泛化误差

给定样本集$S=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}, x_i\in X, y_i\in Y$ ，一个假设$h\in H$, 一个目标概念$c\in C$, 一个未知的分布D , h的泛化误差为：

$$R(h)=\Pr_{x\sim D}[h(x)\neq c(x)]=\mathop{\mathrm{E}}_{x\sim D}\left[1_{h(x)\neq c(x)}\right]$$

经验误差

给定样本集$S=\{(x_1,y_1),(x_2,y_2),\ldots(x_m,y_m)\}, x_i\in X, y_i\in Y$ ，一个假设$h\in H$, 一个目标概念$c\in C$, 一个未知的分布D , h的经验误差为：

$$\widehat{R}(h)=\frac{1}{m}\sum_{i=1}^{m}1_{h(x)\neq c(x)}$$

经验误差

给定样本集$S=\{(x_1,y_1),(x_2,y_2),\dots(x_m,y_m)\}$, $x_i\in X$, $y_i\in Y$ ，一个假设$h\in H$, 一个目标概念$c\in C$，一个未知的分布D，h的经验误差为：

$$\hat{R}(h)=\frac{1}{m}\sum_{i=1}^{m}1_{h(x)\neq c(x)}$$

$$\underset{S\sim D^m}{\mathrm{E}}\left[\hat{R}(h)\right]=\frac{1}{m}\sum_{i=1}^{m}\underset{S\sim D^m}{\mathrm{E}}\left[1_{h(x_i)\neq c(x_i)}\right]=\frac{1}{m}\sum_{i=1}^{m}\underset{S\sim D^m}{\mathrm{E}}\left[1_{h(x)\neq c(x)}\right]$$

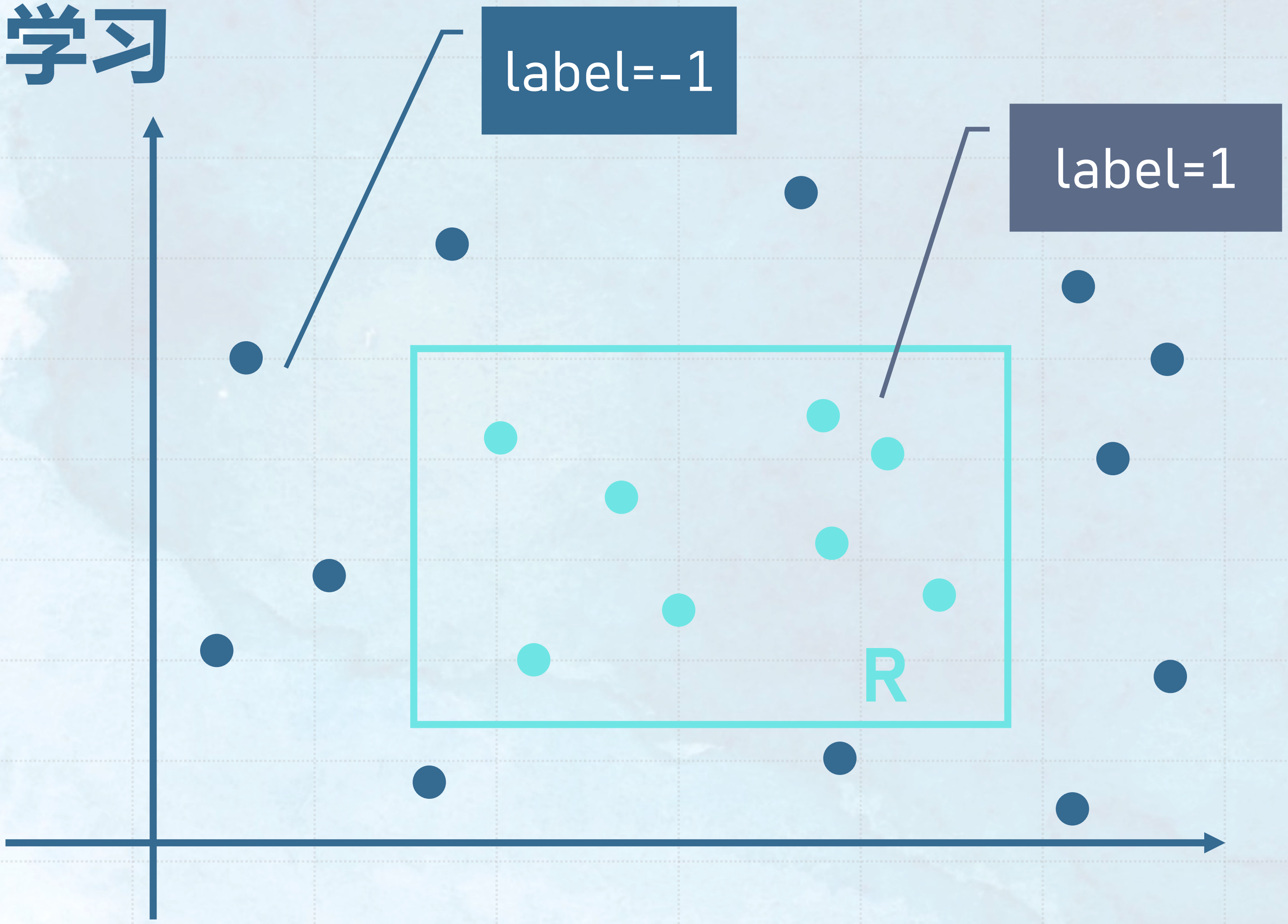
$$=\underset{S\sim D^m}{\mathrm{E}}\left[1_{\{h(x)\neq c(x)\}}\right]=\underset{x\sim D}{\mathrm{E}}\left[1_{\{h(x)\neq c(x)\}}\right]=R(h)$$

PAC学习

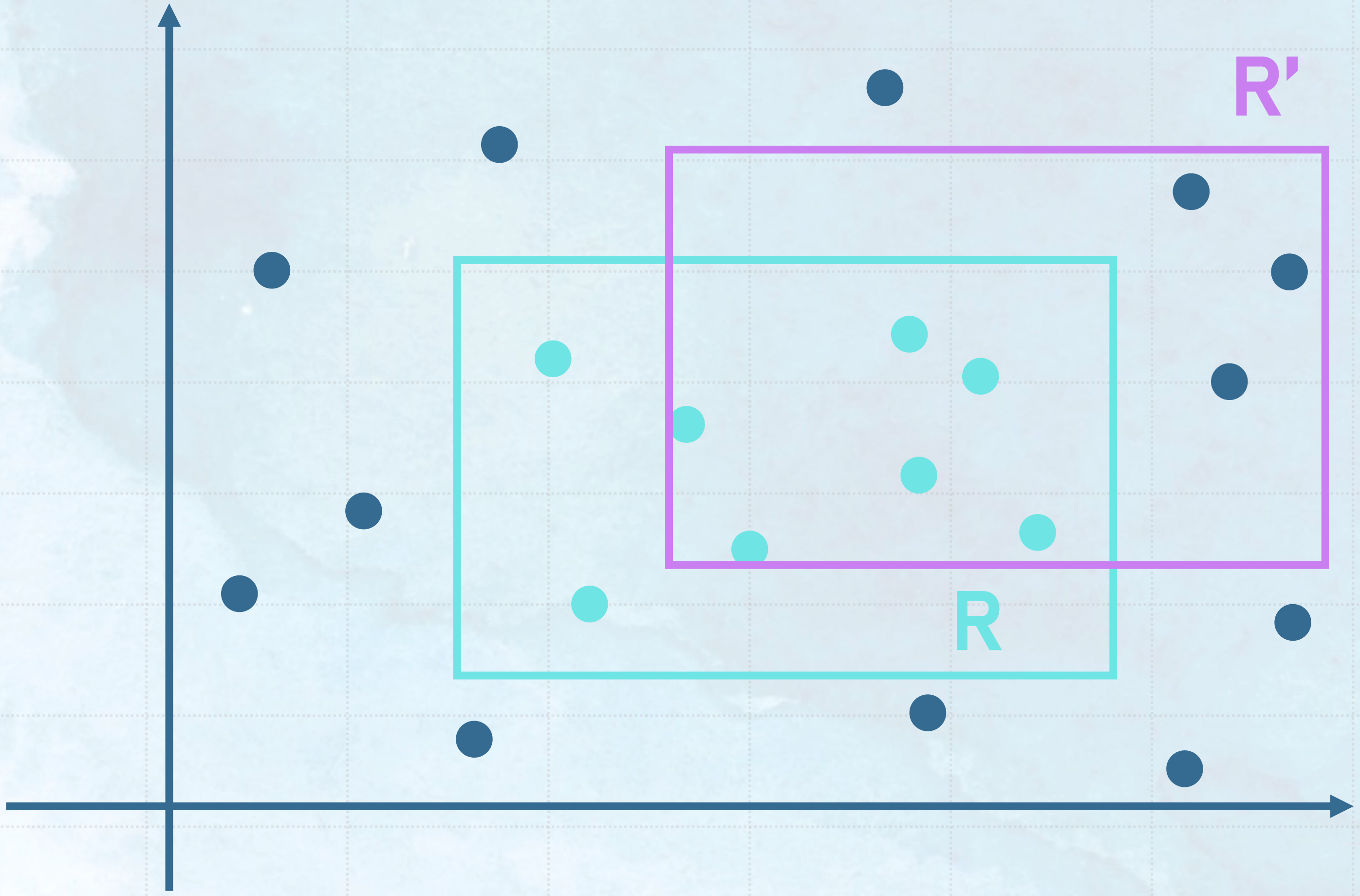
如果存在一个算法A以及一个多项式函数 $poly(\cdot,\cdot,\cdot,\cdot)$，使它能对任意的 $\epsilon > 0$以及任意的$\delta > 0$，对所有在X上的分布D和对任意的目标概念$c \in C$，对任意样本量$m \geq poly(\frac{1}{\epsilon}, \frac{1}{\delta}, n, size(c))$均能让下面的表达式成立，那么我们认为概念类C是PAC可学习的

$$\Pr_{S\sim D^m}[R(h_S) \leq \epsilon] \geq 1 - \delta$$

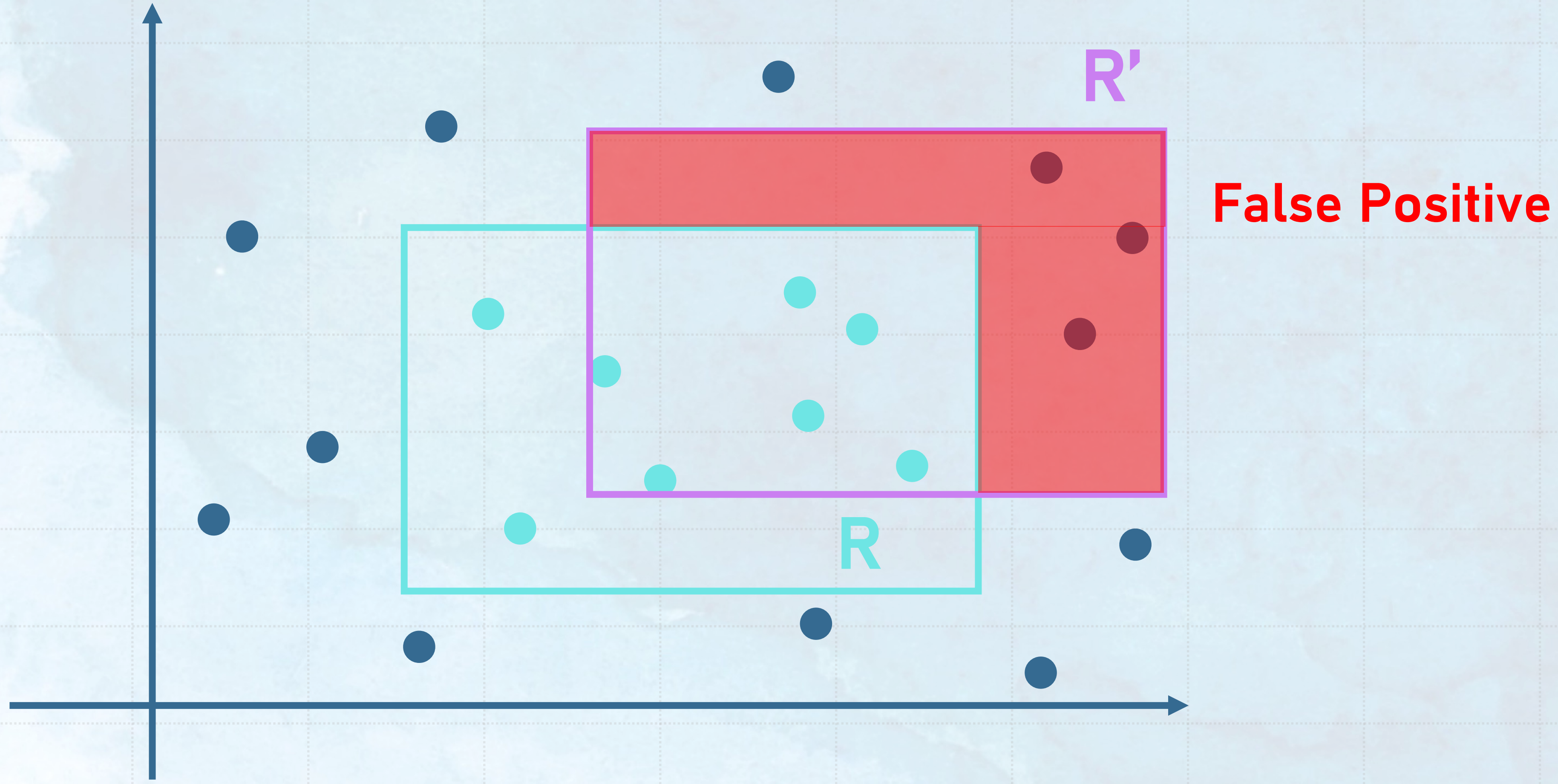
PAC学习



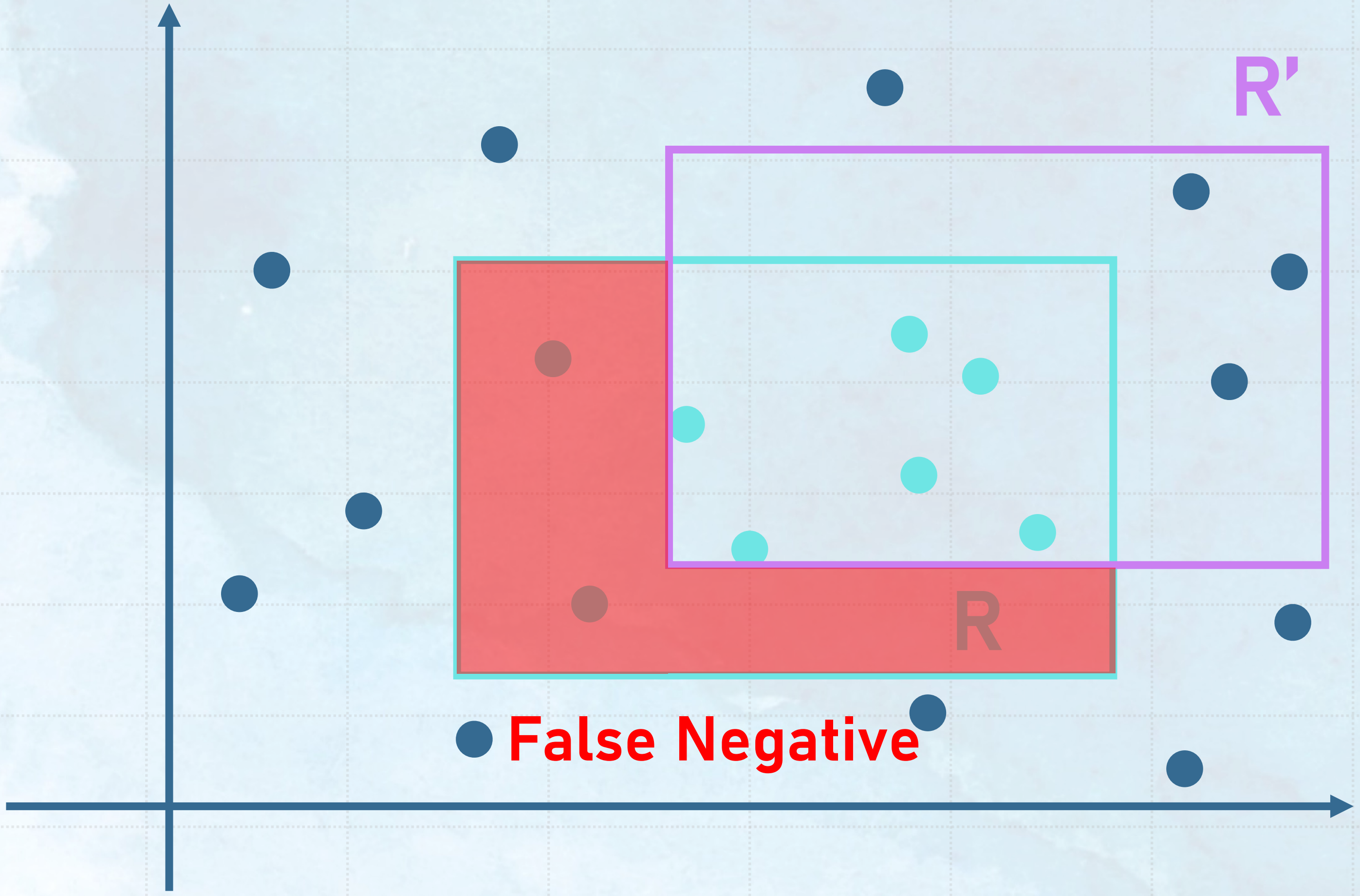
PAC学习



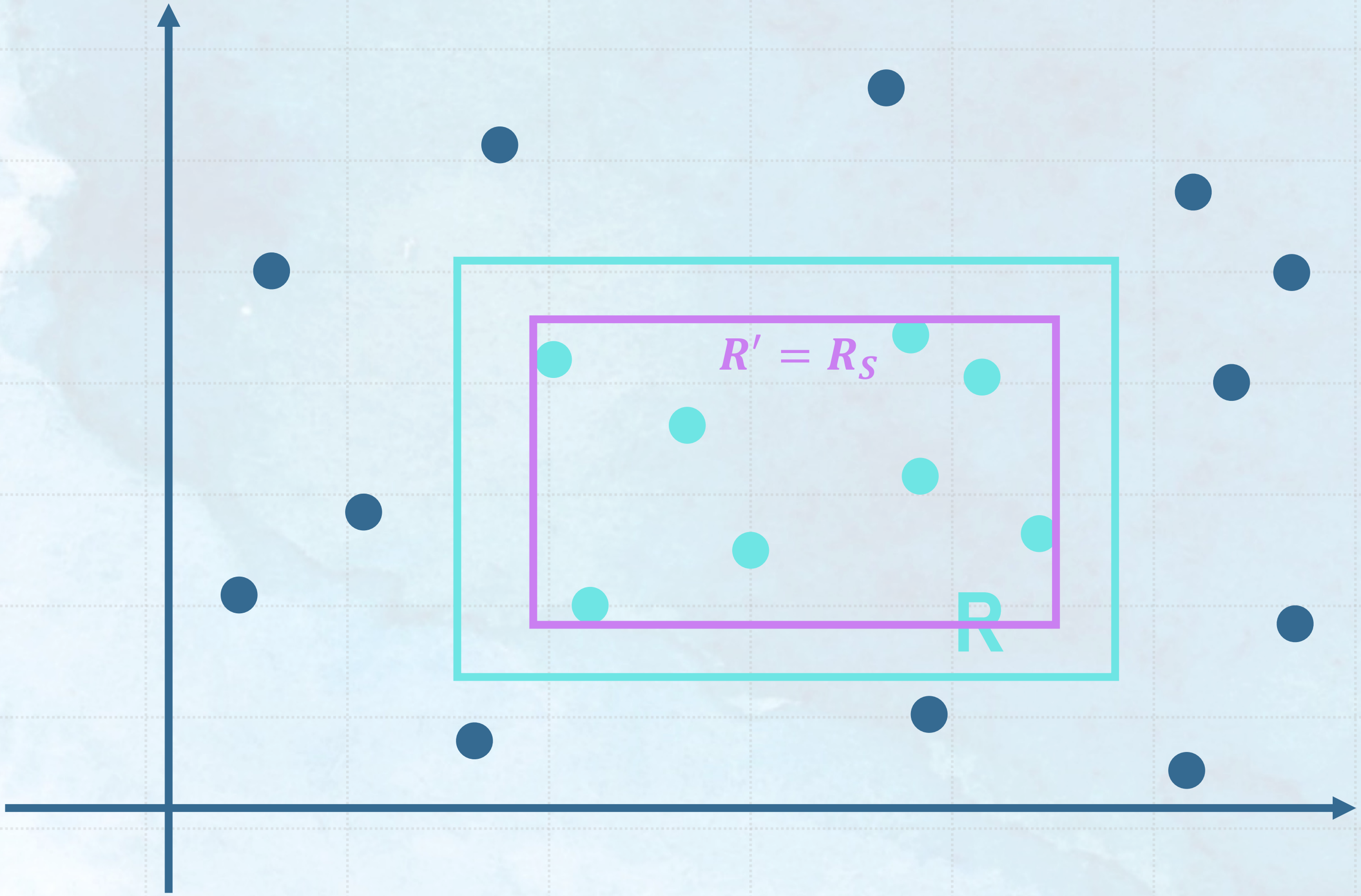
PAC学习



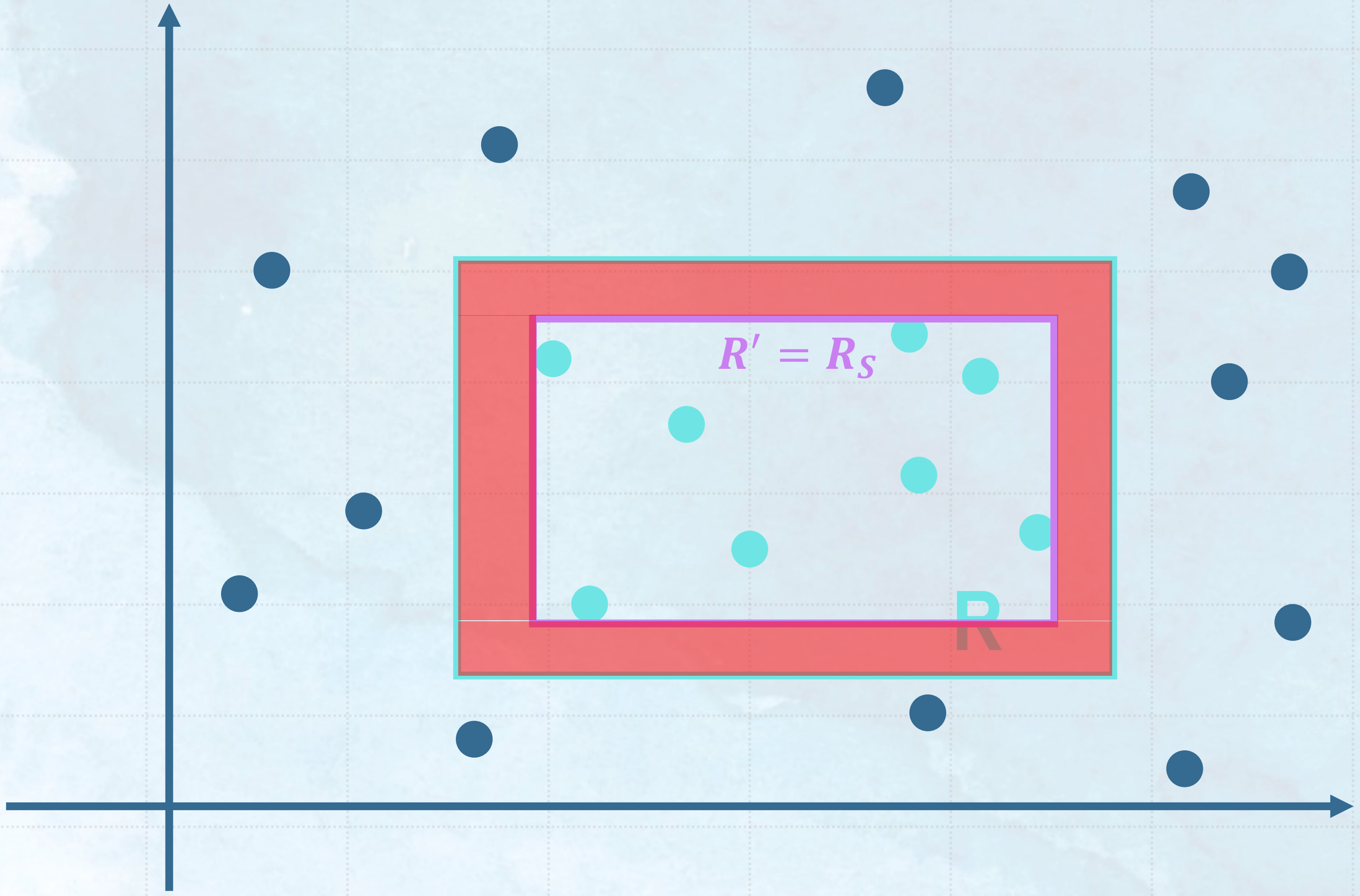
PAC学习



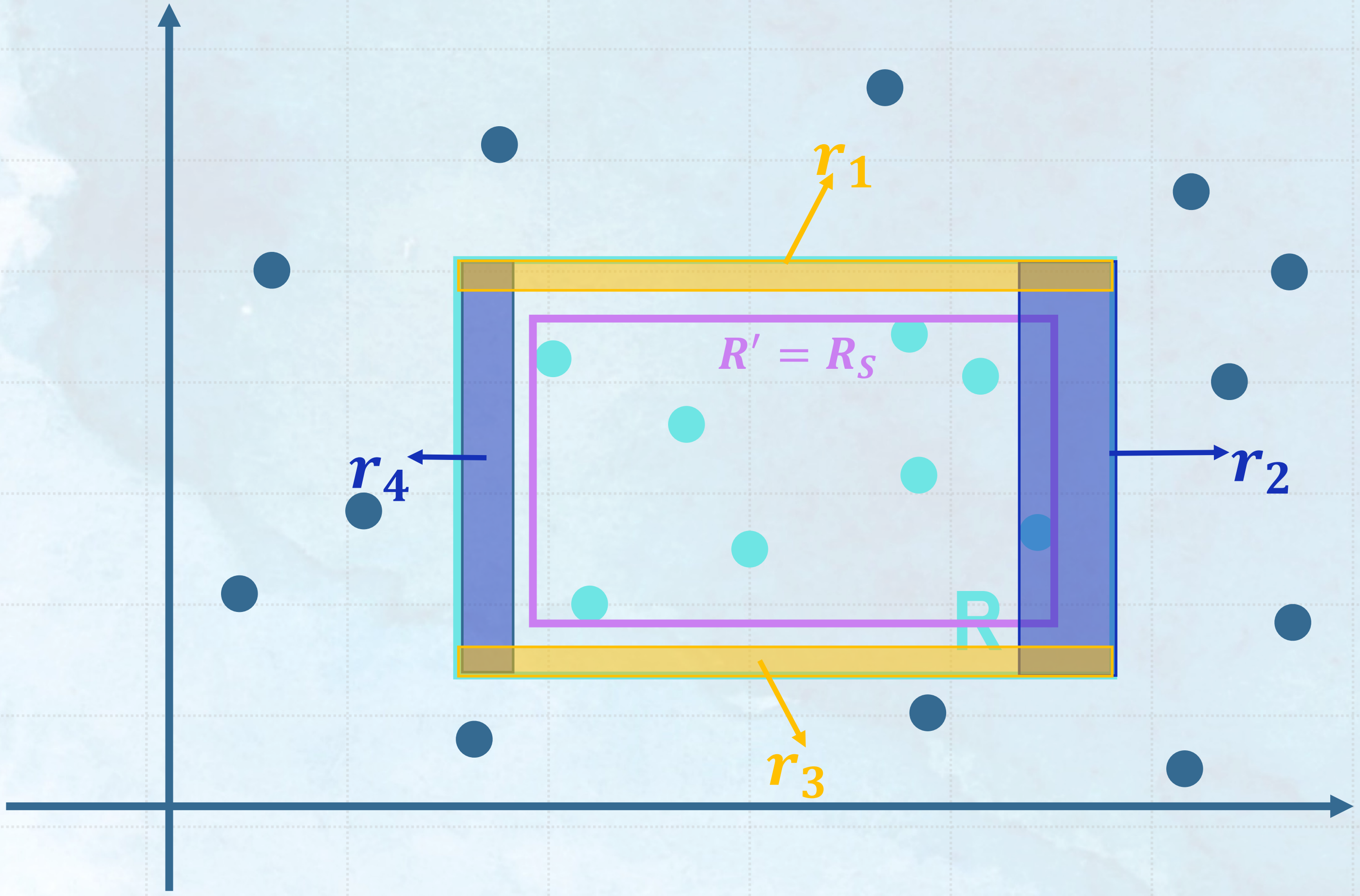
PAC学习



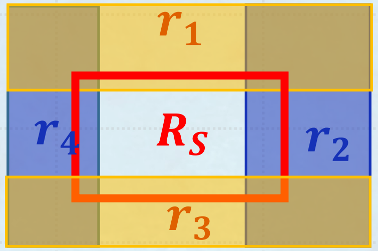
PAC学习



PAC学习

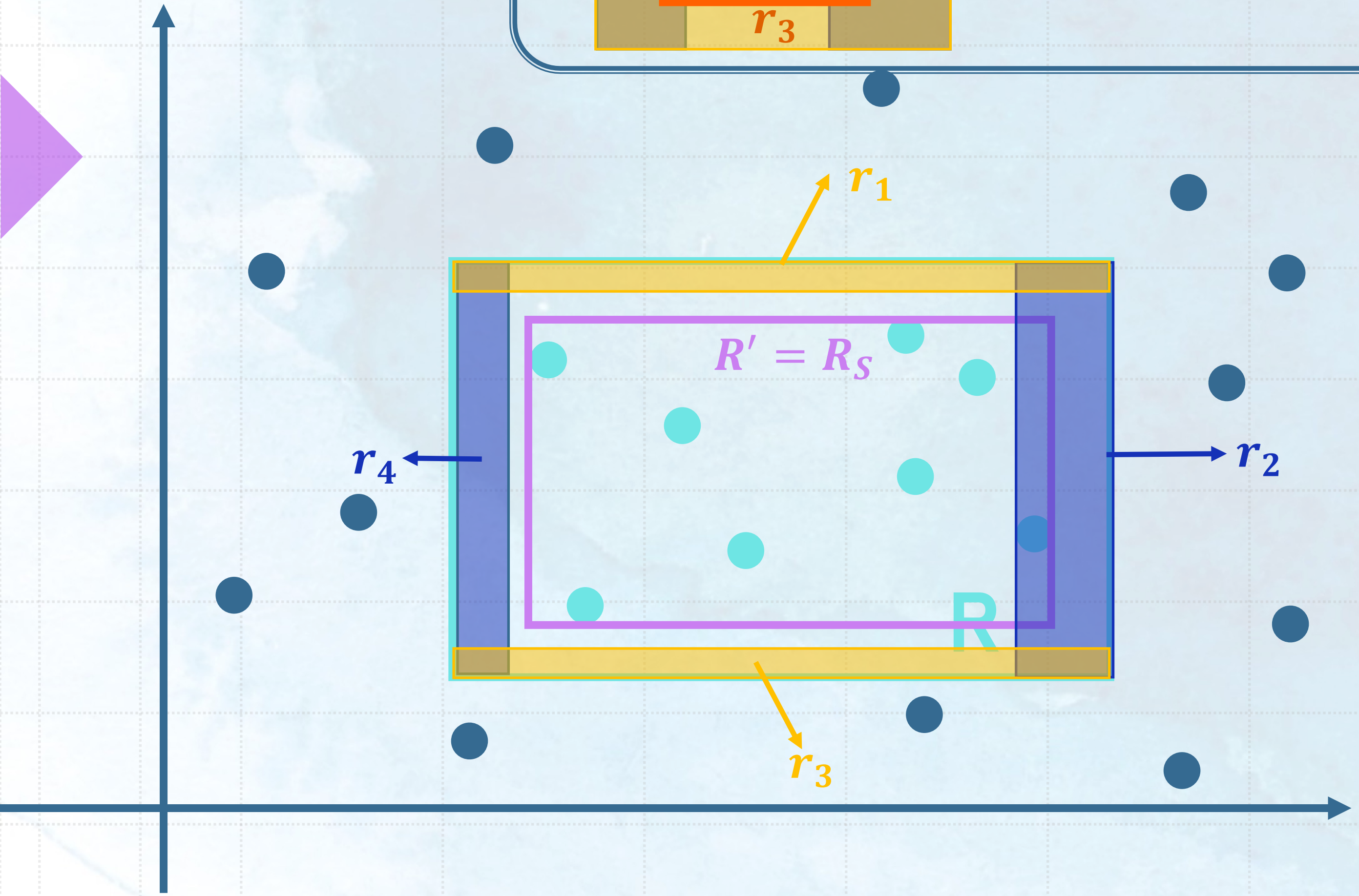


PAC学习

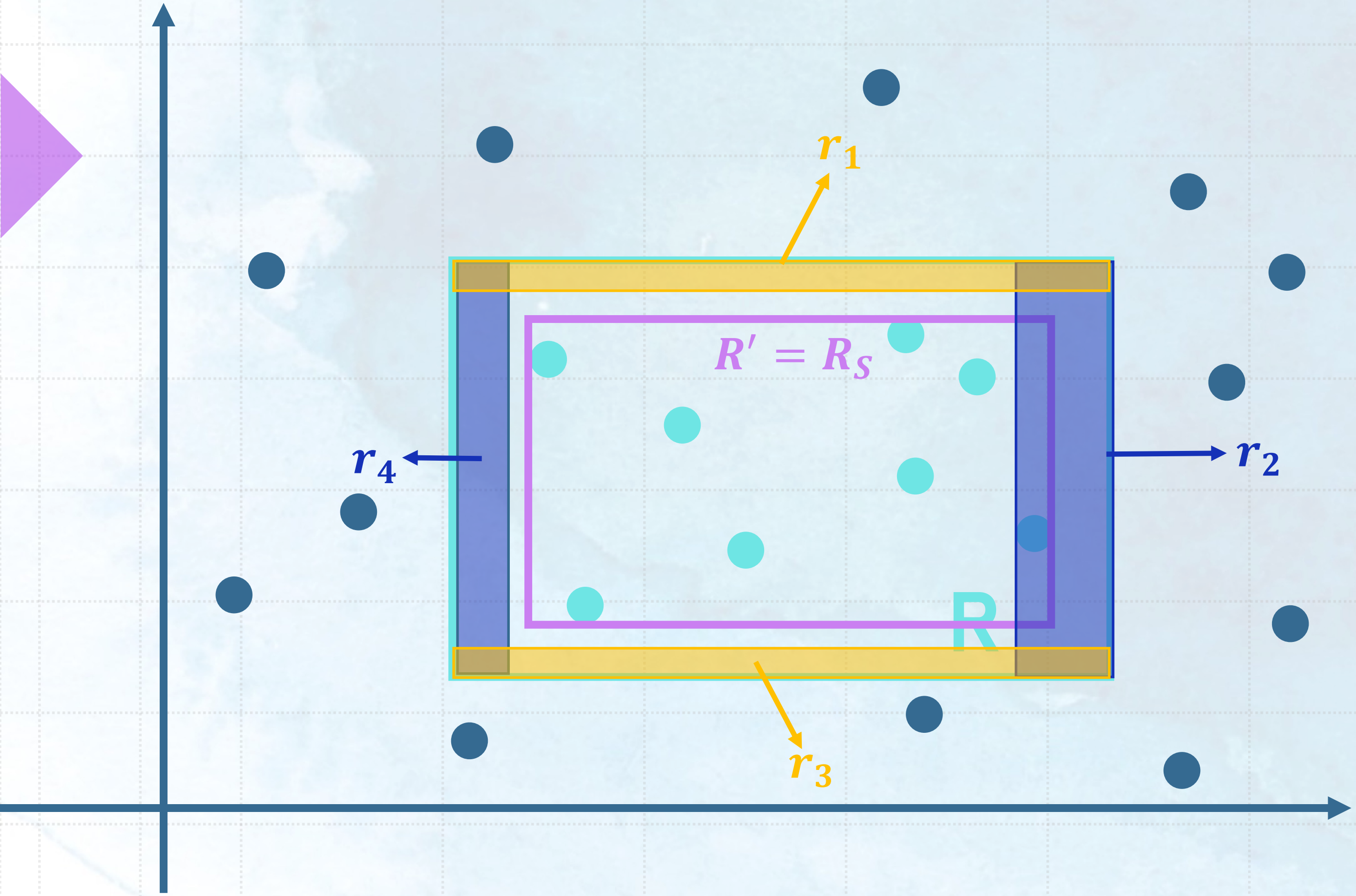


R_S四条边都在r内部

$$\underset{S\sim D^m}{Pr}[R(R_s) > \epsilon] \le \underset{S\sim D^m}{Pr}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \underset{S\sim D^m}{Pr}[R_s \cap r_i = \emptyset]$$

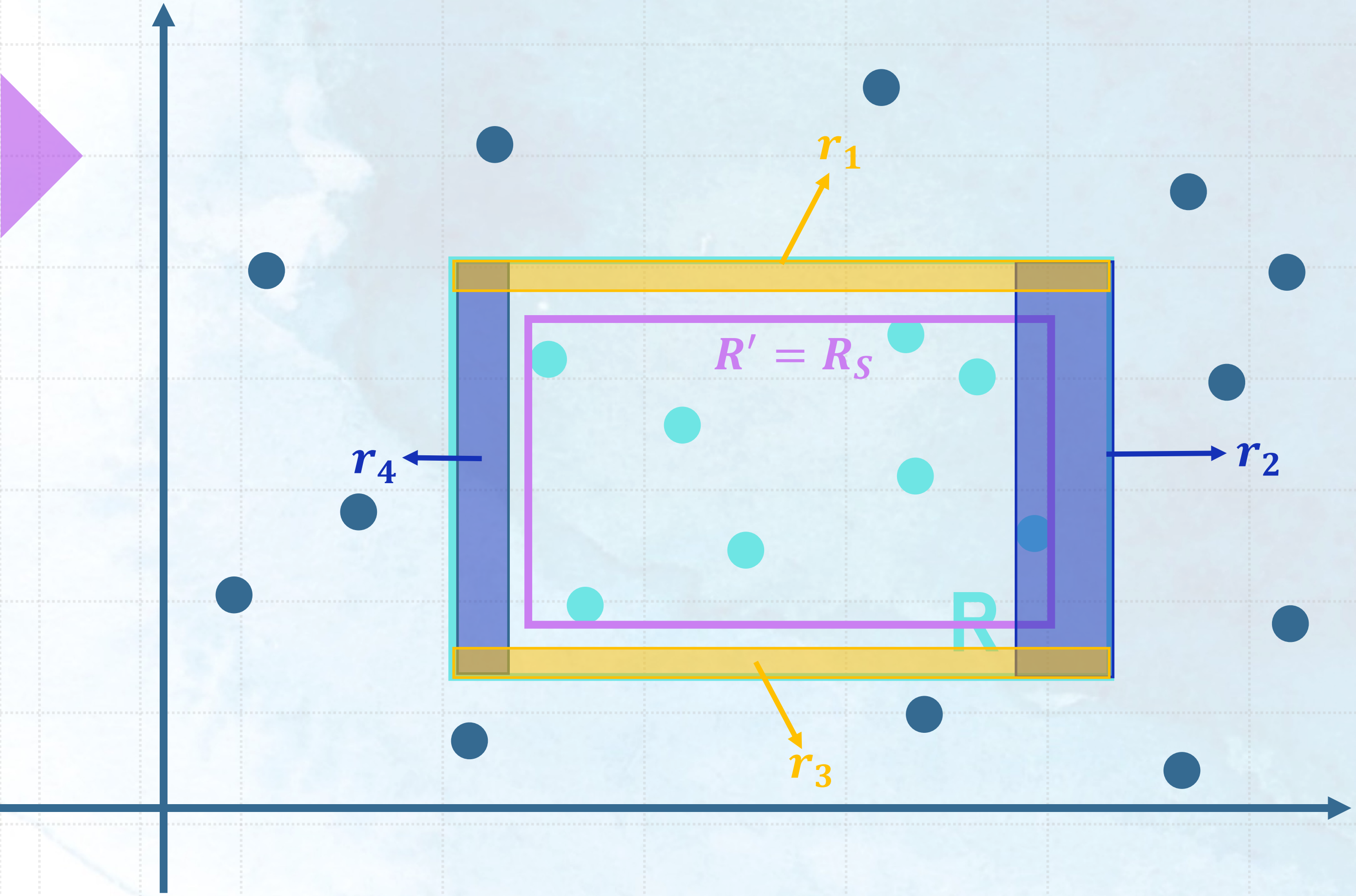


PAC学习



$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4(1-\frac{\epsilon}{4})^m$$

PAC学习



$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$

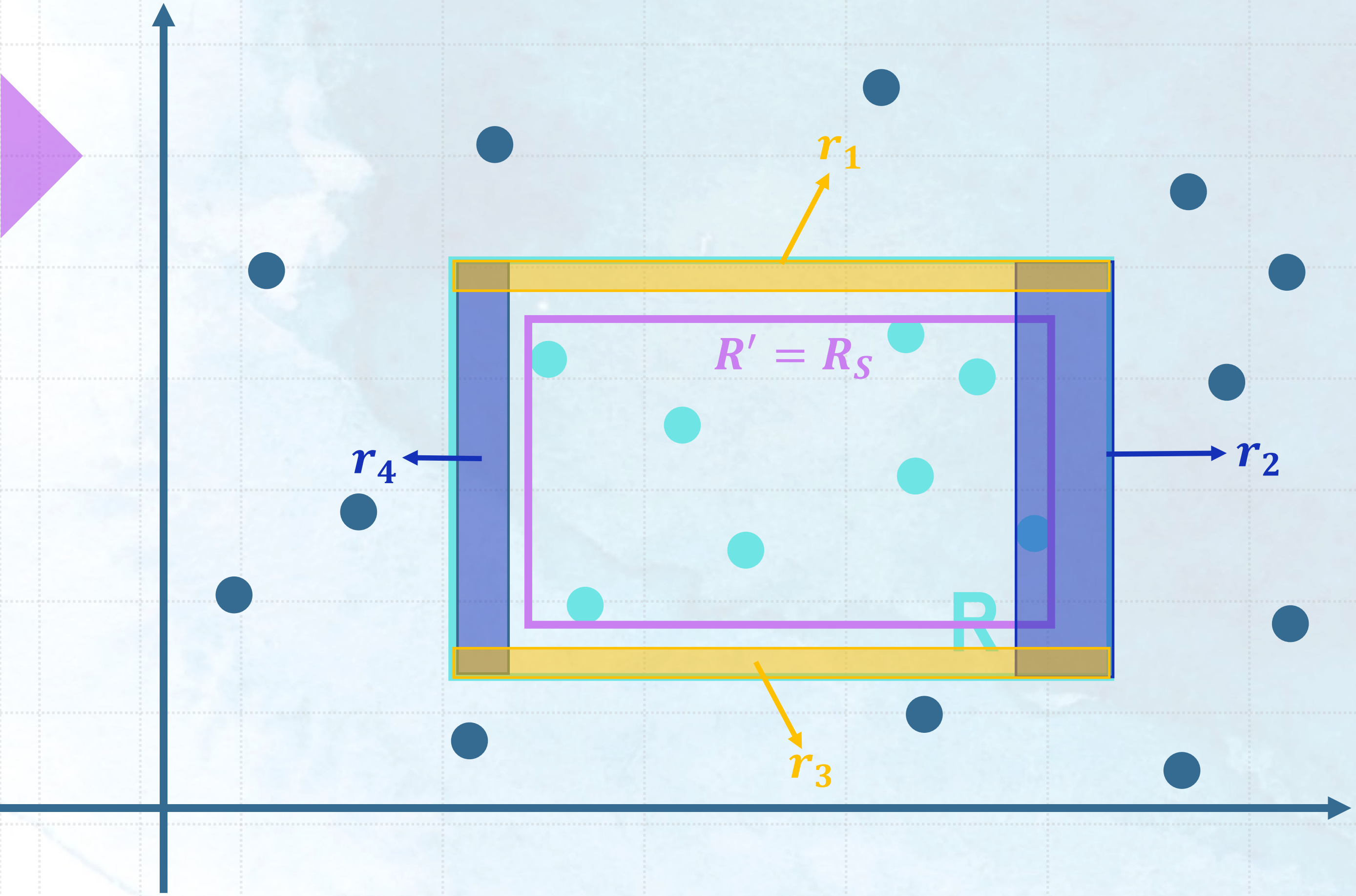
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$

$$\le 4(1 - \frac{\epsilon}{4})^m$$

$$\le 4exp(-\frac{m\epsilon}{4})$$

$$\boxed{1 - x \le \exp(-x)}$$

PAC学习

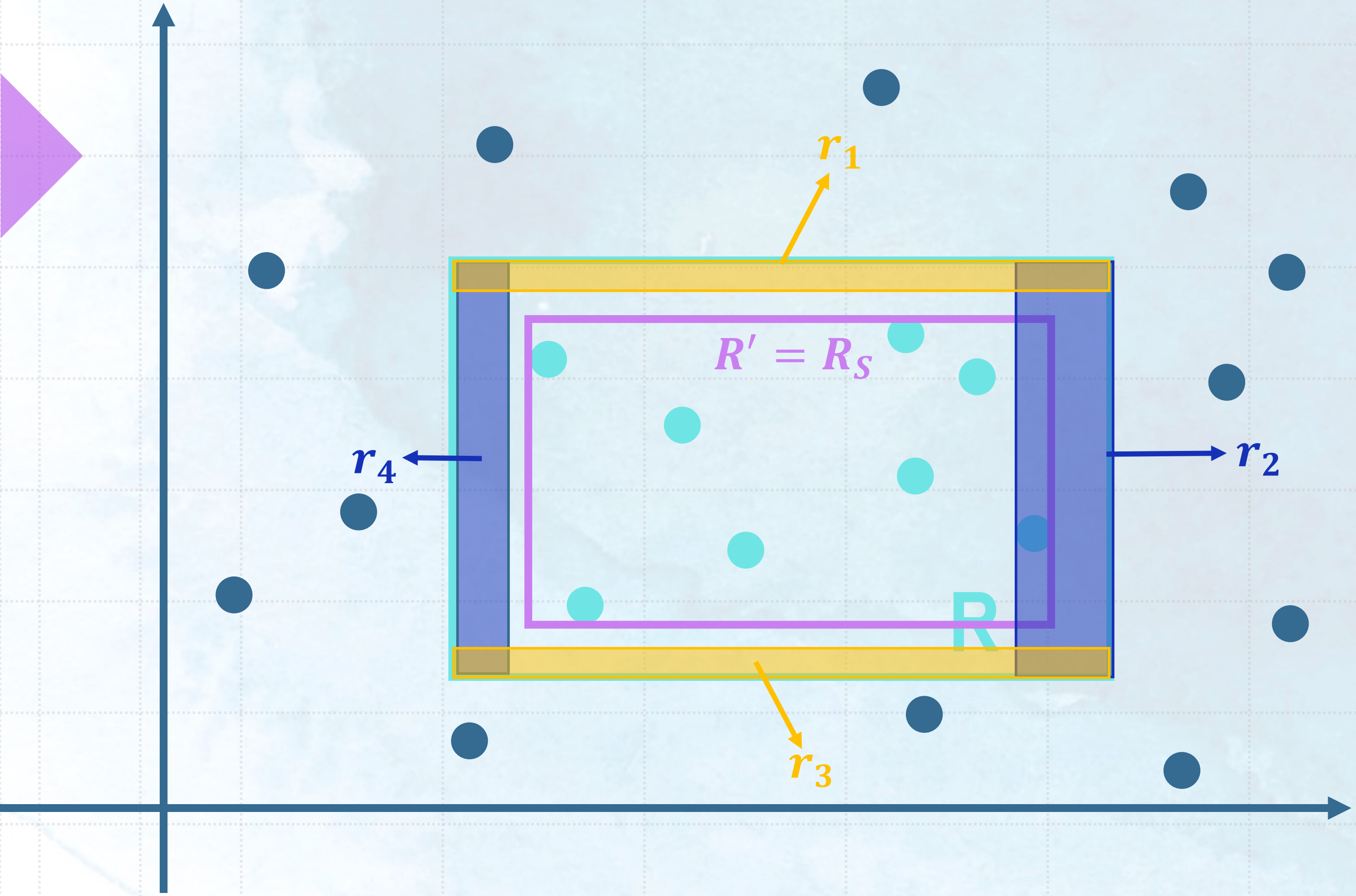


$$\Pr_{S\sim D^m}[R(R_S) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4\left(1-\frac{\epsilon}{4}\right)^m$$
$$\le 4exp\left(-\frac{m\epsilon}{4}\right)$$

$$\boxed{1 - x \le \exp(-x)}$$

$$\Pr_{S\sim D^m}[R(h_S) \le \epsilon] \ge 1-\delta$$

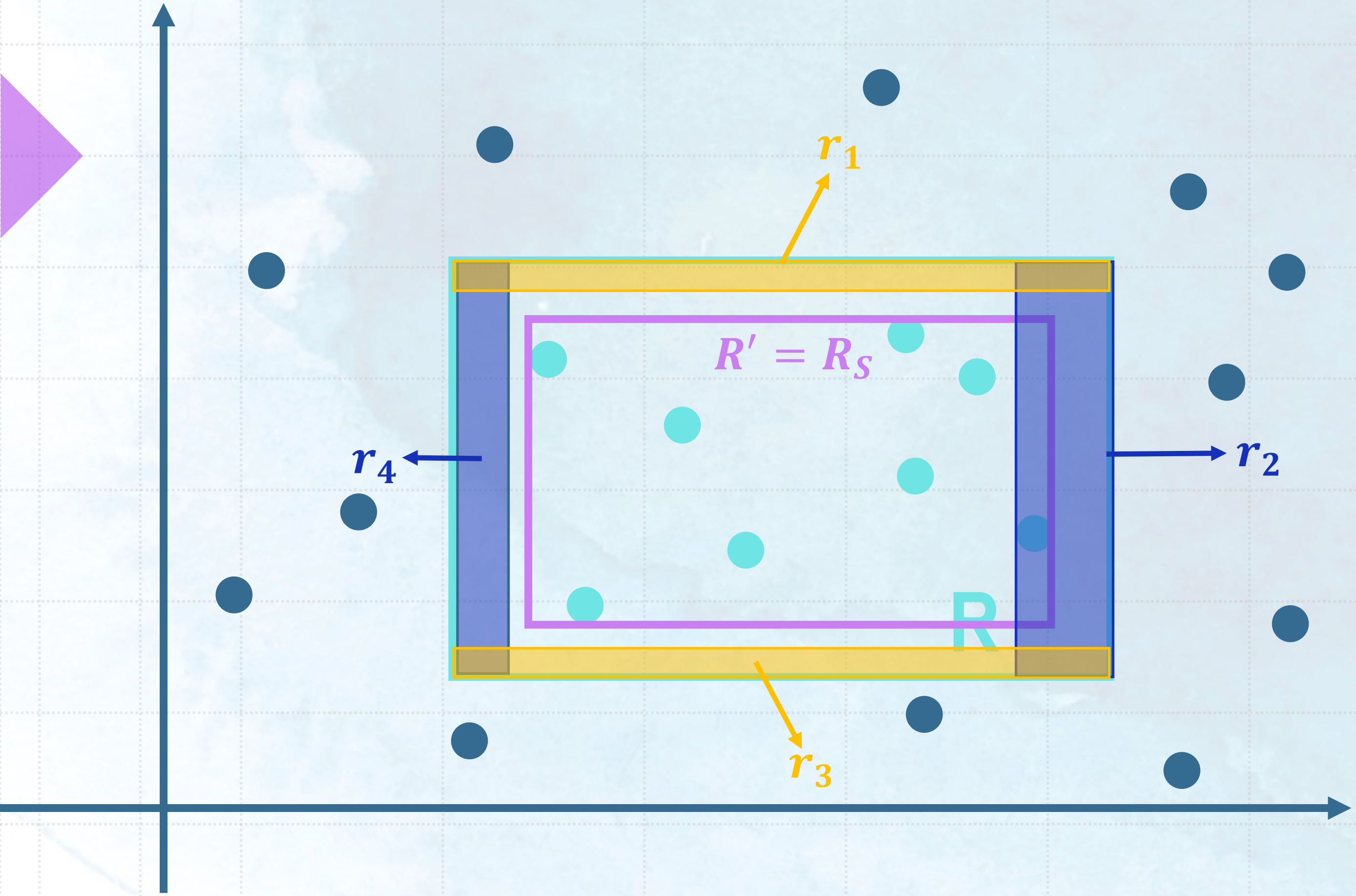
PAC学习



$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4\left(1 - \frac{\epsilon}{4}\right)^m$$
$$\le 4exp\left(-\frac{m\epsilon}{4}\right) \qquad \boxed{1 - x \le \exp(-x)}$$

$$\Pr_{S\sim D^m}[R(h_S) \le \epsilon] \ge 1 - \delta$$
$$\Pr_{S\sim D^m}[R(h_S) > \epsilon] \le \delta$$

PAC学习

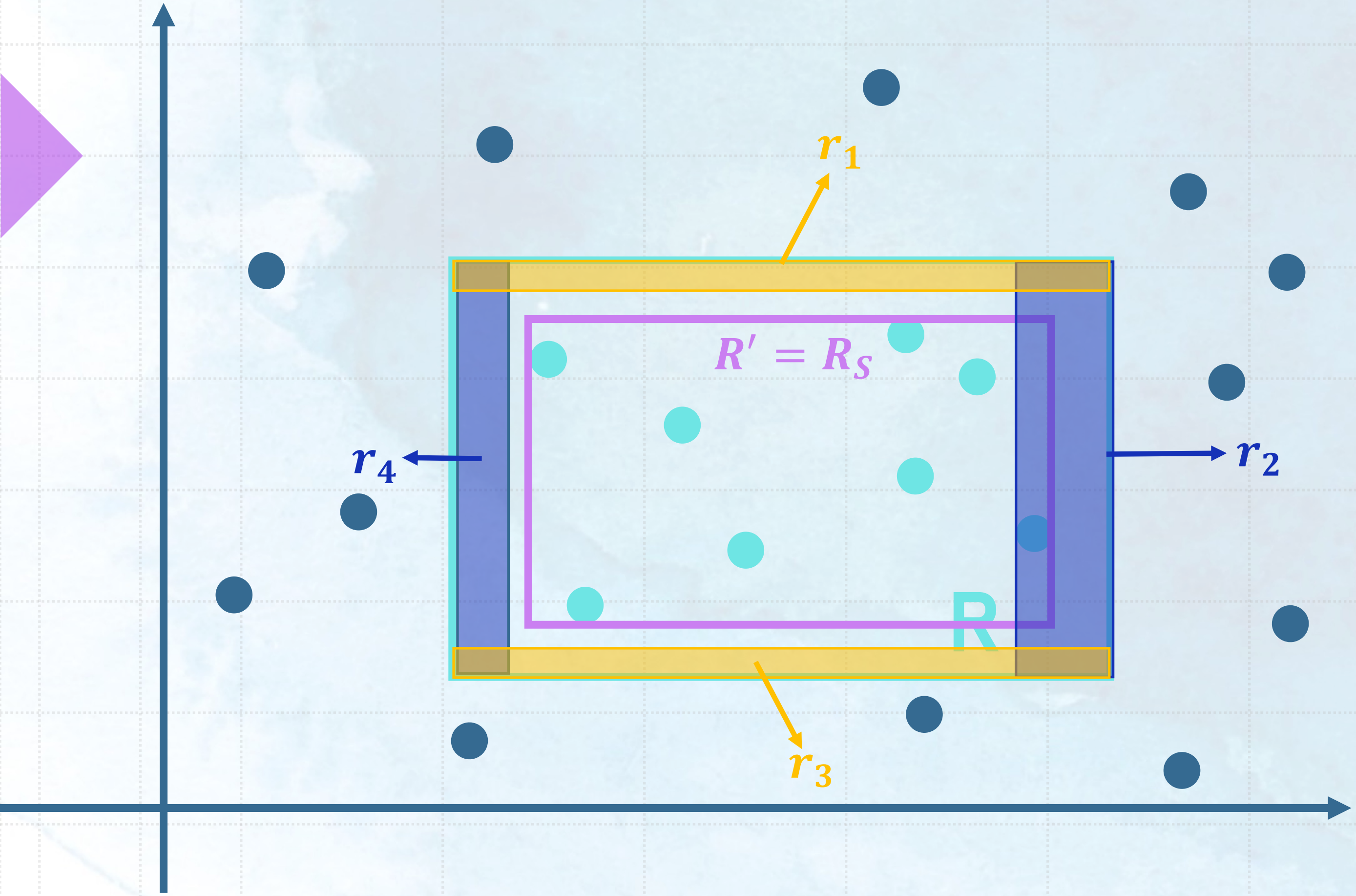


$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4} \{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4(1-\frac{\epsilon}{4})^m$$
$$\le 4exp(-\frac{m\epsilon}{4})$$

$$\boxed{1 - x \le \exp(-x)}$$

$$\Pr_{S\sim D^m}[R(h_S) \le \epsilon] \ge 1 - \delta$$
$$\Pr_{S\sim D^m}[R(h_S) > \epsilon] \le \delta$$
$$4exp(-\frac{m\epsilon}{4}) \le \delta$$

PAC学习

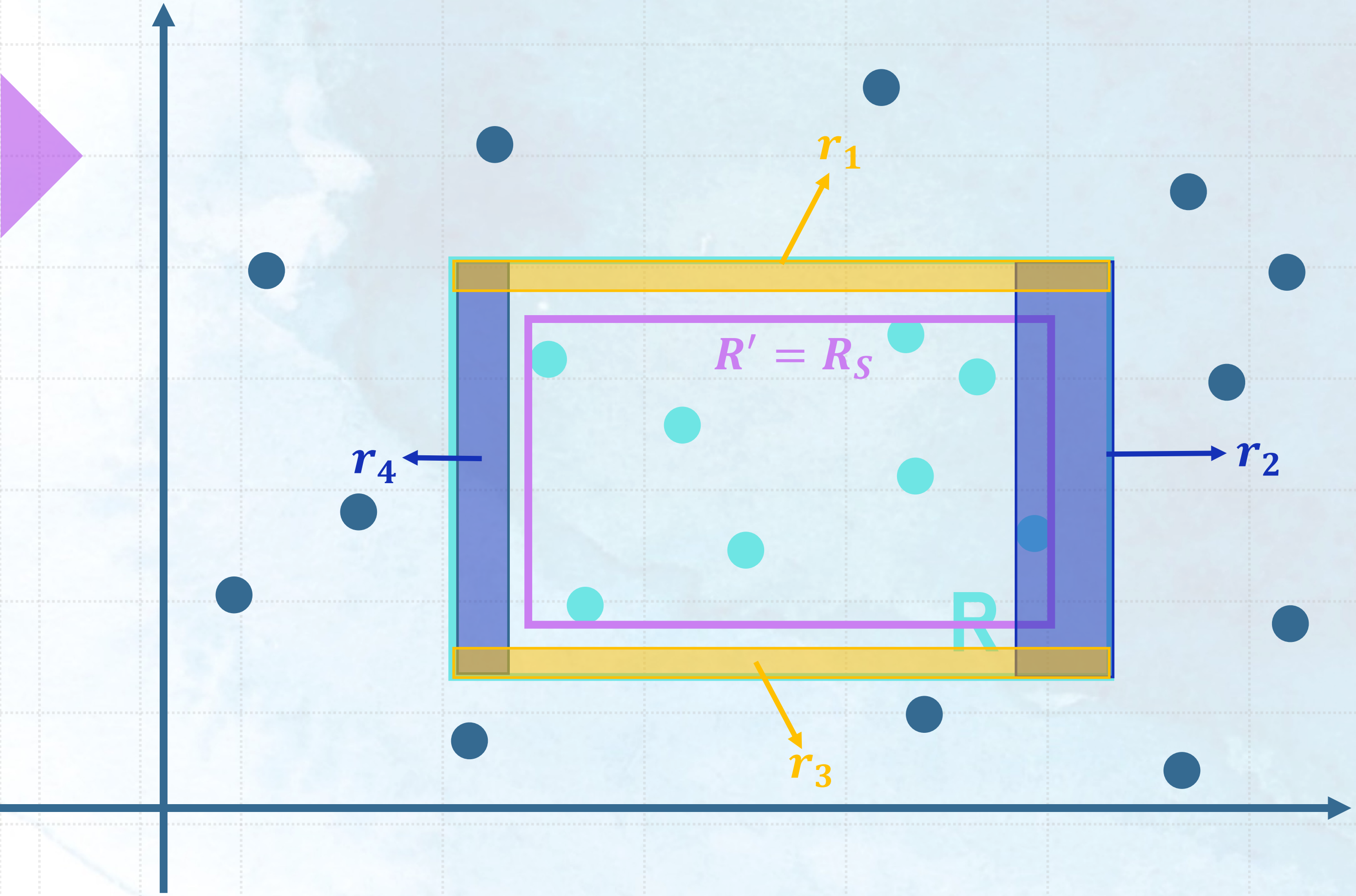


$$\underset{S\sim D^m}{Pr}[R(R_s) > \epsilon] \le \underset{S\sim D^m}{Pr}\left[\cup_{i=1}^{4} \{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \underset{S\sim D^m}{Pr}[R_s \cap r_i = \emptyset]$$
$$\le 4\left(1 - \frac{\epsilon}{4}\right)^m$$
$$\le 4exp\left(-\frac{m\epsilon}{4}\right)$$

$$\boxed{1 - x \le \exp(-x)}$$

$$\underset{S\sim D^m}{Pr}[R(h_S) \le \epsilon] \ge 1 - \delta$$
$$\underset{S\sim D^m}{Pr}[R(h_S) > \epsilon] \le \delta$$
$$4exp\left(-\frac{m\epsilon}{4}\right) \le \delta$$
$$m \ge \frac{4}{\epsilon} log\frac{4}{\delta}$$

PAC学习



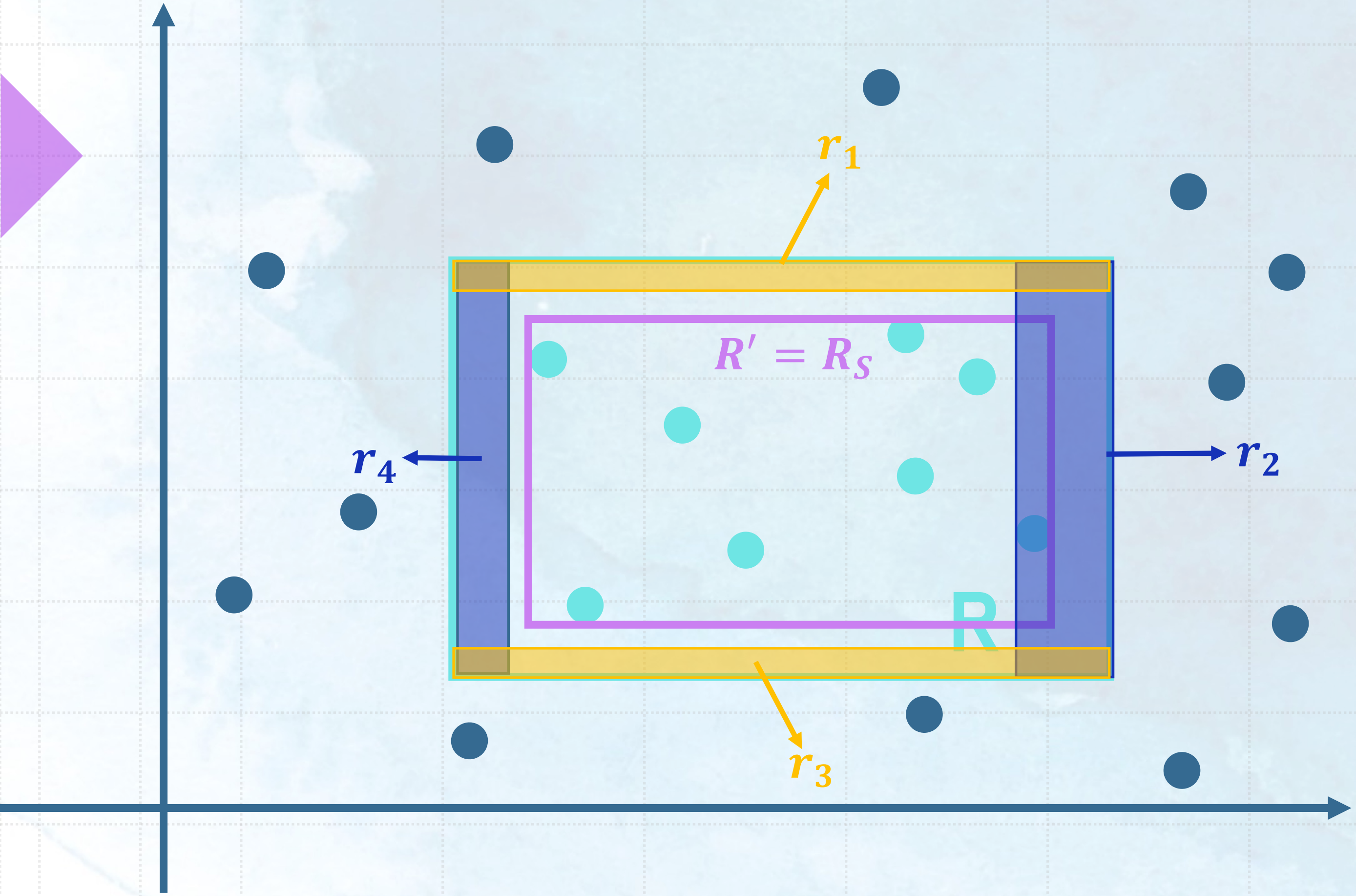
$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4} \{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4(1 - \frac{\epsilon}{4})^m$$
$$\le 4exp(-\frac{m\epsilon}{4})$$

$$\boxed{1 - x \le \exp(-x)}$$

$$\Pr_{S\sim D^m}[R(h_S) \le \epsilon] \ge 1 - \delta$$
$$\Pr_{S\sim D^m}[R(h_S) > \epsilon] \le \delta$$
$$4exp(-\frac{m\epsilon}{4}) \le \delta$$
$$m \ge \frac{4}{\epsilon} log\frac{4}{\delta}$$

样本复杂度: $O(\frac{1}{\epsilon} log\frac{1}{\delta})$

PAC学习

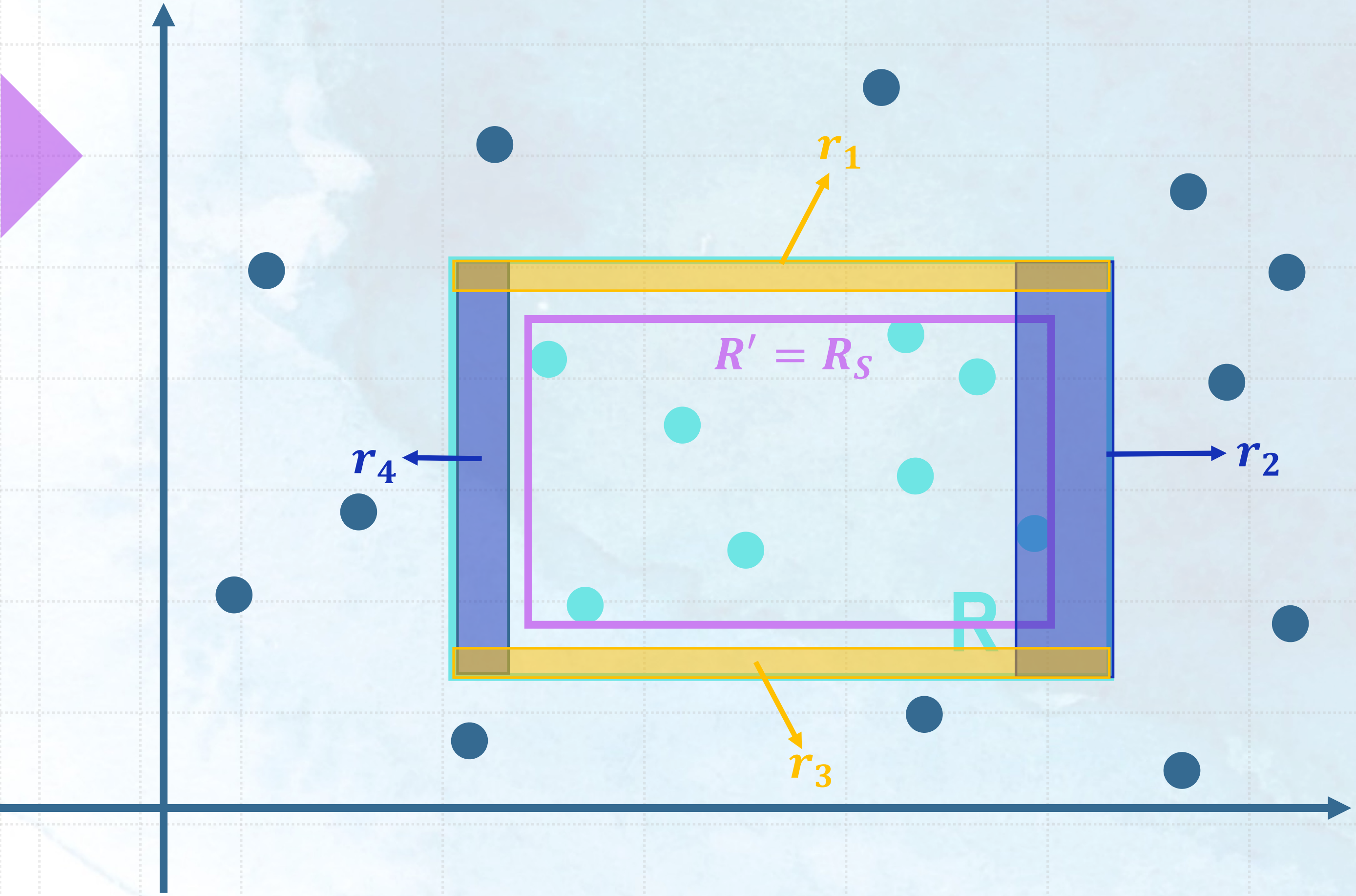


$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4(1-\frac{\epsilon}{4})^m$$
$$\le 4exp(-\frac{m\epsilon}{4})$$

$$\boxed{1 - x \le \exp(-x)}$$

令 $\delta = 4exp(-\frac{m\epsilon}{4})$

PAC学习



$$\Pr_{S\sim D^m}[R(R_s) > \epsilon] \le \Pr_{S\sim D^m}\left[\cup_{i=1}^{4}\{R_s \cap r_i = \emptyset\}\right]$$
$$\le \sum_{i=1}^{4} \Pr_{S\sim D^m}[R_s \cap r_i = \emptyset]$$
$$\le 4\left(1-\frac{\epsilon}{4}\right)^m$$
$$\le 4exp\left(-\frac{m\epsilon}{4}\right)$$

$$\boxed{1 - x \le \exp(-x)}$$

令 $\delta = 4exp\left(-\frac{m\epsilon}{4}\right)$

$$\epsilon = \frac{4}{m}log\frac{4}{\delta}$$
$$R(R_s) \le \frac{4}{m}log\frac{4}{\delta}$$

一致情形(Consistent case)

给定 $\epsilon > 0$, $R(h) > \epsilon$

$$Pr\big(h(x) = c(x)\big) = 1 - Pr\big(h(x) \neq c(x)\big) = 1 - R(h) < 1 - \epsilon$$

一致情形(Consistent case)

给定 $\epsilon > 0$ ，$R(h) > \epsilon$

$$Pr\big(h(x) = c(x)\big) = 1 - Pr\big(h(x) \neq c(x)\big) = 1 - R(h) < 1 - \epsilon$$

h 和S一致的概率为：

$$Pr((h(x_1) = c(x_1)) \wedge (h(x_2) = c(x_2)) \wedge \cdots \wedge (h(x_m) = c(x_m)))$$

$$= (1 - Pr\big(h(x) \neq c(x)\big)^m < (1 - \epsilon)^m$$

一致情形(Consistent case)

给定 $\epsilon > 0$ ，

$Pr\left[\exists h \in H\text{: } (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$

$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$

一致情形(Consistent case)

给定 $\epsilon > 0$,

$$Pr\left[\exists h \in H\colon\ (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$$

$$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) \wedge (R(h) > \epsilon))$$

联合界引理： $Pr(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} Pr(A_i)$

一致情形(Consistent case)

给定 $\epsilon > 0$，

$$Pr\left[\exists h \in H:\ (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$$

$$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) \wedge (R(h) > \epsilon))$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) | (R(h) > \epsilon))$$

$$\boxed{P(B|A) \geq P(AB)}$$

一致情形(Consistent case)

给定 $\epsilon > 0$，

$$Pr\left[\exists h \in H\colon\ (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$$

$$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) \wedge (R(h) > \epsilon))$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) | (R(h) > \epsilon))$$

$$\boxed{Pr((\hat{R}(h) = 0) | (R(h) > \epsilon)) \leq (1 - \epsilon)^m}$$

$$\leq |H|(1 - \epsilon)^m$$

一致情形(Consistent case)

给定 $\epsilon > 0$,

$$Pr\left[\exists h \in H\colon\ (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$$
$$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$$
$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) \wedge (R(h) > \epsilon))$$
$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) | (R(h) > \epsilon))$$
$$\leq |H|(1-\epsilon)^m$$
$$\leq |H| e^{-m\epsilon}$$

$$\boxed{Pr((\hat{R}(h) = 0) | (R(h) > \epsilon)) \leq (1-\epsilon)^m}$$

$$\boxed{1 - x \leq \exp(-x)}$$

一致情形(Consistent case)

给定 $\epsilon > 0$，

$$Pr\left[\exists h \in H \colon (\hat{R}(h) = 0) \wedge (R(h) > \epsilon)\right]$$

$$= Pr[(h_1 \in H, (\hat{R}(h_1) = 0) \wedge (R(h_1) > \epsilon)) \vee (h_2 \in H, (\hat{R}(h_2) = 0) \wedge (R(h_2) > \epsilon)) \vee \cdots]$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) \wedge (R(h) > \epsilon))$$

$$\leq \sum_{h \in H} Pr((\hat{R}(h) = 0) | (R(h) > \epsilon))$$

$$\boxed{Pr((\hat{R}(h) = 0) | (R(h) > \epsilon)) \leq (1 - \epsilon)^m}$$

$$\leq |H|(1 - \epsilon)^m$$

$$\boxed{1 - x \leq \exp(-x)}$$

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

令$|H| e^{-m\epsilon} = \delta$，可得$m \geq \frac{1}{\epsilon}(log|H| + log\frac{1}{\delta})$

一致情形(Consistent case)

令 H表示一个从 X 到 Y的有限映射集合，令A为学习任意目标概念 $c \in H$的算法，并且根据独立同分布样本集S返回一个一致的假设 h_S: $\hat{R}(h_S) = 0$，那么对于任意$\epsilon, \delta > 0$，在以下的条件成立的情况下使得 $\Pr_{S\sim D^m}[R(h_S) \le \epsilon] \ge 1 - \delta$ 成立：

$$m \ge \frac{1}{\epsilon}(log|H| + log\frac{1}{\delta})$$

对于任意$\epsilon, \delta > 0$，至少$1 - \delta$的概率有：

$$R(h_S) \le \frac{1}{m}(log|H| + log\frac{1}{\delta})$$

一致情形(Consistent case)

概念类 C_n ：$x_1, x_2, \ldots, x_n$的合取
目标概念c: $x_1 \wedge \overline{x_2} \wedge x_4$

x_1	x_2	x_3	x_4	$x_1 \wedge \overline{x_2} \wedge x_4$
1	0	0	1	+
1	0	0	0	-

一致情形(Consistent case)

概念类 C_n ：$x_1, x_2, \dots, x_n$的合取
目标概念c: ?

x_1	x_2	x_3	x_4	
1	0	1	0	+

c不包含$\bar{x}_1$

c不包含x_2

c不包含$\bar{x}_3$

c不包含x_4

一致情形(Consistent case)

x_1	x_2	x_3	x_4	x_5	x_6	
0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	

$\rightarrow$ $\overline{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$

一致情形(Consistent case)

x_1	x_2	x_3	x_4	x_5	x_6	
0	1	1	0	1	1	+
0	1	1	1	1	1	+
0	0	1	1	0	1	-
0	1	1	1	1	1	+
1	0	0	1	1	0	-
0	1	0	0	1	1	+
0	1	?	?	1	1	

$\rightarrow$ $\bar{x}_1 \wedge x_2 \wedge x_5 \wedge x_6$

$|H| = |C_n| = 3^n$

样本复杂度：$m \geq \frac{1}{\epsilon}(nlog3 + log\frac{1}{\delta})$

不一致情形(Inconsistent case)

霍夫丁不等式：若$x_1, x_2, \dots, x_m$为m个独立随机变量，且满足$0 \le x_i \le 1$，则对于任意$\epsilon > 0$，有：

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} E(x_i) \ge \epsilon\right] \le \exp\left(-2m\epsilon^2\right),$$

$$Pr\left[\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} E(x_i) \le -\epsilon\right] \le \exp\left(-2m\epsilon^2\right)$$

不一致情形(Inconsistent case)

推论：固定$\epsilon > 0$，则对于任意假设$h: X \to \{1, 0\}$, 有：

$$\Pr_{S\sim D^m}\left(\widehat{R}(h) - R(h) \geq \epsilon\right) \leq \exp\left(-2m\epsilon^2\right),$$

$$\Pr_{S\sim D^m}\left(\widehat{R}(h) - R(h) \leq -\epsilon\right) \leq \exp\left(-2m\epsilon^2\right),$$

$$\Pr_{S\sim D^m}\left(|\widehat{R}(h) - R(h)| \geq \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right)$$

不一致情形(Inconsistent case)

推论：给定$\epsilon > 0$，则对于任意假设$h: X \rightarrow \{1,0\}$, 有：

$$\Pr_{S\sim D^m}\left(\widehat{R}(h) - R(h) \geq \epsilon\right) \leq \exp\left(-2m\epsilon^2\right),$$

$$\Pr_{S\sim D^m}\left(\widehat{R}(h) - R(h) \leq -\epsilon\right) \leq \exp\left(-2m\epsilon^2\right),$$

$$\Pr_{S\sim D^m}\left(|\widehat{R}(h) - R(h)| \geq \epsilon\right) \leq 2\exp\left(-2m\epsilon^2\right)$$

设$\delta = 2\exp\left(-2m\epsilon^2\right)$

$$\epsilon = \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

不一致情形(Inconsistent case)

推论：给定一个假设$h: X \rightarrow \{1,0\}$，对于任意$\delta > 0$，至少$1-\delta$的概率有：

$$R(h) \leq \widehat{R}(h) + \sqrt{\frac{log\frac{2}{\delta}}{2m}}$$

不一致情形(Inconsistent case)

假设抛一个不均匀的硬币，正面朝上的概率是p，而假设h的预测结果总是反面，则真实的误差为$R(h) = p$,训练得到的经验误差为$\hat{R}(h) = \hat{p}$，则有至少$1 - \delta$的概率使得

$$|p - \hat{p}| \leq \sqrt{\frac{log\frac{2}{\delta}}{2m}}$$

不一致情形(Inconsistent case)

- 假设抛一个不均匀的硬币，正面朝上的概率是p，而假设h的预测结果总是反面，则真实的误差为$R(h)=p$,训练得到的经验误差为$\hat{R}(h)=\hat{p}$，则有至少$1-\delta$的概率使得

$$|p-\hat{p}|\le\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

当$\delta=0.02$， $\mathrm{m}=500$时，有至少98%的概率

$$|p-\hat{p}|\le 0.048$$

不一致情形(Inconsistent case)

$$Pr\left[\exists h \in H, |\widehat{R}(h) - R(h)| > \epsilon\right]$$
$$= Pr\left[\left(h_1 \in H, \left|\widehat{R}(h_1) - R(h_1)\right| > \epsilon\right) \vee \left(h_2 \in H, \widehat{R}(h_2) - R(h_2)\right| > \epsilon\right) \vee \cdots\right]$$
$$\leq \sum_{h \in H} Pr[|\widehat{R}(h) - R(h)| > \epsilon]$$
$$\leq 2|\mathrm{H}|\exp(-2m\epsilon^2)$$

不一致情形(Inconsistent case)

$$Pr\left[\exists h \in H, |\hat{R}(h) - R(h)| > \epsilon\right]$$
$$= Pr\left[\left(h_1 \in H, \left|\hat{R}(h_1) - R(h_1)\right| > \epsilon\right) \vee \left(h_2 \in H, \hat{R}(h_2) - R(h_2)\right| > \epsilon\right) \vee \cdots\right]$$
$$\leq \sum_{h \in H} Pr[|\hat{R}(h) - R(h)| > \epsilon]$$
$$\leq 2|\mathrm{H}|\exp(-2m\epsilon^2)$$

令$2|\mathrm{H}|\exp\left(-2m\epsilon^2\right) = \delta$，可得$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2m}}$

不一致情形(Inconsistent case)

令 H表示一个从 X 到 Y的有限映射集，令A为学习任意目标概念 $c \in H$的算法，对于任意$\delta > 0$，有至少$1 - \delta$ 的概率使得：

$$\forall h \in H, R(h) \le \widehat{R}(h) + \sqrt{\frac{log|H| + log\frac{2}{\delta}}{2m}}$$

$$R(h) \le \widehat{R}(h) + O(\sqrt{\frac{log_2|H|}{m}})$$

不可知PAC学习(Agnostic PAC-Learning)

令 H表示一个从 X 到 Y的有限映射集，若存在一个多项式$poly(\cdot,\cdot,\cdot,\cdot)$，对于任意$\epsilon > 0, \delta > 0$，在所有在$X\times Y$上的分布$D$，对于满足$m \geq poly(\frac{1}{\epsilon},\frac{1}{\delta},n,size(c))$，使得下式成立，则$A$是一个不可知PAC学习算法。

$$\Pr_{S\sim D^m}[R(h_S) - \min_{h\in H} R(h) \leq \epsilon] \geq 1 - \delta$$

若 A的运行复杂度在 $poly(\frac{1}{\epsilon},\frac{1}{\delta},n,size(c))$ 内，则A是一个高效不可知PAC学习算法