机器学习

李成龙

安徽大学人工智能学院
“多模态认知计算”安徽省重点实验室
合肥综合性国家科学中心人工智能研究院

内容安排

- 什么是机器学习
- 机器如何学习
- 如何让机器学习的更好
- 为什么机器能学习

内容安排

- 机器如何学习
  - 有监督学习
    - 感知机
    - 支持向量机
    - 朴素贝叶斯分类
    - 决策树
    - 集成学习（Bagging算法与随机森林、Boosting算法）
    - 线性回归
    - 逻辑回归
    - Softmax回归
    - 神经网络与深度学习
  - 无监督学习
    - 聚类
    - 主成分分析

本节目录

- 背景知识
- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类

本节目录

背景知识

- 贝叶斯
  - 贝叶斯(约1701-1761) Thomas Bayes，英国数学家。约1701年出生于伦敦，做过神甫。1742年成为英国皇家学会会员。1761年4月7日逝世。贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》(An essay towards solving a problem in the doctrine of chances)寄给了英国皇家学会，对于现代概率论和数理统计产生了重要的影响

背景知识

- 贝叶斯
  - 贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策
  - 贝叶斯决策理论方法是统计模型决策中的一个基本方法，其基本思想是：
    - 已知类条件概率密度参数表达式和先验概率
    - 利用贝叶斯公式转换成后验概率
    - 根据后验概率大小进行决策分类

背景知识

- 贝叶斯
  - 最早的PathFinder系统，该系统是淋巴疾病诊断的医学系统，它可以诊断60多种疾病，涉及100多种症状;后来发展起来的Internist－I系统，也是一种医学诊断系统，但它可以诊断多达600多种常见的疾病
  - 1995年，微软推出了第一个基于贝叶斯网的专家系统，一个用于幼儿保健的网站OnParent (www.onparenting.msn.com)，使父母们可以自行诊断
  - 故障诊断(diagnose)、专家系统(expert system)、规划(planning)、学习(learning)、分类(classifying)

背景知识

- 概率基础
  - 事件A和B发生的概率

    $P(A)$和$P(B)$
  - B已发生条件下A发生的概率

    $P(A|B)$
  - 贝叶斯公式：根据条件概率的定义和性质，有

    $P(B|A) = P(B)P(A|B)/P(A)$

背景知识

- 概率基础
  - 事件A：机器学习任务中样本的取值状态为X
  - 事件B：机器学习模型参数θ的取值为θ_i
  - 公式可化为：

$$P(\theta_i|X) = P(\theta_i)P(X|\theta_i)/P(X)$$

其中$P(\theta_i|X)$表示在样本取值状态X的情况下模型参数取值为θ_i的条件概率

  - 根据全概率公式可以得到概率$P(X)$

$$P(X) = \sum_k P(X|\theta_k)P(\theta_k)$$

  - 将$P(X)$代入上式可得

$$P(\theta_i|X) = \frac{P(\theta_i)P(X|\theta_i)}{\sum_k P(X|\theta_k)P(\theta_k)}$$

$P(\theta_i)$是一种先验概率，表示参数取值为θ_i的概率

  - 贝叶斯方法的基本求解思路为：后验概率=先验概率×样本信息

背景知识

- **例题：**某抽奖游戏使用三个外观一致碗和三张抽奖券，其中两张1元券和一张1000元券。游戏主持人分别用每个碗盖住一张券且不让抽奖者知道每个碗盖的是几元券，在抽奖者选定一个碗之后翻开剩下两个碗中的一个，使得翻开的碗盖的是1元券。抽奖者如何选择才能以较高的概率获得1000元券

背景知识

解：令A_n表示第n个碗盖有1000元券，则有：$P(A_n)=1/3$，$n=1,2,3$

假设抽奖者选择碗1，下面讨论主持人打开了碗2概率，由于主持人知道哪个碗盖有1000元券。令B表示事件“主持人翻开了碗2”，则：

- 如果碗1盖有1000元券：$P(B|A_1)=1/2$
- 如果碗2盖有1000元券：$P(B|A_2)=0$
- 如果碗3盖有1000元券：$P(B|A_3)=1$

由全概率公式计算$P(B)$，即：

$$P(B)=P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+P(B|A_3)P(A_3)=\frac{1}{2}$$

根据公式$P(A_i|\mathrm{B})=P(B|A_i)\frac{P(A_i)}{P(B)}$，可得：$P(A_1|\mathrm{B})=1/3$，$P(A_2|\mathrm{B})=0$，$P(A_3|\mathrm{B})=2/3$

因此，如果在抽奖者选择碗1的情况下，主持人翻开碗2，则抽奖者应将碗3作为最终选择以期获得1000元券

本节目录

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策考虑如何基于这些概率和误判损失来选择最优的类别标记
  - 假设有N种可能的类别标记, 即 $y = \{c_1, c_2, \ldots, c_N\}$， λ_{ij} 是将一个真实标记为 c_j 的样本误分类为 c_i 所产生的损失。基于后验概率 $P\{c_i \mid \mathbf{x}\}$ 可获得将样本 $\mathbf{x}$ 分类为c_i 所产生的期望损失（expected loss），即在样本上的“条件风险”（conditional risk）

$$R(c_i \mid \mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \mathbf{x})$$

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法
  - 任务是寻找一个判定准则 $h : X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right]$$

  - 对每个样本 $\mathbf{x}$ ，若 h 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$ ，则总体风险 $R(h)$ 也将被最小化
  - 这就产生了贝叶斯判定准则（Bayes decision rule）：为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x)$$

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x)$$

  - 此时，被称为贝叶斯最优分类器(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 称为贝叶斯风险(Bayes risk)
  - $1 - R(h^*)$ 反映了分类起所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法
  - 若目标是最小化分类错误率，则误判损失λ_{ij} 可写为

$$\lambda_{i,j}\begin{cases} 0, & \text{if } i=j; \\ 1, & \text{otherwise}, \end{cases}$$

  - 此时条件风险

$$R(c \mid \mathbf{x}) = 1 - P(c \mid \mathbf{x})$$

  - 最小化分类错误率的贝叶斯最优分类器为

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}} P(c \mid x)$$

    - 即对每个样本$\mathbf{x}$，选择能使后验概率 $P(c \mid \mathbf{x})$ 最大的类别标记

贝叶斯决策论

- 贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法
  - 使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$
  - 然而，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率
  - 主要有两种策略
    - 判别式模型（discriminative models）
      - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$ 来预测
      - 决策树、BP神经网络、支持向量机
    - 生成式模型（generative models）
      - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
      - 生成式模型考虑 $P(c \mid \mathbf{x}) = \dfrac{P(\mathbf{x}, c)}{P(\mathbf{x})}$

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$$

类标记 c 相对于样本 $\mathbf{x}$ 的“类条件概率”（class-conditional probability），或称“似然”。

先验概率
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“证据”（evidence）因子，与类标记无关

贝叶斯决策论

- 估计类条件概率的常用策略：先假定其具有某种确定的概率分布形式，再基于训练样本对概率分布参数估计
- 记关于类别c的类条件概率为 $P(\mathbf{x} \mid c)$
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集D估计参数 $\boldsymbol{\theta}_c$

贝叶斯决策论

- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的方案
  - 频率主义学派（frequentist）认为参数虽然未知，但却存在客观值，因此可通过优化似然函数等准则来确定参数值
  - 贝叶斯学派（Bayesian）认为参数是未观察到的随机变量、其本身也可由分布，因此可假定参数服从一个先验分布，然后基于观测到的数据计算参数的后验分布

贝叶斯决策论

- **例题：**某种细胞A分为正常细胞w_1和异常细胞w_2。已知先验概率$P(A \in w_1) = 0.9$和$P(A \in w_2) = 0.1$，细胞A出现某特征X的条件概率为$P(X|A \in w_1) = 0.2, P(X|A \in w_2) = 0.4$，决策损失函数为$\lambda_{11} = 0, \lambda_{12} = 1, \lambda_{21} = 6, \lambda_{22} = 0$。试用贝叶斯决策方法判别细胞$A$出现特征$X$的情况下是否为正常细胞。

解：根据贝叶斯公式求出后验概率为：

$$P(A \in w_1|X) = \frac{P(A \in w_1)P(X|A \in w_1)}{P(X)} = \frac{P(A \in w_1)P(X|A \in w_1)}{\sum_{j=1}^{2} P(X|A \in w_j)P(A \in w_j)}$$

$$= 0.818$$

$$P(A \in w_2|X) = \frac{P(A \in w_2)P(X|A \in w_2)}{P(X)} = \frac{P(A \in w_2)P(X|A \in w_2)}{\sum_{j=1}^{2} P(X|A \in w_j)P(A \in w_j)}$$

$$= 0.182$$

贝叶斯决策论

- 解：依据上述后验概率和决策损失函数λ_{ij}可求得将细胞A分类为正常细胞w_1和异常细胞w_2的条件期望损失$R(y_i|X)$分别为：

$$R(A \in w_1|X) = \sum_{j=1}^{2} \lambda_{1j} P(A \in w_j|X) = \lambda_{12} P(A \in w_1|X) = 0.818$$

$$R(A \in w_2|X) = \sum_{j=1}^{2} \lambda_{2j} P(A \in w_j|X) = \lambda_{21} P(A \in w_2|X) = 1.092$$

由于$R(A = w_1|\mathrm{X}) < R(A = w_2|\mathrm{X})$，故$A = w_1$成立，即认为$A$是正常细胞。

本节目录

极大似然估计

- 令D_c表示训练集中第c类样本的组合的集合，假设这些样本是独立的，则参数θ_c对于数据集D_c的似然是

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c})$$

  - 对$\boldsymbol{\theta}_c$进行极大似然估计，寻找能最大化似然$P(D_c \mid \theta_c)$的参数值$\hat{\theta}_c$。直观上看，极大似然估计是试图在θ_c所有可能的取值中，找到一个使数据出现的“可能性”最大值

极大似然估计

- 连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$LL(\boldsymbol{\theta_c}) = \log P(D_c \mid \boldsymbol{\theta_c}) = \sum_{\mathbf{x} \in D_c} \log P(\mathbf{x} \mid \boldsymbol{\theta}_c)$$

- 此时参数θ_c 的极大似然估计 $\hat{\theta}_c$ 为

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}} LL(\boldsymbol{\theta}_c)$$

极大似然估计

- 例如，在连续属性情形下，假设概率密度函数 $p(\mathbf{x} \mid c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$ ，则参数$\boldsymbol{\mu}_c$和$\boldsymbol{\sigma}_c^2$的极大似然估计为

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} \mathbf{x}$$

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\mathbf{x} \in D_c} (\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$$

- 也就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是 $(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$ 的均值，这显然是一个符合直觉的结果
- 这种参数化的方法虽能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布

极大似然估计

- EM算法
  - “不完整”的样本：西瓜已经脱落的根蒂，无法看出是“蜷缩”还是“坚挺”，则训练样本的“根蒂”属性变量值未知，如何计算？
  - 未观测的变量称为“隐变量”（latent variable）。令 $\mathbf{X}$ 表示已观测变量集，$\mathbf{Z}$ 表示隐变量集，若预对模型参数 Θ 做极大似然估计，则应最大化对数似然函数

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})$$

  - 由于 $\mathbf{Z}$ 是隐变量，上式无法直接求解。此时我们可以通过对 $\mathbf{Z}$ 计算期望，来最大化已观测数据的对数“边际似然”（marginal likelihood）

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \boldsymbol{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})$$

极大似然估计

- EM算法
  - EM（Expectation-Maximization）算法[Dempster et al., 1977]是常用的估计参数隐变量的利器
    - 当参数Θ已知 —> 根据训练数据推断出最优隐变量$\mathbf{Z}$的值(E步)
    - 当$\mathbf{Z}$已知 —> 对Θ做极大似然估计(M步)
  - 以初始值Θ^0为起点，可迭代执行以下步骤直至收敛
    - 基于Θ^t推断隐变量$\mathbf{Z}$的期望, 记为$\mathbf{Z}^t$
    - 基于已观测到变量$\mathbf{X}$和$\mathbf{Z}^t$对参数Θ做极大似然估计，记为Θ^{t+1}
    - 这就是EM算法的原型
  - 可以证明EM算法是收敛的

极大似然估计

- EM算法
  - 三硬币模型

假设有3枚硬币A, B, C，它们正面朝上的概率分别是π, p, q。先进行如下试验：我们先抛硬币A，根据硬币A的结果决定接下来抛硬币B还是硬币C。如果硬币A正面朝上，我们就抛硬币B，若硬币B正面朝上记y_j=1，若硬币B反面朝上记y_j=0；如果硬币A反面朝上，我们就抛硬币C，若硬币C正面朝上记y_j=1，若硬币C反面朝上记y_j=0。独立的进行n次试验，这里n=10，得到如下观测结果：1, 1, 0, 1, 0, 0, 1, 0, 1, 1根据这组观测结果，如何估计3枚硬币模型正面朝上的概率，即3枚硬币模型的参数

极大似然估计

- EM算法
  - 三硬币模型

解：在这个问题中，实验结果是可观测数据 $Y=(y_1,...,y_N)$，硬币A的结果是不可观测数据 $Z=(z_1,...,z_N)$ 且 z 只有两种可能取值1和0

对于第j次试验，

$$
\begin{aligned}
P(y_j|\theta) &= \sum_z P(y_j, z|\theta) \\
&= \sum_z P(z|\theta)P(y_j|z,\theta) \\
&= P(z=1|\theta)P(y_j|z=1,\theta) + P(z=0|\theta)P(y_j|z=0,\theta) \\
&= \begin{cases} \pi p + (1-\pi)q, & \text{if } y_j = 1; \\ \pi(1-p) + (1-\pi)(1-q), & \text{if } y_j = 0. \end{cases} \\
&= \pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j}
\end{aligned}
$$

则：$P(Y|\theta) = \prod_{j=1}^{N} P(y_j|\theta) = \prod_{j=1}^{N} (\pi p^{y_j}(1-p)^{1-y_j} + (1-\pi)q^{y_j}(1-q)^{1-y_j})$

极大上述似然函数没有解析解，可以通过EM算法求解

极大似然估计

- EM算法
  - 三硬币模型

E步，求期望（Q函数）：

$$
\begin{aligned}
Q(\theta|\theta_n) &= \sum_z P(z|Y,\theta_n)\ln P(Y,z|\theta) \\
&= \sum_{j=1}^{N}\{\sum_z P(z|y_j,\theta_n)\ln P(y_j,z|\theta)\} \\
&= \sum_{j=1}^{N}\{P(z=1|y_j,\theta_n)\ln P(y_j,z=1|\theta) + P(z=0|y_j,\theta_n)\ln P(y_j,z=0|\theta)\}
\end{aligned}
$$

先求

$$
P(z|y_j,\theta_n) = \begin{cases} \dfrac{\pi_n p_n^{y_j}(1-p_n)^{1-y_j}}{\pi_n p_n^{y_j}(1-p_n)^{1-y_j} + (1-\pi_n)q_n^{y_j}(1-q_n)^{1-y_j}} = \mu_{j,n} & \text{if } z=1; \\ 1-\mu_{j,n} & \text{if } z=0. \end{cases}
$$

再求

$$
P(y_j,z|\theta) = \begin{cases} \pi p^{y_j}(1-p)^{1-y_j} & \text{if } z=1; \\ (1-\pi)q^{y_j}(1-q)^{1-y_j} & \text{if } z=0. \end{cases}
$$

因此，Q函数表达式为：

$$
Q(\theta|\theta_n) = \sum_{j=1}^{N}\{\mu_{j,n}\ln[\pi p^{y_j}(1-p)^{1-y_j}] + (1-\mu_{j,n})\ln[(1-\pi)q^{y_j}(1-q)^{1-y_j}]\}
$$

极大似然估计

- EM算法
  - 三硬币模型

M步，求Q函数的极大值：

1、$$\frac{\partial Q(\theta|\theta_n)}{\partial \pi} = \sum_{j=1}^{N}\{\frac{\mu_{j,n}\ln[\pi p^{y_j}(1-p)^{1-y_j}] + (1-\mu_{j,n})\ln[(1-\pi)q^{y_j}(1-q)^{1-y_j}]}{\partial \pi}\}$$

$$= \sum_{j=1}^{N}\{\mu_{j,n}\frac{p^{y_j}(1-p)^{1-y_j}}{\pi p^{y_j}(1-p)^{1-y_j}} + (1-\mu_{j,n})\frac{-q^{y_j}(1-q)^{1-y_j}}{(1-\pi)q^{y_j}(1-q)^{1-y_j}}\}$$

$$= \sum_{j=1}^{N}\{\frac{\mu_{j,n}-\pi}{\pi(1-\pi)}\}$$

$$= \frac{(\sum_{j=1}^{N}\mu_{j,n}) - n\pi}{\pi(1-\pi)} \qquad \frac{\partial Q(\theta_|\theta_n)}{\partial \pi} = 0 \implies \pi = \frac{1}{n}\sum_{j=1}^{N}\mu_{j,n}$$

2、$$\frac{\partial Q(\theta|\theta_n)}{\partial p} = \sum_{j=1}^{N}\{\frac{\mu_{j,n}\ln[\pi p^{y_j}(1-p)^{1-y_j}] + (1-\mu_{j,n})\ln[(1-\pi)q^{y_j}(1-q)^{1-y_j}]}{\partial p}\}$$

$$= \sum_{j=1}^{N}\{\mu_{j,n}\frac{\pi(y_j p^{y_j-1}(1-p)^{1-y_j} + p^{y_j}(-1)(1-y_j)(1-p)^{1-y_j-1})}{\pi p^{y_j}(1-p)^{1-y_j}} + 0\}$$

$$= \sum_{j=1}^{N}\{\frac{\mu_{j,n}(y_j-p)}{p(1-p)}\}$$

$$= \frac{(\sum_{j=1}^{N}\mu_{j,n}y_j) - (p\sum_{j=1}^{N}\mu_{j,n})}{p(1-p)} \qquad \frac{\partial Q(\theta_|\theta_n)}{\partial p} = 0 \implies p = \frac{\sum_{j=1}^{N}\mu_{j,n}y_j}{\sum_{j=1}^{N}\mu_{j,n}}$$

3、$$q = \frac{\sum_{j=1}^{N}(1-\mu_{j,n})y_j}{\sum_{j=1}^{N}(1-\mu_{j,n})}$$

极大似然估计

- EM算法
  - 三硬币模型
    - 对于初值π=0.5, p=0.5, q=0.5，得到参数估计结果分别为π=0.5, p=0.6, q=0.6
    - 对于初值π=0.4, p=0.6, q=0.7，得到参数估计结果分别为π=0.4064, p=0.5368, q=0.6432
    - EM算法与初值的选择有关

本节目录

朴素贝叶斯分类

- 估计后验概率 $P(c \mid \mathbf{x})$ 主要困难：类条件概率 $P(\mathbf{x} \mid c)$ 是所有属性上的联合概率难以从有限的训练样本估计获得
- 朴素贝叶斯分类器(Naïve Bayes Classifier)采用了“属性条件独立性假设”(attribute conditional independence assumption)：每个属性独立地对分类结果发生影响

朴素贝叶斯分类

- 基于属性条件独立性假设，则

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

其中d为属性数目，x_i 为 $\mathbf{x}$ 在第i个属性上的取值

- 由于对所有类别来说$P(x)$ 相同，因此贝叶斯判定准则有

$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

这就是朴素贝叶斯分类器的表达式

朴素贝叶斯分类

- 朴素贝叶斯分类器的训练器的训练过程就是基于训练集D估计类先验概率$P(c)$并为每个属性估计条件概率
  - 令D_c表示训练集D中第i类样本组合的集合，若有充足的独立同分布样本，则可容易地估计出类先验概率

$$P(c) = \frac{|D_c|}{D}$$

  - 对离散属性而言，令D_{c,x_i}表示D_c中在第i个属性上取值为x_i的样本组成的集合，则条件概率$P(x_i \mid c)$可估计为

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D}$$

  - 对连续属性而言可考虑概率密度函数，假定$p(x_i \mid c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$，其中$\mu_{c,i}$和$\sigma_{c,i}^2$分别是第$c$类样本在第$i$个属性上取值的均值和方差，则有

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2{\sigma_{c,i}}^2}\right)$$

朴素贝叶斯分类

- 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

朴素贝叶斯分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 估计类先验概率
  - P(好瓜=是)=8/17=0.471
  - P(好瓜=否)=9/17=0.529

朴素贝叶斯分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 为每个离散属性估计条件概率
  - P(色泽=青绿|好瓜=是)=3/8=0.375　　P(色泽=青绿|好瓜=否)=3/9=0.333
  - P(根蒂=蜷缩|好瓜=是)=5/8=0.625　　P(根蒂=蜷缩|好瓜=否)=3/9=0.333
  - P(敲声=浊响|好瓜=是)=6/8=0.750　　P(敲声=浊响|好瓜=否)=4/9=0.444
  - P(纹理=清晰|好瓜=是)=7/8=0.875　　P(纹理=清晰|好瓜=否)=2/9=0.222
  - P(脐部=凹陷|好瓜=是)=5/8=0.625　　P(脐部=凹陷|好瓜=否)=2/9=0.222
  - P(触感=硬滑|好瓜=是)=6/8=0.750　　P(触感=硬滑|好瓜=否)=6/9=0.667

朴素贝叶斯分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 为每个连续属性估计条件概率
  - $P(\text{密度}=0.697 \mid \text{好瓜}=\text{是}) = \frac{1}{\sqrt{2\pi}\times 0.129}\exp\left(\frac{(0.697-0.574)^2}{2\times 0.129^2}\right) = 1.959$
  - $P(\text{密度}=0.697 \mid \text{好瓜}=\text{否}) = \frac{1}{\sqrt{2\pi}\times 0.195}\exp\left(\frac{(0.697-0.496)^2}{2\times 0.195^2}\right) = 1.203$
  - $P(\text{含糖量}=0.460 \mid \text{好瓜}=\text{是}) = \frac{1}{\sqrt{2\pi}\times 0.101}\exp\left(\frac{(0.460-0.279)^2}{2\times 0.101^2}\right) = 0.788$
  - $P(\text{含糖量}=0.460 \mid \text{好瓜}=\text{否}) = \frac{1}{\sqrt{2\pi}\times 0.108}\exp\left(\frac{(0.460-0.154)^2}{2\times 0.108^2}\right) = 0.066$

朴素贝叶斯分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 计算联合概率，判断结果
  - P(好瓜=是)×P(色泽=青绿|好瓜=是)×P(根蒂=蜷缩|好瓜=是)×P(敲声=浊响|好瓜=是) P(纹理=清晰|好瓜=是)×P(脐部=凹陷|好瓜=是)×P(密度=0.697|好瓜=是)×P(含糖量=0.460|好瓜=是)=0.063
  - P(好瓜=否)×P(色泽=青绿|好瓜=否)×P(根蒂=蜷缩|好瓜=否)×P(敲声=浊响|好瓜=否) P(纹理=清晰|好瓜=否)×P(脐部=凹陷|好瓜=否)×P(密度=0.697|好瓜=否)×P(含糖量=0.460|好瓜=否)=0.068×10^{-5}
  - 因此，测1被判断成为好瓜

朴素贝叶斯分类

- 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题，比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理

朴素贝叶斯分类

- 例子：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）
  - 令N表示训练集D中可能的类别数，N_i 表示第i个属性可能的取值数，则概率公式修正为

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \qquad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i}$$

朴素贝叶斯分类

- 其他改进
  - 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设进行一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器”（semi-na ïve Bayes classifiers）

总结

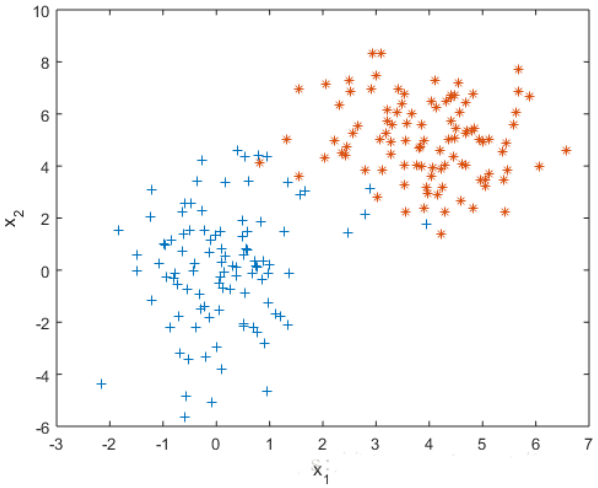
- **背景知识**
  - 贝叶斯
  - 概率基础
- **贝叶斯决策论**
  - 贝叶斯判定准则
  - 贝叶斯最优分类器
  - 判别式模型
  - 生成式模型
- **极大似然估计**
  - EM算法
- **朴素贝叶斯分类**
  - 属性条件独立性假设
  - 拉普拉斯修正

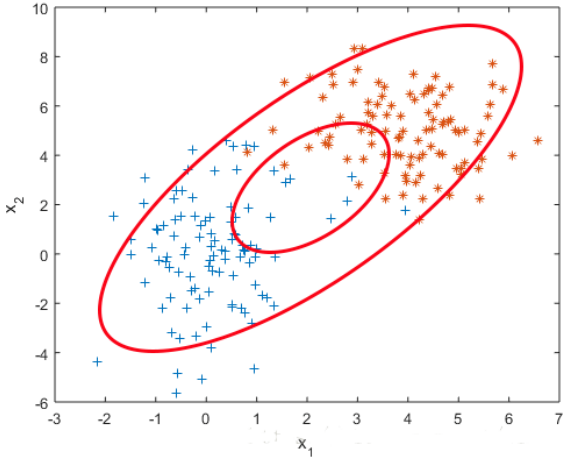
思考题

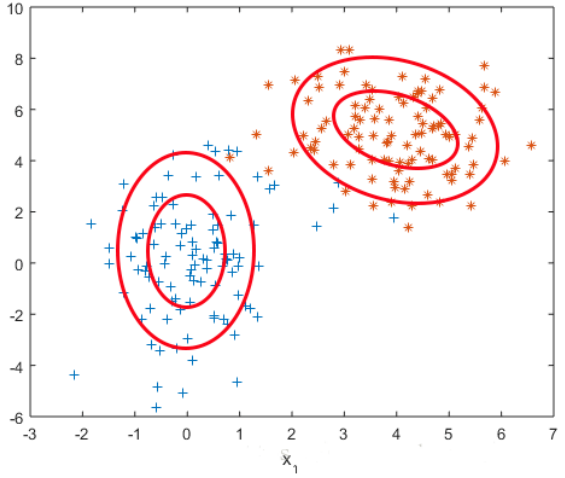
- EM算法求解高斯混合模型
  - 单高斯模型

$$f(X)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp[-\frac{1}{2}(X-u)^T\Sigma^{-1}(X-u)], X=(x_1,x_2...x_n)$$

  - 混合高斯模型
    - 越靠近椭圆的中心样本出现的概率越大，这是由概率密度函数决定的
    - 样本服从单高斯分布的假设并不合理，即单高斯模型无法产生这样的样本
    - 把两个高斯模型的线性加权融合成一个模型，可以产生这样的样本







https://blog.csdn.net/lin_limin/article/details/81048411

思考题

- EM算法求解高斯混合模型
  - 混合高斯模型
    - 假设混合高斯模型由K个高斯模型组成（即数据包含K个类），则GMM的概率密度函数如下

$$p(x)=\sum_{k=1}^{K}p(k)p(x|k)=\sum_{k=1}^{K}\pi_k N(x|u_k,\Sigma_k)$$

$p(x|k)=N(x|u_k,\Sigma_k)$ 是第k个高斯模型的概率密度函数，可以看成选定第k个模型后，该模型产生x的概率。$p(k)=\pi_k$ 是第k个高斯模型的权重，称作选择第k个模型的先验概率，且满足 $\sum_{k=1}^{K}\pi_k=1$ 。

    - 混合高斯模型的本质就是融合几个单高斯模型，来使得模型更加复杂，从而产生更复杂的样本。理论上，如果某个混合高斯模型融合的高斯模型个数足够多，它们之间的权重设定得足够合理，这个混合模型可以拟合任意分布的样本
    - 已知混合模型中各个类的分布模型（都是高斯分布）和对应的采样数据，而不知道这些采样数据分别来源于哪一类（隐变量），那这时候就需要使用EM算法
    - EM算法可以用于解决数据缺失的参数估计问题（隐变量的存在实际上就是数据缺失问题，缺失了各个样本来源于哪一类的记录）

练习题

- 试编程实现拉普拉斯修正的朴素贝叶斯分类器，并以西瓜数据集3.0为训练集，对例子中的“测1”样本进行判别