机器学习
Machine Learning

第5讲：计算学习理论简介

Brief Computational Learning Theory

1. PAC，概率近似正确

- PAC：Probably Approximately Correct。
- 主要由Valiant（1984）等发展的一种理论，2010年获图灵奖，第一位机器学习领域的图灵奖得主。
- Vapnik和Chervonenkis提出VC维并研究了无限假设空间的计算理论。
- 参考书：
  - Shai Shalev-Shwartz & Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge Press, 2014.（中文版，张文生等，“理解机器学习”，机械工业出版社，2016）
  - M. Mohri, et al. Foundation of Machine Learning, MIT Press, 2012, (中文版，张文生译，机器学习基础，机械出版社)

Chernoff bound

2. 一个引理　Hoeffding不等式

设　$Z_1, Z_2, \cdots, Z_N$

是N个独立同分布的随机变量，均服从伯努利分布，且

$$P(Z_i = 1) = \mu$$

定义样本均值为

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Z_i$$

令：$\varepsilon > 0$　为一个固定值，则

$$P(|\mu - \hat{\mu}| > \varepsilon) \leq 2\exp\left(-2\varepsilon^2 N\right)$$

3. 二元分类问题

$$y \in \{0, 1\}$$

训练集来自概率 $\mathcal{D}$

$$\boldsymbol{D} = \left\{\left(\boldsymbol{x}_i, y_i\right)\right\}_{i=1}^{N}$$

训练误差定义

empirical risk or **empirical error**

$$\hat{R}(h) = \frac{1}{N}\sum_{i=1}^{N} I\left(h(\boldsymbol{x}_i) \neq y_i\right)$$

二元分类问题（续）

泛化误差

$$R(h) = P_{(\boldsymbol{x},y)\sim p_{\mathscr{D}}}\left(h(\boldsymbol{x}) \neq y\right)$$

假设集 **hypothesis class $\mathcal{H}$**

$$h^* = \arg\min_{h\in\mathscr{H}} R(h)$$ 使得泛化误差最小的假设

实际中

$$\hat{h} = \arg\min_{h\in\mathscr{H}} \hat{R}(h)$$

是通过训练集按ERM得到最优假设

empirical risk minimization (ERM)

二元分类问题（续）

假设集的例子

2分类假设 $h(\boldsymbol{x}) = I\left(\overline{\boldsymbol{w}}^{\mathrm{T}}\overline{\boldsymbol{x}} \geq 0\right)$

$$\mathscr{H} = \left\{ h_{\overline{\boldsymbol{w}}} \,\middle|\, h_{\overline{\boldsymbol{w}}}(\boldsymbol{x}) = I\left(\overline{\boldsymbol{w}}^{\mathrm{T}}\overline{\boldsymbol{x}} \geq 0\right), \overline{\boldsymbol{w}} \in \mathbf{R}^{K+1} \right\}$$

ERM解转化为对如下参数的求解

$$\hat{h} = \arg\min_{\overline{\boldsymbol{w}} \in \mathbf{R}^{K+1}} \hat{R}\left(h_{\overline{\boldsymbol{w}}}\right)$$

4. 假设集为有限集合的界

$\mathcal{H}$ 为有界集 $\mathcal{H}=\{h_1,h_2,\cdots,h_K\}$ $K=|\mathcal{H}|$

1.对一个任意 h，比较其训练误差和泛化误差关系

取一个 $h_i \in \mathcal{H}$

定义伯努利变量 $Z=I\left(h_k(\boldsymbol{x})\neq y\right)$

$$Z_i=I\left(h_k(\boldsymbol{x}_i)\neq y_i\right)$$

训练误差 $\hat{R}\left(h_k\right)=\frac{1}{N}\sum_{i=1}^{N}Z_j$

假设集为有限集合的界（续）

利用引理的Hoeffding不等式，得

$$P\left(\left|R\left(h_k\right)-\hat{R}\left(h_k\right)\right|>\varepsilon\right)\leq 2\exp\left(-2\varepsilon^2 N\right)$$

用 A_k 表示 $\left|R\left(h_k\right)-\hat{R}\left(h_k\right)\right|>\varepsilon$

故：

$$P\left(A_k\right)\leq 2\exp\left(-2\varepsilon^2 N\right)$$

假设集为有限集合的界（续）

$$P\left(\exists h\in\mathcal{H},\left|R(h)-\hat{R}(h)\right|>\varepsilon\right)=P\left(A_1\cup A_2\cup\cdots\cup A_K\right)$$

$$\le\sum_{k=1}^{|\mathcal{H}|}P\left(A_k\right)$$

$$\le\sum_{k=1}^{|\mathcal{H}|}2\exp\left(-2\varepsilon^2N\right)$$

$$=2\left|\mathcal{H}\right|\exp\left(-2\varepsilon^2N\right)$$

故有

$$P\left(\left|R(h)-\hat{R}(h)\right|\le\varepsilon,\ \ \forall h\in\mathcal{H}\right)\ge 1-2\left|\mathcal{H}\right|\exp\left(-2\varepsilon^2N\right)$$

假设集为有限集合的界（续）

取 $\delta = 2|\mathcal{H}|\exp\left(-2\varepsilon^2 N\right)$

得 $$N = \frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}$$ （#1）

对于任意 $h \in \mathcal{H}$，以概率至少，$1 - \delta$ 满足 $|R(h) - \hat{R}(h)| \leq \varepsilon$ 需要的样本集满足（#1），且有

$$\left|R(h) - \hat{R}(h)\right| \leq \sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}$$

误差界定理

定理对于假设空间$\mathscr{H}$，固定δ, N，则以概率不小于$1-\delta$，泛化误差与训练误差满足

$$\|R(h)-\hat{R}(h)| \le \sqrt{\frac{1}{2N}\ln\frac{2|\mathscr{H}|}{\delta}}$$

或固定δ, ε，若样本数目取

$$N \ge \frac{1}{2\varepsilon^2}\ln\frac{2|\mathscr{H}|}{\delta}$$

则，以概率不小于$1-\delta$满足$\left|R(h)-\hat{R}(h)\right| \le \varepsilon$。

假设集为有限集合的界（续）

2. 经验误差最小假设和泛化误差最小假设的界

$$\hat{h} = \arg\min_{h \in \mathscr{H}} \hat{R}(h)$$ 经验误差最小假设

$$h^* = \arg\min_{h \in \mathscr{H}} R(h)$$ 泛化误差最小假设

得到泛化误差

$$\begin{aligned} R(\hat{h}) &\le \hat{R}(\hat{h}) + \varepsilon \\ &\le \hat{R}(h^*) + \varepsilon \\ &\le R(h^*) + 2\varepsilon \end{aligned}$$

假设集为有限集合的界（续）

泛化误差界的基本定理

定理 对于假设空间 $\mathcal{H}$，固定 δ, N，则以概率不小于 $1-\delta$，泛化误差满足如下不等式

$$R\left(\hat{h}\right) \leq \min_{h \in \mathcal{H}} R\left(h\right) + 2\sqrt{\frac{1}{2N} \ln \frac{2\left|\mathcal{H}\right|}{\delta}}$$

假设集为有限集合的界（续）

（样本复杂度的基本推论）

或固定δ, ε，若样本数目取

$$N \geq \frac{1}{2\varepsilon^2} \ln \frac{2|\mathcal{H}|}{\delta}$$

则，以概率不小于$1-\delta$满足$R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + 2\varepsilon$。

可记为：

$$N \sim O_{\varepsilon,\delta}\left(\ln|\mathcal{H}|\right)$$

5. 假设集为无限集合的界　$|H| = \infty$

VC维的定义

打散：对于一个包含 d 点的集合 $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d\}$，

其中 $\boldsymbol{x}_i \in \mathcal{X}$，称 $\mathcal{H}$ 可打散 S 是指：集合 S 对应加

上一个任意标注集 $\{y_1, y_2, \cdots, y_d\}$，则一定存在 $h \in \mathcal{H}$，

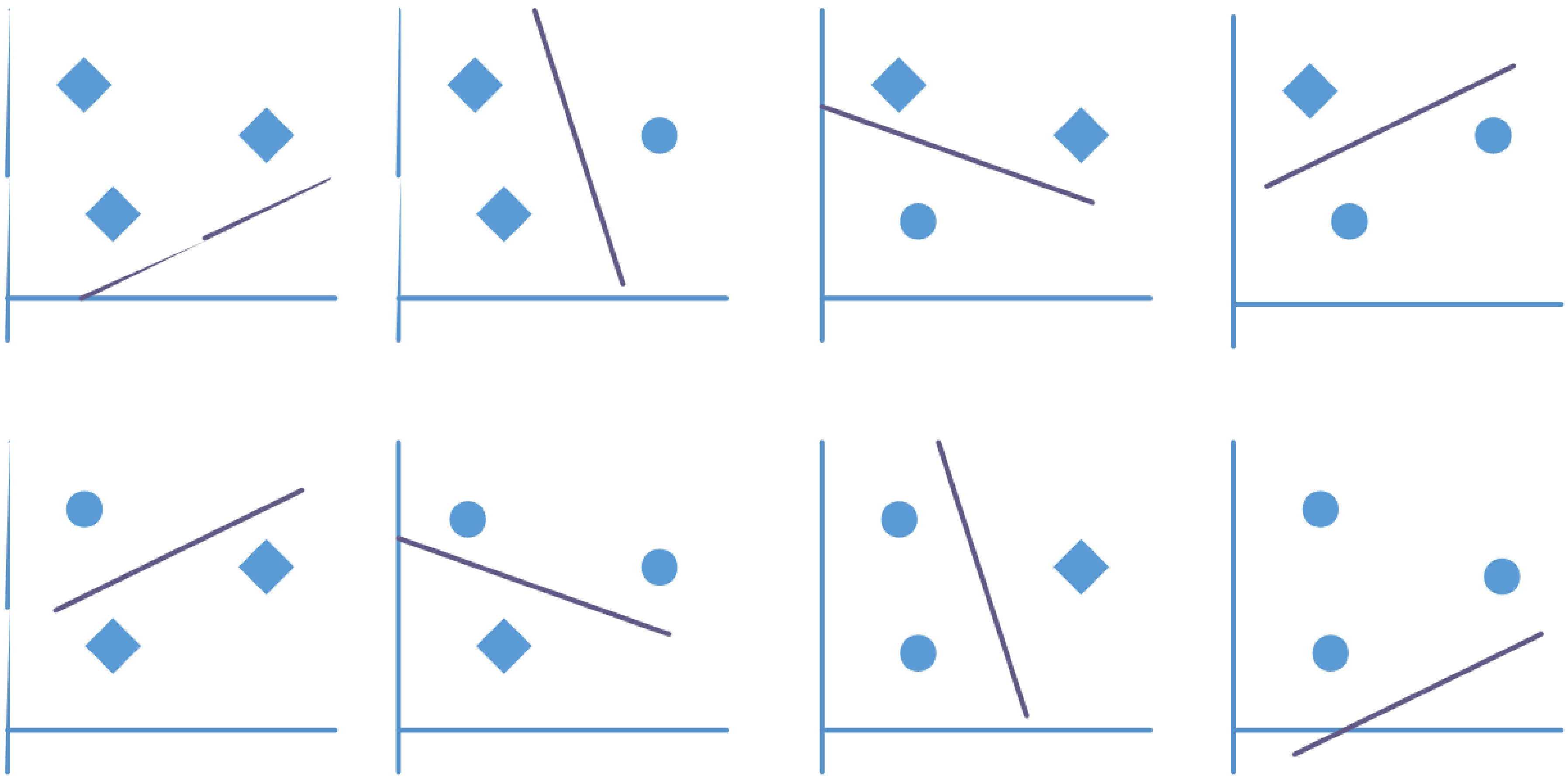
使得 $h(\boldsymbol{x}_i) = y_i$，　$i = 1, 2, \cdots, d$。

VC维的定义

VC 维的定义：对于一个假设空间 $\mathcal{H}$，至少存在一个最大元素数为 d 的点集合 S，$\mathcal{H}$ 可打散 S，则 $\mathcal{H}$ 的 VC 维为 d，记为 $VC(\mathcal{H})=d$。

这里 d 是最大能被 $\mathcal{H}$ 打散集合的元素数，对于有 $d+1$ 元素的点集合，$\mathcal{H}$ 均不可能打散它。

VC维的例子：二维线性分类器 $\mathrm{VC}(\mathcal{H}) = 3$



假设集为无限集合的界（续）

无限集合的界受VC维控制

定理： 对于假设空间 $\mathcal{H}$，若其 VC 维为 $d = VC(\mathcal{H})$，

则对于所有 $h \in \mathcal{H}$，以概率不小于 $1-\delta$，有如下不等式

$$\left|R(h) - \hat{R}(h)\right| \le O\left(\sqrt{\frac{d}{N}\ln\frac{N}{d} + \frac{1}{N}\ln\frac{1}{\delta}}\right)$$

对于 $\hat{h}$ 有不等式

$$R(\hat{h}) \le \min_{h \in \mathcal{H}} R(h) + O\left(\sqrt{\frac{d}{N}\ln\frac{N}{d} + \frac{1}{N}\ln\frac{1}{\delta}}\right)$$

假设集为无限集合的界（续）

样本复杂度

对于以概率不小于$1-\delta$满足$R(\hat{h}) \le \min_{h\in\mathcal{H}} R(h)+2\varepsilon$，

样本数目需要求

$$N \sim O_{\varepsilon,\delta}(d)$$