第四讲 概率近似正确学习框架

牟克典

2021年4月7日

概要

Chernoff界(1)

给定随机变量X和$\epsilon > 0$, 对任意$t > 0$,

$$Pr[X \geq \epsilon] = Pr[e^{tX} \geq e^{t\epsilon}]$$

由Morkov不等式可知

$$Pr[e^{tX} \geq e^{t\epsilon}] \leq \frac{1}{e^{t\epsilon}} E[e^{tX}].$$

因此

$$Pr[X \geq \epsilon] \leq e^{-t\epsilon} E[e^{tX}].$$

Chernoff界(2)

- 基于不等式
$$Pr[X \geq \epsilon] \leq e^{-t\epsilon} E[e^{tX}],$$
  - 寻找$E[e^{tX}]$的上界$g(t)$;
  - 选择t使得$e^{-t\epsilon} g(t)$最小.
- 如何寻找$E[e^{tX}]$的上界$g(t)$?

Hoeffding引理

设X为满足$E[X]=0$且$a \leq X \leq b$的随机变量，其中$a<b$，则对任意$t>0$，下列不等式成立：

$$E[e^{tX}] \leq e^{\frac{t^2(b-a)^2}{8}}.$$

Hoeffding不等式(1)

Hoeffding不等式

设$X_1, \ldots, X_m$为一组独立随机变量且对任一$i \in [1, m]$, X_i在$[a_i, b_i]$中取值. 令$S_m = \sum_{i=1}^{m} X_i$, 则对任意$\epsilon > 0$, 下列不等式成立:

$$(1.)\quad Pr[S_m - E[S_m] \geq \epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

$$(2.)\quad Pr[S_m - E[S_m] \leq -\epsilon] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

Hoeffding不等式(2)

证明：应用Chernoff界技术，可得

$$Pr[S_m - E[S_m] \geq \epsilon] \leq e^{-t\epsilon} E[e^{t(S_m - E[S_m])}].$$

考虑到$X_1, \ldots, X_m$的相互独立性，则

$$E[e^{t(S_m - E[S_m])}] = \prod_{i=1}^{m} E[e^{t(X_i - E[X_i])}].$$

由Hoeffding引理可得 $E[e^{t(X_i - E[X_i])}] \leq e^{\frac{t^2(b_i - a_i)^2}{8}}$,则

$$Pr[S_m - E[S_m] \geq \epsilon] \leq \exp\left(-t\epsilon + t^2 \left(\sum_{i=1}^{m} \frac{(b_i - a_i)^2}{8}\right)\right).$$

Hoeffding不等式(3)

选择$t=\frac{4\epsilon}{\sum_{i=1}^{m}(b_i-a_i)^2}$则得到$\exp\left(-t\epsilon+t^2\left(\sum_{i=1}^{m}\frac{(b_i-a_i)^2}{8}\right)\right)$的最小值

$\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\right)$. 因此可以得到

$$(1.)\quad Pr[S_m-E[S_m]\geq\epsilon]\leq\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\right).$$

同样方式可证明(2.) □

Hoeffding不等式(4)

Hoeffding不等式：关于$|S_m - E[S_m]| \geq \epsilon$的概率界

设$X_1, \ldots, X_m$为一组独立随机变量且对任一$i \in [1, m]$, X_i在$[a_i, b_i]$中取值. 令$S_m = \sum_{i=1}^{m} X_i$, 则对任意$\epsilon > 0$，下列不等式成立:

$$Pr[|S_m - E[S_m]| \geq \epsilon] \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right).$$

Hoeffding不等式

Hoeffding不等式：关于$\frac{1}{m}S_m - E[\frac{1}{m}S_m] \geq \epsilon$的概率界

设$X_1, \ldots, X_m$为一组独立随机变量且满足 $0 \leq X_i \leq 1$, 则对任意$\epsilon > 0$, 下列不等式成立:

(1.) $Pr[\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i] \geq \epsilon] \leq \exp(-2m\epsilon^2)$.

(2.) $Pr[|\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i]| \geq \epsilon] \leq 2\exp(-2m\epsilon^2)$.

概要

- 输入空间 $\mathcal{X}$, 输出空间$\mathcal{Y} = \{0, 1\}$.
- 概念$c : \mathcal{X} \to \mathcal{Y}$是从$\mathcal{X}$到$\mathcal{Y}$的映射.
- 概念类 C: 我们希望学习的概念的集合.
- 输入空间$\mathcal{X}$中的所有样本相互独立且服从同一固定但未知的分布D.

学习问题, 对学习者(学习算法)$\mathcal{A}$:

- 它所考虑的所有可能的概念的集合为假设空间H (未必与C相同).
- 训练数据集$S = \{x_i\}_{i=1}^m$中的每个样例x_i都是依分布D独立同分布产生的，其标记$y_i = c(x_i)$.

学习算法的任务就是

- 就是利用S选择出一个假设$h_S \in H$使其具有相对于目标概念c的较小泛化误差.

泛化误差

给定假设$h \in H$, 目标概念$c \in C$, 隐含分布D, 则h的泛化误差或者风险定义为

$$R(h) = Pr_{x \sim D}[h(x) \neq c(x)] = E_{x \sim D}[I(h(x) \neq c(x))].$$

经验误差

给定假设$h \in H$, 目标概念$c \in C$, 样本$S = \{x_i\}_{i=1}^{m}$, 则h的经验误差或者风险定义为

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} [I(h(x_i) \neq c(x_i))].$$

经验误差与泛化误差的关系：

$$E_{S \sim D^m}[\hat{R}(h)] = R(h).$$

- $O(n)$: $x \in \mathcal{X}$的表示代价上界.
- $size(c)$: 概念c的表示最大代价.

PAC学习(Probably Approximately Correct (PAC)) Learning

- 目标概念类C是PAC可学习的,是指存在一个学习算法$\mathcal{A}$和一个多项式函数$\mathrm{poly}(\cdot,\cdot,\cdot,\cdot)$使得对任意的$1 > \epsilon, \delta > 0$、$\mathcal{X}$上任一分布$D$、任一目标概念$c \in C$,下列不等式

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

对任意的样本容量$m \geq \mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \mathrm{size}(c))$都成立.
- 进一步,如果$\mathcal{A}$的运行时间也是$\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \mathrm{size}(c))$,则称概念类$C$是高效PAC可学习的,且称$\mathcal{A}$为概念类$C$的PAC学习算法.

概要

- 我们考虑$|H| < \infty$，即有限假设集的情形.
- 如果算法$\mathcal{A}$基于训练数据集S选择出的假设h_S在S上预测误差为0:

$$\hat{R}(h_S) = 0$$

  我们就说h_S与S是一致的.
- 对一致的假设集H而言，我们就可以假定目标概念$c \in H$.

定理 1.1

设H是从$\mathcal{X}$到$\mathcal{Y}$的有限的映射集. 设算法$\mathcal{A}$对任一目标概念$c \in H$和独立同分布样本集S, 返回一致假设h_S: $\hat{R}_S = 0$. 则对任意$\epsilon, \delta > 0$, 如果$m \geq \frac{1}{\epsilon}(\log|H| + \log\frac{1}{\delta})$, 不等式

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

成立.
同时,此样本复杂度保证对任意$\delta > 0$,

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

以至少$1 - \delta$的概率成立.

证明： 固定$\epsilon > 0$, 考虑下列概率

$$
\begin{aligned}
& Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
= & Pr\left[\vee_{i=1}^{|H|} (h_i \in H : \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon)\right]
\end{aligned}
$$

证明： 固定$\epsilon > 0$, 考虑下列概率

$$
\begin{aligned}
& Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
= Pr & \left[\vee_{i=1}^{|H|} (h_i \in H : \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon)\right] \\
& \leq \sum_{h \in H} Pr\left[\hat{R}(h) = 0 \wedge R(h) > \epsilon\right]
\end{aligned}
$$

证明： 固定$\epsilon > 0$, 考虑下列概率

$$
\begin{aligned}
& Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
= Pr & \left[\vee_{i=1}^{|H|} (h_i \in H : \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon) \right] \\
& \leq \sum_{h \in H} Pr\left[\hat{R}(h) = 0 \wedge R(h) > \epsilon \right] \\
& \leq \sum_{h \in H} Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon \right]
\end{aligned}
$$

证明： 固定$\epsilon > 0$, 考虑下列概率

$$
\begin{aligned}
& Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
= & Pr\left[\vee_{i=1}^{|H|}(h_i \in H : \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon)\right] \\
\le & \sum_{h \in H} Pr\left[\hat{R}(h) = 0 \wedge R(h) > \epsilon\right] \\
\le & \sum_{h \in H} Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon\right]
\end{aligned}
$$

对任一$R(h) > \epsilon$的$h \in H$来说,

$$
Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon\right] \le (1 - \epsilon)^m.
$$

证明： 固定$\epsilon > 0$，考虑下列概率

$$Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon]$$
$$= Pr\left[\vee_{i=1}^{|H|}(h_i \in H : \hat{R}(h_i) = 0 \wedge R(h_i) > \epsilon)\right]$$
$$\leq \sum_{h \in H} Pr\left[\hat{R}(h) = 0 \wedge R(h) > \epsilon\right]$$
$$\leq \sum_{h \in H} Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon\right]$$

对任一$R(h) > \epsilon$的$h \in H$来说，

$$Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon\right] \leq (1-\epsilon)^m.$$

应用不等式$1 - x \leq e^{-x}$ 可得到

$$\sum_{h \in H} Pr\left[\hat{R}(h) = 0 | R(h) > \epsilon\right] \leq |H|(1-\epsilon)^m \leq |H|e^{-m\epsilon}.$$

进一步得到

$$Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \leq |H| e^{-m\epsilon}.$$

令

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

则

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

此时

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

令 $|H| e^{-m\epsilon} = \delta$, 解得$1-\delta$的概率成立的泛化误差界

$$R(h_S) \leq \frac{1}{m}(\log |H| + \log \frac{1}{\delta}). \qquad \square$$

例：布尔文字的合取

- 布尔变量x_i: 取值要么为0要么为1.
- 布尔文字：要么为x_i，要么为$\bar{x}_i$.
- 目标概念集C_n: 至多n个布尔文字$x_1/\bar{x}_1, \ldots, x_n/\bar{x}_n$的合取式的集合.
- 比如对$n = 4$来说，若目标概念为$x_1 \wedge x_2 \wedge \bar{x}_4$，则$(1, 1, 0, 0)$为正样本，$(1, 1, 1, 1)$为负样本.
- 对正样本$s = (s^{(1)}, s^{(2)}, \cdots, s^{(n)})$来说，
  - 若$s^{(i)} = 1$，则目标概念中不可能含有$\bar{x}_i$;
  - 若$s^{(i)} = 0$，则目标概念中不可能含有x_i.

例：布尔文字的合取

构造如下选择一致假设的算法$\mathcal{A}$:

(1) 对每个正样本 $s = (s^{(1)}, s^{(2)}, \cdots, s^{(n)})$，对$i \in [1, n]$,
- 若$s^{(i)} = 1$，则将$\bar{x}_i$排除在目标概念类的可能文字之外;
- 若$s^{(i)} = 0$，则将x_i排除在目标概念类的可能文字之外.

(2) 没有被排除的所有文字的合取就是和目标概念一致的假设.

例如考虑$S = (s_1, s_2, s_3)$，其中
- 正样本:
  - $s_1 = (1, 0, 1, 0)$: 排除$\bar{x}_1$, x_2, $\bar{x}_3$, x_4;
  - $s_2 = (1, 1, 0, 0)$: 排除$\bar{x}_1$, $\bar{x}_2$, x_3, x_4.
- 负样本:
  - $s_3 = (0, 0, 1, 0)$.

由此返回假设$x_1 \wedge \bar{x}_4$.

例：布尔文字的合取

- 我们以至多由n个布尔文字的合取式的集合作为假设集H.
- $|H| = |C_n| = 3^n$.
- 则由上述定理可知对任意$\epsilon, \delta > 0$, 得到如下样本复杂度的界：

$$m \geq \frac{1}{\epsilon}(n \log 3 + \log \frac{1}{\delta}).$$

因此，至多由n个布尔文字的合取式构成的概念类是PAC可学习的。

概要

1. Hoeffding不等式
2. PAC学习模型
3. 一致有限假设集情形
4. 不一致有限假设集情形

- 考虑$|H| < \infty$, 即有限假设集的情形.
- 对更一般的情形来说, H中未必有假设和训练样本集一致,即
$$\hat{R}(h_S) = 0$$
未必满足.
- 其实不一致($\hat{R}(h_S) > 0$)是更典型的情形.

Hoeffding不等式回顾

Hoeffding不等式：关于$\frac{1}{m}S_m - E[\frac{1}{m}S_m] \geq \epsilon$的概率界

设$X_1, \ldots, X_m$为一组独立随机变量且满足 $0 \leq X_i \leq 1$, 则对任意$\epsilon > 0$, 下列不等式成立:

(1.) $Pr[\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i] \geq \epsilon] \leq \exp(-2m\epsilon^2)$.

(2.) $Pr[|\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m} E[X_i]| \geq \epsilon] \leq 2\exp(-2m\epsilon^2)$.

推论 1.1

固定$\epsilon > 0$,且设S为容量为m的独立同分布样本集， 则对任意假设$h : \mathcal{X} \to \{0, 1\}$, 下列不等式成立

- $Pr_{S\sim D^m}[\hat{R}(h) - R(h) \geq \epsilon] \leq e^{-2m\epsilon^2}$.
- $Pr_{S\sim D^m}[\hat{R}(h) - R(h) \leq -\epsilon] \leq e^{-2m\epsilon^2}$.
- $Pr_{S\sim D^m}[|\hat{R}(h) - R(h)| \geq \epsilon] \leq 2e^{-2m\epsilon^2}$.

推论 1.2单个假设的泛化误差界

固定一个假设$h:\mathcal{X}\to\{0,1\}$，则对任意$\delta>0$，下列不等式至少以$1-\delta$的概率成立:

$$R(h)\le\hat{R}(h)+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

证明概要： 令

$$\delta\ge 2e^{-2m\epsilon^2},$$

则

$$\begin{aligned}
& Pr_{S\sim D^m}[R(h)\le\hat{R}(h)+\epsilon] \\
\ge\ & Pr_{S\sim D^m}[|\hat{R}(h)-R(h)|\le\epsilon] \\
\ge\ & 1-\delta.
\end{aligned}$$

我们考虑一个硬币投掷问题：

- 正面出现的概率为p,
- 假设h的预测总是反面：
  - h的泛化误差$R(h) = p$;
  - h的经验误差$\hat{R}(h) = \hat{p}$.
- 由上面的推论可知

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

  至少以概率$1 - \delta$成立.
- 我们选择$\delta = 0.02$, $m = 500$, 则

$$|p - \hat{p}| \leq \sqrt{\frac{\log \frac{2}{0.02}}{2 \times 500}} \approx 0.068$$

  至少以概率98%成立.

定理 1.2

设H为有限假设集， 则对任意$\delta > 0$, 下列不等式至少以概率$1-\delta$成立：

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

证明概要：

$$
\begin{aligned}
& Pr[\exists h \in H \ s.t. \ |\hat{R}(h) - R(h)| > \epsilon] \\
= \ & Pr[\vee_{h \in H} \left(|\hat{R}(h) - R(h)| > \epsilon \right)] \\
\leq \ & \sum_{h \in H} Pr[|\hat{R}(h) - R(h)| > \epsilon] \\
\leq \ & 2|H|e^{-2m\epsilon^2}.
\end{aligned}
$$

令

$$\delta = 2|H|e^{-2m\epsilon^2},$$

得到

$$\epsilon = \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2m}}. \qquad \square$$

- 对有限假设集H来说，

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2m}}.$$

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{\log_2|H|}{m}}\right).$$

- 对有限假设集H的一致情形来说，

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

$$R(h_S) \leq O\left(\frac{\log_2|H|}{m}\right).$$

概要

- 在更一般的情形中，通常假定$S=\{(x_i,y_i)\}_{i=1}^m$依$\mathcal{X}\times\mathcal{Y}$上的分布$D$独立采样得到.
- 这种情形通常被称为随机情形(stochastic scenarios).
- 任一$h\in H$, h的泛化误差定义为

$$
\begin{aligned}
R(h) &= Pr_{(x,y)\sim D}[h(x)\neq y] \\
&= E_{(x,y)\sim D}[I(h(x)\neq y)].
\end{aligned}
$$

不可知PAC学习(Agnostic PAC-Learning)

- 设H为假设集.$\mathcal{A}$称为一个不可知PAC学习算法,是指存在一个多项式函数$\mathrm{poly}(\cdot,\cdot,\cdot,\cdot)$使得对任意的$1 > \epsilon, \delta > 0$、$\mathcal{X} \times \mathcal{Y}$上任一分布$D$，下列不等式

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

  对任意的样本容量$m \geq \mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \mathrm{size}(c))$都成立.
- 进一步，如果$\mathcal{A}$的运行时间也是$\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, n, \mathrm{size}(c))$， 则称$\mathcal{A}$为高效不可知PAC学习算法.

- 对于确定(deterministic)的情形来说，一个数据点的标记由可测函数c唯一确定（以概率1）.
- 因此， 只考虑按照$\mathcal{X}$上的分布D来抽样数据

$$S = \{x_i\}_{i=1}^{m}.$$

- x_i的标记通过

$$c(x_i) = y_i$$

得到。

- 对目标函数c来说，

$$R(c) = 0.$$

贝叶斯误差(Bayes Error)

- 给定$\mathcal{X}\times\mathcal{Y}$上的分布$D$，贝叶斯误差$R^*$被定义为$\mathcal{X}$到$\mathcal{Y}$的可测函数的误差的下确界：

$$R^* = \inf_{h:\mathcal{X}\to\mathcal{Y}} R(h).$$

- 如果假设h满足$R(h)=R^*$，则称h为贝叶斯假设或者贝叶斯分类器，通常表示为 h_{Bayes}.
- $\forall x\in\mathcal{X}$, $h_{\text{Bayes}}(x) = \operatorname{argmax}_{y\in\{0,1\}} Pr[y|x]$.
- h_{Bayes}在$x\in\mathcal{X}$的预测误差为$\min\{Pr[0|x], Pr[1|x]\}$.
- 给定$\mathcal{X}\times\mathcal{Y}$上的分布$D$，在点$x\in\mathcal{X}$的噪声定义为

$$noise(x) = \min\{Pr[1|x], Pr[0|x]\}.$$

- $noise = E[noise(x)] = R^*$

- 给定假设集H, 则对$h \in H$来说,

$$R(h) - R^* = R(h) - R(h^*) + R(h^*) - R^*,$$

这里$R(h^*) = \inf_{h \in H} R(h)$, h^*是H中具有最小误差的假设。

- $R(h^*) - R^*$为逼近误差(Approximation error), 刻画了假设集H逼近贝叶斯误差的程度.
- $R(h) - R(h^*)$为估计误差(Estimation error), 刻画了假设h相对于 h^*的质量.
- 不可知PAC可学习就是基于估计误差的.

- h_S^{ERM}为依据经验风险最小化(Empirical Risk Minimization)策略基于训练集S选择的假设:

$$h_S^{ERM} = \operatorname*{argmin}_{h \in H} \hat{R}_S(h).$$

- 则

$$\begin{aligned} R(h_S^{ERM}) - R(h^*) &= R(h_S^{ERM}) - \hat{R}(h_S^{ERM}) + \hat{R}(h_S^{ERM}) - R(h^*) \\ &\leq R(h_S^{ERM}) - \hat{R}(h_S^{ERM}) + \hat{R}(h^*) - R(h^*) \\ &\leq 2 \sup_{h \in H} |R(h) - \hat{R}(h)| \end{aligned}$$

- 回顾 $R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{\log_2 |H|}{m}}\right)$, 则随着$|H|$的增大， 右端增大， 而$R(h^*)$减少。

小结

1. Hoeffding不等式
2. PAC学习模型
3. 一致有限假设集情形
4. 不一致有限假设集情形
5. 不可知PAC学习