

第七章 概率基础

第 22 讲 概率不等式、随机变量的收敛和随机过程

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

- ① 22.1 概率不等式
- ② 22.2 随机变量的收敛
- ③ 22.3 随机过程

- 1 22.1 概率不等式
- 2 22.2 随机变量的收敛
- 3 22.3 随机过程

引言

- 概率不等式对于一些很难计算的量很有用，也常用于随机变量的收敛定理。
- 统计机器学习方法的泛化能力是指由该方法学习到的模型对未知数据的预测能力，是机器学习方法重要的基本性质。
- 在统计学习理论中，学习方法的泛化能力往往是通过研究泛化误差的概率上界进行的，简称泛化误差上界。而这个泛化误差上界的获得也需要用到概率不等式。

22.1.1 概率不等式：马尔可夫不等式

定理 1

(马尔可夫不等式) 令 X 为一非负随机变量, 假设 $E(X)$ 存在, 对任意 $t > 0$ 有

$$P(X > t) \leq \frac{E(X)}{t}$$

证明.

因为 $X \geq 0$, 所以

$$\begin{aligned} E(X) &= \int_0^{\infty} xf(x) dx = \int_0^t xf(x) dx + \int_t^{\infty} xf(x) dx \\ &\geq \int_t^{\infty} xf(x) dx \geq t \int_t^{\infty} f(x) dx = tP(X > t) \end{aligned}$$

由上述不等式即得定理结论。 □

切比雪夫不等式

定理 2

(切比雪夫不等式) 令 $\mu = E(X)$, $\sigma^2 = V(X)$, 则

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}, \quad P(|Z| \geq k) \leq \frac{1}{k^2}$$

其中, $Z = (x - \mu)/\sigma$, 特别地, $P(|Z| > 2) \leq 1/4$, $P(|Z| > 3) \leq 1/9$

证明.

利用马尔可夫不等式可得

$$P(|X - \mu| \geq t) = P(|X - \mu|^2 \geq t^2) \leq \frac{E(X - \mu)^2}{t^2} = \frac{\sigma^2}{t^2}$$

第二部分令 $t = k\sigma$ 即得。



例 1

假设检验一种预测方法, 涉及 n 种检验情形, 以神经网络为例。如果预测错误则令 $X_i = 1$, 反之则令 $X_i = 0$ 。从而 $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ 是观察到的误差率, 每个 X_i 可认为服从未知均值 p 的伯努利分布。要想知道——但是不知道——真实误差率 p 。从直觉上判断, \bar{X}_n 应与 p 非常接近, \bar{X}_n 不在 p 附近 ϵ 的范围内的概率为多少? 已知

$V(\bar{X}_n) = V(X_1)/n = p(1-p)/n$, 从而

$$P(|\bar{X}_n - p| > \epsilon) \leq \frac{V(\bar{X}_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2}$$

上式利用了不等式 $p(1-p) \leq 1/4$ 。对于 $\epsilon = 2$ 和 $n = 100$, 所求的界为 0.0625。

霍夫丁不等式

定理 3

(霍夫丁 (Hoeffding) 不等式) 令 Y_1, \dots, Y_n 为独立观察值, 满足 $E(Y_i) = 0$, 且 $a_i \leq Y_i \leq b_i$. 令 $\epsilon > 0$, 则对于任意 $t > 0$ 有

$$P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) \leq e^{-te} \prod_{i=1}^n e^{t^2(b_i - a_i)^2/8}$$

证明: 对任意 $t > 0$, 由马尔可夫不等式得

$$\begin{aligned} P\left(\sum_{i=1}^n Y_i \geq \epsilon\right) &= P\left(t \sum_{i=1}^n Y_i \geq t\epsilon\right) = P\left(e^{t \sum_{i=1}^n Y_i} \geq e^{t\epsilon}\right) \\ &\leq e^{-t\epsilon} E\left(e^{t \sum_{i=1}^n Y_i}\right) = e^{-t\epsilon} \prod_i E\left(e^{tY_i}\right) \end{aligned}$$

因为 $a_i \leq Y_i \leq b_i$, 可将 Y_i 写成 a_i, b_i 的凸组合, 即 $Y_i = \alpha b_i + (1 - \alpha)a_i$, 其中, $\alpha = (Y_i - a_i) / (b_i - a_i)$.

所以根据 e^{ty} 的凸性得到

$$e^{tY_i} \leq \frac{Y_i - a_i}{b_i - a_i} e^{tb_i} + \frac{b_i - Y_i}{b_i - a_i} e^{ta_i}$$

两边取期望并利用 $E(Y_i) = 0$ 得

$$E(e^{tY_i}) \leq -\frac{a_i}{b_i - a_i} e^{tb_i} + \frac{b_i}{b_i - a_i} e^{ta_i} = e^{g(\omega)}$$

其中, $u = t(b_i - a_i)$, $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$, $\gamma = -a_i / (b_i - a_i)$ 注意到 $g(0) = g'(0) = 0$ 且对所有 $u > 0$, $g''(u) < 1/4$, 根据泰勒定理, 存在 $\xi \in (0, \mu)$ 满足

$$\begin{aligned} g(u) &= g(0) + ug'(0) + \frac{u^2}{2} g''(\xi) \\ &= \frac{u^2}{2} g''(\xi) \leq \frac{u^2}{8} = \frac{t^2 (b_i - a_i)^2}{8} \end{aligned}$$

因此

$$Be^{(Y_1)} \leq e^{g(\omega)} \leq e^{t^2(b_4 - a_4)^2/8}$$

定理 4

令 X_1, \dots, X_n 服从参数为 p 的伯努利分布, 则对于任意 $\epsilon > 0$ 有

$$P(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

其中, $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$

证明.

令 $Y_i = (1/n)(X_i - p)$, 则 $E(Y_i) = 0$, 令 $a = -p/n, b = (1-p)/n$, 则 $a \leq Y_i \leq b$ 且 $(b-a)^2 = 1/n^2$, 根据霍夫丁不等式得

$$P(\bar{X}_n - p > \epsilon) = P\left(\sum_i Y_i > \epsilon\right) \leq e^{-te} e^{t^2/(8n)}$$



霍夫丁不等式应用

例 2

令 X_1, \dots, X_n 服从参数为 p 的伯努利分布, 令 $n = 100, \epsilon = 0.2$, 由切比雪夫不等式可得

$$P(|\bar{X}_n - p| > \epsilon) \leq 0.0625$$

由霍夫丁不等式得

$$P(|\bar{X}_n - p| \leq 0.2) \leq 2e^{-2(100)(0.2)^2} = 0.00067$$

这比 0.0625 要小很多。

霍夫丁不等式提供了一种建立在参数 p 的二项式分布置信区间的简单方法。

固定 $\alpha > 0$ 并令 $\epsilon_n = \sqrt{\frac{1}{2n} \log(\frac{2}{\alpha})}$. 由霍夫丁不等式可知 $P(|\bar{X}_n - p| > \epsilon_n) \leq 2e^{-2n\epsilon_n^2} = \alpha$

令 $C = (\bar{X}_n - \epsilon_n, \bar{X}_n + \epsilon_n)$, 则 $P(p \notin C) = P(|\bar{X}_n - p| > \epsilon_n) \leq \alpha$. 因此, $P(p \in C) \geq 1 - \alpha$, 也即随机区间 C 包括真实参数 p 的概率为 $1 - \alpha$; 称 C 为 $1 - \alpha$ 置信区间。

Mill 不等式

下面的不等式对于正态分布随机变量的概率范围确定非常有用。

定理 5

(Mill 不等式) 令 $Z \sim N(0, 1)$, 则:

$$P(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$$

22.1.2 有关期望的不等式：柯西-施瓦兹不等式

接下来介绍有关期望的两个不等式。

定理 6

(柯西-施瓦兹不等式) 如果 X 和 Y 具有有限方差, 则

$$E|XY| \leq \sqrt{E(X^2) E(Y^2)}$$

詹森不等式

定理 7

(詹森不等式) 如果 g 为凸函数, 则

$$Eg(X) \geq g(EX)$$

如果 g 为凹函数, 则

$$Eg(X) \leq g(EX)$$

证明.

令直线 $L(x) = a + bx$ 与 $g(x)$ 相切于点 $E(X)$, 因为 g 是凸函数, 它位于直线 $L(x)$ 的上方, 所以

$$Eg(X) \geq EL(X) = E(a + bX) = a + bE(X) = L(E(X)) = g(EX)$$

由詹森不等式可知 $E(X^2) \geq (EX)^2$; 如果 X 为正, 则 $E(1/X) \geq 1/E(X)$; 因为对数函数是凹函数, 所以 $E(\log X) \leq \log E(X)$. □

22.1.3 概率不等式在统计机器学习中的应用：泛化能力分析

泛化误差

如果学到的模型是 \hat{f} , 那么用这个模型对未知数据预测的误差即为泛化误差 (generalization error):

$$R_{exp}(\hat{f}) = E_p[L(Y, \hat{f}(X))] = \int_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{f}(x)) P(x, y) dx dy$$

泛化误差上界

泛化误差的概率上界称为泛化误差上界。具体来说就是通过比较两种学习方法的泛化误差上界的大小来比较它们的优劣。泛化误差上界通常具有以下性质:

- 它是样本容量的函数, 单样本容量增加时, 泛化上界趋于 0
- 它是假设空间容量的函数, 假设空间容量越大, 模型就越难学, 泛化误差上界就越大

二分类问题的泛化误差上界

考虑二分类问题, 已知训练数据集 $\mathbb{T} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, N 是样本容量, \mathbb{T} 是从联合概率分布 $P(X, Y)$ 独立同分布产生的, $X \in \mathbb{R}^n$, $Y \in \{-1, +1\}$ 。假设空间是函数的有限集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$, d 是函数个数。设 f 是从 \mathcal{F} 中选取的函数。损失函数是 0-1 损失。关于 f 的期望风险和经验风险分别是

$$R(f) = E[L(Y, f(X))]$$

$$\hat{R}(f) = \frac{1}{N} \sum_{i=1}^N L(y_i, f(x_i))$$

经验风险最小化函数是

$$f_N = \arg \min_{f \in \mathcal{F}} \hat{R}(f)$$

f_N 依赖训练数据集的样本容量 N 。我们更关心 f_N 的泛化能力

$$R(f_N) = E[L(Y, f_N(X))]$$

接下来我们讨论从有限集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 中任意选出的函数 f 的泛化误差上界。

泛化误差上界

定理 8

[泛化误差上界] 对于二分类问题, 当假设空间是有限个函数的集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 时, 对于任意函数 $f \in \mathcal{F}$, 至少以概率 $1 - \delta$, $0 \leq \delta \leq 1$, 以下不等式成立:

$$R(f) \leq \hat{R}(f) + \epsilon(d, N, \delta) \quad (1)$$

其中

$$\epsilon(d, N, \delta) = \sqrt{\frac{1}{2N}(\log d + \log \frac{1}{\delta})} \quad (2)$$

不等式1左端 $R(f)$ 是泛化误差, 右端即泛化误差的上界。在泛化误差上界中, 第 1 项是训练误差, 训练误差越小, 泛化误差也越小。第 2 项 $\epsilon(d, N, \delta)$ 是 N 的单调递减函数, 当 N 趋于无穷时趋于 0; 同时它也是 $\sqrt{\log d}$ 阶的函数, 假设空间包含的函数越多, 其值越大。

证明.

对任意函数 $f \in \mathcal{F}$, $\hat{R}(f)$ 是 N 个独立的随机变量 $L(Y, f(X))$ 的样本均值, $R(f)$ 是随机变量 $L(Y, f(X))$ 的期望值。如果损失函数取值于区间 $[0, 1]$, 即对所有 i , $[a_i, b_i] = [0, 1]$, 那么由 Hoeffding 不等式不难得知, 对 $\epsilon > 0$, 以下不等式成立:

$$P(R(f) - \hat{R}(f) \geq \epsilon) \leq \exp(-2N\epsilon^2)$$

由于 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ 是一个有限集合, 故

$$\begin{aligned} P(\exists f \in \mathcal{F} : R(f) - \hat{R}(f) \geq \epsilon) &= P\left(\bigcup_{f \in \mathcal{F}} \{R(f) - \hat{R}(f) \geq \epsilon\}\right) \\ &\leq \sum_{f \in \mathcal{F}} P(R(f) - \hat{R}(f) \geq \epsilon) \\ &\leq d \exp(-2N\epsilon^2) \end{aligned}$$

令 $\delta = d \exp(-2N\epsilon^2)$, 则等价地, 对任意 $f \in \mathcal{F}$, 有 $P(R(f) - \hat{R}(f) \geq \epsilon) \leq \delta$

或者有 $P(R(f) < \hat{R}(f) + \epsilon) \geq 1 - \delta$

即至少以概率 $1 - \delta$ 有 $R(f) < \hat{R}(f) + \epsilon$, 其中 ϵ 可从 $\delta = d \exp(-2N\epsilon^2)$ 中反解, 即定理中的表达式1。 □

- 1 22.1 概率不等式
- 2 22.2 随机变量的收敛
- 3 22.3 随机过程

引言

- 概率论最重要的一方面就是关注随机变量序列的趋势，这部分内容称为大样本理论或极限理论或渐进理论. 最基本的问题是：关于随机变量序列 X_1, X_2, \dots 的极限性质可以作何论断？因为统计与数据挖掘涉及大量数据，自然而然地，也会关心当搜集到越来越多的数据时会发生什么样的情况.
- 在概率论中，随机变量序列收敛的概念比实数序列 x_n 收敛于极限 x 更加深奥。例如，假设 X_1, X_2, \dots 为独立同分布随机序列，服从 $N(0, 1)$ 分布，因为所有变量具有相同的分布，所以可以尝试着称 X_n “收敛于” $X \sim N(0, 1)$ ，但这种描述并不十分精确，因为对所有 n , $P(X_n = X) = 0$ (两个连续随机变量相同的概率为 0).
- 很明显，需要其他工具来讨论更严格意义下的随机变量的收敛，主要有两种方法：
 1. **大数定律**说明样本均值 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ 依概率收敛于期望 $\mu = E(X_i)$ ，意味着 \bar{X}_n 以很高的概率趋于 μ .
 2. **中心极限定理**说明 $\sqrt{n}(\bar{X}_n - \mu)$ 依分布收敛于正态分布，意味着对很大的 n ，样本均值渐进服从正态分布.
- 统计机器学习中经验风险收敛到期望风险依赖于大数定律在函数空间中的推广。

22.2.1 收敛的类型：依概率收敛和依分布收敛

定义 1

令 X_1, X_2, \dots , 为随机变量序列, X 为另一随机变量, 用 F_n 表示 X_n 的 CDF, 用 F 表示 X 的 CDF

1. 如果对任意 $\epsilon > 0$, 当 $n \rightarrow \infty$ 时有

$$P(|X_n - X| > \epsilon) \rightarrow 0$$

则 X_n 依概率收敛于 X , 记为 $X_n \xrightarrow{P} X$ 。

2. 如果对所有的 F 的连续点 t , 有

$$\lim_{n \rightarrow \infty} F_n(t) = F(t)$$

则 X_n 依分布收敛于 X , 记为 $X_n \rightsquigarrow X$

当求 X 服从点分布时, 需要改变一下符号, 如果 $P(X = c) = 1$ 且 $X_n \xrightarrow{P} X$, 则记 $X_n \xrightarrow{P} c$, 类似地, 如果 $X_n \rightsquigarrow X$, 则记 $X_n \rightsquigarrow c$.

这里再介绍另外一种形式的收敛，这种收敛对证明概率中的收敛很有用.

定义 2

如果 $n \rightarrow \infty$ 时有

$$E(X_n - X)^2 \rightarrow 0,$$

则称 X_n 均方意义下收敛于 X (也称 L_2 收敛), 记为 $X_n \xrightarrow{qm} X$.

同上面类似, 如果 X 服从在 c 点的点分布, 则用 $X_n \xrightarrow{qm} c$ 代替 $X_n \xrightarrow{qm} X$.

随机变量概率收敛举例

例 3

令 $X_n \sim N(0, 1/n)$, 从直觉上判断, 当 n 很大时, X_n 集中在 0 附近, 所以就希望称 X_n 以概率收敛于 0, 那么现在来看一下是否正确?

令 F 为在零点的点分布的分布函数, 注意到 $\sqrt{n}X_n \sim N(0, 1)$, 令 Z 表示标准正态分布随机变量.

对于 $t < 0$, 因为 $\sqrt{nt} \rightarrow -\infty$, 所以 $F_n(t) = P(X_n < t) = P(Z < \sqrt{nt}) \rightarrow 0$;

对于 $t > 0$, 因为 $\sqrt{nt} \rightarrow \infty$, 所以 $F_n(t) = P(X_n < t) = P(Z < \sqrt{nt}) \rightarrow 1$.

因此, 对所以 $t \neq 0$ 有 $F_n(t) \rightarrow F(t)$, 所以 $X_n \rightsquigarrow 0$. 注意 $F_n = 1/2 \neq F(1/2) = 1$, 所以在 $t = 1$ 处收敛不成立.

这并不影响结果, 因为 $t = 0$ 不是 F 的连续点, 而分布收敛的定义仅需连续的点收敛即可.

现在再考察概率收敛, 对于任意 $\epsilon > 0$, 使用马尔科夫不等式, 当 $n \rightarrow \infty$ 时有

$$P(|X_n| > \epsilon) = P(|X_n|^2 > \epsilon^2) \leq \frac{E(X_n^2)}{\epsilon^2} = \frac{1/n}{\epsilon} \rightarrow 0$$

因此 $X_n \xrightarrow{P} 0$

各种收敛间的关系

定理 9

(a) $X_n \xrightarrow{qm} X$ 意味着 $X_n \xrightarrow{P} X$

(b) $X_n \xrightarrow{P} X$ 意味着 $X_n \rightsquigarrow X$

(c) 如果 $X_n \rightsquigarrow X$ 且对于实数 c 有 $P(X = c) = 1$, 则 $X_n \xrightarrow{P} X$.

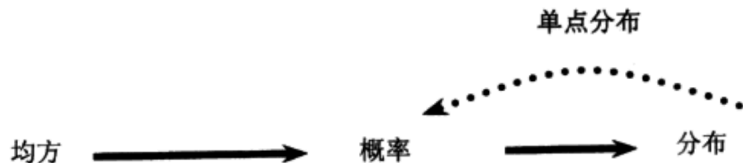


图 1: 各种收敛间的关系

下面来说明反向并不成立.

依概率收敛不能推出均方意义下收敛

令 $U \sim \text{Uniform}(0, 1)$, $X_n = \sqrt{n}I_{0,1/n}(U)$, 则

$$P(|X_n| > \epsilon) = P(\sqrt{n}I_{0,1/n}(U) > \epsilon) = P(0 \leq U < 1/n) = 1/n \rightarrow 0$$

. 因此 $X_n \xrightarrow{P} 0$, 但是对所有 n 有 $E(X_n^2) = n \int_0^{1/n} du = 1$, 所以均方意义下不收敛.

依分布收敛不能推出依概率收敛

令 $X \sim N(0, 1)$, $X_n = -X$, 其中 $n = 1, 2, 3, \dots$;

因此, $X_n \sim N(0, 1)$, 即对所有 n , X_n 与 X 同分布,

所以对所有 x , $\lim_n F_n(x) = F(x)$, 也就是说 $X_n \rightsquigarrow X$

但是 $P(|X_n - X| > \epsilon) = P(|2X| > \epsilon) = P(|X| > \frac{\epsilon}{2}) \neq 0$, 也即 X_n 不依概率收敛于 X .

22.2.2 大数定律

大数定律可以称为是概率论中最伟大的成果，它指出大量样本的均值近似于分布的均值。例如，无数次投掷硬币出现正面的概率趋近于 $1/2$ 。下面对该定律简要描述。

令 X_1, X_2, \dots 为 IID 样本，令 $\mu = E(X_1), \sigma^2 = V(X_1)$ 。则样本均值为 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ ， $E(\bar{X}_n) = \mu, V(\bar{X}_n) = \sigma^2/n$ 。

定理 10

(弱大数定律) 若 X_1, X_2, \dots, X_n 为 IID 样本, 则 $\bar{X}_n \xrightarrow{P} \mu$ 。

弱大数定律的含义: 当 n 逐渐变大时, \bar{X}_n 的分布靠近 μ 。称 \bar{X}_N 为 μ 的一致估计 (一致性)。在定理条件下, 当样本数目 N 无限增加时, 随机样本均值将几乎变成一个常量, 样本方差也依概率收敛于方差 σ^2 。

证明.

假设 $\sigma < \infty$, 该假设并不是必需的, 但有利于简化证明, 利用切比雪夫不等式得:

$$P(|X_n - \mu| > \epsilon) \leq \frac{V(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon}$$

当 $n \rightarrow \infty$ 时, 上式趋于 0。



例 4

假定抛一枚硬币, 出现正面的概率是 p , 令 X_i 表示每次结果, 因此 $p = P(X_i = 1) = E(X_i)$, 当抛 n 次后正面次数所占比例为 \bar{X}_n , 根据大数定律 X_n 以概率收敛于 p , 它意味着当 n 很大时, X_n 的分布会紧密围绕在 p 的附近. 假设 $p = 1/2$, 需要多大的 n 才能使得 $P(0.4 \leq \bar{X}_n \leq 0.6) = 0.7$ 呢?

首先, $E(\bar{X}_n) = p = 1/2$ 且 $V(\bar{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$, 有切比雪夫不等式

$$\begin{aligned} P(0.4 \leq \bar{X}_n \leq 0.6) &= P(|\bar{X}_n - \mu| \leq 0.1) \\ &= 1 - P(|\bar{X}_n - \mu| \geq 0.1) \\ &\geq 1 - \frac{1}{4n(0.1)^2} = 1 - \frac{25}{n} \end{aligned}$$

当 $n = 84$ 时就能保证上式大于 0.7.

22.2.3 大数定律的推广和应用

我们前面介绍过在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上最小化风险泛函的问题

$$R(\alpha) = \int Q(z, \alpha) dF(z), \alpha \in \Lambda$$

其中, 分布函数 $F(z)$ 是未知的, 但给定了依据分布函数抽取的独立同分布数据 z_1, \dots, z_t 。为了求解上述问题, 我们提出了经验风险最小化原则。根据这一原则, 我们用最小化经验风险泛函

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{t=1}^t Q(z_t, \alpha), \alpha \in \Lambda$$

来代替最小化泛函。设

$$Q(z, \alpha_t) = Q(z, \alpha(z_1, \dots, z_t))$$

为最小化泛函的一个函数。经验风险最小化理论的基本问题是描述经验风险最小化原则一致性的条件。下面我们给出一致性的经典定义。

定义 3

对于函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和概率分布函数 $F(z)$, 如果下面两个序列依概率收敛于同一极限

$$R(\alpha_t) \xrightarrow[t \rightarrow \infty]{P} \inf R(\alpha) \quad (3)$$

$$R_{emp}(\alpha_t) \xrightarrow[t \rightarrow \infty]{P} \inf R(\alpha) \quad (4)$$

则, 我们称经验风险最小化原则 (方法) 是一致的。

换句话说, 如果经验风险最小化方法能够提供一个函数序列 $Q(z, \alpha_t), \alpha_t \in \Lambda$ 使得期望风险和经验风险依概率收敛于 (对于给定的函数集) 最小的可能风险值, 则经验风险最小化方法是一致的。方程(3)表明, 对于给定的函数集, 所得风险值序列收敛于最小的可能风险; 方程(4)表明, 经验风险序列的极限估计出风险的最小可能值。

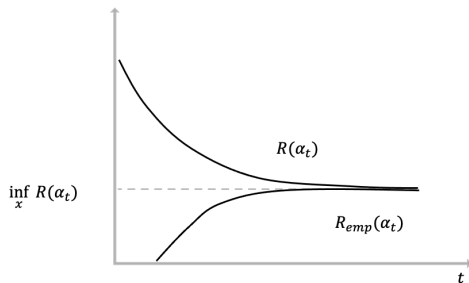


图 2: 如果期望风险 $R(\alpha_t)$ 和经验风险 $R_{emp}(\alpha_t)$ 都收敛于风险最小可能值 $\inf_{\alpha \in \Lambda} R(\alpha)$, 则学习过程是一致的

统计学习理论的核心问题之一是找到经验风险最小化方法的一致性条件。而经验风险最小化的一致性分析在本质上是与两种经验过程的收敛性分析相联系的。

- 设概率分布函数 $F(z)$ 定义在空间 $z \in \mathbb{R}^n$ 上, $Q(z, \alpha), \alpha \in \Lambda$ 为一个 (关于分布 $F(z)$ 的) 可测函数集, 又设 z_1, \dots, z_t, \dots 为一个独立同分布的向量序列。
- 考虑随机变量序列

$$\xi^l = \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^l Q(z_i, \alpha) \right|, l = 1, 2, \dots \quad (5)$$

- 我们将这一依赖于测度 $F(z)$ 和函数集 $Q(z, \alpha), \alpha \in \Lambda$ 的随机变量序列称为双边经验过程。我们的问题是要描述一组条件, 在这组条件下上述经验过程依概率收敛于零。

换句话说, 我们的问题是描述一组条件, 使得对于任意正的 ϵ 下列收敛性成立

定义 4

假设测度 $F(z)$ 和函数集 $Q(z, \alpha), \alpha \in \Lambda$, 对于任意正的 ϵ 有

$$P \left\{ \sup_{\alpha \in \Lambda} \left| \int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^t Q(z, \alpha) \right| > \epsilon \right\} \xrightarrow[t \rightarrow \infty]{} 0$$

我们称上述关系式为给定函数集上均值到数学期望的一致收敛性, 或者简称为一致收敛性。

定义 5

与经验过程 ξ^l 一起, 我们考虑由随机值序列给出的单边经验过程

$$\xi_+^l = \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{l} \sum_{i=1}^t Q(z, \alpha) \right), l = 1, 2, \dots$$

我们称上述关系式为在给定的函数集上均值到数学期望的一致单边收敛性, 或者简称为一致单边收敛性。

关于大数定律及其推广的评述

大数定律

值得注意的是如果函数集 $Q(z, \alpha), \alpha \in \Lambda$ 仅包含一个元素, 则由式5定义的随机变量序列 ξ^l 永远依概率收敛于零。这一事实构成了统计学的主要定律, 即大数定律。随着 l 的增加, 均值序列收敛于随机变量的期望 (如果期望存在)。

大数定律的推广

- 将大数定律推广到函数集 $Q(z, \alpha), \alpha \in \Lambda$ 包含有限个元素的情况是容易的。与包含有限个元素情况相比, 对于包含无穷多个元素的函数集 $Q(z, \alpha), \alpha \in \Lambda$ 随机变量序列 ξ^l 不一定收敛于零, 那么就出现一个问题:
描述函数集 $Q(z, \alpha), \alpha \in \Lambda$ 和概率测度 $F(z)$ 的性质, 使得在这些性质下随机变量序列 ξ^l 依概率收敛于零。
- 在这种情况下, 我们称大数定律在函数空间 (函数 $Q(z, \alpha), \alpha \in \Lambda$ 的空间) 中成立, 或者给定函数集上存在均值到期望的一致 (双边) 收敛性。
- 因此, 函数空间中的大数定律 (均值到期望的一致双边收敛性) 的存在性问题可以看成经典大数定律的推广。

- 应该注意的是，在经典统计学中，并没有考虑一致单边收敛性的存在性问题。由于关键定理指出了分析 ERM 归纳原则一致性问题方法，一致单边收敛性的存在性问题变得重要起来了。
- 一致收敛性意味着，对于充分大的 l ，在给定函数集的所有函数上，经验风险泛函一致地逼近于风险泛函。在前面，我们已经证明，当存在一致收敛性时，最小化经验风险的函数给出了接近于最小可能风险的风险值。
- 所以，一致收敛性给出了经验风险最小化方法一致性的充分条件。在这种情况下，就会出现一个新问题：
是否有可能认为一致收敛性的要求太强？是否存在这样一种情况，经验风险最小化方法是一致的，但一致收敛性不成立？
- 事实上，这样的情况是不可能出现的。可以证明一致单边收敛性不但构成了经验风险最小化方法一致性的充分条件，而且还构成了他的必要条件。

定理 11

设存在常数 α 和 A , 使得对于函数集 $Q(z, \alpha), \alpha \in \Lambda$ 中的所有函数和给定的分布函数 $F(z)$, 有下列不等式成立:

$$\alpha \leq \int Q(z, \alpha) dF(z) \leq A, \alpha \in \Lambda$$

则, 下面两种表述方式是等价的:

1. 对于给定分布函数 $F(z)$, 经验风险最小化方法在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上是严格一致的。
2. 对于给定分布函数 $F(z)$, 在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上出现均值到数学期望的一致单边收敛性的。

推论 1

假设存在常数 α 和 A , 使得对于函数集 $Q(z, \alpha), \alpha \in \Lambda$ 中的所有函数和集合 \mathcal{P} 中的所有分布函数 $F(z)$, 有下列不等式成立:

$$\alpha \leq \int Q(z, \alpha) dF(z) \leq A, \alpha \in \Lambda$$

则, 下面两种表述方式是等价的:

1. 对于集合 \mathcal{P} 中的任意分布函数 $F(z)$, 经验风险最小化方法在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上是严格一致的。
2. 对于集合 \mathcal{P} 中的任意分布函数 $F(z)$, 在函数集 $Q(z, \alpha), \alpha \in \Lambda$ 上出现均值到数学期望的一致单边收敛性的。

表 1: 经典统计学体系和统计学习理论体系的结构

	经典统计学体系	统计学习理论体系
问题的表达	函数的参数估计	利用经验数据最小化期望风险
问题的解决方法	ML 法	ERM 或 SRM 方法
证明	参数估计的有效性	一致大数定律的存在性

22.2.4 中心极限定理

大数定律指出 X_n 的分布会聚集在 μ 附近, 这还不能描述 X_n 的概率性质, 为此还需要中心极限定理.

假设 X_1, X_2, \dots, X_n 为均值 μ , 方差 σ^2 的 IID 序列, 中心极限定理 (CLT) 指出

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

近似服从期望为 μ , 方差为 σ^2/n 的正态分布, 这一结论非常卓越, 因为除了 X_i 的分布的均值和方差需要存在外, 没有其他别的条件.

定理 12

(中心极限定理 (CLT)) 令 X_1, \dots, X_n 的均值为 μ , 方差为 σ^2 的 IID 序列, 令 $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, 则

$$Z_n = \frac{X_n - \mu}{\sqrt{V(\bar{X}_n)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

其中 $Z \sim N(0, 1)$, 换句话说, 下式成立:

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

定理的含义: 有关 \bar{X}_n 概率陈述可以利用正态分布来近似, 注意这仅仅是概率陈述上的近似, 而并不是随机变量本身.

除了 $Z_n \rightsquigarrow N(0, 1)$ 外, 还有其他几个符号可以表示 Z_n 的分布收敛于正态分布, 他们表达的含义本质是一样的, 具体形式如下:

$$Z_n \approx N(0, 1),$$

$$\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n}),$$

$$\bar{X}_n - \mu \approx (0, \frac{\sigma^2}{n}),$$

$$\sqrt{n}(\bar{X}_n - \mu) \approx N(0, \sigma^2),$$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \approx N(0, 1),$$

例 5

假设每个计算机程序产生误差的数量服从均值为 5 的泊松分布, 有 125 个程序, 令 X_1, \dots, X_{125} 分别表示程序中的误差数量, 求 $P(\bar{X}_n < 5.5)$. 令 $\mu = E(X_1) = \lambda = 5, \sigma^2 = V(X_1) = \lambda = 5$ 则

$$P(\bar{X}_n < 5.5) = P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(5.5 - \mu)}{\sigma}\right) \approx P(Z < 2.5) = 0.9938$$

中心极限定理说明 $Z_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ 近似服从 $(0, 1)$, 然而却很少知道 σ , 后面将介绍可以用 X_1, \dots, X_n 的函数

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

去估计 σ^2 . 这又产生了另外一个问题: 如果用 S_n 去代替 σ , 中心极限定理还成立吗? 答案是肯定的.

中心极限定理相关定理

定理 13

假设跟 CLT 相同条件, 则

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \rightsquigarrow N(0, 1)$$

正态近似的精度由 Berry-Essèen 定理给出。

定理 14

(Berry-Essèen 定理) 假设 $E|X_1|^3 < \infty$, 则

$$\sup_z |P(Z_n \leq z) - \Phi(z)| \leq \frac{33}{4} \frac{E|X_1 - \mu|^3}{\sqrt{n}\sigma^3}$$

定理 15

多元中心极限定理 (多元中心极限定理) 令 X_1, \dots, X_n 为 IID 随机向量, 其中 $(X_{1i}, X_{2i}, \dots, X_{ki})^T$ 其均值为

$$\mu = (\mu_1, \mu_2, \dots, \mu_n)^T = (E(X_{1i}), E(X_{2i}), \dots, E(X_{ki}))^T$$

方差矩阵为 σ , 令

$$\bar{X} = (\bar{X}_1, \bar{X}_1, \dots, \bar{X}_1^T)$$

其中 $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ji}$, 则 $\sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \Sigma)$

- 1 22.1 概率不等式
- 2 22.2 随机变量的收敛
- 3 22.3 随机过程**

22.3.1 随机过程

定义 6

一个随机过程 $\{X_t : t \in T\}$ 是一个随机变量集合，通常写成 $X(t)$ 而不是 X_t ，其中变量 X_t 在一个被称作状态空间的集合 \mathcal{X} 里取值，集合 T 被称作指标集，通常可以视为时间。指标集可以是离散的 $T = \{0, 1, 2, \dots\}$ 或者连续的 $T = [0, \infty)$ 。

日常生活中的很多例子包括股票的波动、语音信号、身高的变化都可以看作是随机过程。常见的和时间相关的随机过程模型包括伯努利过程，随机游走、马尔可夫过程等，和空间相关的随机过程通常称为随机场。比如一张二维的图片，每个像素点（变量）通过空间的位置进行索引，这些像素就组成了一个随机过程。

例 6

(IID 观测) 一个 IID 随机变量序列可以写作 $\{X_t : t \in T\}$ ，其中 $T = \{1, 2, 3, \dots\}$ 。因此，一个 IID 随机变量序列就是一个随机过程。

例 7

(天气) 令 $\mathcal{X} = \{\text{晴}, \text{多云}\}$ 。一个典型的序列 (依赖于你住在哪里) 为
晴, 晴, 多云, 晴, 多云, 多云...
该过程具有一个离散的状态空间和一个离散的指标集。

例 8

(经验分布函数) 令 $X_1, \dots, X_n \sim F$ 其中 F 为 $[0, 1]$ 上的某个 CDF. 令

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq t)$$

为经验 CDF. 对于任意固定值 t , $\hat{F}_n(t)$ 是一个随机变量。但是整个经验 CDF

$$\hat{F}_n(t) : t \in [0, 1]$$

为一个具有连续状态空间和连续指标集的随机过程。

22.3.2 马尔可夫链

一个马尔可夫链就是 X_t 的分布只依赖于 X_{t-1} 的随机过程。本节假设状态空间是离散的，或者为 $\mathcal{X} = \{1, \dots, N\}$ 或者 $\mathcal{X} = \{1, 2, \dots\}$ ，且其指标集为 $T = \{0, 1, 2, \dots\}$ 。

定义 7

若

$$P(X_n = x | X_0, \dots, X_{n-1}) = P(X_n = x | X_{n-1})$$

对于所有的 n 和对所有的 $x \in \mathcal{X}$ 成立，则称过程 $\{X_n : n \in T\}$ 是一个马尔可夫链。

马尔可夫链可以用下面的 DAG 来表示：

$$X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n \rightarrow \dots$$

每个变量具有单个母节点，即前一个观测。马尔可夫链理论非常丰富且复杂，主要涉及如下问题：

- 一个马尔可夫链何时“安定”为某种平稳态？
- 如何估计一个马尔可夫链的参数？
- 如何构造一个收敛到既定平稳分布的马尔可夫链和为什么想要那样做？

转移概率

一个马尔可夫链的重要的量为一个状态到另一个状态的概率。一个马尔可夫链是时齐的, 若 $P(X_{n+1} = j | X_n = i)$ 不随着时间而变化。因此, 对于一个时齐马尔可夫链, $P(X_{n+1} | X_n = i) = P(X_1 = j | X_0 = i)$ 。下面只讨论时齐马尔可夫链

定义 8

称

$$p_{ij} = P(X_{n+1} = j | X_n = i)$$

为转移概率, 第 (ij) 个元素为 p_{ij} 的矩阵 P 称作转移矩阵。

注意到 P 具有两个性质 (i) $p_{ij} \geq 0$ 且 (ii) $\sum_i p_{ij} = 1$ 。每行可以看作一个概率密度函数。

例 9

(带吸收壁的随机游动) 令 $\mathcal{X} = \{1, \dots, N\}$ 。假设你正站在这些点中的一个点上, 以 $P(\text{正面朝上}) = p$ 且 $P(\text{反面朝上}) = q = 1 - p$ 的概率投掷一枚硬币。若是正面朝上, 向右走一步, 若是反面朝上, 向左走一步。若你碰上某个终点, 停止。转移矩阵为

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ q & 0 & p & 0 & \dots & 0 & 0 \\ 0 & q & 0 & p & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & p \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}$$

例 10

假设状态空间为 $\mathcal{X} = \{\text{晴}, \text{多云}\}$, 则 X_1, X_2, \dots 表示一系列日子的天气。今天的天气还明显依赖于昨天的天气。它还可能依赖于前两天的天气, 但是作为第一个近似, 可以假设依赖性只倒退一天。在这种情况下, 天气为一个马尔可夫链且一个典型的转移矩阵为

$$P = \begin{pmatrix} 0.4 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}$$

例如, 若今天是晴天, 则明天有 60% 的可能性是多云。

定义 9

令

$$p_{ij}(n) = P(X_{m+n} = j | X_m = i)$$

为在 n 步中从状态 i 转移到状态 j 的概率。令 P_n 表示第 (i, j) 个元素为 $p_{ij}(n)$ 的元素。这些被称为 n 步转移概率。

定理 16

(Chapman-Kolmogorov 方程) n 步概率满足

$$p_{ij}(m+n) = \sum_k p_{ik}(m) p_{kj}(n)$$

仔细观察上述方程，这只不过是矩阵乘法公式。因此证明了

$$P_{m+n} = P_m P_n$$

由定义, $P_1 = P$ 由上述定理, $P_2 = P_{1+1} = P_1 P_1 = PP = P^2$ 按该方法继续下去, 可以看到

$$P_n = P^n = P \times P \times \cdots \times P$$

令 $\mu_n = (\mu_n(1), \dots, \mu_n(N))$ 为行向量, 其中,

$$\mu_n(i) = P(X_n = i)$$

为该链在时刻 n 时出于状态 i 的边际概率。特别地, μ_0 被称作初始分布。为了模拟一个马尔可夫链, 所要知道的就是 μ_0 和 P 。模拟步骤应如下:

- **第一步**产生 $X_0 \sim \mu_0$, 因此 $P(X_0 = i) = \mu_0(i)$
- **第二步**用 i 表示第一步的输出。产生 $X_1 \sim P$ 换句话说, $P(X_1 = j | X_0 = i) = p_{ij}$
- **第三步**假设第二步的输出为 j 。产生 $X_2 \sim P$ 。换句话说 $P(X_2 = k | X_1 = j) = p_{jk}$
- 继续下去。

理解 μ_n 的含义可能比较困难。想象模拟该链许多次, 将所有的链在时刻 n 的输出收集起来。该直方图会近似于 μ_n

定理 17

边际概率可由下式给出

$$\mu_n = \mu_0 P^n$$

状态分类

定义 10

i 到达 j (或 j 从 i 是可达的) 若对于某个 n 有 $p_{ij}(n) > 0$, 且记作 $i \rightarrow j$ 。若 $i \rightarrow j$ 且 $j \rightarrow i$ 则记作 $i \leftrightarrow j$, 并且称 i 和 j 互通。

定理 18

互通关系满足下面的性质

- $i \leftrightarrow i$
- 若 $i \leftrightarrow j$ 则 $j \leftrightarrow i$
- 若 $i \leftrightarrow j$ 且 $j \leftrightarrow k$ 则 $i \leftrightarrow k$
- 状态集 \mathcal{X} 可以写作不相交的类的并 $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \dots$, 其中, 两个状态之间互通当且仅当它们在同一个类中。

状态分类

按互通关系是等价关系，可以把状态空间 S 划分为若干个不相交的集合（或者说等价类），并称之为状态类。

若两个状态互通，则这两个状态属于同一类。

任意两个类或不相交或者相同。

若所有的状态之间是互通的，则该链被称作不可约的。一个状态集是闭的，若一旦进入该状态集则永不出来。只含有单个状态的闭集称作一个吸收态。

例 11

令 $\mathcal{X} = \{1, 2, 3, 4\}$ 且

$$P = \begin{pmatrix} 1/2 & 2/3 & 0 & 0 \\ 2/3 & 1/3 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

类为 $\{1, 2\}, \{3\}, \{4\}$ 。状态 4 为一个吸收态。

假设从状态 i 开始一个链。该链会返回状态 i 吗？若如此，称状态 i 为持久的或常返的。

定义 11

状态 i 为持久的或常返的，若

$$P(X_n = i \text{ 对于某个 } n \geq 1 | X_0 = i) = 1$$

否则，状态 i 为瞬过的。

定理 19

一个状态 i 为常返的当且仅当

$$\sum_n p_{ii}(n) = \infty$$

一个状态为瞬过的当且仅当

$$\sum_n p_{ii}(n) < \infty$$

定理 20

关于常返性的事实

- 若状态 i 为常返的且 $i \leftrightarrow j$, 则 j 是常返的。
- 若状态 i 为瞬过的且 $i \leftrightarrow j$, 则 j 是瞬过的。
- 一个有限马尔可夫链必然至少有一个常返态。
- 一个有限的不可约马尔可夫链的状态都是常返的。

定理 21

(分解定理) 状态空间 \mathcal{X} 可以写成不相交集的并

$$\mathcal{X} = \mathcal{X}_T \cup \mathcal{X}_1 \cup \mathcal{X}_2 \dots$$

其中 \mathcal{X}_T 为瞬过态, 且每个 \mathcal{X}_i 为一个闭的, 不可约的常返态集。

马尔可夫链的收敛性

为了讨论马尔可夫链的收敛性，需要一些定义。

定义 12

假设 $X_0 = i$ 定义常返时间

$$T_{ij} = \min\{n > 0 : X_n = j\}$$

假设 X_n 可返回状态 i ，否则定义 $T_{ij} = \infty$ 一个常返态 i 的平均常返时间为

$$m_i = E(T_{ii}) = \sum_n n f_{ii}(n)$$

其中

$$f_{ij}(n) = P(X_1 \neq j, X_2 \neq j, \dots, X_{n-1} \neq j, X_n = j | X_0 = i)$$

若 $m_i = \infty$ 称一个常返态是零的，否则称之为非零的或正的。

定理 22

若一个状态是零的且是常返的, 则 $p_{ii}^n \rightarrow 0$

定理 23

在一个有限的状态的马尔可夫链里, 所有的常返态都是正的。

周期

例 12

考虑具有三个状态的马尔可夫链，其转移矩阵为

$$\begin{pmatrix} 0 & 1 & 0 \\ & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

假设该链的初始状态为 1，那么将在时刻 3, 6, 9, ... 到达状态 3，这是一个周期链的例子。

若 $p_{ii}(n) = 0$ ，其中 n 不能被 d 整除且 d 是满足该性质的最大整数，则称状态 i 的周期为 d 。因此， $d = \gcd\{n : p_{ii}(n) > 0\}$ ，其中 \gcd 的意思为“最大公约数”。若 $d(i) > 1$ ，则称该链的状态 i 是周期的。若 $d(i) = 1$ 是非周期的。周期为 1 的一个状态被称作非周期的。

定理 24

若状态 i 具有周期 d 且 $i \leftrightarrow j$ ，则 j 也具有周期 d

定义 13

若一个状态是常返的，非零的且周期的，则称这个状态 i 是遍历的。若其所有状态是遍历的，则称这一个链是遍历的。

令 $\pi = (\pi_i : i \in \mathcal{X})$ 为一个非负数向量，且分量和为 1。因此 π 可以视为一个概率密度函数。

定义 14

若 $\pi = \pi \mathbf{P}$ ，则称 π 是一个平稳（或不变）分布。

这里给出直观的思路。 X_0 服从 π 分布并且假设 π 是一个平稳分布。现在根据马尔可夫链的转移概率来抽取 X_1 ，得到 X_1 的分布为 $\mu_1 = \mu_0 \mathbf{P} = \pi \mathbf{P} = \pi$ 。 X_2 的分布为 $\pi \mathbf{P}^2 = (\pi \mathbf{P}) \mathbf{P} = \pi \mathbf{P} = \pi$ ，如此继续下去，会看到 X_n 的分布为 $\pi \mathbf{P}^n = \pi$ 换句话说，若该链在任何时候都具有分布 π ，则它将持续具有分布 π

定义 15

称一个链具体极限分布 π ，若

$$\mathbf{P}^n \rightarrow (\pi, \pi, \dots, \pi)^T$$

对于某个 π ，即 $\pi_j = \lim_{n \rightarrow \infty} P_{ij}^n$ 存在与 i 是独立的。

这里给出收敛性的主要定理, 该定理表明一个遍历链收敛到它的平稳分布。而且, 样本均值收敛到它的平稳分布下的理论期望。

定理 25

一个不可约, 遍历的马尔可夫链具有唯一的平稳分布 π . 极限分布存在且等于 π . 若 g 是任意一个有界函数, 则以概率 1

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N g(X_n) \rightarrow E_{\pi}(g) = \sum_j g(j) \pi_j$$

细致平稳条件

定义 16

若

$$\pi_i p_{ij} = p_{ij} \pi_j$$

则 π 满足细致平衡

细致平衡保证了 π 是一个平稳分布。

定理 26

若 π 满足细致平衡, 则 π 是一个平稳分布。

注意 仅仅因为一个链有一个平稳分布并不意味着它收敛。

例 13

令

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

令 $\pi = 1/3, 1/3, 1/3$ 则 $\pi P = \pi$ 所以 π 是一个平稳分布。若该链是从分布 π 开始的, 它将停留在该分布里。想象模拟许多链且在每个时刻 n 去验证其边际分布。它将永远为均匀分布 π 但是该链没有极限。它将继续循环下去。

例 14

令 $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, 令

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

则 $C_1 = \{1, 2\}$ 且 $C_2 = \{5, 6\}$ 是不可约的闭集。状态 3 和状态 4 是暂留的因为路径为 $3 \rightarrow 4 \rightarrow 6$ 且一旦到达状态 6 就不能返回 3 或 4。因为 $p_{ii}(1) > 0$, 所有的状态都是非周期

例 15

(Hardy-Weinberg) 这里有一个著名的遗传学的例子。假设一个基因可以为 A 型或 a 型。有三种类型的人 (称作基因型): AA, Aa 和 aa 。令 (p, q, r) 表示每种基因型的人的比例。假设每个人将其每个基因型的两个基因复本之一随机地传给其子女。还假设配偶也是被随机选择的。第二个假设在现实生活中是不实际的。然而并不是基于 AA, Aa 或者 aa 来选择配偶又是合理的。想象若将每个人的基因汇聚起来。 A 基因的比例为 $P = p + (q/2)$ 且 a 基因的比例为 $Q = r + (q/2)$ 。一个孩子有基因 AA 的概率为 P^2 , 为 aA 的概率为 $2PQ$, 而且为 aa 的概率为 Q^2 因此 A 基因在这代的比例为

$$P^2 + PQ = (p + \frac{q}{2})^2 + (p + \frac{q}{2})(r + \frac{q}{2})$$

然而, $r = 1 - q - p$ 将其带入上面的方程会得到 $P^2 + PQ = P$ 相似的计算表明 a 基因的比例为 Q 。已经证明 A 型和 a 型的比例为 P 和 Q , 且这将在一代后保持稳定。从第二代开始, 类型为 AA, Aa, aa 的人的比例则为 $(P^2, 2PQ, Q^2)$, 如此下去, 这被称作 Hardy-Weinberg 定律。

假设每个人恰好有一个孩子。现在考虑一个固定的人且令 X_n 为他们的第 n 代的基因型。这是一个马尔可夫链且其状态空间为 $\mathcal{X} = \{AA, Aa, aa\}$ 。一些简单的计算将其表明其转移矩阵为

$$\begin{pmatrix} P & Q & 0 \\ P/2 & (P+Q)/2 & Q/2 \\ 0 & P & 0 \end{pmatrix}$$

例 16

(马尔可夫链蒙特卡罗) 这里简述一种叫马尔可夫链蒙特卡罗 (MCMC) 的模拟方法。令 $f(x)$ 为实轴上的一个概率密度函数且假设 $f(x) = cg(x)$ 其中 $g(x)$ 是一个已知函数且 $c > 0$ 是未知的。原则上可以计算出 c , 因为 $\int f(x)dx = 1$ 意味着 $c = 1/\int g(x)dx$ 。然而, 计算该积分可能行不通, 而且 c 对下面的计算也没有必要。令 X_0 为一个任意开始值。给定 X_0, \dots, X_i 按下面的方法产生 X_{i+1} 。首先, 选取 $W \sim N(X_i, b^2)$ 其中 $b > 0$ 是一个固定的常数。令

$$r = \min \left\{ \frac{g(W)}{g(X_i)}, 1 \right\}$$

选取 $U \sim U(0, 1)$ 且设定

$$X_{i+1} = \begin{cases} W, & U < r \\ X_i & U \geq r \end{cases}$$

在弱条件下, X_0, X_1, \dots 是以一个遍历的马尔可夫链且平稳分布为 f 。因此, 可以将选取出来的变量看作来自 f 的一个样本。

本讲小结

概率不等式

- 马尔可夫不等式
- 切比雪夫不等式
- 霍夫丁不等式
- Mill 不等式
- 关于期望的不等式：柯西-施瓦兹不等式和詹森不等式

收敛性和随机过程

- 收敛类型：依概率收敛和依分布收敛
- 大数定律
- 中心极限定理
- 随机过程
- 马尔可夫链

概率不等式和大数定律在统计学习理论中具有重要应用：PAC 可学性、复杂度、泛化误差界的证明、稳定性、经验风险收敛到期望风险的一致性理论基础等！