Ch 7 集中不等式 (Concentration)

回顾前一次课

在$Y=y$条件下X的条件期望 $E(X|y)=\int_{-\infty}^{+\infty}xf_{X|Y}(x|y)dx$

在$Y=y$条件下X的条件期望 $E(X|y)=\sum_i x_iP(X=x_i|Y=j)$

条件期望的性质、$E(g(X)|Y)$、$E(X)=E_Y\big(E(X|Y)\big)$

全期望公式：$E(X)=E(X|A)P(A)+E\big(X\big|\bar{A}\big)P\big(\bar{A}\big)$

随机向量期望$E(X)=\big(E(X_1),E(X_2),\cdots,E(X_n)\big)^\top$和协方差矩阵

$$\mathrm{Cov}(X)=\Sigma=\begin{pmatrix}\mathrm{Cov}(X_1,X_1) & \cdots & \mathrm{Cov}(X_1,X_n)\\ \mathrm{Cov}(X_2,X_1) & \cdots & \mathrm{Cov}(X_2,X_n)\\ \vdots & & \vdots \\ \mathrm{Cov}(X_n,X_1) & \cdots & \mathrm{Cov}(X_n,X_n)\end{pmatrix}$$

协方差矩阵的半正定性，多维正态分布$(X_1,X_2,\ldots,X_n)^\top\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$

回顾前一次课

Markov不等式：对随机变量$X \geq 0$和$\epsilon > 0$, 有$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$

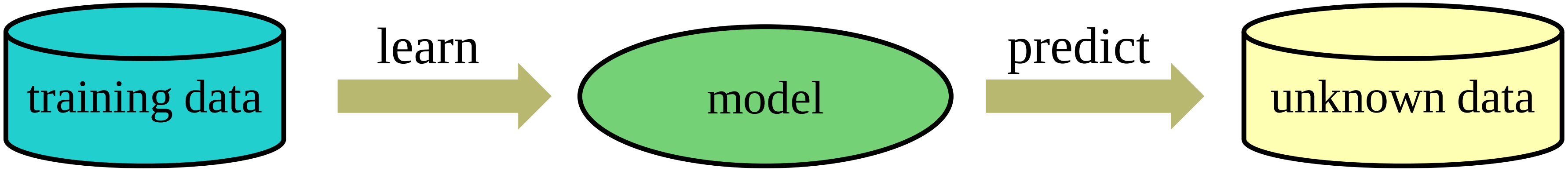
Chebyshev不等式：$P(|X - \mu| > \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$

Hölder不等式：$E(|XY|) \leq \left(E(|X|^p)\right)^{\frac{1}{p}}\left(E(|Y|^q)\right)^{\frac{1}{q}}$

单边Chebyshev不等式[Cantelli不等式]：随机变量X的均值$\mu > 0$, 方差σ^2, 则对任意$\epsilon > 0$有

$$P(X - \mu \geq \epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2} \qquad P(X - \mu \leq -\epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}$$

机器学习



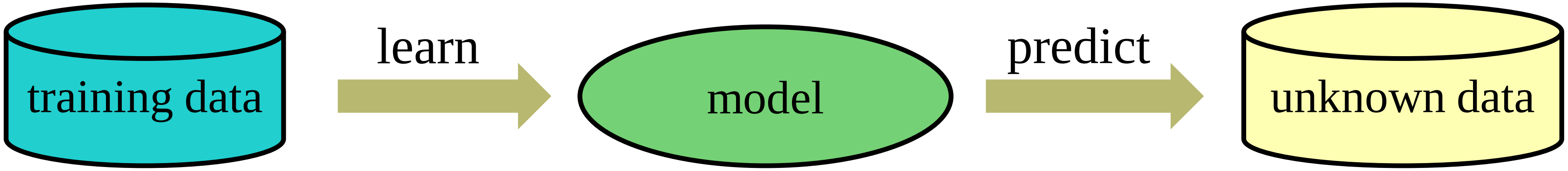
未见数据: 在空间$\mathcal{X} \times \mathcal{Y}$的未知分布$\mathcal{D}$

$\mathcal{X}$: 特征空间　　$\mathcal{Y}$: 标记空间

训练数据: $S_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

经典假设: 数据集S_n中数据(x_i, y_i)是根据分布$\mathcal{D}$独立同分布采样

机器学习



机器学习: 训练数据S_n学习函数$f: \mathcal{X} \to \mathcal{Y}$, 在分布$\mathcal{D}$分类效果好

训练错误率: 函数f在训练数据S_n的分类错误率

$$\hat{R}(f, S_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[f(x_i) \neq y_i] = \frac{1}{n}\sum_{i=1}^{n} X_i$$

这里$\mathbb{I}[\cdot]$为指示函数, 论断为真返回值为1, 否则为0

泛化错误率: 函数f在未见数据分布$\mathcal{D}$的分类错误率

$$R(f) = E_{(x,y)\sim\mathcal{D}}\big[\mathbb{I}[f(x) \neq y]\big] = E[X]$$

机器学习的根本问题

由于分布$\mathcal{D}$不可知, 不能直接计算$R(f)$

已知训练数据集S_n和训练错误率$\hat{R}(f,S_n)$

如何基于训练错误率$\hat{R}(f,S_n)$来有效估计$R(f)$?

根本问题可归纳为

$$\mathrm{P}_{S_n}\left[\left|\hat{R}(f,S_n)-R(f)\right|\geq\ t\right]\text{是否足够小?}$$

即能否以很大的概率保证

$$\left|\hat{R}(f,S_n)-R(f)\right|<\ t$$

从理论上保证$\hat{R}(f,S_n)$是$R(f)$的一个有效估计

例子

假设训练数据集$S_n = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$根据分布$\mathcal{D}$独立采样所得，分类器$f$在训练集$S_n$的错误率为零(全部预测正确)，求分类器$f$在分布$\mathcal{D}$上的错误率介于0和$\epsilon$之间的概率($\epsilon > 0$)

问题归纳

设随机变量

$$X_i = \mathbb{I}[f(x_i) \neq y_i]$$

问题归纳： 已知n个独立同分布随机变量$X_1, X_2, \ldots, X_n$，如何以很大概率获得期望$E[X]$的一个估计, 即

$$\mathrm{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E(X_i)\right| > \epsilon\right] < \text{非常小?}$$

Chernoff方法

给定任意随机变量X和任意$t > 0$和$\epsilon > 0$, 利用Markov不等式有

$$P[X \geq \epsilon] = P[e^{tX} \geq e^{t\epsilon}] \leq e^{-t\epsilon}E[e^{tX}]$$

特别地, 有

$$P[X \geq \epsilon] \leq \min_{t>0}\{e^{-t\epsilon}E[e^{tX}]\}$$

Chernoff方法

对任意$\epsilon > 0$和$t < 0$有

$$P[X \le -\epsilon] = P[tX \ge -t\epsilon] \le e^{t\epsilon} E[e^{tX}]$$

同理有

$$P[X \le -\epsilon] \le \min_{t<0}\{e^{t\epsilon} E[e^{tX}]\}.$$

上述方法称为Chernoff方法, 是证明集中不等式最重要的方法之一

二值Chernoff界

设随机变量$X_1, X_2, \cdots, X_n$相互独立、并且满足$X_i \sim \mathrm{Ber}(p_i)$，令$\mu = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p_i$. 对任意$\epsilon > 0$有

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

对任意$0 < \epsilon < 1$有

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

定理

设随机变量$X_1, X_2, \cdots, X_n$相互独立、而且满足$X_i \sim \mathrm{Ber}(p_i)$，令$\mu = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p_i$. 对任意$0 < \epsilon < 1$有

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

Rademacher随机变量

若随机变量$X \in \{+1, -1\}$满足

$$P(X = +1) = P(X = -1) = 1/2$$

则称X为Rademacher随机变量

定理： 对n个独立的Rademacher随机变量$X_1, X_2, \cdots, X_n$, 有

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geq \epsilon\right] \leq e^{-n\epsilon^2/2} \quad P\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leq -\epsilon\right] \leq e^{-n\epsilon^2/2}$$

推论

独立同分布随机变量$X_1, \cdots, X_n$满足$P(X_1 = 0) = P(X_1 = 1) = 1/2$,则有

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2} \geq \epsilon\right] \leq e^{-2n\epsilon^2}$$

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2} \leq -\epsilon\right] \leq e^{-2n\epsilon^2}$$

有界的Chernoff不等式

研究有界的随机变量$X_i \in [a,b]$的Chernoff不等式

Chernoff引理: 设随机变量$X \in [0,1]$期望$\mu = E[X]$. 对$\forall t > 0$有

$$E[e^{tX}] \leq \mathrm{e}^{t\mu + \frac{t^2}{8}}$$

推论: 随机变量$X \in [a,b]$的期望$\mu = E[X]$, 对任$\forall\ t > 0$有

$$E[e^{tX}] \leq \mathrm{e}^{t\mu + \frac{t^2(b-a)^2}{8}}$$

Chernoff不等式

设$X_1, \ldots, X_n$是n独立的随机变量且$X_i \in [a, b]$. 对任意$\epsilon > 0$有

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \geq \epsilon\right] \leq e^{-2n\epsilon^2/(b-a)^2}$$

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i] \leq -\epsilon\right] \leq e^{-2n\epsilon^2/(b-a)^2}$$