由此可得

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}(X_n-\mu)\geqslant\epsilon\right]\leqslant=\exp\left(-nt\epsilon+\frac{t^2n\sigma^2}{2(1-bt)}\right)$$

取 $t=\epsilon/(\sigma^2+b\epsilon)$ 完成证明.

例 6.7 给出 Bernstein 不等式的 $1-\delta$ 表述.

6.4 应用: 随机投影 (Random Projection)

设高维空间 $\mathbb{R}^d$ 有 n 个点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ (d 非常大, 如 100 万或 1 亿). 处理这样一个高维的问题很难, 实际中的一种解决方案是能否找到一个保距变换: $f:\mathbb{R}^d\to\mathbb{R}^k$ ($k\ll d$), 使得以较大概率有

$$(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\leqslant\|f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\|_2^2\leqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2.$$

随机投影广泛应用于高维的机器学习问题, 例如最近邻、k-近邻、降维、聚类等问题.

随机投影可以简单的表示为

$$f(\boldsymbol{x})=\boldsymbol{x}P/c,$$

其中 P 是一个 $d\times k$ 的随机矩阵, 其每个元素之间相互独立, c 为一常数 (根据随机矩阵 P 确定). 下面介绍三种常见的随机矩阵:

- $P=(p_{ij})_{d\times k}\in\mathbb{R}^{d\times k}$, $p_{ij}\sim\mathcal{N}(0,1)$, 此时 $c=\sqrt{k}$;
- $P=(p_{ij})_{d\times k}\in\{-1,1\}^{d\times k}$, p_{ij} 为 Rademacher 随机变量, 即 $\Pr(p_{ij}=1)=\Pr(p_{ij}=-1)=1/2$, 此时 $c=\sqrt{k}$;
- $P=(p_{ij})_{d\times k}\in\{-1,0,1\}^{d\times k}$, 满足 $\Pr(p_{ij}=1)=\Pr(p_{ij}=-1)=1/6$ 和 $\Pr(p_{ij}=0)=2/3$, 此时 $c=\sqrt{k/3}$. 【主要用于sparse 投影, 减少计算量】

下面我们重点理论分析 Gaussian 随机变量, 其它随机变量可参考相关资料, 对 Gaussian 随机变量, 这里介绍著名的 Johnson – Lindenstrauss 引理, 简称 JL 引理.

引理 6.5 设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 为 $\mathbb{R}^d$ 空间的 n 个点, 随机矩阵 $P=(p_{ij})_{d\times k}\in\mathbb{R}^{d\times k}$, $p_{ij}\sim\mathcal{N}(0,1)$ 且每个元素相互独立, 令

$$\boldsymbol{y}_i=f(\boldsymbol{x}_i)=\boldsymbol{x}_iP/\sqrt{k},\qquad i\in[n]$$

将 d 维空间中 n 个点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 通过随机矩阵 P 投影到 k 维空间. 对任意 $\epsilon\in(0,1/2)$, 当 $k\geqslant 8\log 2n/(\epsilon^2-\epsilon^3)$ 时至少以 $1/2$ 的概率有

$$(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\leqslant\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\qquad(i,j\in[n]).$$

证明 下面分三步证明 J-L 引理.

第一步: 对任意非零 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)\in\mathbb{R}^d$, 首先证明

$$E\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\right]=\|\boldsymbol{x}\|_2^2,$$

即在期望的情况下, 随机投影变换前后的点到原点的距离相同. 根据 $P=(p_{ij})_{d\times k}$ ($p_{ij}\sim\mathcal{N}(0,1)$) 有

$$\begin{aligned}E\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\right] &= E\left[\sum_{j=1}^{k}\left(\sum_{i=1}^{d}\frac{x_ip_{ij}}{\sqrt{k}}\right)^2\right]=\sum_{j=1}^{k}\frac{1}{k}E\left[\left(\sum_{i=1}^{d}x_ip_{ij}\right)^2\right]\\ &= \sum_{j=1}^{k}\frac{1}{k}\sum_{i=1}^{d}x_i^2=\frac{1}{k}\sum_{j=1}^{k}\|\boldsymbol{x}\|_2^2=\|\boldsymbol{x}\|_2^2.\end{aligned}$$

第二步: 对任意非零 $\boldsymbol{x}=(x_1,x_2,\cdots,x_d)\in\mathbb{R}^d$, 证明

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\exp(-(\epsilon^2-\epsilon^3)k/4).$$

将矩阵 P 表示为 $P=(P_1,P_2,\ldots,P_k)$, 其中 P_i ($i\in[d]$) 是一个 $d\times1$ 的列向量, 令 $v_j=\boldsymbol{x}P_j/\|\boldsymbol{x}\|_2$, 即

$$(v_1,v_2,\ldots,v_k)=\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_1,\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_2,\ldots,\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}P_k\right).$$

根据 Gaussian 分布的性质有 $v_j\sim\mathcal{N}(0,1)$, 且 $v_1,v_2,\ldots,v_k$ 是 k 个独立的随机变量. 对任意 $t\in(0,1/2)$, 根据 Chernoff 方法有

$$\begin{aligned}&\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]=\Pr\left[\left\|\frac{\boldsymbol{x}P}{\|\boldsymbol{x}\|_2}\right\|_2^2\geqslant(1+\epsilon)k\right]\\ =\ &\Pr\left[\sum_{j=1}^{k}v_j^2\geqslant(1+\epsilon)k\right]\leqslant e^{-(1+\epsilon)kt}\left(E[e^{t\sum_{j=1}^{k}v_j^2}]\right)^k=e^{-(1+\epsilon)kt}\left(E[e^{tv_1^2}]\right)^k.\end{aligned}$$

对标准 Gaussian 分布有

$$E[e^{tv_1^2}] = \int_{-\infty}^{+\infty}\frac{e^{tu^2}}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du=\int_{-\infty}^{+\infty}\frac{e^{-\frac{u^2}{2}(1-2t)}}{\sqrt{2\pi}}du=\frac{1}{\sqrt{1-2t}},$$

代入可得

$$\Pr\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\left(\frac{e^{-2(1+\epsilon)t}}{1-2t}\right)^{k/2}.$$

上式右边对 t 求最小解得 $t_{\min}=\frac{\epsilon}{2(1+\epsilon)}$, 代入可得

$$\Pr\left[\left\|\boldsymbol{x}P/\sqrt{k}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant\left((1+\epsilon)e^{-\epsilon}\right)^{k/2}.$$

设 $f(\epsilon)=\ln(1+\epsilon)$, 根据 $\epsilon\in(0,1/2)$ 有

$$f^{'}(\epsilon)=\frac{1}{1+\epsilon},f^{''}(\epsilon)=-\frac{1}{(1+\epsilon)^2},f^{'''}(\epsilon)=\frac{2}{(1+\epsilon)^3}\leqslant 2.$$

根据泰勒中值定理有

$$f(\epsilon)=f(0)+f^{'}(0)\epsilon+\frac{f^{''}(0)\epsilon^2}{2!}+\frac{f^{'''}(\xi)\epsilon^3}{3!}\leqslant\epsilon-\frac{\epsilon^2}{2}+\frac{\epsilon^2}{3}\leqslant\epsilon-\frac{\epsilon^2-\epsilon^3}{2}.$$

于是得到

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

同理可证

$$\Pr\left[\left\|\frac{\boldsymbol{x}P}{\sqrt{k}}\right\|_2^2\geqslant(1-\epsilon)\|\boldsymbol{x}\|_2^2\right]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

第三步: 对任意给定 $i\neq j$, 根据第二步的结论可知

$$\Pr[\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4},$$
$$\Pr[\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2]\leqslant e^{-k(\epsilon^2-\epsilon^3)/4}.$$

由于 $i,j\in[n]$, 因此共有 $n(n-1)$ 对 (i,j), 根据 Union 不等式有

$$\Pr\left[\exists i\neq j\colon\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\geqslant(1+\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\quad\text{或}\quad\|\boldsymbol{y}_i-\boldsymbol{y}_j\|_2^2\leqslant(1-\epsilon)\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\right]\leqslant 2n^2e^{-k(\epsilon^2-\epsilon^3)/4},$$

设 $2n^2e^{-k(\epsilon^2-\epsilon^3)/4}\leqslant 1/2$, 求解 $k\geqslant 8\log 2n/(\epsilon^2-\epsilon^3)$. 引理得证.

第 7 章　大数定律及中心极限定理

7.1 大数定律

给定随机变量 $X_1, X_2, \cdots, X_n$, 这些随机变量的均值 (算术平均值) 为

$$\frac{1}{n}\sum_{i=1}^{n} X_i.$$

当 n 非常大时, 大数定律考虑随机变量的均值是否具有稳定性.

定义 7.1 (依概率收敛) 设 $X_1, X_2, \cdots, X_n, \cdots$ 是一随机变量序列, a 是一常数, 如果对任意 $\epsilon > 0$ 有

$$\lim_{n\to\infty} \Pr\{|X_n - a| < \epsilon\} = 1 \quad 或 \quad \lim_{n\to\infty} \Pr\{|X_n - a| > \epsilon\} = 0,$$

则称随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 依概率收敛于 a, 记 $X_n \xrightarrow{P} a$.

问题: 与数列极限的区别? 下面我们给出依概率的性质:

1) 若 $X_n \xrightarrow{P} a$ 且函数 $g\colon \mathbb{R} \to \mathbb{R}$ 在 $X = a$ 点连续, 则$g(X_n) \xrightarrow{P} g(a)$.

2) 若 $X_n \xrightarrow{P} a, Y_n \xrightarrow{P} b$, 函数 $g\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ 在点 $(X, Y) = (a, b)$ 处连续, 则 $g(X_n, Y_n) \xrightarrow{P} g(a, b)$.

例如: 如果 $X_n \xrightarrow{P} a$ 和 $Y_n \xrightarrow{P} b$, 那么 $X_n + Y_n \xrightarrow{P} a + b$ 和 $X_nY_n \xrightarrow{P} ab$.

定理 7.1 (大数定律) 若随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n} E[X_i],$$

则称 $\{X_n\}$ 服从大数定律.

大数定理刻画了随机变量的均值 (算术平均值) 依概率收敛于期望的均值 (算术平均值). 下面介绍几种大数定律:

定理 7.2 (马尔可夫 Markov 大数定律) 如果随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足

$$\frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0 \qquad n \to \infty,$$

则 $\{X_n\}$ 服从大数定理.

马尔可夫大数定律不要求随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立或同分布, 其证明直接通过 Chebyshev 不等式有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right] \leqslant \frac{1}{n^2\epsilon^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \to 0 \qquad n \to \infty.$$

定理 7.3 (切比雪夫 Chebyshev 大数定律) 设随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 相互独立, 且存在常数 $c > 0$ 使得 $\mathrm{Var}(X_n) \leqslant c$, 则 $\{X_n\}$ 服从大数定律.

此处独立的随机变量可以修改为'不相关随机变量'. 证明直接通过切比雪夫不等式

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right| \geqslant \epsilon\right] \leqslant \frac{1}{\epsilon^2 n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \leqslant \frac{c}{n\epsilon^2} \to 0 \qquad n \to \infty.$$

定理 7.4 (辛钦 Khintchine 大数定律) 设 $X_1, X_2, \cdots, X_n, \cdots$ 为独立同分布随机变量序列, 且每个随机变量的期望 $E[X_i] = \mu$ 存在, 则 $\{X_n\}$ 服从大数定律.

辛钦大数定律不要求方差一定存在, 其证明超出了本书范围.

定理 7.5 (Bernoulli大数定律) 设随机变量序列 $X_n \sim B(n, p)$ $(p > 0)$, 对任意 $\epsilon > 0$ 有

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right] = 0,$$

即 $X_n/n \xrightarrow{P} p$.

定理的证明依据二项分布的性质: 独立同分布随机变量 $Y_1, Y_2, \ldots, Y_n$ 满足 $Y_i \sim \mathrm{Ber}(p)$, 则

$$X_n = \sum_{i=1}^{n} Y_i \sim \mathrm{B}(n, p).$$

于是得到

$$\lim_{n\to\infty} \Pr\left[\left|\frac{X_n}{n} - p\right| \geqslant \epsilon\right] = \lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} Y_i - E[Y_i]\right| \geqslant \epsilon\right] \leqslant \frac{1}{\epsilon^2 n^2}\mathrm{Var}(\sum_{i=1}^{n} Y_i) = \frac{p(1-p)}{\epsilon^2 n} \to 0.$$

如何判断随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足大数定律:

- 若随机变量独立同分布, 则利用辛钦大数定律查看期望是否存在;
- 对非独立同分布随机变量, 则利用 Markov 大数定律判断方差是否趋于零.

例 7.1 独立的随机变量序列 $X_1, X_2, \cdots, X_n, \cdots$ 满足 $\Pr\{X_n = n^{1/4}\} = \Pr\{X_n = -n^{1/4}\} = 1/2$. 证明 $\{X_n\}$ 服从大数定律.

证明 根据题意可得 $E[X_i]=0$, 以及 $\mathrm{Var}(X_i)=E[X_i^2]=i^{1/2}$, 根据 Chebysheve 不等式和独立性有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right|\geqslant\epsilon\right]\leqslant\frac{1}{n^2\epsilon^2}\mathrm{Var}(\sum_{i=1}^{n}X_i)=\frac{1}{n^2\epsilon^2}\sum_{i=1}^{n}Var(X_i)=\frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n}i^{1/2}\leqslant\frac{1}{\epsilon^2\sqrt{n}}$$

再根据

$$\sum_{i=1}^{n}i^{1/2}\leqslant\sum_{i=1}^{n}\int_i^{i+1}i^{1/2}dx\leqslant\sum_{i=1}^{n}\int_i^{i+1}x^{1/2}dx=\int_1^{n+1}x^{1/2}dx=2((n+1)^{3/2}-1)/3$$

由此可得当 $n\to+\infty$ 时有

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right|\geqslant\epsilon\right]\leqslant\frac{2((n+1)^{3/2}-1)/3}{\epsilon^2n^2}\to0$$

大数定律小结:

- Markov 大数定律: 若随机变量序列 $\{X_i\}$ 满足 $\mathrm{Var}(\sum_{i=1}^{n}X_n)/n^2\to0$, 则满足大数定律;
- Chebyshev 大数定律: 若独立随机变量序列 $\{X_i\}$ 满足 $\mathrm{Var}(X_i)\leqslant c$, 则满足大数定律;
- Khintchine 大数定律: 若独立同分布随机变量序列 $\{X_i\}$ 期望存在, 则满足大数定律;
- Bernoulli 大数定律: 对二项分布 $X_n\sim\mathrm{B}(n,p)$, 有 $X_n/n\xrightarrow{P}p$.

7.1.1 中心极限定理

对独立的随机变量序列 $X_1,X_2,\cdots,X_n,\cdots$, 我们考虑标准化后随机变量

$$Y_n=\frac{\sum_{i=1}^{n}X_i-\sum_{i=1}^{n}E(X_i)}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n}X_i)}}$$

的极限分布是否为服从正态分布. 首先介绍依分布收敛.

定义 7.2 设随机变量 Y 的分布函数为 $F_Y(y)=\Pr(Y\leqslant y)$, 以及随机变量序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 的分布函数分别为 $F_{Y_n}(y)=\Pr(Y_n\leqslant y)$, 如果

$$\lim_{n\to\infty}\Pr[Y_n\leqslant y]=\Pr[Y\leqslant y],\quad\text{即}\quad\lim_{n\to\infty}F_{Y_n}(y)=F_Y(y),$$

则称随机变量序列 $Y_1,Y_2,\cdots,Y_n,\cdots$ 依分布收敛于 Y, 记 $Y_n\xrightarrow{d}Y$.

下面介绍独立同分布中心极限定理, 又被称为林德贝格-勒维 (Lindeberg-Lévy) 中心极限定理":