\varepsilon\\&\le\varepsilon\end{aligned}$$

以至少$1-\delta$的概率成立。由

$\sqrt{\frac{\left(\ln 2/\delta^{'}\right)}{2m}}=\frac{\varepsilon}{2}$ 和 $\sqrt{\frac{8d\ln\frac{2em}{d}+8\ln\frac{4}{\delta^{'}}}{m}}=\frac{\varepsilon}{2}$可以解出$m$，再由 H 的任意性可知该定理得证。

7. 参考文献

[1] Abu-Mostafa Y S, Magdon-Ismail M, Lin H T. Learning from data[M]. New York, NY, USA:: AMLBook, 2012.

[2] Vapnik V N, Chervonenkis A Y. Theory of uniform convergence of frequencie of appearance of attributes to their probabilities and problems of defining optimal solution by empiric data[J]. Avtomatika i Telemekhanika, 1971, 2: 42-53.

[3] 周志华. 机器学习[M]. Qing hua da xue chu ban she, 2016.