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

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。
- 在计算学习理论中，会涉及大量的概率不等式。
- 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：联合界不等式

定理 1

(联合界不等式)

$$P(X \cup Y) \leq P(X) + P(Y)$$

马尔可夫不等式

定理 2

(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

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

证明.

因为 $X>0$，所以

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

由上述不等式即得定理结论。 □

切比雪夫不等式

定理 3

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

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

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

证明.

利用马尔可夫不等式可得

$$P(|X-\mu| \geqslant t) = P\left(|X-\mu|^2 \geqslant t^2\right) \leqslant \frac{E(X-\mu)^2}{t^2} = \frac{\sigma^2}{t^2}$$

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

$$P(|X_n - p| > \varepsilon) \leqslant \frac{V\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}$$

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。

Cantelli 不等式

定理 4

(Cantelli 不等式) $\forall \epsilon > 0$ 有

$$P(X - E(X) \geq \epsilon) \leq \frac{V(X)}{V(X) + \epsilon^2}$$

$$P(X - E(X) \leq -\epsilon) \leq \frac{V(X)}{V(X) + \epsilon^2}$$

Chernoff 不等式

定理 5

(Chernoff 不等式)$\forall t > 0$ 有

$$P(X \geq \epsilon) = P(e^{tX} \geq e^{t\epsilon}) \leq \frac{E(e^{tX})}{e^{t\epsilon}}$$

$\forall t < 0$ 有

$$P(X \leq \epsilon) = P(e^{tX} \geq e^{t\epsilon}) \leq \frac{E(e^{tX})}{e^{t\epsilon}}$$

Chernoff 不等式还有另外一种乘积的形式:

对 m 个独立同分布的随机变量 $X_i \in [0,1], i = 1, 2, \ldots, m$，令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$，对 $r \in [0,1]$ 有

$$P(\bar{X} \geq (1+r)E(\bar{X})) \leq e^{-mr^2 E(\bar{X})/3}$$

$$P(\bar{X} \leq (1-r)E(\bar{X})) \leq e^{-mr^2 E(\bar{X})/2}$$

霍夫丁不等式

定理 6

(霍夫丁 (Hoeffding) 不等式) 对 m 个独立随机变量 $X_i \in [0,1], i = 1, 2, \ldots, m$，令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ 有

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-2m\epsilon^2}$$

常用到 Hoeffding 不等式的另一种表达形式，令 $\delta = e^{-2m\epsilon^2}$，则至少以 $1-\delta$ 的概率有

$$\bar{X} \leq E(\bar{X}) + \sqrt{\frac{1}{2m}\ln\frac{1}{\delta}}$$

若考虑 $X_i \in [a, b]$，则得到 Hoffing 不等式更一般的形式：

$$P(\bar{X} - E(\bar{X}) \geq \epsilon) \leq e^{-2m\epsilon^2/(b-a)^2}$$

$$P(\bar{X} - E(\bar{X}) \leq -\epsilon) \leq e^{-2m\epsilon^2/(b-a)^2}$$

霍夫丁不等式应用

定理 7

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

$$P\left(\left|\bar{X}_n - p\right| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}$$

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

例 2

令 $X_1, ..., X_n$ 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

$$P\left(\left|\bar{X}_n - p\right| > \varepsilon\right) \leqslant 0.0625$$

由霍夫丁不等式得

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

这比 0.0625 要小很多。

霍夫丁不等式应用

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

$$\varepsilon_n = \sqrt{\frac{1}{2n} \log\left(\frac{2}{\alpha}\right)}.$$

由霍夫丁不等式可知

$$P\left(\left|\bar{X}_n - p\right| > \varepsilon_n\right) \leqslant 2e^{-2n\varepsilon_n^2} = \alpha$$

令 $C = \left(\bar{X}_n - \varepsilon_n, \bar{X}_n + \varepsilon_n\right)$，则 $P(p \notin C) = P\left(\left|\bar{X}_n - p\right| > \varepsilon_n\right) \leqslant \alpha$。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。

McDiarmid 不等式

Hoeffding 不等式是 McDiarmid 不等式的特例。

定理 8

(McDiarmid 不等式) 对 m 个独立随机变量 $X_i \in \mathcal{X}, i = 1, 2, \ldots, m$，若 $f: \mathcal{X}^m \to \mathbb{R}$ 是关于 X_i 的实值函数且 $\forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$ 都有

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i$$

则 $\forall \epsilon > 0$

$$P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m)) \geq \epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^m c_i^2}$$

$$P(f(X_1, \ldots, X_m) - E(f(X_1, \ldots, X_m)) \leq -\epsilon) \leq e^{-2\epsilon^2 / \sum_{i=1}^m c_i^2}$$

Bennett 不等式

定理 9

(Bennett 不等式) 对 m 个独立同分布随机变量 $X_i \in \mathcal{X}, i = 1, 2, \ldots, m$，令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$，若 $X_i - E(X_i) \leq 1$，则

$$P(\hat{X} \geq E(\hat{X}) + \epsilon) \leq \exp(\frac{-m\epsilon^2}{2\,V(X_1) + 2\epsilon/3})$$

在机器学习研究中常用到 Bennett 不等式的另一种形式：

若

$$P(\hat{X} \geq E(\hat{X}) + \epsilon) \leq \exp(\frac{-m\epsilon^2}{2\,V(X_1) + 2\epsilon/3}) = \delta$$

则下式至少以 $1 - \delta$ 的概率成立

$$\bar{X} \leq E(\bar{X}) + \epsilon \leq E(X) + \frac{2\ln 1/\delta}{3m} + \sqrt{\frac{2\,V(X_1)}{m}\ln\frac{1}{\delta}}$$

Bernstein 不等式

定理 10

(Bernstein 不等式) 对 m 个独立同分布随机变量 $X_i \in \mathcal{X}, i = 1, 2, \ldots, m$，令 $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$，若存在 $b > 0$ 使得 $\forall k \geq 2$ 有 $E(|X_i|^k) \leq k! b^{k-2} V(X_1)/2$ 成立，则有

$$P(\bar{X} \geq E(\bar{X}) + \epsilon) \leq \exp\left(\frac{-m\epsilon^2}{2V(X_1) + 2b\epsilon}\right)$$

Azuma 不等式

定理 11

(Azuma 不等式) 对于均值为 μ 的鞅 *(martingale)*$\{Z_m, m \geq 1\}$，令 $Z_0 = \mu$，若 $-c_i \leq Z_i - Z_{i-1} \leq c_i$，则 $\forall \epsilon > 0$ 有

$$P(Z_m - \mu \geq \epsilon) \leq e^{-\epsilon^2/2\sum_{i=1}^m c_i^2}$$

$$P(Z_m - \mu \leq -\epsilon) \leq e^{-\epsilon^2/2\sum_{i=1}^m c_i^2}$$

令 $X_i = Z_i - Z_{i-1}$，可以得到鞅差序列（*martingale difference sequence*）$X_1, X_2, \ldots, X_m$，于是有

$$P(\sum_{i=1}^m X_i \geq \epsilon) \leq e^{-\epsilon^2/2\sum_{i=1}^m c_i^2}$$

$$P(\sum_{i=1}^m X_i \leq -\epsilon) \leq e^{-\epsilon^2/2\sum_{i=1}^m c_i^2}$$

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 12

(Mill 不等式) 令 $Z \sim N(0,1)$，则：

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

22.1.2 有关期望的不等式：詹森不等式

定理 13

(詹森不等式) 如果 g 为凸函数，则

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

如果 g 为凹函数，则

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

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

$$\mathrm{Eg}(X) \geqslant EL(X) = E(a + bX) = a + bE(X) = L(E(X)) = g(EX)$$

由詹森不等式可知 $E\left(X^2\right) \geqslant (EX)^2$; 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 $E(\log X) \leqslant \log E(X)$。 □

柯西 -施瓦兹不等式和 Hölder 不等式

定理 14

(柯西 -施瓦兹不等式) 如果 X 和 Y 具有有限方差，则

$$E|XY| \leqslant \sqrt{E\left(X^2\right) E\left(Y^2\right)}$$

定理 15

(Hölder 不等式) 如果 X 和 Y 具有有限方差，对 $p, q \in \mathbb{R}_+$ 且 $\frac{1}{p}+\frac{1}{q}=1$，则有

$$E|XY| \leqslant (E[|X|^p])^{\frac{1}{p}} (E[|Y|^q])^{\frac{1}{q}}$$

Lyapunov 不等式、Minkowski 不等式和 Bhatia-Davis 不等式

定理 16

(Lyapunov 不等式) 对于 $0 < r \leq s$，有

$$\sqrt[r]{E(|X|^r)} \leq \sqrt[s]{E(|X|^s)}$$

定理 17

(Minkowski 不等式) 对于 $1 \leq p$，有

$$\sqrt[p]{E(|X+Y|^p)} \leq \sqrt[p]{E(|X|^p)} + \sqrt[p]{E(|Y|^p)}$$

定理 18

(Bhatia-Davis 不等式) 对 $X \in [a, b]$，有

$$V(X) \leq (b - E(X))(E(X) - a) \leq \frac{(b-a)^2}{4}$$

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):

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

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 $\mathbb{T}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$N$ 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 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$，$d$ 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 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$ 中任意选出的函数 f 的泛化误差上界。

泛化误差上界

定理 19

[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 $\mathcal{F}=\{f_1,f_2,\cdots,f_d\}$ 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：

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

其中

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

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

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

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)。 □

引言

- 机器学习和数据挖掘涉及大量数据，一个核心问题是当搜集到越来越多的数据时会发生什么样的情况？这涉及到研究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 $X_1, X_2, \cdots$ 的极限性质可以作何论断？因此需要讨论严格意义下随机变量的收敛问题，主要有两种思想：

1. **大数定律**说明样本均值 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 依概率收敛于期望 $\mu = E(X_i)$, 意味着 $\overline{X}_n$ 以很高的概率趋于 μ.
2. **中心极限定理**说明 $\sqrt{n}(\overline{X}_n - \mu)$ 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 $X_1, X_2, \cdots,$ 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 $n \to \infty$ 时有

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

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

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

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 $n \to \infty$ 时有

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

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 依概率收敛于 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 $\sqrt{n}t \to -\infty$，所以 $F_n(t) = P(X_n < t) = P(Z < \sqrt{n}t) \to 0$；

对于 $t>0$，因为 $\sqrt{n}t \to \infty$，所以 $F_n(t) = P(X_n < t) = P(Z < \sqrt{n}t) \to 1$。

因此，对所有 $t \neq 0$ 有 $F_n(t) \to F(t)$，所以 $X_n \rightsquigarrow 0$。注意 $F_n(0) = 1/2 \neq F(1/2) = 1$，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 $n \to \infty$ 时有

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

因此 $X_n \xrightarrow{P} 0$

各种收敛间的关系

定理 20

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

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

. 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 $n = 1, 2, 3, \cdots$;

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

令 $X_1, X_2, \cdots$ 为 *IID* 样本，令 $\mu = E(X_1), \sigma^2 = V(X_1)$. 则样本均值为 $\overline{X}_n = \frac{1}{n}\sum_{i=1} X_i$, $E(\overline{X}_n) = \mu$, $V(\overline{X}_1) = \sigma^2/n$。

定理 21

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

弱大数定律的含义: 当 n 逐渐变大时，X_n 的分布靠近 μ. 称 $\overline{X}_N$ 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:

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

当 $n \to 0$ 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 $\overline{X}_n$，根据大数定律 X_n 依概率收敛于 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 $P(0.4 \le \overline{X}_n \le 0.6) = 0.7$ 呢?

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

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

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 $z_1,\ldots,z_t$。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

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

来代替最小化泛函。设

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

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

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

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

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。

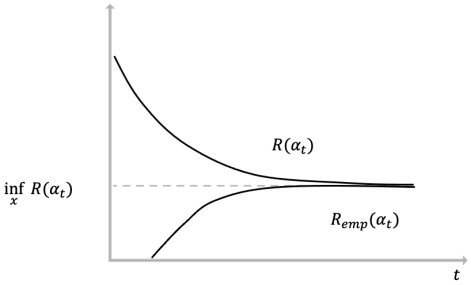


图 2: 如果期望风险 $R(\alpha_t)$ 和经验风险 $R_{emp}(\alpha_t)$ 都收敛于风险最小可能值 $\inf_{\alpha\in\Lambda} R(\alpha)$, 则学习过程是一致的

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双边经验过程和单边经验过程两种经验过程的收敛性分析相联系的。

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。

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。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。

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。

Table 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 $X_1, X_2, ..., X_n$ 为均值 μ, 方差 σ^2 的 IID 序列, 中心极限定理 (CLT) 指出

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

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 22

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

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

其中 $Z \sim N(0,1)$，换句话说，下式成立:

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

定理的含义: 有关 $\overline{X}_n$ 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:

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

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

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

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

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

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 $X_1, \cdots, X_{125}$ 分别表示程序中的误差数量，求 $P(\overline{X}_n < 5.5)$。

令 $\mu = E(X_1) = \lambda = 5, \sigma^2 = V(X_1) = \lambda = 5$ 则

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

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

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

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 23

假设跟 CLT 相同条件，则

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

正态近似的精度由 Berry-Essèen 定理给出。

定理 24

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

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

定理 25

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

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

方差矩阵为 Σ, 令

$$\overline{X} = \left(\overline{X_1}, \overline{X_2}, \ldots, \overline{X_k}\right)$$

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

22.3.1 随机过程

定义 4

一个随机过程 $\{X_t : t \in T\}$ 是一个随机变量集合，通常写成 $X(t)$，其中变量 X_t 在状态空间（每个时刻可能结果的集合）$\mathcal{X}$ 里取值，集合 T 被称作指标集，通常表示不同的时间瞬间。若指标集是离散的自然数集 $T = \{0, 1, 2, \dots\}$ 或者连续的实数集 $T = [0, \infty)$，则分别称为离散时间随机过程或连续时间随机过程。

因为状态空间 $\mathcal{X}$ 可以分为离散状态空间和连续状态空间，所以更进一步可把随机过程分为：

- 离散时间离散空间随机过程
- 离散时间连续空间随机过程
- 连续时间离散空间随机过程
- 连续时间连续空间随机过程

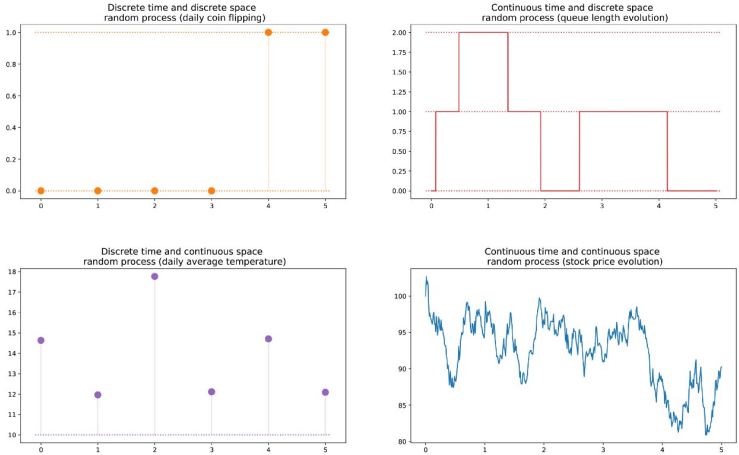

Discrete time and discrete space
random process (daily coin flipping)
Continuous time and discrete space
random process (queue length evolution)
Discrete time and continuous space
random process (daily average temperature)
Continuous time and continuous space
random process (stock price evolution)

图 3: 四种类型随机过程

例 6

(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...

该过程是具有一个离散的状态空间和一个离散的指标集离散随机过程。

例 7

(经验分布函数) 令 $X_1, \ldots, X_n \sim F$ 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。

机器学习、数据科学领域会用到很多常见的随机过程：

- 马尔可夫链
- 高斯过程
- 泊松过程
- 自回归模型
- 移动平均模型
- $\cdots$

22.3.2 马尔可夫链：马尔可夫性质

定义 5

当一个随机过程在给定现在状态及所有过去状态情况下，其未来状态的条件概率分布仅依赖于当前状态；换句话说，在给定现在状态时，它与过去状态（即该过程的历史路径）是条件独立的，那么称此随机过程具有马尔可夫性质（*Markov property*）。

马尔可夫性质也被称为“无记忆性（memory lessness）”，即 $t+1$ 步的随机变量在给定第 t 步随机变量后与其余的随机变量条件独立。

具有马尔可夫性质且定义在离散指标集和状态空间中的随机过程称为离散时间马尔可夫链（Discrete-Time MC, DTMC）；具有马尔可夫性质且定义在连续指标集中的随机过程称为连续时间马尔可夫链（Continuous-Time MC, CTMC）。通常把离散时间马尔可夫链简称为马尔可夫链，把连续时间马尔可夫链称为马尔可夫过程（Markov process）。本讲我们主要讲述离散时间马尔可夫链。

马尔可夫链的数学定义

马尔可夫链是指 X_t 的分布只依赖于 X_{t-1} 的随机过程。通常假设状态空间是离散的，记为 $\mathcal{X} = \{1, \ldots, N\}$ 或者 $\mathcal{X} = \{1, 2, \ldots\}$，且其指标集为 $T = \{0, 1, 2, \ldots\}$。

定义 6

若

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

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（有向非循环图）来表示：

$$X_0 \to X_1 \to X_2 \to \cdots \to X_n \to \ldots$$

每个变量具有单个母节点，即前一个观测。

一个常见的解释性例子：简化的股票涨跌模型

简化的股票涨跌模型：若一天中某股票上涨，则明天该股票有概率 p 开始下跌，$1-p$ 继续上涨；若一天中该股票下跌，则明天该股票有概率 q 开始上涨，$1-q$ 继续下跌。该股票的涨跌情况是一个马尔可夫链，且定义中各个概念在例子中有如下对应：

- 随机变量：第 t 天该股票的状态；状态空间：“上涨”和“下跌”；指数集：天数。
- 条件概率关系：按定义，即便已知该股票的所有历史状态，其在某天的涨跌也仅与前一天的状态有关。
- 无记忆性：该股票当天的表现仅与前一天有关，与其他历史状态无关（定义条件概率关系的同时定义了无记忆性）。
- 停时前后状态相互独立：取出该股票的涨跌记录，然后从中截取一段，我们无法知道截取的是哪一段，因为截取点，即停时 t 前后的记录（$t-1$ 和 $t+1$）没有依赖关系。

马尔可夫链的理论性质

马尔可夫链理论非常丰富且复杂，主要涉及：

- 转移理论：用于马尔可夫链的结构表征或者动态性表示问题。给定时间的概率分布可能取决于一个或过去和/或未来时间的多重性，所有这些可能的时间，使得描述随机过程的动态性变得困难，根据马尔可夫性质，马尔可夫链的随机动态非常容易定义。实际上，我们只需要知道初始概率分布（即时刻 $n = 0$ 的概率分布）表示和随机变量的状态随时间步的变化的概率，也即转移概率即可。
- 收敛性分析：
  - 一个马尔可夫链何时达到某种平稳态？
  - 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？

1、转移理论

马尔可夫链中随机变量的状态随时间步的变化被称为演变（evolution）或转移（transition）。下面介绍描述马尔可夫链结构的两种途径：即转移矩阵和转移图，并定义马尔可夫链在转移过程中表现出的性质。

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，$P(X_{n+1}=j|X_n=i)=P(X_1=j|X_0=i)$。下面只讨论时齐马尔可夫链。

转移概率和转移矩阵

定义 7

称

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

为转移概率，第 (i, j) 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 8

(带吸收壁的随机游动) 令 $\mathcal{X}=\{1,\ldots,N\}$。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

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

例 9

假设状态空间为 $\mathcal{X} = \{\text{晴，多云}\}$，则 $X_1, X_2, \ldots$ 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 8

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 26

(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

$$\boldsymbol{P}_{m+n} = \boldsymbol{P}_m \boldsymbol{P}_n$$

由定义，$\boldsymbol{P}_1 = P$ 由上述定理，$\boldsymbol{P}_2 = \boldsymbol{P}_{1+1} = \boldsymbol{P}_1 \boldsymbol{P}_1 = \boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}^2$ 按该方法继续下去，可以看到

$$\boldsymbol{P}_n = \boldsymbol{P}^n = \boldsymbol{P} \times \boldsymbol{P} \times \cdots \times \boldsymbol{P}$$

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

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

为该链在时刻 n 时处于状态 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 $\boldsymbol{P}$。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 27

边际概率可由下式给出

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

转移图：可达与互通

马尔可夫链的演变可以按图（graph）结构，表示为转移图（transition graph），图中的每条边都被赋予一个转移概率。通过转移图可引入“可达”和“互通”的概念，用来表示马尔可夫链中状态之间的关系。

定义 9

若对于马尔可夫链中的状态 i 和 j, 对于某个 n 有 $p_{ij}(n) > 0$，即采样路径上的所有转移概率不为 0，则称状态 j 是状态 i 的可达状态（或 j 从 i 是可达的），在转移图中表示为有向连接，记作 $i \to j$。若 $i \to j$ 且 $j \to i$，则记作 $i \leftrightarrow j$，并且称 i 和 j 互通。

定理 28

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 $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots$，其中，两个状态之间互通当且仅当它们在同一个类中。

注：按互通关系是等价关系，可以把状态空间 $\mathcal{X}$ 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。

状态分类：闭集和不可约

定义 10

设 C 为状态空间 $\mathcal{X}$ 的一个子集，若对任意的 $i \in C$ 和 $j \notin C$ 有 $p_{ij} = 0$ 则称 C 为闭集。

注：若 C 为闭集，则表示自 C 内任意状态 i 出发，始终不能到达 C 以外的任何状态 j。显然，整个状态空间构成一个闭集。

定义 11

只含有单个状态的闭集称作为吸收态。

注：若状态空间含有吸收状态，那么这个吸收状态构成一个最小的闭集。

定义 12

若除整个状态空间 $\mathcal{X}$ 以外没有其它的闭集，则称此马氏链是不可约的。

如果闭集 C 的状态都是互通的，则称闭集 C 是不可约的。

例 10

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。

状态分类：常返态和瞬时态

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 13

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 29

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

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

状态分类：分解定理

定理 30

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 31

(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 X_i 为一个闭的，不可约的常返态集。

2、马尔可夫链的收敛性：常返性

接下来讨论马尔可夫链的收敛性，首先引入一些定义，然后介绍马尔可夫链的长时间尺度行为，即平稳分布和极限分布，并定义平稳马尔可夫链。

定义 14

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

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

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 32

若一个状态是零的且是常返的，则 $p_{ii}(n) \to 0$

定理 33

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。

例 11

考虑具有三个状态的马尔可夫链，其转移矩阵为

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

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

周期性和遍历性

定义 15

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 34

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

定义 16

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。

平稳分布和极限分布

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 17

若 $\pi = \pi \boldsymbol{P}$，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 $\mu_1 = \mu_0 \boldsymbol{P} = \pi \boldsymbol{P} = \pi$。$X_2$ 的分布为 $\pi \boldsymbol{P}^2 = (\pi \boldsymbol{P}) \boldsymbol{P} = \pi \boldsymbol{P} = \pi$，如此继续下去，会看到 X_n 的分布为 $\pi \boldsymbol{P}^n = \pi$ 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 18

称一个链具体极限分布 π，若 $\boldsymbol{P}^n \to (\pi, \pi, \ldots, \pi)^T$ 对于某个 π，即 $\pi_j = \lim_{n \to \infty} P_{ij}^n$ 存在与 i 是独立的。

收敛性定理

下面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 35

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*

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

细致平稳

定义 19

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 36

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 12

令

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

令 $\pi=1/3,1/3,1/3$ 则 $\pi\boldsymbol{P}=\pi$ 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。

3、马尔可夫链的例子

例 13

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 $3\to 4\to 6$ 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的，总之 *3* 和 *4* 是暂留的，而 *1*，*2*，*5* 和 *6* 是遍历的。

例 14

伯努利过程（*Bernoulli process*） 伯努利过程也被称为二项马尔可夫链（*Binomial Markov chain*），其建立过程如下：给定一系列相互独立的“标识”，每个标识都是二项的，按概率 $r, r-1$ 取正和负。令正随机过程 $\{X_n : n \geq 0\}$ 表示 n 个标识中有 k 个正标识的概率，则其是一个伯努利过程，其中的随机变量服从二项分布（*binomial distribution*）：

$$p(X_n = k) = \binom{n}{k} r^k (1-r)^{n-k}, k \leq n$$

由建立过程可知，增加的新标识中正标识的概率与先前正标识的数量无关，具有马尔可夫性质，因此伯努利过程是一个马尔可夫链。

例 15

随机游走（*random walk*）

定义一系列独立同分布（*independent identically distributed, iid*）的整数随机变量 $\{Y_n\}$，并定义如下随机过程：

$$X_0 = 0, X_n = \sum_{i=1}^{n} Y_i$$

则随机过程 $\{X_n : n \geq 0\}$ 是整数集内的随机游走，而 $\{Y_n\}$ 是步长。由于步长是独立同分布的，因此当前步与先前步相互独立，该随机过程是马尔可夫链。伯努利过程是随机游走的特例。

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）的模拟方法的基本思想。

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。原则上讲可以计算出 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 $X_0, \ldots, X_i$ 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，$X_0, X_1, \ldots$ 是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。

4、马尔可夫链的推广

- 马尔可夫过程：连续时间马尔可夫链，是马尔可夫链或离散时间马尔可夫链的推广，其状态空间是可数集，但一维指数集不再有可数集的限制，可以表示连续时间。
- 马尔可夫模型：马尔可夫链或马尔可夫过程不是唯一以马尔可夫性质为基础建立的随机过程，事实上，隐马尔可夫模型、马尔可夫决策过程、马尔可夫随机场等随机过程/随机模型都具有马尔可夫性质并被统称为马尔可夫模型。
  - 隐马尔可夫模型（Hidden Markov Model, HMM）：HMM 是一个状态空间不完全可见，即包含隐藏状态（hidden status）的马尔可夫链。
  - 马尔可夫决策过程（Markov decision process, MDP）：MDP 是在状态空间的基础上引入了“动作”的马尔可夫链，即马尔可夫链的转移概率不仅与当前状态有关，也与当前动作有关。
  - 马尔可夫随机场（Markov Random Field, MRF）：MRF 是马尔可夫链由一维指数集向高维空间的推广。
- 哈里斯链（Harris chain）：哈里斯链是马尔可夫链从可数状态空间向连续状态空间的推广。

5、马尔可夫链链在 PageRank 算法解释中的应用

PageRank 要解决的问题

我们如何通过使用给定集合的页面之间的现有链接对它们进行排序。

PageRank 算法假设

初始网页是在所有网页中一个随机网页上。然后，开始通过点击当前页面上的一个链接来随机地访问网页。对于给定页面，所有允许的链接都有相同的机会被点击。

PageRank 中的马尔可夫链

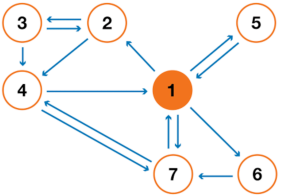
我们来考虑 PageRank 中的马尔可夫链中的设定。每个页面都是一个不同的可能状态，其中的转换概率是由页面到页面的链接定义的。

如果我们假设定义的链是循环正和非周期性的，那么在很长一段时间后，“当前页面”概率分布会收敛到静止分布。因此，无论起始页面如何，经过很长一段时间后，如果我们选择随机时间步长，每个页面都有一个概率作为当前页面。

PageRank 认为静态分布中最可能的页面也必须是最重要的。所以，静态概率分布为每个状态定义了 PageRank 的值。

例 17

给定如下 7 个页面以及它们之间的连接关系。各个节点的概率通过转移矩阵 p 给出。其中 0 值用 . 来替代。



$$p = \begin{pmatrix} . & 0.25 & . & . & 0.25 & 0.25 & 0.25 \\ . & . & 0.5 & 0.5 & . & . & . \\ . & 0.5 & . & 0.5 & . & . & . \\ 0.5 & . & . & . & . & . & 0.5 \\ 1.0 & . & . & . & . & . & . \\ . & . & . & . & . & . & 1.0 \\ 0.5 & . & . & 0.5 & . & . & . \end{pmatrix}$$

图 4: 页面以及页面之间的连接关系以及性质转移矩阵

在进一步计算之前，我们可以注意到这个马尔可夫链是不可约的以及非周期性的，因此，在长期运行之后，系统收敛到静止分布。正如我们已经看到的，我们可以通过求解下面的左特征向量问题来计算这个静态分布

$$\pi = \pi p$$

这样就能得到每页的 PageRank 值（静态分布的值）

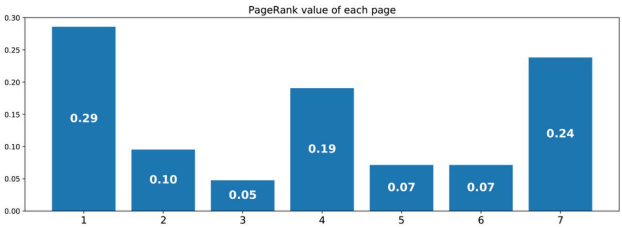


图 5: 页面以及页面之间的连接关系以及性质转移矩阵

这样我们最终得到的 PageRank 排名是 1> 7> 4> 2> 5 = 6> 3。

22.3.3 高斯过程

高斯过程也是一种应用广泛的随机过程模型。假设有一组连续随机变量 $X_0, X_1, \cdots, X_T$，如果由这组随机变量构成的任一有限集合

$$X_{t_1,\ldots,t_N} = [X_{t_1}, \cdots, X_{t_N}]^\top, \quad 1 \leq N \leq T$$

都服从一个多元正态分布，那么这组随机变量为一个随机过程，称为高斯过程。高斯过程也可以定义为：如果 $X_{t_1,\ldots,t_N}$ 的任一线性组合都服从一元正态分布，那么这组随机变量称为一个高斯过程。

- 高斯过程由其数学期望和协方差函数完全决定，并继承了正态分布的诸多性质。
- 高斯过程的例子包括维纳过程、奥恩斯坦 -乌伦贝克过程等。
- 对高斯过程进行建模和预测是机器学习、信号处理等领域的重要内容，其中常见的模型包括高斯过程回归（Gaussian Process Regression, GPR）和高斯过程分类（Gaussian Process Classification, GPC）。

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！