大数据算法-2025 春

Lecture 14: VC Dimension

2025.4.10

Lecturer: 丁虎 *Scribe:* 张嘉贤

VC Dimension（Vapnik-Chervonenkis Dimension）是衡量一个假设类（hypothesis class）表达能力和复杂度的重要工具。它由统计学习理论的奠基人 Vladimir Vapnik 和 Alexey Chervonenkis 提出，用于分析机器学习模型的泛化能力。对于 SVM 问题，我们希望找到一个具有最大间隔的分类超平面，从而将正负样本分离开，但是，我们在训练数据集上得到的一个良好的分类超平面，如何能在新的测试数据点上较好的分类呢？对于图片识别或者人脸识别任务，我们的训练数据集总归是有限的，但是现实世界的图片或者人脸却可以是无穷多的，那么我们怎么能保证在有限数据集上训练的模型能很好的针对无穷的现实世界的数据呢？VC dimension 为以上这些问题提供了理论保证，它描述了一个模型能够 “完美拟合” 的最复杂数据集的大小。VC dimension 越高，模型的拟合能力越强，但也可能更容易过拟合；VC dimension 越低，模型更简单，但可能欠拟合。

1 基础知识

Definition 1.1 (Range Space)**.** Range Space 是一个二元组 $\Sigma = (X, R)$，其中 X 是一个有限或者无限的集合，称为 Ground Set，$R = \{r \mid r$ 是X 的一个子集$\}$，其中一个 r 可以称为一个 Range

比如说，X 可以指一个二维平面 R^2，r 可以是二维平面里的一个圆。那么 R 就是二维平面上所有圆的集合，$\Sigma = (X, R)$ 就构成了一个 Range Space

Definition 1.2 (投影)**.** 给定一个 Range Space $\Sigma = (X, R)$ 和 X 的一个子集 $Y \subseteq X$，那么 $P_R(Y) = \{r \cap Y \mid r \subseteq R\}$ 称为 Y 在 R 上的投影

比如说，对于上面举例的那个 Range Space $\Sigma = (X, R)$，给定 Y 是二维平面上的一个半空间 $x > 0$，那么这个半空间和所有 r 的交集便是 Y 在这个 Range Space 的一个投影

Definition 1.3 (打散)**.** 如果 Y 是一个有限集合，并且 $|P_R(Y)| = 2^{|Y|}$，则称 Y 能被 R 打散 (shattered)

比如说，在上面的例子里，如果 Y 包含三个不共线的点，那么它可以被 R 打散；如果 Y 包含四个点，那么它不可以被 R 打散

Definition 1.4 (VC Dimension)**.** 给定一个 Range Space $\Sigma = (X, R)$，它的 VC Dimension $VC(\Sigma)$ 为 X 上能被打散的最大子集的大小

对于不同的 Range Space 的 $\Sigma = (X, R)$，给出几个 VC Dimension 的例子

1. X 是二维平面 R^2，r 是二维平面里的一个圆，那么 $VC(\Sigma) = 3$
2. X 是三维平面 R^3，r 是三维平面里的一个球，那么 $VC(\Sigma) = 4$
3. X 是二维平面 R^2，r 是二维平面里的一个多边形，那么 $VC(\Sigma) = +\infty$

一个 Range Space 的 VC Dimension 由他的 Ground Set 和 Range 决定。当给定 Ground Set 时，VC Dimension 是衡量 Range 的复杂程度的量，从直觉上看，如果 Range 越复杂，那么 Range Space 的 VC Dimension 就越大

Theorem 1.5. *如果有一个 Range Space $\Sigma = (X, R)$，它的 VC Dimension $VC(\Sigma) = d$，那么定义两个不同的集合：*

$$R^{\cap k} = \{R\text{中}k\text{个 Range 的并集}\}$$

$$R^{\cup k} = \{R\text{中}k\text{个 Range 的交集}\}$$

令 $\Sigma_1 = (X, R^{\cap k})$，$\Sigma_2 = (X, R^{\cup k})$，则这两个 Range Space 的 VC Dimension 在 dk 和 $dk \log k$ 之间，即

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

2 两个重要结论

Definition 2.1. （ε-net 和 ε-sample）对于一个 $\varepsilon \in (0, 1)$，一个 Range Space $\Sigma = (X, R)$ 和 X 的一个子集 $Q \subseteq X$

1. （ε-net）如果对于任意 $r \in R$ 和 $\frac{|X \cap r|}{|X|} > \varepsilon$，有 $Q \cap r \neq \emptyset$，则 Q 是 Σ 的一个 ε-net
2. （ε-sample）如果对于任意 $r \in R$ 有 $|\frac{|X \cap r|}{|X|} - \frac{|Q \cap r|}{|Q|}| < \varepsilon$，则 Q 是 Σ 的一个 ε-sample

如果 Q 是 Σ 的一个 ε-sample，那么对于任意 $r \in R$ 和 $\frac{|X \cap r|}{|X|} > \varepsilon$，由于

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

则有

$$\frac{|Q \cap r|}{|Q|} > 0$$

于是 $Q \cap r \neq \emptyset$，则 Q 是 Σ 的一个 ε-net。可以发现，ε-sample 是一个比 ε-net 更强的条件，如果 Q 是 Σ 的一个 ε-sample，那么它就一定是 Σ 的一个 ε-net。

Theorem 2.2. 假设一个 *Range Space* $\Sigma = (X, R)$ 的 *VC Dimension = d,* Q 是 X 的均匀采样，$\varepsilon, \delta \in (0, 1)$ 是两个参数，如果

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

$$\approx \widetilde{\Theta}\left(\frac{1}{\varepsilon^2} d\right) \text{(与} X \text{ 无关)}$$

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

Theorem 2.3. 假设一个 *Range Space* $\Sigma = (X, R)$ 的 *VC Dimension = d,* Q 是 X 的均匀采样，$\varepsilon, \delta \in (0, 1)$ 是两个参数，如果

$$|Q| = \max\left\{\frac{4}{\varepsilon} \log \frac{2}{\delta}, \frac{8d}{\varepsilon} \log \frac{8d}{\varepsilon}\right\}$$

$$\approx \widetilde{\Theta}\left(\frac{d}{\varepsilon}\right)$$

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

假如说有一个巨大的包含正负样本的数据集，数据集的数据维度为 d，在该数据集上存在一个良好的分类超平面能很好的分开正样本和负样本，那么我们根据这两个定理，可以得出，当我们的 SVM 分类器的训练数据集的大小大概为 $\widetilde{\Theta}\left(\frac{d}{\varepsilon}\right)$ 时，我们可以保证，对任何一个新来的数据点，分类出错的概率小于 ε