计算学习理论

Outline

Topic

简介

- 学习器（机器的或非机器的）应遵循什么样的规则？
- 学习器所考虑的假设空间的大小和复杂度
- 目标概念须近似到怎样的精度
- 学习器输出成功的假设的可能性
- 训练样例提供给学习器的方式

目标是为了回答以下的问题：

- 样本复杂度（Sample complexity）。学习器要收敛到成功假设（以较高的概率），需要多少训练样例？
- 计算复杂度 (Computational complexity)。学习器要收敛到成功假设（以较高的概率）需要多大的计算量？
- 出错界限（Mistake bound）。在成功收敛到一个假设前，学习器对训练样例的误分类有多少次？

概念学习任务

已知：

- 实例集合 X: 实例以其属性表示：
  $\{Sky, AirTemp, Humidity, Wind, Water, Forecast\}$
- 目标函数 c: $EnjoySport : X \to \{0, 1\}$
- 假设空间 H: 布尔文字合取，如

$$\langle ?, Cold, High, ?, ?, ? \rangle.$$

- 训练样例集合 D: 目标函数的正/负样例

$$\langle x_1, c(x_1) \rangle, \ldots \langle x_m, c(x_m) \rangle$$

求：

- H 中的假设 h，使 D 中的所有 x 满足 $h(x) = c(x)$
- H 中的假设 h，使 X 中的所有 x 满足 $h(x) = c(x)$

Topic

样本复杂度

需要多少样例才能学习目标概念？

- 学习器提出实例，询问施教者
  - 学习器提出实例 x, 施教者提供 $c(x)$
- 施教者提供训练样例
  - 施教者提供样例序列：$\langle x, c(x) \rangle$
- 以随机过程提出实例
  - 随机产生实例 x , 施教者提供 $c(x)$

样本复杂度 1

学习器提出实例 x, 施教者提供 $c(x)$
(假定 c 在学习器的假设空间 H 中)
最优询问策略: play 20 questions

- 选取实例 x，使得 VS 中一半的假设将 x 分类为正，一半分类为负
  - 可行时，需要 $\lceil \log_2 |H| \rceil$ 次查询可学习 c
  - 不可行时，需要更多样例

样本复杂度 2

施教者提供样例序列：$\langle x, c(x)\rangle$
(假定 c 在学习器的假设空间 H 中)
最优施教策略: 依赖学习器使用的 H

- 考虑 $H =$ 最多 n 个布尔文字及其否定的合取
  - 如：$(AirTemp = Warm) \wedge (Wind = Strong)$，其中 $AirTemp, Wind, \ldots$ 各有两个可能取值。
- 若 H 中有 n 个可能布尔属性 H, 只需 $n + 1$ 个样例

样本复杂度 3

已知:

- 实例集合 X
- 假设集合 H
- 目标概念集合 C
- 训练实例由一个 X 上的固定但未知的概率分布 $\mathcal{D}$ 产生

针对一些目标概念 $c \in C$，学习器观察到训练样例（形如 $\langle x, c(x) \rangle$ ）序列 D，

- 实例 x 从分布 $\mathcal{D}$ 中抽取
- 施教者提供 $c(x)$

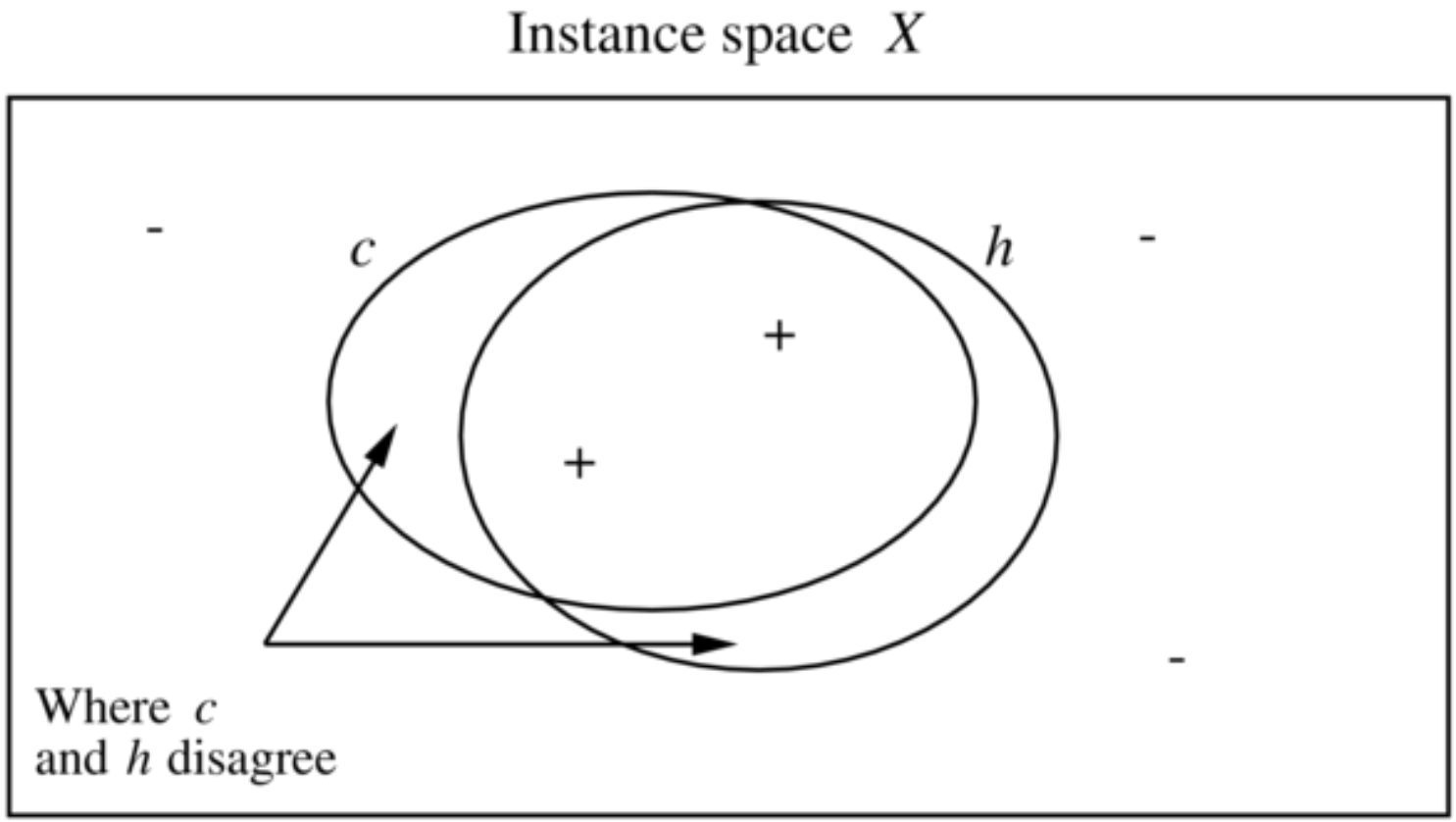
学习器必须输出一个假设 h 来估计 c

- 在后续按 $\mathcal{D}$ 抽取的实例上评估 h 的性能

注意: 随机抽取实例，类别无噪声

Topic

假设的错误率



假设的错误率

定义：假设 h 关于目标概念 c 和分布 $\mathcal{D}$ 的真实错误率 (true error) 为 h 误分类实例（按 $\mathcal{D}$ 随机抽取）的概率。

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

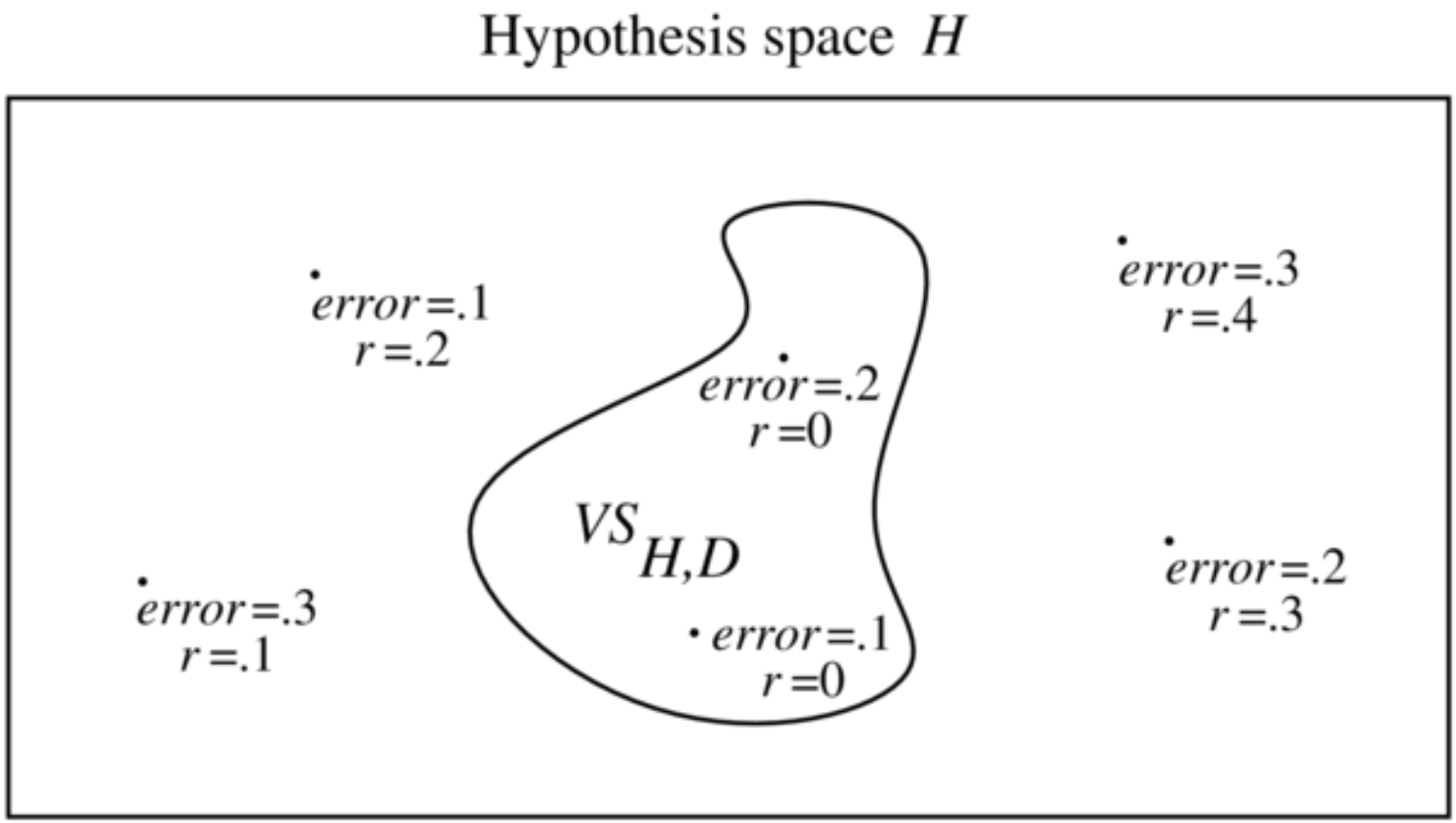
这里符号 $\Pr_{x \in \mathcal{D}}$ 代表在实例分布 $\mathcal{D}$ 上计算概率。

两种错误率

- 针对目标概念 c，假设 h 的训练错误率
  - 在训练实例中 $h(x) \neq c(x)$ 的比例
- 针对目标概念 c，假设 h 的真实错误率
  - 未来抽取的随机实例中 $h(x) \neq c(x)$ 的比例
- 给定 h 的训练错误率是否可确定 h 的真实错误率的界限
  - 先考虑 h 的训练错误率为 0 的情况 (如，$h \in VS_{H,D}$)

Topic

变型空间详尽化


Hypothesis space H
error =.1
r =.2
error =.2
r =0
error =.3
r =.4
$VS_{H,D}$
error =.3
r =.1
error =.1
r =0
error =.2
r =.3

变型空间详尽化

定义：考虑一假设空间 H，目标概念 c，实例分布 $\mathcal{D}$ 以及 c 的一组训练样例 D。当 $VS_{H,D}$ 中每个假设 h 关于 c 和 $\mathcal{D}$ 错误率小于 ϵ 时，变型空间被称为关于 c 和 $\mathcal{D}$ 是 ϵ - 详尽的（ϵ -exhausted）。

多少样例可使变型空间 ϵ - 详尽化

{**Theorem:**} **[Haussler, 1988].**

变型空间的 ϵ -详尽化 (ϵ ***-exhausting the version space)****：若假设空间 H 有限，且 D 为目标概念 c 的一系列 $m \geq 1$ 个独立随机抽取的样例，那么对于任意 $0 \leq \epsilon \leq 1$ ，变型空间 $VS_{H,D}$ 不是 ϵ -详尽（关于 c ）的概率小于或等于：*

$$|H|e^{-\epsilon m}$$

多少样例可使变型空间 ϵ - 详尽化

- 定理限定了任何一致学习器输出的假设 h 满足 $error(h) \geq \epsilon$ 的概率。
- 若要使此概率小于 δ

$$|H|e^{-\epsilon m} \leq \delta$$

则

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

学习布尔文字的合取

- 需要多少样例足够保证至少以概率 $(1-\delta)$ 使得
  - 每个 $VS_{H,D}$ 中的 h 满足 $error_{\mathcal{D}}(h) \leq \epsilon$
- 使用定理:

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

假设空间 H 定义为 n 个布尔文字的合取，则假设空间 H 的大小为 $|H| = 3^n$, 则

$$m \geq \frac{1}{\epsilon}(\ln 3^n + \ln(1/\delta))$$

或

$$m \geq \frac{1}{\epsilon}(n \ln 3 + \ln(1/\delta))$$

EnjoySport?

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

若 H 象 *EnjoySport* 中那样，则 $|H| = 973$, 且

$$m \geq \frac{1}{\epsilon}(\ln 973 + \ln(1/\delta))$$

若想保证以 95% 的概率使变型空间 *VS* 只包含 $error_{\mathcal{D}}(h) \leq .1$ 的假设, 需要 m 个样例：

$$m \geq \frac{1}{.1}(\ln 973 + \ln(1/.05))$$

$$m \geq 10(\ln 973 + \ln 20)$$

$$m \geq 10(6.88 + 3.00)$$

$$m \geq 98.8$$

PAC Learning

定义：考虑一概念类别 C 定义在长度为 n 的实例集合 X 上，学习器 L 使用假设空间 H。称 C 是使用 H 的 L 可 PAC 学习的，若对所有 $c \in C$，X 上的分布 $\mathcal{D}$，ϵ 满足 $0 < \epsilon < 1/2$，以及 δ 满足 $0 < \delta < 1/2$，学习器 L 将以至少 $1 - \delta$ 的概率输出一假设 $h \in H$，使 $error_{\mathcal{D}}(h) \leq \epsilon$，所使用的时间为 $1/\epsilon, 1/\delta, n$ 以及 $size(c)$ 的多项式函数。

Agnostic Learning

- 目前为止，假定 $c \in H$
- 不可知学习设定：不假定 $c \in H$
- 寻找具有最小训练错误率的 h
- 此时的样本复杂度从 Hoeffding 边界

$$Pr[error_{\mathcal{D}}(h) > error_{D}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

求得

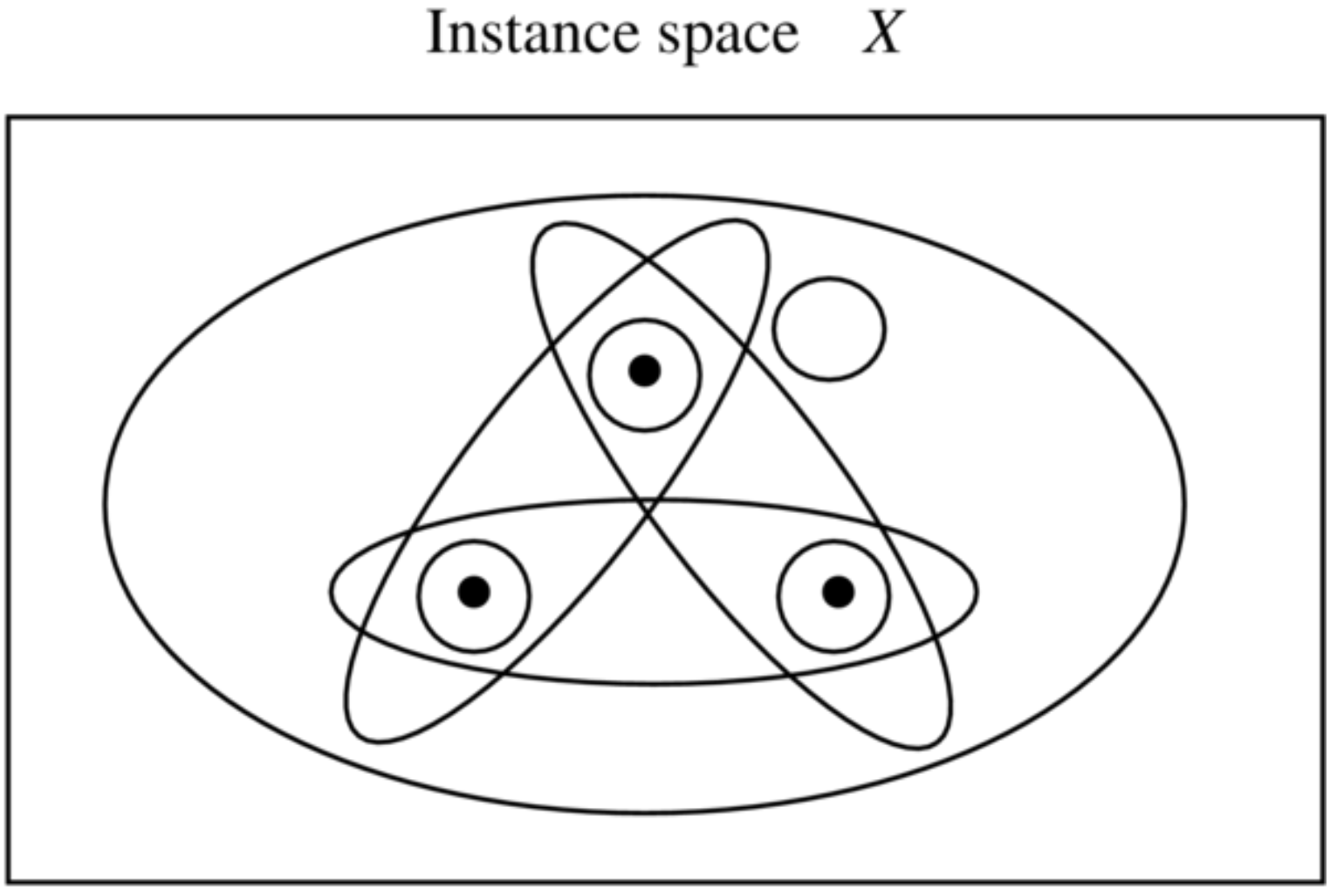
$$m \geq \frac{1}{2\epsilon^2}(\ln |H| + \ln(1/\delta))$$

Shattering a Set of Instances

定义：将集合 S 分成不相交的两个子集的划分称为集合 S 的二分法划分（*dichotomy*）

定义：一实例集 S 被假设空间 H 拆散 *(shatter)*，当且仅当对 S 的每个二分法划分，存在 H 中的某假设与此划分一致。

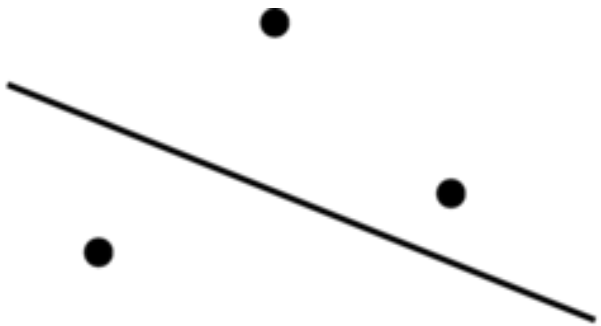
Three Instances Shattered



Vapnik-Chervonenkis Dimension

定义：定义在实例空间 X 上的假设空间 H 的 Vapnik-Chervonenkis 维，或 $VC(H)$，是可被 H 拆散的 X 的最大有限子集的大小。如果 X 的任意有限大的子集可被 H 拆散，那么 $VC(H) \equiv \infty$。

VC Dim. of Linear Decision Surfaces



(a)

(b)

Sample Complexity from VC Dimension

需要多少样例足够以至少 $(1-\delta)$ 的概率 ϵ -详尽 $VS_{H,D}$ ？

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

Topic

出错界限 (Mistake Bounds)

- 已分析: 需要多少样本来学习？
- 问题: 出多少次错误才能收敛？
- 与 PAC 问题框架相同，考虑：
  - 实例依分布 $\mathcal{D}$ 从 X 中随机抽取
  - 学习器必须先预测目标值 c(x)，之后再由施教者给出正确的目标值。
  - 在学习器学习到目标概念前，它的预测会有多少次出错？

Mistake Bounds: Find-S

考虑 Find-S 算法，$H =$ 布尔文字合取

Find-S：

- 将 h 被始化为最特殊假设 $l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \ldots l_n \wedge \neg l_n$
- 对每个正例 x
  - 从 h 中移去任何不满足 x 的文字
- 输出假设 h

收敛到正确的 h 前出错几次？

Mistake Bounds: Halving Algorithm

考虑 Halving 算法:

- 使用变型空间候选消除（Candidate-Elimination）算法学习概念
- 变型空间成员投票分类新样例

收敛到正确的 h 前出错几次？

- $\lfloor \log_2 |H| \rfloor$
- 0

Optimal Mistake Bounds

对任意学习算法 A 和任意目标概念 c，令 $M_A(C)$ 代表 A 为了确切学到 c，在所有可能训练样例序列中出错的最大值。现在对于任意非空概念类 C，令

$$M_A(C) \equiv \max_{c \in C} M_A(c)$$

定义：令 C 为任意非空概念类。C 的最优出错界限（optimal mistake bound）$Opt(C)$，是所有可能学习算法 A 中 $M_A(C)$ 的最小值。

$$Opt(C) \equiv \min_{A \in learning\ algorithms} M_A(C)$$

对任意概念类 C，C 的最优出错边界，Halving 算法出错边界和 C 的 VC 维之间关系：

$$VC(C) \leq Opt(C) \leq M_{Halving}(C) \leq log_2(|C|)$$