机器学习理论基础

Foundations of Machine Learning Theory

基本理论：

可学性，复杂性，泛化界，模型选择，稳定性 ...

基本算法：

SVM，Kernel Method，AdaBoost，Ridge Regression，

LASSO，MaxEnt，PCA，Perceptron，Decision Tree ...

dwx

机器学习理论基础

Foundation of Machine Learning Theory

参考书：

- 《机器学习》周志华　西瓜书
- 《统计学习方法》李航
- 进阶
  - 《Pattern Recognition and Machine Learning》M. Bishop　PRML　贝叶斯学派
  - 《The Elements of Statistical Learning》T. Hastie et al.　ESL　频率学派
  - 《Machine Learning》Tom. Mitchell.　1997
  - 《Foundations of Machine Learning》M. Mohri et al.　FML
- 《机器学习理论导引》周志华 et. al.　宝箱书

前序数学知识：

- Calculus.
- Linear Algebra.
- Probability Theory and Mathematical Statistics ☆　测度集中 ☆ (Concentration of measure)
- ☆ Mathematical Optimization　优化
- Information Theory

注意：这是一个理论系列！

1. Introduction

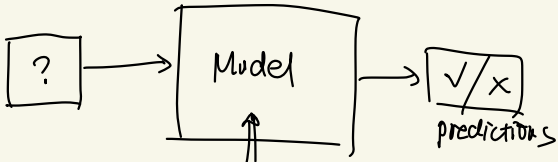
① 定义

Machine Learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions.

天气数据:

$\underbrace{37^\circ C \quad 35^\circ C \quad \cdots \quad 27^\circ C \quad 33^\circ C}_{\text{past information}}$ 明天?

experience → $34^\circ C$



algorithms : time & space complexity,

sample complexity (样本复杂度) → guarantee

Computer scieme
statistics
probablity
optimization

Machine Learning

② 解决什么问题?

- NLP 文本分类 (zero-shot) 情绪分类. 文本生成、名词检测 (ner).
- CV (computer vision): 图象分类. 识别 (object detection) . 分割. 生成
- Speech Recognition.
- Computational Biology
- Applications: 游戏 Go

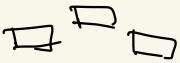
③ 任务？

= 续 < 正 positive / 反 negative (binary ~)

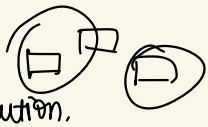
- 分类 Classification → discrete

- 回归 Regression → continuous

- 排名 Ranking



- 聚类 Clustering



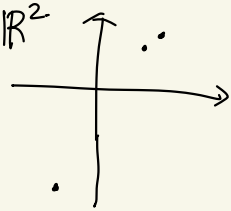
- 降维 Dimensionality Reduction.
  (流形学习)

Vocabulary 10000

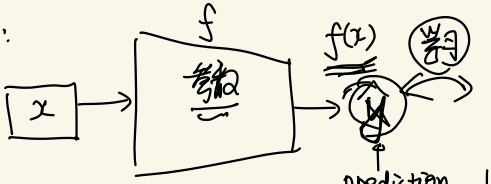
词语 (token) $\mathbb{R}^{10000}$ [0.0, 1, 0.0, ---, 0]

↓

$\mathbb{R}^2$



④ 名词：



prediction　label (标记)

天气

	湿度	天气状况	PM2.5	湿度≥34℃？
昨天	78%	☀	78	✓

features, 特征　　label 标记

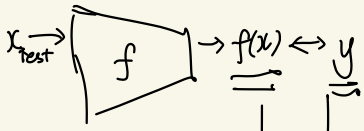
example / instance / item 样例

$f(x) = wx + b \leftrightarrow y$

true label

学习

测试



accuracy 70%

数据集 (X, Y)

训练集 (40%)	验证集 (30%)	测试集 (40%)
training	validation (cross-validation)	test

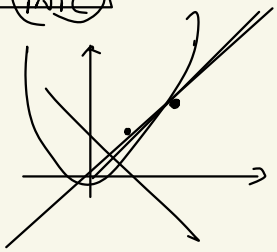
训练：　$x \to \hat{y} = f(x) \sim y$.　$L(y, \hat{y})$: loss function (损失函数)

↳ 得到一组"不错"的 f 函数 ⓗ.

minimize $L(y, \hat{y})$ 拟合

$f(x) = wx + b$

$f(x) = wx^2 + b$



测试：　$x_{test} \to \hat{y}_{test} = f(x_{test}) \sim y_{test}$　score：

→ 反馈给外界而成长

模型选择

→ 用来训练

$x' = x - \alpha \cdot \nabla f(x)$

超参数 (Hyperparameters) → 人为设定的

$\alpha = 0.1, 0.01$

参数 (Learnable Params) → 模型学习得到的

机器学习:

① partitioning

60%	20%	20%

② feature select 特征选择 特征工程 特征学习

③ training → 参数

④ validating → 超参数 (prior knowledge)

⑤ testing (performance)

⑤ 学习场景 Learning Scenarios.

- Supervised learning (监督学习): x < feature, labels

  $y = +/-$
  $y = 0.97$
  $y = 3rd$

  → 类 回归 排名

- Unsupervised learning (无~): x feature.

  clustering

  dimensionality reduction

  $\mathbb{R}^{100} \to \mathbb{R}^2$

- Semi-supervised learning (半监督)

  x x'

  ↳ 弱监督

- 在线 Online learning: training / testing intermixed

  遗憾 (regret)

  $$\sum_{t=1}^{T} l(w_t, (x_t, y_t)) - \min_{w \in W} \sum_{t=1}^{T} l(w, (x_t, y_t))$$

- 强化 Reinforcement Learning.

  "exploration 探索 vs. exploitation 利用 dilemma"

  environment → rewards

  action

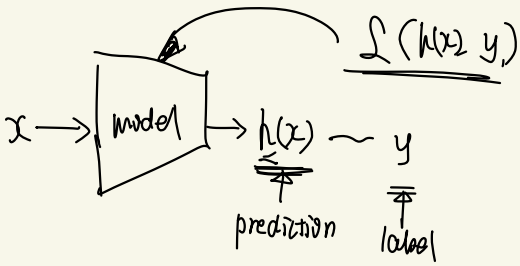
  player

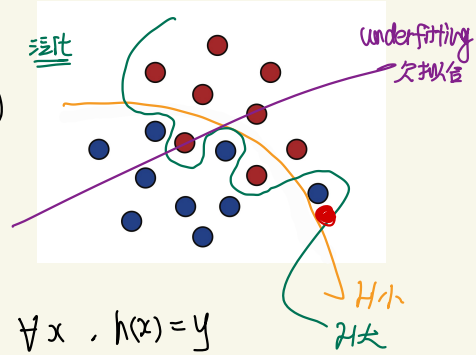
  states

2. The PAC Learning Framework.

minimize $\mathcal{L}$

学习过程:





$\forall x,\ h(x) = y$

h是否存在? ⇒ 可学习的? (learnable) ☆☆ $\boxed{\text{Overfitting}}$

可高效学习? (efficiently learnable) 过拟合 "强解"

二分类

$x \in \mathcal{X}$: input space (the set of all possible examples/instances) e.g. 温度 湿度 pH值 {0%~100%, 0~100}

$y \in \mathcal{Y} = \{0,1\}$ output space (~ labels/target value) {晴, 阴, 雨, 多云, ...}

$\mathcal{Y} = \{-1, 1\}$

$\mathcal{X} \to \mathcal{Y}$ mapping (映射)

- concept 概念 (想要的) "函数" $c \in C$ (Concept Class)
  - $c \subseteq \mathcal{X}$
- hypothesis 假设 (实际可能学到的) $h \in \mathcal{H}$ (Hypothesis Set)

$\mathcal{H}$ 越大越好吗? $\boxed{\text{trade-off}}$

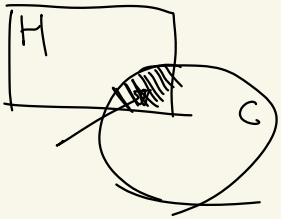
consistent v.s. computation cost.

generalization 泛化 ☆ (易)

↕

specification 特化 (针对于某个特定的训练集)

↑



"consistent"

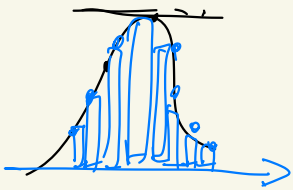
$H \cap C \neq \phi$

"inconsistent"

$H \cap C = \phi$

assumption: distribution theory

$x^{(1)}, \cdots, x^{(m)}$ 抽样(独立同分布抽样) $\sim \mathcal{D}$



i.i.d : independent & identical sampling

估计分布 (贝叶斯学习)

$\Downarrow$

误差(error) 风险(risk)

def (generalization error 泛化误差) $R(h) = \Pr_{x\sim\mathcal{D}}\left(h(x) \neq c(x)\right) = \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{h(x)\neq c(x)}\right]$

def (empirical error 经验误差) $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i)\neq c(x_i)}$

$S = (x_1, \cdots, x_m)$: sample 样本(集) by i.i.d sampling

x_i : example 样例

$$\boxed{\mathbb{E}_{S\sim\mathcal{D}^m}\left[\hat{R}_S(h)\right] = R(h)}$$

Proof: $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\hat{R}_S(h)\right] = \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x_i\sim\mathcal{D}}\left[\mathbb{1}_{h(x_i)\neq c(x_i)}\right] \overset{\text{i.i.d}}{=} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{1}_{h(x)\neq c(x)}\right]$

$= R(h)$ $\square$

(sample mean $\xrightarrow[m\to\infty]{}$ expectation

$\bar{X} \to \mu$ (Law of Large Number))

def (PAC-learning) A concept class $\mathcal{C}$ is PAC-learnable:

☆ if exist an algorithm $\mathcal{A}$. $\forall \varepsilon>0, \delta>0, \forall \mathcal{D}$ (distribution) $\forall c\in\mathcal{C}$.

$$\forall m \geq \text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, \text{size}(c)\right), \quad \Pr_{S\sim\mathcal{D}^m}\left(R(h_S) \leq \varepsilon\right) \geq 1-\delta$$

Probably Approximation Correct

概率近似正确

表示样例 $x \in \mathbb{R}^n$

表示概念 c

近似正确

高概率 (at least $1-\delta$)

Furthermore. $\mathcal{A}$ runs in $\text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, \text{size}(c)\right)$: efficiently PAC-learnable

approximately correct. (error $\le \varepsilon$) $\quad 1-\varepsilon$: accuracy

with high probability ($\ge 1-\delta$) $\quad 1-\delta$: confidence (置信度)

① distribution-free

② training / testing sampling from same distribution

③ concept class C , not particular concept $c \in C$

known $\qquad$ unknown

Example 1. Unbiased Estimator.

$X_1, \cdots, X_m$. $\quad \mathbb{E}[X_i] = \mu$. $\quad Var[X_i] = \sigma^2$ $\quad$ i.i.d. r.v.

h_S: $\quad \bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i \Rightarrow \mathbb{E}[\bar{X}] = \mu$. $\quad Var[\bar{X}] = \frac{\sigma^2}{m}$.

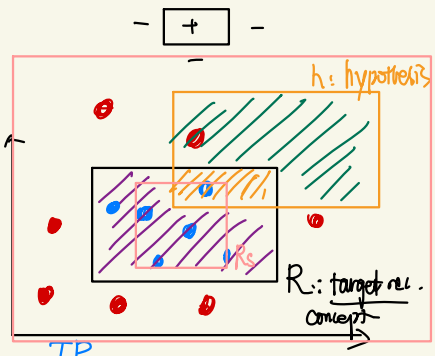
Chebyshev's inequality: $\quad Pr\left(|\bar{X} - \mu| \ge \varepsilon\mu\right) \le \frac{Var[\bar{X}]}{\varepsilon^2\mu^2} = \frac{\sigma^2}{m\varepsilon^2\mu^2} \triangleq \delta$

($\bar{X}$: h_S, μ: c)

$$\Rightarrow m \ge \frac{\sigma^2}{\varepsilon^2\mu^2\delta} = \text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}\right)$$

Concentration Inequality

"集中"



Example 2. Learning Axis-Aligned Rectangles

$\mathcal{X} = \mathbb{R}^2 \quad \mathcal{Y} = \{0, 1\}$

		hypothesis +	hypothesis −
label	+	TP	FN ✓
	−	FP ✗	TN

Precision 查准 $= \frac{TP}{TP+FP}$

Recall 查全 $= \frac{TP}{TP+FN}$

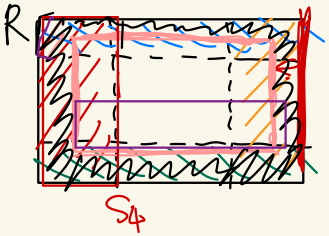
Prove that the concept class C is PAC-learnable.

A: "tightest" rectangle

反例的概率 $\Pr(R_i) > \varepsilon$ 所有矩形

(otherwise, $\Pr(R) \le \varepsilon$. 无需担心情况)

$R = [l, r] \times [b, t]$



$r_4 = [l, S_4] \times [b, t]$

$S_4 = \inf\{s : \Pr([l, s] \times [b, t]) \ge \frac{\varepsilon}{4}\}$

$\overline{r_4} = [l, S_4) \times [b, t]$

tightest R_S if: $R_S \cap r_i \ne \phi \quad \forall i = 1, 2, 3, 4$

error area

otherwise: $\underset{S \sim D^m}{\Pr}(R(R_S) > \varepsilon) \le \underset{S \sim D^m}{\Pr}\left(\bigcup_{i=1}^{4} \{R_S \cap r_i = \phi\}\right)$

一个不在, Upper Bound 至少一个不在

☆(Union Bound) $\le \sum_{i=1}^{4} \underset{S \sim D^m}{\Pr}(\{R_S \cap r_i = \phi\})$

$\le 4\left(1 - \frac{\varepsilon}{4}\right)^m$ $\quad \Pr(r_i) \ge \frac{\varepsilon}{4}$

$\le 4 \exp\left(-\frac{m\varepsilon}{4}\right)$ $\quad 1 + x \le e^x$

$\le \delta$

用 ε 和 δ 表示 m

$\Leftrightarrow \boxed{m \ge \frac{4}{\varepsilon} \log \frac{4}{\delta} = \mathrm{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}\right)}$ → sample complexity 样本复杂度 ☆

$\Rightarrow \forall \varepsilon > 0, \delta > 0, \Pr(R(R_S) > \varepsilon) \le \delta$: PAC-learnable

w.p. $\ge 1 - \delta$

$\boxed{R(R_S) \le \varepsilon = \frac{4}{m} \log \frac{4}{\delta}}$ ← 用 m 和 δ 表示 ε

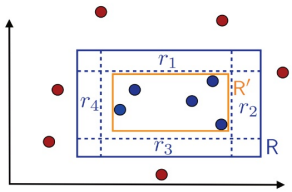
Generalization Bound 泛化界

* Note for the sample problem:

[Learning the axi-aligned rectangles]

令 R 为我们要学得的目标概念（concept），即矩形 R 中的点为正例，R 外的点为反例. 随机采样（取点），$\Pr(R)$ 表示区域 R 的概率质量（几何概型，即点落在区域 R 中的概率）. 给定 $\epsilon > 0$，可以假设 $\Pr(R) > \epsilon$，**否则**，落在 R 外的点的概率 $1 - \Pr(R) > 1 - \epsilon$，无论给定什么样本，我们都认为其是反例，泛化误差将会 $< \epsilon$（且是以概率 1 的，就自然满足PAC可学习了）.

由于 $\Pr(R) > \epsilon$，我们可以从矩形的四条边四个方向出发向内延伸，得到四个矩形区域 r_1, r_2, r_3, r_4，点落在每一个区域的概率都至少为 $\epsilon/4$（延伸到概率**恰好**大于等于 $\epsilon/4$ 的位置），如下图所示. 同时，我们假设 $\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4$ 为矩形区域 r_1, r_2, r_3, r_4 去掉最靠"内部"的一条边所得到的区域，例如：假设 $r_4 = [l, s_4] \times [b, t]$，则 $\bar{r}_4 = [l, s_4) \times [b, t]$，且 $\Pr(r_4) \geq \epsilon/4$，$\Pr(\bar{r}_4) \leq \epsilon/4$（由于区域的构造方式是"恰好"的）.



我们学得的假设对应的区域是矩形 R_S，即 R_S 中的点为正例，R_S 的点为反例. 我们的算法使得 R_S 为"包围"住所有样本点 $\{x_i\}_{i=1}^m \subset \mathbb{R}^2$ 的最小（tight）矩形. 易知 R_S 一定包含在 R 内，因为算法接触到的样本 $S = \{x_i\}_{i=1}^m$ 一定包含在样例 x 所有可能的取值中.

注意到，由于 R_S 为矩形，如果其和 r_1, r_2, r_3, r_4 四个区域都有交集，那么它的四条边应该落在 r_1, r_2, r_3, r_4 四个区域中. 这样一来，误差的区域（error area，即在 R 中但不在 R_S 中的区域）就一定在 $\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4$ 四个区域的**并集**中. 将 $\bar{r}_1, \bar{r}_2, \bar{r}_3, \bar{r}_4$ 的概率相并，通过Union Bound不等式，就可以得到 $\Pr(\bar{r}_1 \cup \bar{r}_2 \cup \bar{r}_3 \cup \bar{r}_4) \leq \sum_{i=1}^4 \Pr(\bar{r}_i) \leq \epsilon$，也即点落在误差区域中的概率将小于等于 ϵ.

考虑**逆否命题**（contraposition），如果点落在误差区域中的概率大于 ϵ，即泛化误差 $R(R_S) > \epsilon$，则 R_S 这个矩形一定会有至少一条边跟 r_1, r_2, r_3, r_4 中的某个区域**没有**交集，因此，"泛化误差大于 ϵ" 的概率就小于等于事件 $\{R_S \cap r_i = \varnothing\}_{i=1}^4$ 的并（事件的蕴含关系，implication）. 这样一来，通过Union Bound，就有

$$\Pr(R(R_S) > \epsilon) \leq \Pr(\bigcup_{i=1}^4 \{R_S \cap r_i = \varnothing\}) \leq \sum_{i=1}^4 \Pr(\{R_S \cap r_i = \varnothing\}) \leq 4\left(1 - \frac{\epsilon}{4}\right)^m \leq 4\exp\left(-\frac{m\epsilon}{4}\right),$$

其中我们用到了 S 中的样本是独立同分布（i.i.d）的假设和放缩不等式 $1 + x \leq \exp(x)$. 令不等式右边为 δ，即可以解出

$$4\exp\left(-\frac{m\epsilon}{4}\right) \leq \delta \quad \Leftrightarrow \quad m \geq \frac{4}{\epsilon}\log\frac{4}{\delta} = \mathrm{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right).$$

又因为表示样本所需的代价 n 为常数（因为样本 $x \in \mathbb{R}^2$，表示两个实数即可），表示概念 c 所需的代价 $size(c)$ 也为常数（因为矩形只需四个点，表示四个实数即可），所以这个问题是PAC可学习的，该算法也是PAC可学习算法.

• def (PAC-learning) A concept class $\mathcal{C}$ is PAC-learnable:

☆ if exist an algorithm $\mathcal{A}$. $\forall \varepsilon>0$. $\delta>0$. $\forall \mathcal{D}$ (distribution) $\forall c\in\mathcal{C}$.

$$\forall m \geq \text{poly}\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, n, \text{size}(c)\right), \quad \Pr_{S\sim\mathcal{D}^m}\left(R(h_S)\leq\varepsilon\right) \geq 1-\delta$$

Probability Approximation Correct

概率 近似 正确

表示 样例 x

表示 概念 c

近似正确

高概率 (at least $1-\delta$)

⟺ Sample complexity (样本复杂度)

generalization bound (泛化误差界): $R(h_S) \leq \text{poly}\left(\frac{1}{m}, \frac{1}{\delta}, n, \text{size}(c)\right)$ w.p. $\geq 1-\delta$

$\mathcal{H}$: infinite → $\boxed{\text{finite } |\mathcal{H}|}$

(Chapter 3) (Chapter 2)

① Consistent case $\quad \mathcal{H}\cap C \neq \phi \Rightarrow \forall h\in\mathcal{H}.\ \hat{R}_S(h)=0$

Thm 1. finite $\mathcal{H}$. consistent case: $\Pr_{S\sim\mathcal{D}^m}\left(R(h_S)\leq\varepsilon\right)\geq 1-\delta$ holds

☆ if $m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right) \Leftrightarrow \boxed{R(h_S) \leq \frac{1}{m}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)}$ w.p. $\geq 1-\delta$.

"uniform convergence bound"

Proof: h_S unknown

Consider $\mathcal{H}_\varepsilon = \{h\in\mathcal{H} : R(h) > \varepsilon\}$ △

probability that hypothesis in $\mathcal{H}_\varepsilon$ is consistent on S drawn i.i.d is

$\forall h\in\mathcal{H}_\varepsilon$: $\quad \Pr(\hat{R}_S(h)=0) \leq (1-\varepsilon)^m$ "consistent"

$$\Pr(\exists h\in\mathcal{H}_\varepsilon : \hat{R}_S(h)=0) = \Pr(\hat{R}_S(h_1)=0 \vee \hat{R}_S(h_2)=0 \vee \cdots \vee \hat{R}_S(h_{|\mathcal{H}_\varepsilon|})=0)$$

$$\leq \sum_{h\in\mathcal{H}_\varepsilon} \Pr(\hat{R}_S(h)=0) \quad \text{(Union Bound)}$$

↑ 泛化误差 $>\varepsilon$

经验误差 $=0$ (一致)

$$\le \sum_{h\in\mathcal{H}_\varepsilon}(1-\varepsilon)^m \le |\mathcal{H}_\varepsilon|\cdot(1-\varepsilon)^m \le |\mathcal{H}|\cdot(1-\varepsilon)^m \le |\mathcal{H}|\,e^{-m\varepsilon} \triangleq \delta .$$

$$\Rightarrow m \ge \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right)$$

$$\Rightarrow \Pr\left(R(h_S)\le\varepsilon\right) \ge 1-\delta \quad \Rightarrow \text{PAC-learnable !} \qquad \square$$

Under finite $|\mathcal{H}|$: consistent algorithm $\mathcal{A}$ is PAC-learnable algorithm.

$$R(h_S) \le \frac{1}{m}\left(\log|\mathcal{H}| + \log\frac{1}{\delta}\right) \quad \text{w.p.} \ge 1-\delta$$

- m 越大越好.
- $|\mathcal{H}|$ 越小越好. (简洁的 $\mathcal{H}$ 更好)

$\log_2|\mathcal{H}|$: # bits to represent hypothesis space

Example. Conjunction of Boolean Literals.

合取范式 (and 合) $\quad x_1 \wedge \bar{x}_2 \wedge x_4 \quad (n=4)$

$(1,0,0,1)$ positive ✓ → $(b_1,\cdots,b_n)$ $\begin{cases} b_i=1. & \text{对 conjunction 中一定无 } \bar{x}_i. \\ b_i=0. & \text{对 conjunction 中一定无 } x_i \end{cases}$

$(1,0,0,0)$ negative → 无用!

给定一堆 P.N examples 训练. 再给你一个 instance → 判定 P/N?

$\mathcal{A}$: 只考察所有 positive examples. $\begin{cases} b_i=1. \to \text{rule out } \bar{x}_i \\ b_i=0 \to \text{rule out } x_i. \end{cases}$

对于每一个 literal → 最终有3种情况 $+$, $-$, $?$ (unknown)

x_1, x_4 $\quad x_2 \quad x_3$

$|\mathcal{H}| = 3^n = |C_n|$

$\Rightarrow$ by Thm, $m \ge \frac{1}{\varepsilon}\left((\log 3)\cdot n + \log\frac{1}{\delta}\right)$

PAC learnable.

PAC efficient learnable ✓

处理每个样本的时间 $O(n)$

Example 2. Universal concept class.

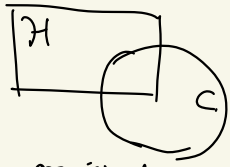
$\mathcal{X} = \{0,1\}^n$ $\quad U_n$: concept class by all subsets of $\mathcal{X}$

$$|\mathcal{X}| = 2^n \Rightarrow |U_n| = 2^{2^n} \le |\mathcal{H}| \Rightarrow m \ge \frac{1}{\varepsilon}\left((\log 2)\cdot 2^n + \log\frac{1}{\delta}\right)$$

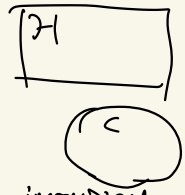
✗ PAC-learnable

② finite $|\mathcal{H}|$ inconsistent case

估计一部分error



consistent



inconsistent

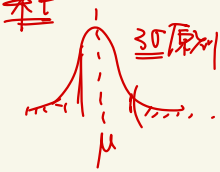
Concentration Inequality

Markov. Chebyshev. Chernoff. $\boxed{\text{Chernoff-Hoeffding.}}$ McDiarmid. Azuma

Hoeffding: $X_1, \cdots, X_m$. $X_i \in [a_i, b_i]$ independent. $S_m = \sum_{i=1}^{m} X_i$

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

"集中"



inconsitent: $\hat{R}_S(h_S) \neq 0$. $\mathbb{E}_S[\hat{R}_S(h_S)] = R(h_S)$.

by Hoeffding's Bound. $\Pr_{S\sim D^m}\left(|\hat{R}_S(h_S) - R(h)| \geq \varepsilon\right) \leq 2\exp(-2m\varepsilon^2) = \delta$

$\Rightarrow$ single hypothesis: $h: \mathcal{X} \to \{0,1\}$ $\forall \delta > 0$. w.p. $\geq 1-\delta$.

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \quad \to \text{对 } h$$

Uniform convergence bound. $\forall h \in \mathcal{H}$

$\boxed{\text{Thm 2 finite } \mathcal{H}. \text{ inconsistent case. } \forall \delta > 0. \text{ w.p.} \geq 1-\delta. \quad \forall h \in \mathcal{H}. \ R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}}$ $\to$ 对 $\mathcal{H}$.

Proof: by Union Bound.

$$\Pr\left(\exists h \in \mathcal{H}, |\hat{R}_S(h) - R(h)| > \varepsilon\right) = \Pr\left(\left(|\hat{R}_S(h_1) - R(h_1)| > \varepsilon\right) \vee \cdots \vee \left(|\hat{R}_S(h_{|\mathcal{H}|}) - R(h_{|\mathcal{H}|})| > \varepsilon\right)\right)$$

$$\leq \sum_{h \in \mathcal{H}} \Pr\left(|\hat{R}_S(h) - R(h)| > \varepsilon\right) \leq 2\exp(-2m\varepsilon^2)|\mathcal{H}| \triangleq \delta.$$

$$\Rightarrow \Pr\left(\forall h \in \mathcal{H}. |\hat{R}_S(h) - R(h)| \leq \varepsilon\right) \geq 1-\delta. \quad \varepsilon = \sqrt{\left(\log|\mathcal{H}| + \log\frac{2}{\delta}\right)/2m} \qquad \square$$

finite $|\mathcal{H}|$ $\begin{cases} \text{consistent: } R(h) \le \dfrac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m} & \text{w.p.} \ge 1-\delta \\ \text{inconsistent: } R(h) \le \hat{R}_S(h) + \sqrt{\dfrac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}} & \text{w.p.} \ge 1-\delta \end{cases}$

$\mathcal{H}$. m $\Rightarrow$ ratio: $\dfrac{\log_2|\mathcal{H}|}{m}$ $\begin{cases} m: \text{inconsistent case 需要更多的 samples (quadratically)} \\ |\mathcal{H}|: \text{simpler } \mathcal{H}. \end{cases}$

☆ $\boxed{\text{Occam's Razor}}$ 奥卡姆剃刀原则. $\rightarrow$ "inductive bias" 归纳偏好

Plurality should not be posited without necessity.

"离散而封闭"　　"坐活必需品"

The simplest implementation is the best! (简单为王)

Deterministic V.S. Stochastic Scenarios

learning: $R(h) = \Pr_{(x,y)\sim D}(h(x) \ne y) = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{h(x)\ne y}]$

$S = ((x_1, y_1), \cdots, (x_m, y_m)) \sim D^m$. $\rightarrow$ deterministic

output label y: probabistic distribution ("输出有多种可能") $\rightarrow$ stochastic

def (Agnostic PAC learning) $D: \mathcal{X} \times \mathcal{Y}$

不可知PAC学习

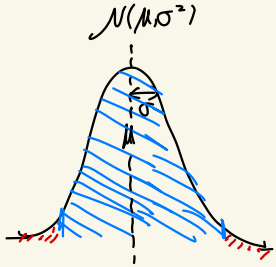
$$\Pr_{S\sim D^m}\left(R(h_S) - \min_{h\in\mathcal{H}} R(h) \le \varepsilon\right) \ge 1-\delta \quad \triangle$$

"没有完美的假设，但有最好的假设"

PAC-learning framework + 2 guarantee

浓度集中 Concentration of Measure

Concentration Inequality,



① Markov inequality: $\Pr(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$ (一阶矩)

② Chebyshev inequality: $\Pr(|X-\mathbb{E}[X]| \geq t) \leq \frac{Var[X]}{t^2}$ (二阶矩)

$\Pr(|f(x)-y| \geq \varepsilon) \leq \delta$

$\hat{R}_S(h)$ $R(h)$

$R(h)=\mathbb{E}[\hat{R}_S(h)]$

Chernoff Bound

"Multiplicative"

$X_1, \cdots, X_n$ independent trials $\in \{0,1\}$

$(p_1, \cdots, p_n)$

$X=\sum_{i=1}^{n} X_i$. $\mu=\mathbb{E}[X]$

$\mu = p_1+\cdots+p_n$

for any $\delta>0$: $\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$

$0<\delta<1$: $\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}$

Proof: $\Pr(X \geq (1+\delta)\mu) = \Pr(e^{tX} \geq e^{t(1+\delta)\mu})$ $(\forall t>0)$

$\leq e^{-t(1+\delta)\mu} \cdot \mathbb{E}[e^{tX}]$

Moment Generating Fmution (MGF)

矩母函数

$\mathbb{E}[e^{tX}] = \mathbb{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] \overset{indep}{=} \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}]$

$= \prod_{i=1}^{n}(\underbrace{e^t \cdot p_i}_{(e^t-1)p_i} + \underbrace{1-p_i}) \underset{(1+x\leq e^x)}{\leq} \prod_{i=1}^{n} e^{(e^t-1)p_i} = e^{(e^t-1)\cdot\sum_{i=1}^{n} p_i} = e^{(e^t-1)\mu}$.

$\Pr(X \geq (1+\delta)\mu) \leq e^{-t(1+\delta)\mu} \cdot e^{(e^t-1)\mu} = e^{(e^t-1-t(1+\delta))\mu}$

$= \left(\frac{e^{e^t-1}}{e^{t(1+\delta)}}\right)^{\mu} = \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$. $t=\ln(1+\delta)$ □

$\delta>0$: $\Pr(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} \leq \begin{cases} e^{-\frac{\mu\delta^2}{3}}, & 0<\delta<1 \\ 2^{-(1+\delta)\mu}, & 1+\delta \geq 2e \end{cases}$

$0<\delta<1$: $\Pr(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \leq e^{-\frac{\mu\delta^2}{2}}$

Multiplicative Chernoff Bound

Chernoff-Hoeffding Bound (Hoeffding's Inequality) ☆☆☆

$\Pr(S_n - \mathbb{E}[S_n] \geq t) \leq \exp(-\frac{2t^2}{n})$

$\Pr(S_n - \mathbb{E}[S_n] \leq -t) \leq \exp(-\frac{2t^2}{n})$

$X_1, \cdots, X_n \in \{0,1\}$. $S_n = \sum_{i=1}^{n} X_i$. $\forall t>0$. $\boxed{\Pr(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{n}\right)}$

$\Rightarrow$ General: $X_1, \cdots, X_n$ independent. $X_i \in [a_i, b_i]$

独立 有界

$$\boxed{\Pr(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)}$$

若 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. X_i i.i.d

$\bar{X}$ 服从二项分布

t大

$\Pr(\bar{X} - p > \varepsilon)$

$= \sum_{k=\lceil (p+\varepsilon)n \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k} \leq e^{-2n\cdot\varepsilon^2}$

二项分布的解释 $\frac{X_i}{n} \in [0, \frac{1}{n}]$

others: Bernstein

Bennett

Sanov.

Hoeffding Lemma

$Y \in [a,b]$. $\mathbb{E}[Y] = 0$ $\Rightarrow$ $\boxed{\mathbb{E}[e^{tY}] \leq e^{\frac{t^2(b-a)^2}{8}}}$ $M_Y(t)$

Proof: $\Phi_Y(t) = \ln \mathbb{E}[e^{tY}]$

只需证 $\Phi_Y(t) \leq \frac{t^2(b-a)^2}{8}$

Taylor's Expansion, $\exists \xi \in [0,t]$.

$\Phi_Y(t) = \underbrace{\Phi_Y(0)}_{0} + t\cdot\underbrace{\Phi_Y'(0)}_{0} + \frac{t^2}{2}\cdot\Phi_Y''(\xi) \leq \frac{t^2(b-a)^2}{8}$

$M_Y(0) = \mathbb{E}[e^{0Y}] = 1$

$M_Y'(0) = \mathbb{E}[Ye^{0Y}] = \mathbb{E}[Y] = 0$

$\Rightarrow \Phi_Y(0) = \ln M_Y(0) = 0$

$\Phi_Y'(0) = \frac{M_Y'(0)}{M_Y(0)} = 0$

只需证 $\Psi_Y''(\xi) \le \frac{(b-a)^2}{4}$

$$\Psi_Y''(\xi) = \left(\frac{M_Y'(\xi)}{M_Y(\xi)}\right)' = \frac{M_Y''(\xi)}{M_Y(\xi)} - \frac{M_Y'(\xi)^2}{M_Y(\xi)^2}$$

$$M_Y(\xi) = \mathbb{E}[e^{\xi Y}]$$

$$= \mathbb{E}\left[\frac{Y^2 e^{\xi Y}}{M_Y(\xi)}\right] - \mathbb{E}\left[\frac{Y e^{\xi Y}}{M_Y(\xi)}\right]^2$$

$$= \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = Var[Z]$$

$Z \in [a,b]$

for a new r.v. Z with $F_Z(z) = \int_{-\infty}^{z} \frac{e^{\xi y}}{\mathbb{E}[e^{\xi Y}]} dF_Y(y)$

"normalize"

$$\Rightarrow Var[Z] = Var\left[Z - \frac{a+b}{2}\right]$$

$$\le \mathbb{E}\left[\left(Z - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}.$$

End of Proof □

$$\Pr\left(|S_n - \mathbb{E}[S_n]| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$ (two-sided)

Proof: $Y_i = X_i - \mathbb{E}[X_i] \Rightarrow Y = S_n - \mathbb{E}[S_n] = \sum_{i=1}^{n} Y_i \Rightarrow \mathbb{E}[Y] = \mathbb{E}[Y_i] = 0$

$$\Pr(S_n - \mathbb{E}[S_n] \ge t) = \Pr(Y \ge t) = \Pr(e^{\lambda Y} \ge e^{\lambda t}) \quad \forall \lambda > 0$$

$$\le e^{-\lambda t} \cdot \mathbb{E}[e^{\lambda Y}] = e^{-\lambda t} \cdot \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda Y_i}\right]$$

$$= e^{-\lambda t} \cdot \prod_{i=1}^{n} \mathbb{E}[e^{\lambda Y_i}] \le e^{-\lambda t} \cdot \prod_{i=1}^{n} e^{\frac{\lambda^2 (b_i-a_i)^2}{8}}$$

Hoeffding Lemma

$X_i \in [a_i, b_i]$

$Y_i \in [a_i - \mu_i, b_i - \mu_i]$

$$= \exp\left(-\lambda t + \frac{\lambda^2}{8} \cdot \sum_{i=1}^{n}(b_i-a_i)^2\right)$$

$$\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$$

$$= \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$ (one-sided)

Similarly, $\Pr\left(S_n - \mathbb{E}[S_n] \le -t\right) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$ □

e.g. Controlling a Fair Voting

n people.

- isolated → independent
- neutral → uniform.

How many peoples are there enough to manipulate the result of a majority vote with $1-\delta$ certainty ?

$S_n = X_1 + \cdots + X_n$. X_i: i.i.d Bernoulli with $p = \frac{1}{2}$,

$\mathbb{E}[S_n] = n/2$.

$\Delta \le t$ 才能控制

$$\Pr\left(\left|S_n - (n - S_n)\right| \ge t\right) = \Pr\left(\left|S_n - \mathbb{E}[S_n]\right| \ge \frac{t}{2}\right)$$

赞方 反方

不能控制的概率

$$\le 2\exp\left(-\frac{2\left(\frac{t}{2}\right)^2}{n \cdot 1}\right) = 2\exp\left(-\frac{t^2}{2n}\right) \le \delta \Rightarrow t \ge \sqrt{2n \log \frac{2}{\delta}}.$$

Chernoff-Hoeffding Bound

$$* \quad \Pr\left(\left|S_n - (n - S_n)\right| \ge t\right) = \Pr\left(S_n \ge \frac{n}{2} + \frac{t}{2}\right) + \Pr\left(S_n \le \frac{n}{2} - \frac{t}{2}\right)$$

$$= \sum_{k \ge \frac{n+t}{2}} \binom{n}{k} \left(\frac{1}{2}\right)^n + \sum_{k \le \frac{n-t}{2}} \binom{n}{k} \left(\frac{1}{2}\right)^n$$

$$= 2^{1-n} \cdot \sum_{k \le \frac{n-t}{2}} \binom{n}{k} \le 2^{1-n} \cdot 2^{n H\left(\frac{1}{2} - \frac{t}{2n}\right)} \approx 2\exp\left(-\frac{t^2}{n}\right)$$

$H(x) = -x \lg x - (1-x) \lg (1-x)$

(entropy bound of a Hamming Ball)

McDiarmid's Inequality.

$X_1, \cdots, X_n$ independent. $X_i \in \mathcal{X}_i \quad \forall i \in [n]$

$f: \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ bounded difference property

$$\forall i \quad \sup_{\substack{x_1 \in \mathcal{X}_1, \cdots, x_n \in \mathcal{X}_n \\ x_i, x_i' \in \mathcal{X}_i}} \left|f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)\right| \le c_i$$

Lipschitz

$$\Rightarrow \forall t > 0 \quad \Pr\left(\left|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

Chernoff-Hoeffding Bound : f : sum of $[a_i, b_i]$ bounded variables

Every Lipschitz function is approximately a constant function in product measure.

the Method of Bounded Difference.

Doob Sequence

$Y_0, Y_1, \cdots, Y_n$ of n-variate function $f: \mathbb{R}^n \to \mathbb{R}$ on r.v. $X_1, \cdots, X_n$.

$$Y_0 = \mathbb{E}[f(X_1, \cdots, X_n)] = \text{const}$$

$$Y_1 = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1]$$

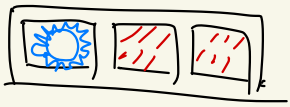
$$Y_2 = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, X_2]$$

$$\vdots$$

$$Y_n = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_n] = f(X_1, \cdots, X_n)$$

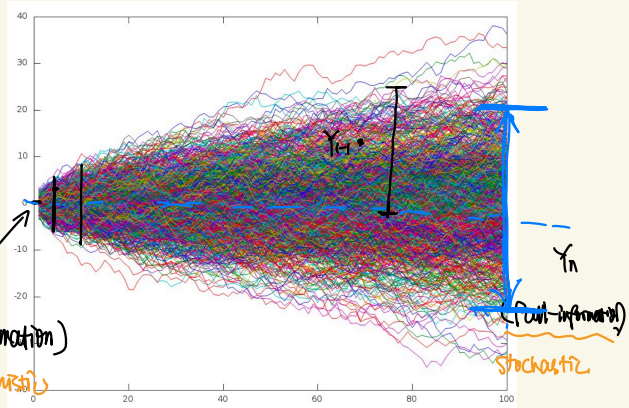
$$Y_i = \mathbb{E}[f(\underbrace{X_1, \cdots, X_i}_{\text{已知}}, \underbrace{X_{i+1}, \cdots, X_n}_{\text{未知}}) \mid X_1, \cdots, X_i]$$

Y_i 随机性的来源



无信息

no-information → full-information



Martingale-property:

$$\mathbb{E}[Y_i \mid X_1, \cdots, X_{i-1}] = Y_{i-1}$$

"期望由前一项决定"

Proof: $\mathbb{E}[Y_i \mid X_1, \cdots, X_{i-1}]$

$$= \mathbb{E}\Big[\mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_i] \mid X_1, \cdots, X_{i-1}\Big]$$

$$= \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_{i-1}]$$

$$= Y_{i-1} \qquad \square$$

$$\mathbb{E}\big[\mathbb{E}[Z \mid Y, X] \mid X\big] = \mathbb{E}[Z \mid X]$$

$$\Pr(|Y_n - Y_0| \geq t)$$

$$\Pr(|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]| \geq t)$$

$\boxed{\text{Martingale}}$ 鞅

def (Martingale) A sequence $\{Y_n : n \geq 0\}$ of r.v. is a martingale w.r.t. $\{X_n : n \geq 0\}$

if $\forall n \geq 0$: $\mathbb{E}[|Y_n|] < \infty$

$$\mathbb{E}[Y_{n+1} \mid X_0, \cdots, X_n] = Y_n \quad \text{(martingale property)}$$

Y_n is a function of $X_0, \cdots, X_n$.

fair gambling game.

e.g. Doob Sequence.

1D-Random Walk: $Y_n = \sum_{i=1}^{n} X_i$, $X_i \in \{-1, +1\}$

① Concentration: $\Pr(|Y_n - Y_0| \geq t) \leq ?$

*② Optional Stopping Theorem (OST)

stopping time τ : $\mathbb{E}[Y_\tau] = \mathbb{E}[Y_t]$

Azuma's Inequality (martingale tail inequality)

if a sequence $\{Y_n : n \geq 0\}$ is a martingale w.r.t. $\{X_n : n \geq 0\}$,

$\forall n \geq 1$. $|Y_n - Y_{n-1}| \leq C_n$.

then $\forall n \geq 1$, $t > 0$.

① $$\Pr(|Y_n - Y_0| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} C_i^2}\right)$$

② Doob Sequence $Y_i = \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_i]$ is a martingale. !

$\Downarrow$

Bounded Difference : $f(X_1, \cdots, X_n)$ bounded.

$\forall i \geq 1$ $\left|\mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_i] - \mathbb{E}[f(X_1, \cdots, X_n) \mid X_1, \cdots, X_{i-1}]\right| \leq C_i$ (average-case)

$\Rightarrow \forall t > 0$. $$\Pr\left(\left|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} C_i^2}\right)$$

McDiarmid's Inequality.

$\sup_{x_1, \cdots, x_n, x_i'} |f(x_1, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq C_i$ (worst-case)

Random Walk (Fair Gambling Game)

Your capital does not change too fast

if { the game is fair (martingale.) / the change to the capital is bounded.

Proof of Azuma's Inequality.

$$\boxed{|Y_n - Y_{n-1}| \leq C_n. \Rightarrow \Pr(|Y_n - Y_0| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} C_i^2}\right)}$$

difference $D_i = Y_i - Y_{i-1}$. $\Rightarrow$ sum $S_n = \sum_{i=1}^{n} D_i = Y_n - Y_0$ 只需证 $\Pr(|S_n| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum C_i^2}\right)$

$\{Y_n : n \geq 0\}$ martingale: $\mathbb{E}[D_n \mid X_0, \cdots, X_{n-1}] = \mathbb{E}[Y_n - Y_{n-1} \mid X_0, \cdots, X_{n-1}]$.

$= \underbrace{\mathbb{E}[Y_n \mid X_0, \cdots, X_{n-1}]}_{\text{(by Martingale Property)}} - \underbrace{\mathbb{E}[Y_{n-1} \mid X_0, \cdots, X_{n-1}]}_{\text{(constant)}} = Y_{n-1} - Y_{n-1} = 0$

plus that: $D_n = Y_n - Y_{n-1} \in [-a_n, b_n]$. s.t. $b_n - a_n = C_n$.

$(|Y_n - Y_{n-1}| \leq C_n)$

by Hoeffding's lemma, $\begin{cases} \mathbb{E}[D_n \mid X_0, \cdots, X_{n-1}] = 0 \\ D_n \in [-a_n, b_n], \text{ s.t. } b_n - a_n = c_n. \end{cases}$ $\Rightarrow \mathbb{E}[e^{\lambda D_n} \mid X_0, \cdots, X_{n-1}] \le e^{\frac{\lambda^2 c_n^2}{8}}$

$$\Rightarrow \mathbb{E}[e^{\lambda S_n}] = \mathbb{E}\left[\mathbb{E}[e^{\lambda S_n} \mid X_0, \cdots, X_{n-1}]\right] \quad (\text{重期望法则})$$

$$= \mathbb{E}\left[\mathbb{E}[e^{\lambda(S_{n-1}+D_n)} \mid X_0, \cdots, X_{n-1}]\right] \quad S_{n-1}\text{确定}$$

$$= \mathbb{E}\left[e^{\lambda S_{n-1}} \cdot \mathbb{E}[e^{\lambda D_n} \mid X_0, \cdots, X_{n-1}]\right]$$

$$\le \mathbb{E}[e^{\lambda S_{n-1}} \cdot e^{\lambda^2 c_n^2/8}] = e^{\frac{\lambda^2 c_n^2}{8}} \cdot \mathbb{E}[e^{\lambda S_{n-1}}]$$

telescoping: $\Rightarrow \mathbb{E}[e^{\lambda S_n}] \le \prod_{i=1}^{n} e^{\frac{\lambda^2 c_i^2}{8}} = \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2\right)$

$$\Pr(Y_n - Y_0 \ge t) = \Pr(S_n \ge t) \le \Pr(e^{\lambda S_n} \ge e^{\lambda t}) \quad \forall \lambda > 0$$

$$\le e^{-\lambda t} \cdot \mathbb{E}[e^{\lambda S_n}] \le \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n} c_i^2\right) = \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

$$\Pr(Y_n - Y_0 \le -t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right) \quad \square$$

$$\lambda = \frac{4t}{\sum_{i=1}^{n} c_i^2}$$