目录

计算学习理论

0.1 基础知识

0.1.1 Jensen 不等式

对任意凸函数 $f(x)$ 有:

$$f(\mathbb{E}(x)) \leq \mathbb{E}(f(x))$$

先证明这个结论: $f(x)$ 是凸函数 ($f^{''}(x) \geq 0$), 证明:

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) \qquad \text{s.t. } p_1 + p_2 = 1, 0 < p_1 < 1$$

证明: 当 $p1 = p2$ 时, 等号成立。不妨设 $x_2 > x_1$, 则:

$$p_1 x_1 + p_2 x_2 - x_1 = (p_2 x_2) - (1 - p_1) x_1 = p_2(x_2 - x_1) > 0 \tag{1}$$

所以: $p_1 x_1 + p_2 x_2 > x_1$, 所以有: $x_1 < z_1 < p_1 x_1 + p_2 x_2$。
由拉格朗日中值定理:

$$\frac{f(p_1 x_1 + p_2 x_2) - f(x_1)}{(p_1 x_1 + p_2 x_2) - x_1} = f'(z_1)$$

即:

$$f(p_1 x_1 + p_2 x_2) - f(x_1) = [(p_1 x_1 + p_2 x_2) - x_1] f'(z_1) = p_2(x_2 - x_1) f'(z_1) \tag{2}$$

同理:

$$p_1 x_1 + p_2 x_2 - x_2 = (p_1 x_1) - (1 - p_2) x_2 = p_1(x_1 - x_2) < 0 \tag{3}$$

所以: $p_1 x_1 + p_2 x_2 < x_2$, 所以有: $p_1 x_1 + p_2 x_2 < z_2 < x_2$ 。
由拉格朗日中值定理:

$$\frac{f(x_2) - f(p_1 x_1 + p_2 x_2)}{x_2 - (p_1 x_1 + p_2 x_2)} = f'(z_2)$$

即:

$$f(x_2) - f(p_1 x_1 + p_2 x_2) = [x_2 - (p_1 x_1 + p_2 x_2)] f'(z_2) = p_1(x_2 - x_1) f'(z_2) \tag{4}$$

因为 $z_1 < z_2$ 且 $f(x)$ 为凸函数, 所以 $f''(x) \geq 0$, 所以: $f'(z_1) \leq f'(z_2)$。

将 $(2) \times p_1 - (4) \times p_2$ 可得:

$$p_1 p_2 (x_2 - x_1) f'(z_1) - p_1 p_2 (x_2 - x_1) f'(z_2) \leq 0$$

所以:

$$p_1 [f(p_1 x_1 + p_2 x_2) - f(x_1)] \leq p_2 [f(x_2) - f(p_1 x_1 + p_2 x_2)]$$

$$f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)$$

接着用数学归纳法证明最开始的 Jensen 不等式。不妨将原式子转化为:

$$f(\sum_{i=1}^{k} p_i x_i) \leq \sum_{i=1}^{k} p_i f(x_i)$$

当 $k+1$ 时:

$$\begin{aligned}\sum_{i=1}^{k+1} p_i f(x_i) &= p_{k+1} f(x_{k+1}) + z_k \sum_{i=1}^{k} \frac{p_i}{z_k} f(x_i) \qquad z_k = \sum_{i=1}^{k} p_i \\ &\geq p_{k+1} f(x_{k+1}) + z_k f(\sum_{i=1}^{k} \frac{p_i}{z_k} p_{x_i}) \\ &\geq f(p_{k+1} x_{k+1} + z_k \sum_{i=1}^{k} \frac{p_i}{z_k} p_{x_i}) \qquad z_k + p_{k+1} = 1 \\ &= f(\sum_{i=1}^{k+1} p_i x_i)\end{aligned}$$

0.1.2 Hoeffding 不等式

若 $x_1, x_2, \cdots, x_m$ 为 m 个独立随机变量，且满足 $0 \leq x_i \leq 1$, 则对任意 $\epsilon > 0$，有:

$$P\left(\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}(x_i) \geq \epsilon\right) \leq \exp(-2m\epsilon^2)$$

$$P\left(|\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}(x_i)| \geq \epsilon\right) \leq 2\exp(-2m\epsilon^2)$$

0.1.3 McDiarmid

不等式若 $x_1, x_2, \cdots, x_m$ 为 m 个独立随机变量，且对任意 $1 \leq i \leq m$，函数 f 满足:

$$\sup_{x_1, \cdots, x_m, x_i'} |f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_m)| \leq c_i$$

则对任意 $\epsilon > 0$，有:

$$P(f(x_1, \cdots, x_m) - \mathbb{E}(f(x_1, \cdots, x_m)) \leq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_i c_i^2})$$

$$P(|f(x_1, \cdots, x_m) - \mathbb{E}(f(x_1, \cdots, x_m))| \leq \epsilon) \leq 2\exp(\frac{-2\epsilon^2}{\sum_i c_i^2})$$

0.2 VC 维

VC 维 (Vapnik-Chervonenkis dimension) 中的一些概念。

增长函数 (growth function)$\Pi_{\mathcal{H}}(m)$ 表示假设空间 $\mathcal{H}$ 对 m 个示例所能赋予标记的最大可能结果数。如果第 i 个示例有 c_i 中标记，那么答案就是 $\prod c_i$。增长函数描述了假设空间 $\mathcal{H}$ 的表示能力，由此反映出假设空间的复杂度。

对分 (dichotomy)：就是对示例集赋予标记的每一种可能结果。

打散 (shattering)：假设空间可以实现示例集上的所有对分，那么称示例集能被假设空间打散。

假设空间 $\mathcal{H}$ 的 VC 维是能被 $\mathcal{H}$ 打散的最大示例集的大小。VC 维理解为给定一些已知位置标签未知的点集的集合，取其中满足这样条件的点集的最大基数作为 VC 维，条件是：对于这个固定已知位置的点集的任何一种标签赋予方式总是线性可分的。也因此二维实平面上所有线性划分构成的假设空间的 VC 维是 3。

VC 维的泛化误差界是分布无关 (distribution-free)、数据独立 (data-independent) 的。

证明：$\mathbb{R}^d$ 空间中线性超平面构成的假设空间的 VC 维是 $d+1$.
先证 VC 维 $\geq d+1$，即是证 VC 维 $= d+1$ 是可以成立的。

设数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$，因为是 $\mathbb{R}^d$ 空间中，所以 $\boldsymbol{x}_i$ 中有 d 个属性，即 $\boldsymbol{x}_i \in \mathbb{R}^{1\times d}$，令 $\boldsymbol{X} = [\boldsymbol{x}_1; \boldsymbol{x}_2; \cdots, \boldsymbol{x}_n], \boldsymbol{Y} = [y_1, y_2, \cdots, y_n]$，存在线性划分即是存在 $\boldsymbol{w}, \boldsymbol{B}$ 使得如下成立：

$$\boldsymbol{X}\boldsymbol{w} + \boldsymbol{B} = \boldsymbol{Y}$$

其中 $\boldsymbol{B} = [b_1; b_2; \cdots; b_n]$。如果令 $\boldsymbol{X} = [[\boldsymbol{x}_1, 1]; [\boldsymbol{x}_2, 1]; \cdots; [\boldsymbol{x}_n, 1];]$；上式可以改写成：

$$\boldsymbol{X}\boldsymbol{W} = \boldsymbol{Y}$$

VC 维 $= d+1$ 成立，即是 $n = d+1$ 时上式可以成立，注意到此时 $\boldsymbol{X} \in \mathbb{R}^{(d+1)\times(d+1)}$，所以只要令 $rank(\boldsymbol{X}) = d+1$，此时 $\boldsymbol{W}$ 一定有解：

$$\boldsymbol{W} = \boldsymbol{X}^{-1}\boldsymbol{Y}$$

所以 VC 维 $= d+1$ 是成立的，也就是 VC 维 $\geq d+1$ 是成立的。

下面证明 VC 维 $\leq d+1$，也就是 VC 维 $= d+2$ 是不成立的。当 VC 维 $= d+2$ 时，即 $n = d+2$，此时 $\boldsymbol{X} \in \mathbb{R}^{(d+2)\times(d+1)}$，所以 $rank(\boldsymbol{X}) \leq d+1$，所以一定存在矩阵 $\boldsymbol{X}$ 中的一行是其他行的线性组合，即：

$$\boldsymbol{x}_j = \sum_{i, i\neq j} a_i \boldsymbol{x}_i$$

此时根据 $\boldsymbol{X}\boldsymbol{W}$ 计算 $\boldsymbol{x}_j$ 的解已经由其他的 $\boldsymbol{x}_i$ 决定了：

$$\boldsymbol{x}_j W = \sum_{i, i\neq j} a_i \boldsymbol{x}_i \boldsymbol{W} = \sum_{i, i\neq j} a_i y_i$$

那么只要令 $y_j \neq \text{sign}(\sum_{i, i\neq j} a_i y_i)$，$\boldsymbol{X}\boldsymbol{W} = \boldsymbol{Y}$ 就一定无解。也就是对任意的 $\boldsymbol{X}$，总能找到相应的标签 $\boldsymbol{Y}$ 使其无解。所以 VC 维不能等于 $d+2$。

综上：$\mathbb{R}^d$ 空间中线性超平面构成的假设空间的 VC 维是 $d+1$。

0.3 Rademacher 复杂度

Rademacher 复杂度 (Rademacher complexity) 是另一种刻画假设空间复杂度的途径，与 VC 维不同的是，它在一定程度上考虑了数据分布。

Rademacher 随机变量 σ_i：它以 0.5 的概率取值为-1，以 0.5 的概率取值为 +1.

函数空间 $\mathcal{F}$ 关于 Z 的**经验 Rademacher 复杂度**为：

$$\hat{\mathrm{R}}_Z(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\delta}}\left[\sup_{f\in\mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \delta_i f(\boldsymbol{z}_i)\right]$$

经验 Rademacher 复杂度衡量了函数空间 $\mathcal{F}$ 与随机噪声在集合 Z 中的相关性。

函数空间 $\mathcal{F}$ 关于 $\mathcal{Z}$ 上**分布 $\mathcal{D}$ 的 Rademacher 复杂度**：

$$\mathrm{R}_m(\mathcal{F}) = \mathbb{E}_{Z\subseteq\mathcal{Z}:|Z|=m}\left[\hat{\mathrm{R}}_Z(\mathcal{F})\right]$$

基于 Rademacher 复杂度可得关于函数空间 $\mathcal{F}$ 的泛化误差界。

计算学习理论领域最好的学术会议是：国际计算学习理论会议 (COLT).