第五章 分类

贝叶斯决策论 (Bayesian decision theory)

概率框架下实施决策的基本理论

给定C个类别，令λ_{ij}代表将第j类样本误分为第i类所产生的损失，则基于后验概率将样本$\boldsymbol{x}$分到第i类的条件风险为：

$$R(c_i|\boldsymbol{x})=\sum_{j=1}^{C}\lambda_{ij}P(c_j|\boldsymbol{x})$$

贝叶斯判定准则 (Bayes decision rule)：

$$h^*(\boldsymbol{x}) = \arg\min_{c\in\mathcal{Y}} R(c \mid \boldsymbol{x})$$

- h^* 称为贝叶斯最优分类器(Bayes optimal classifier)，其总体风险称为贝叶斯风险 (Bayes risk)
- 反映了学习性能的理论上限

判别式 vs. 生成式

问题：后验概率$P(c|\boldsymbol{x})$在现实中通常难以直接获得

目标：基于有限的训练样本 尽可能准确地估计出后验概率

两种基本策略：

判别式 (discriminative) 模型

思路：直接对$P(c|\boldsymbol{x})$ 建模。

代表：

- 决策树
- BP 神经网络
- SVM

生成式 (generative) 模型

思路：先对联合概率分布$P(\boldsymbol{x}, c)$建模，
再由贝叶斯定理获得$P(c|\boldsymbol{x})$。

代表：贝叶斯分类器

贝叶斯定理

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

Thomas Bayes
(1701?-1761)

根据贝叶斯定理，有

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

先验概率（prior），
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（大数定律）

证据（evidence），
与类别无关

样本相对于类标记的类条件概率
(class-conditional probability)，
亦称 似然（likelihood）

主要困难在于估计似然

极大似然估计

先假设某种概率分布形式，再基于训练样例对参数进行估计

假定 $P(\boldsymbol{x} \mid c)$具有确定的概率分布形式，且被参数 $\boldsymbol{\theta}_c$ 唯一确定，则任务就是利用训练集 D 来估计参数 $\boldsymbol{\theta}_c$

$\boldsymbol{\theta}_c$对于训练集 D 中第 c 类样本组成的集合 D_c 的似然(likelihood)为

$$P(D_c \mid \boldsymbol{\theta}_c) = \prod_{\boldsymbol{x} \in D_c} P(\boldsymbol{x} \mid \boldsymbol{\theta}_c)$$

连乘易造成下溢，因此通常使用对数似然 (log-likelihood)

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

于是， $\boldsymbol{\theta}_c$的极大似然估计为 $\hat{\boldsymbol{\theta}}_c = \arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta}_c)$

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分布

朴素贝叶斯分类器 (naïve Bayes classifier)

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

主要障碍：所有属性上的联合概率难以从有限训练样本估计获得，存在样本稀疏问题，类似“维度灾难”。

基本思路：假定对于给定的类属性之间相互独立？

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

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

$P(\boldsymbol{x})$对所有类别相同，于是

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

朴素贝叶斯分类器

- 估计 $P(c)$：$P(c) = \frac{|D_c|}{|D|}$
- 估计 $P(\boldsymbol{x}|c)$：
  - 对离散属性，令D_{c,x_i}表示 D_c 中在第i个属性上取值为x_i 的样本集合，则

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

  - 对连续属性，考虑概率密度函数，假定$p(x_i|c) \sim \mathcal{N}(\mu_{c,i}, \sigma_{c,i}^2)$

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

- 得到类别

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

例子：朴素贝叶斯分类

编号	色泽	纹理	密度	标签
1	青绿	清晰	0.5	是
2	浅白	清晰	0.3	是
3	青绿	模糊	0.6	否
4	青绿	清晰	0.4	否
5	浅白	模糊	0.4	是
6	浅白	清晰	0.5	是
7	青绿	模糊	0.6	否
8	浅白	清晰	0.3	否

色泽	纹理	密度	标签
青绿	清晰	0.6	?

对每个类别，计算以下两步
第一步：先估计 $P(c)$ 以及 $P(x_i|c)$
第二步：计算$\mathrm{P}(c|x) = P(c)\prod_i P(x_i|c)$
最后：选择$\mathrm{P}(c|x)$值最大的类

对类别“是”计算前两步

$$P(c = 是) = \frac{4}{8}$$

$$P(色泽 = 青绿|c = 是) = \frac{1}{4}$$

$$P(纹理 = 清晰|c = 是) = \frac{1}{4}$$

好瓜密度：μ=0.425，　σ =0.0957

$$P(0.6|c = 是) = \frac{1}{\sqrt{2\pi}\cdot 0.0957}\exp\left(-\frac{(0.6-0.425)^2}{2\cdot 0.0957^2}\right) = 0.7832$$

按第二步得到 $P(c = 是|\boldsymbol{x}) = P(c = 是)\times P(青绿|是)\times P(清晰|是)\times P(0.6|是) = 0.0245$

例子：朴素贝叶斯分类

编号	色泽	纹理	密度	标签
1	青绿	清晰	0.5	是
2	浅白	清晰	0.3	是
3	青绿	模糊	0.6	否
4	青绿	清晰	0.4	否
5	浅白	模糊	0.4	是
6	浅白	清晰	0.5	是
7	青绿	模糊	0.6	否
8	浅白	清晰	0.3	否

色泽	纹理	密度	标签
青绿	清晰	0.6	?

对每个类别，计算以下两步
第一步：先估计 $P(c)$ 以及 $P(x_i|c)$
第二步：计算$\mathrm{P}(c|x) = P(c)\prod_i P(x_i|c)$
最后：选择$\mathrm{P}(c|x)$值最大的类

对类别“否”计算前两步

$$P(c = \text{否}) = \frac{4}{8}$$

$$P\,(\text{色泽} = \text{青绿}|c = \text{否}) = \frac{3}{4}$$

$$P\,(\text{纹理} = \text{清晰}|c = \text{否}) = \frac{2}{4}$$

坏瓜密度：μ=0.475，　σ =0.15

$$P\,(0.6|c = \text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.15}\exp\left(-\frac{(0.6-0.475)^2}{2\cdot 0.15^2}\right) = 1.8794$$

按第二步得到 $P\,(c = \text{否}|\text{该样本}) = P(c = \text{否}) \times P(\text{青绿}|\text{否}) \times P(\text{清晰}|\text{否}) \times P(0.6|\text{否}) = 0.3524$

最后：0.3524>0.0245，所以把该样本预测为坏瓜

拉普拉斯修正 (Laplacian correction)

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，因为概率连乘将“抹去”其他属性提供的信息

例如，若训练集中未出现“敲声=清脆”的好瓜，则
模型在遇到“敲声=清脆”的测试样本时 ……

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

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

假设了属性值与类别的均匀分布，这是额外引入的 bias

拉普拉斯修正

编号	色泽	纹理	密度	标签
1	青绿	清晰	0.5	是
2	浅白	清晰	0.3	是
3	青绿	模糊	0.6	否
4	青绿	清晰	0.4	否
5	浅白	模糊	0.4	是
6	浅白	清晰	0.5	是
7	青绿	模糊	0.6	否
8	浅白	清晰	0.3	否

色泽	纹理	密度	标签
青绿	稍糊	0.6	?

先估计$P(c)$ 以及 $P(x_i|c)$,

再计算$P(c|x) = P(c)\prod_i P(x_i|c)$

$$P(c = 是) = \frac{4+1}{8+2}$$

类别个数

$$P(纹理 = 稍糊|c = 是) = \frac{0+1}{4+3}$$

纹理取值个数

朴素贝叶斯分类器的使用

- 若对预测速度要求高

  → 预计算所有概率估值，使用时“查表”

- 若数据更替频繁

  → 不进行任何训练，收到预测请求时再估值

  (惰性学习, lazy learning)

- 若数据不断增加

  → 基于现有估值，对新样本涉及的概率估值进行修正

  (增量学习, incremental learning)

前面例子用的是哪种方式?

总结

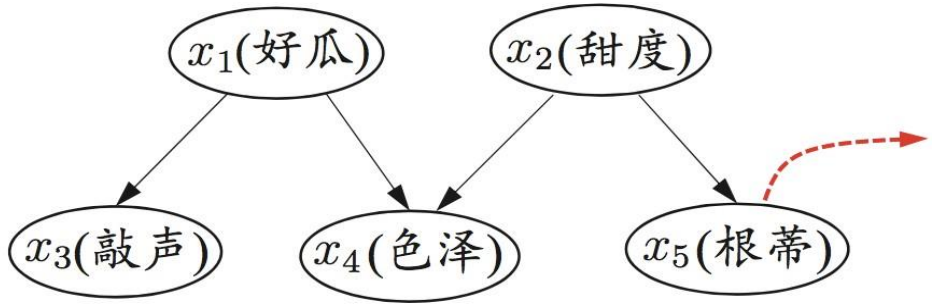
- 分类器的训练可以被认为是再估计后验概率$P(c|\boldsymbol{x})$;
- 贝叶斯分类器利用贝叶斯公式先估计联合概率再得到后验概率，而对联合概率估计的关键是求出似然$P(\boldsymbol{x}|c)$；
- 朴素贝叶斯假设属性之间独立，使得$P(\boldsymbol{x}|c) = \prod_{i=1}^{d} P(x_i|c)$。

贝叶斯网* （Bayesian network; Bayes network）

亦称“信念网”（brief network）

有向无环图

(Directed Acyclic Graph)



条件概率表

(Conditional Probability Table)

		根蒂：硬挺	根蒂：蜷缩
甜度	高	0.1	0.9
甜度	低	0.7	0.3

贝叶斯网 $B = \langle G, \Theta \rangle$

结构　　参数

1985年 J. Pearl 命名为贝叶斯网，为了强调：

- 输入信息的主观本质
- 对贝叶斯条件的依赖性
- 因果与证据推理的区别

Judea Pearl
2011 图灵奖

概率图模型(Probabilistic graphical model)

- 有向图模型 → 贝叶斯网
- 无向图模型 → 马尔可夫网

贝叶斯网（Bayesian network; Bayes network）

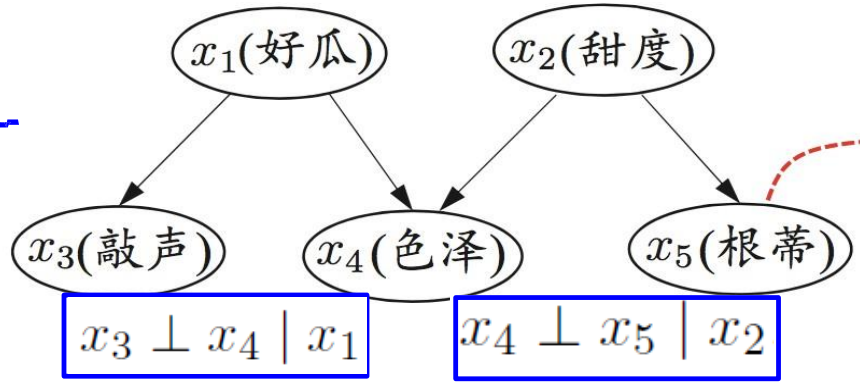
亦称“信念网”（brief network）

有向无环图

(Directed Acyclic Graph)

条件概率表

(Conditional Probability Table)



		根蒂：硬挺	根蒂：蜷缩
甜度	高	0.1	0.9
甜度	低	0.7	0.3

给定父结点集，贝叶斯网假设每个属性与其非后裔属性独立

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

π_i：父结点集

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$