Ch 6 集中不等式 (Concentration)

回顾前一次课

条件概率密度: $f_{X|Y}(x|y) = f(x,y)/f_Y(y)$

条件分布函数: $F_{X|Y}(x|y) = \int_{-\infty}^{x} f_{X|Y}(v|y)dv$

乘法公式：$f(x,y) = f_X(x)f_{Y|X}(y|x), \quad (f_X(x) > 0)$

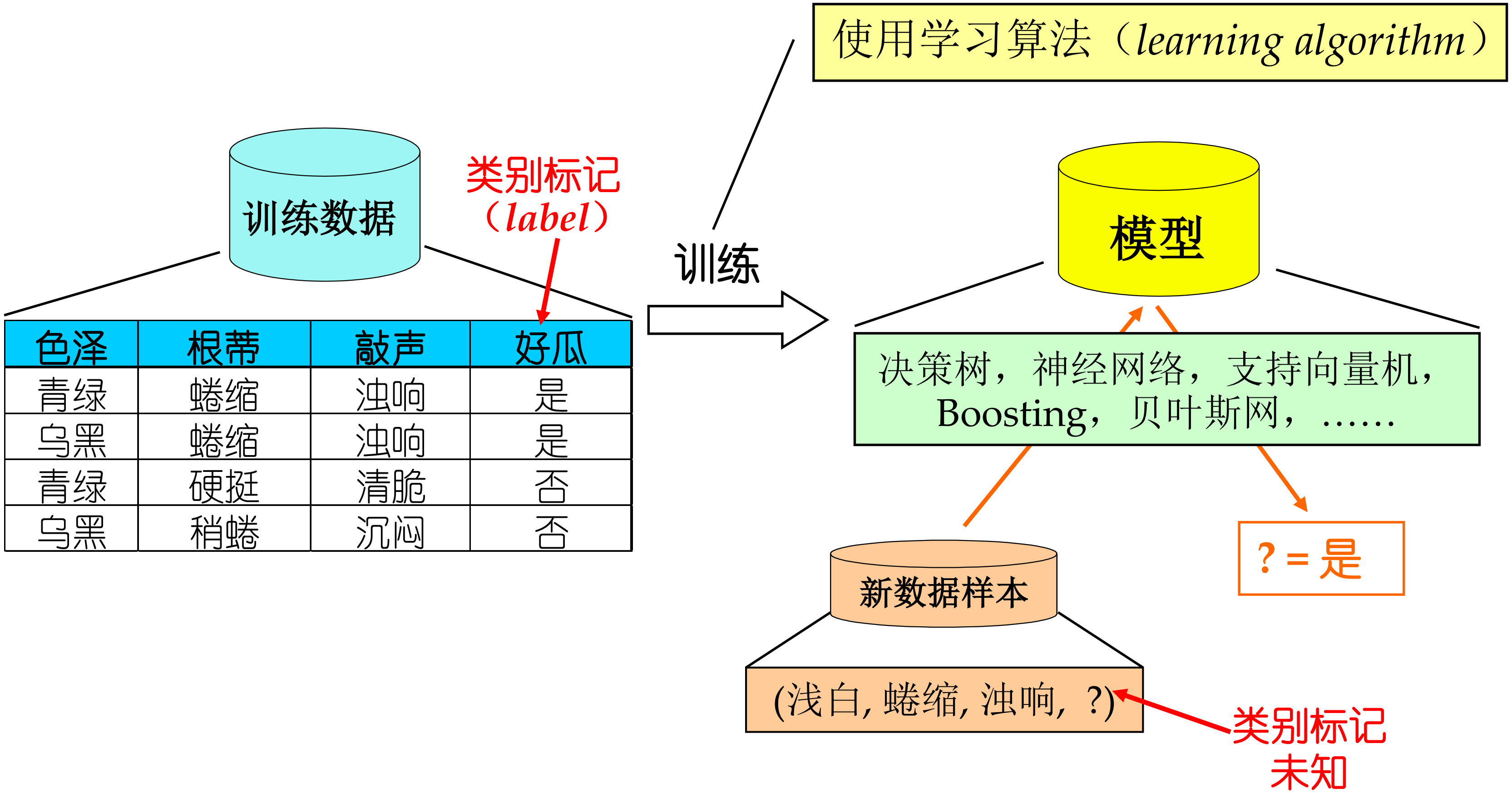
独立性：$f_{Y|X}(y|x) = f_Y(y)$

多维正太分布的条件分布是正太分布

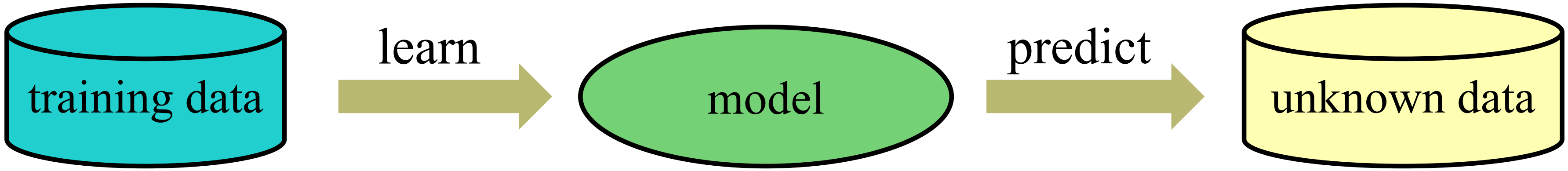
随机变量X的期望为$E[X|Y=y] = \int_{-\infty}^{+\infty} x\, f(x|y)dx$

全期望公式：$E[X] = E[X|A]P(A) + E[X|\bar{A}]\big(1 - P(A)\big)$

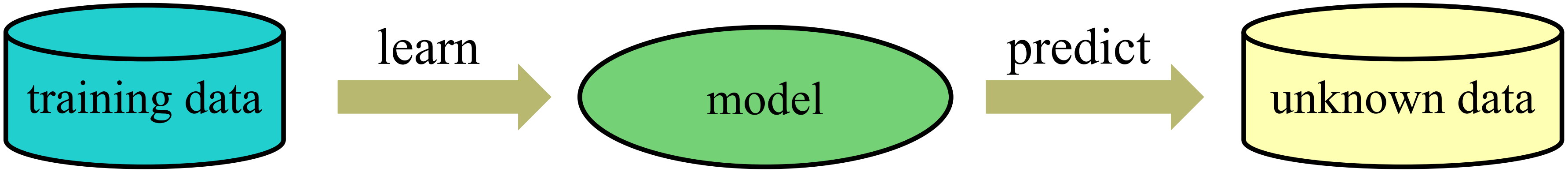
典型的机器学习过程



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否



机器学习形式化



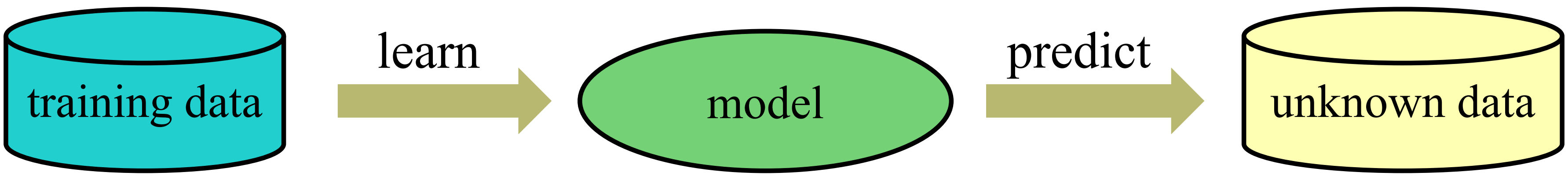
未见数据: 在空间$\mathcal{X} \times \mathcal{Y}$的未知分布$\mathcal{D}$

$\mathcal{X}$: 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: 训练数据集S_n中每个数据(x_i, y_i)是根据分布$\mathcal{D}$独立同分布采样所得

机器学习形式化



机器学习: 通过对训练数据S_n学习分类器$f: \mathcal{X} \to \mathcal{Y}$

分类器f在未见数据分布$\mathcal{D}$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

泛化错误率: 函数f在未见数据分布$\mathcal{D}$的分类错误率

机器学习形式化

训练错误率：函数或分类器$f: \mathcal{X} \to \mathcal{Y}$在训练数据$S_n$上的分类错误率为

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

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

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

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$S_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$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：假设n个独立同分布随机变量$X_1, X_2, \ldots, X_n$，如何从n个独立同分布的随机变量中以很大概率获得期望$E[X]$的一个估计, 即

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

Markov不等式

Markov不等式： 对任意随机变量$X \geq 0$和$\epsilon > 0$, 有

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

推论： 对任意随机变量X和$\epsilon \geq 0$，及单调递增的非负函数$g(x)$, 有

$$P(X \geq \epsilon) \leq \frac{E(g(X))}{g(\epsilon)}$$

Chebyshev不等式

Chebyshev不等式： 设随机变量X的均值为μ, 则有

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

例题

设随机变量$X \sim N(-1,2)$和$Y \sim N(1,8)$，且X和Y的相关系数为-1, 利用Chebyshev不等式估计

$$P(|X+Y| \geq 6) \leq ????$$

课堂练习： 随机变量X和Y满足$E(X)=2$，$E(Y)=2$, $\mathrm{Var}(X)=1, \mathrm{Var}(Y)=4, \rho_{XY}=-1/2$. 利用Chebyshev不等式估计P$(|X-Y| \geq 6)$的上界.

单边Chebyshev不等式

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

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

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

Chebyshev不等式推论

推论： 设独立同分布的随机变量$X_1, X_2, \ldots, X_n$满足$E(X_i) = \mu$和$\mathrm{Var}(X_i) \le \sigma^2$, 对任意$\epsilon > 0$有

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \ge \epsilon\right) \le \frac{\sigma^2}{n\epsilon^2}$$

例题

分类器f在训练集$S_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$的错误率为$\hat{p} > 0$，求分类器$f$在分布$\mathcal{D}$上的错误率在$(9\hat{p}/10, 11\hat{p}/10)$ 之间的概率.

Young不等式

Young不等式： 给定常数$a > 0, b > 0$，对满足$1/p + 1/q = 1$的实数$p > 0, q > 0$有

$$ab \leq \frac{1}{p} a^p + \frac{1}{q} b^q.$$

Hölder不等式

Hölder不等式： 对任意随机变量X和Y以及实数$p > 0$和$q > 0$满足$1/p + 1/q = 1$, 有

$$E(|XY|) \le \left(E(|X|^p)\right)^{\frac{1}{p}}\left(E(|Y|^q)\right)^{\frac{1}{q}}.$$

特别地，当$p = q = 2$时Hölder不等式成为Cauchy-Schwartz不等式

随机变量的矩生成函数(Moment Generating Function)

定义： 定义随机变量X的矩生成函数为

$$M_X(t) = E[e^{tX}].$$

定理： 随机变量X的矩生成函数为$M_X(t)$，对任意$n \geq 1$有

$$E[X^n] = M_X^{(n)}(0)$$

这里$M_X^{(n)}(0)$表示矩生成函数在$t = 0$的n阶导数，而$E[X^n]$被称为随机变量X的n阶矩 (moment).