Ch 6.4 随机投影 (random projection)

回顾前一次课

Sub-Gaussian随机变量: $E\left[e^{(X-E[X])t}\right] \leq e^{\frac{bt^2}{2}}$

有界随机变量、Gaussian随机变量

Bennet不等式

$$P\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2+2\epsilon/3}\right)$$

Bernstein不等式

$$P\left[\frac{1}{n}\sum_{i=1}^{n}X_i-\mu \geq \epsilon\right] \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2+2b\epsilon}\right)$$

Bernstein不等式

$$P\left[\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\geq\epsilon\right]\leq\exp\left(-\frac{n\epsilon^2}{2\sigma^2+2b\epsilon}\right)$$

令$\delta=\exp\left(-\frac{n\epsilon^2}{2\sigma^2+2b\epsilon}\right)$，至少以$1-\delta$的概率有

$$\frac{1}{n}\sum_{i=1}^{n}X_i\leq\mu+\frac{2\mathrm{b}}{n}\ln\frac{1}{\delta}+\sqrt{\frac{2\sigma^2}{n}\ln\frac{1}{\delta}}$$

问题

问题： 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。

$$\boldsymbol{x}_1 = (x_{11}, x_{12}, \cdots, x_{1d})$$

$$\boldsymbol{x}_2 = (x_{21}, x_{22}, \cdots, x_{2d})$$

$$\vdots$$

$$\boldsymbol{x}_n = (x_{n1}, x_{n2}, \cdots, x_{nd})$$

保距变换

保距变换: $f: \mathbb{R}^d \to \mathbb{R}^k \quad (k \ll d)$ 使得以较大概率有

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

随机投影广泛应用于高维的机器学习，例如

- 最近邻
- k-近邻
- 降维
- 聚类

随机投影 (Random projection)

随机投影：

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

其中P是$d \times k$的随机矩阵，其每个元素之间相互独立，c为常数 (根据随机矩阵P确定)

- $P = \left(p_{ij}\right)_{d \times k}$，$p_{ij} \sim N(0,1)$，此时$c = \sqrt{k}$;
- $P = \left(p_{ij}\right)_{d \times k}$，$p_{ij}$为Rademacher随机变量，此时$c = \sqrt{k}$;
- $P = \left(p_{ij}\right)_{d \times k}$， $P(p_{ij} = 1) = P(p_{ij} = -1) = 1/6$ 和 $P\big(p_{ij} =$

Johnson–Lindenstrauss引理

JL-引理： 设$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{\boldsymbol{n}}$为$d$维空间的$n$个点, 随机矩阵$P = \left(p_{ij}\right)_{d\times k}$, 每个元素相互独立且$p_{ij} \sim N(0,1)$, 令

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

将d维空间中n个点通过随机矩阵P投影到k维空间.

对任意$\epsilon \in (0, 1/2)$, 当$k \geq 8 \ln 2n\,/(\epsilon\hat{}2 - \epsilon\hat{}3)$时, 对任意$i \neq j$, 以至少1/2的概率有

$$(1-\epsilon)|\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{j}}|_2^2 \leq\ |\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}|_2^2 \leq\ (1+\epsilon)|\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{j}}|_2^2$$

证明

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

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

期望的情况下, 随机投影前、后到原点的距离相同.

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

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

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

证明

第三步：对任意$i \neq j$, 根据第二步的结论可知

$$P\left[(1-\epsilon)|\boldsymbol{x_i}-\boldsymbol{x_j}|_2^2 \leq \ |\boldsymbol{y_i}-\boldsymbol{y_i}|_2^2 \leq \ (1+\epsilon)|\boldsymbol{x_i}-\boldsymbol{x_j}|_2^2\right]$$
$$\geq 1-2n^2\exp(-(\epsilon^2 \ -\epsilon^3)k/4)$$

Ch 7大数定律及中心极限定理

大数定律的问题

问题： 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

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

$$\lim_{n\to\infty} P[|X_n - a| < \epsilon] = 1$$

$$\lim_{n\to\infty} P[|X_n - a| > \epsilon] = 0$$

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

$$X_n \xrightarrow{P} a$$

依概率收敛性质

若$X_n \xrightarrow{P} a$,函数$g: R \to R$在$X = a$点连续, 则

$$g(X_n) \xrightarrow{P} g(a)$$

若$X_n \xrightarrow{P} a,\, Y_n \xrightarrow{P} b$函数$g: R \times R \to R$在$(a, b)$点连续, 则

$$g(X_n, Y_n) \xrightarrow{P} g(a, b)$$

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)

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

不要求随机变量$X_1, X_2, \cdots, X_n, \cdots$相互独立或同分布

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$\mathrm{Var}(X_n) \le c$, 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**

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.

不要求方差一定存在, 其证明超出了本书范围

Bernoulli大数定律

设随机变量序列$X_n \sim B(n,p)$, 对任意$\epsilon > 0$有

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

即$X_n/n \xrightarrow{P} p$.

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.

习题

独立的随机变量序列$X_1, X_2, \cdots, X_n, \cdots$满足

$$P\left[X_n = n^{1/4}\right] = P[X_n = -n^{1/4}] = 1/2$$

证明$\{X_n\}$服从大数定律

大数定律总结

Markov大数定律：若随机变量序列$\{X_i\}$满足$\mathrm{Var}(\sum_{i=1}^{n} X_n)/n^2 \to 0$, 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$\mathrm{Var}(X_i) \le c$, 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$X_n/n \xrightarrow{P} p$