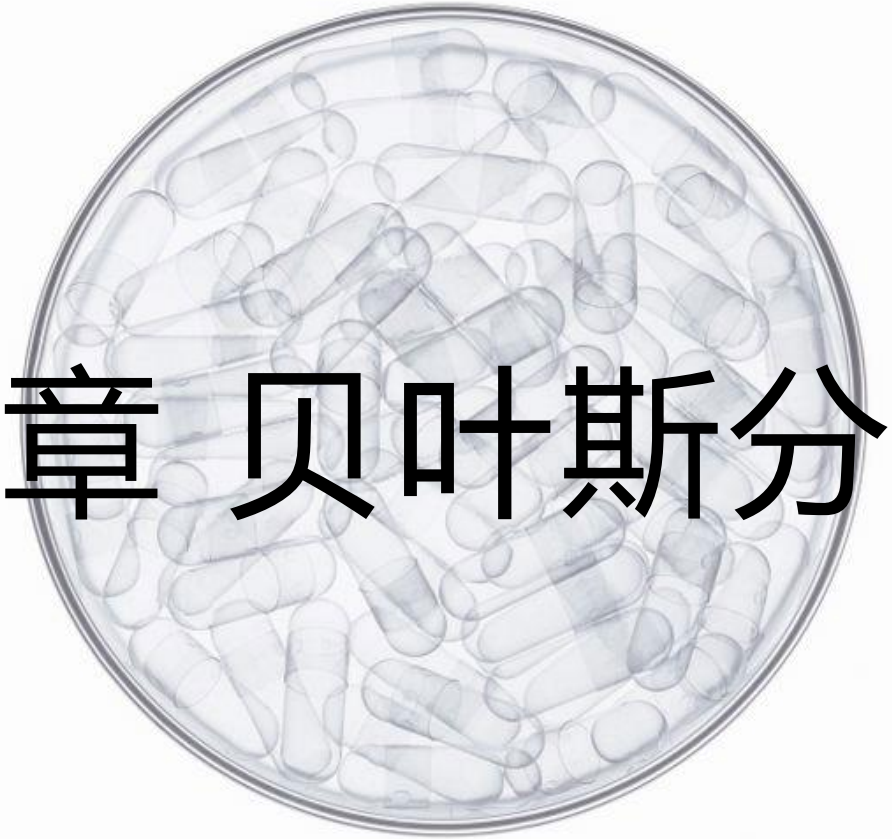
机器学习



第七章 贝叶斯分类器

汇报人：杜昭武

目录

一、贝叶斯决策论

概率框架下实施决策的基本理论

给定N个类别，令 λ_{ij} 代表第j类样本误分类为第i类所产生的损失，则

基于 后验概率将样本x分类到第i类的**条件风险**为：

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

贝叶斯判定准则：

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

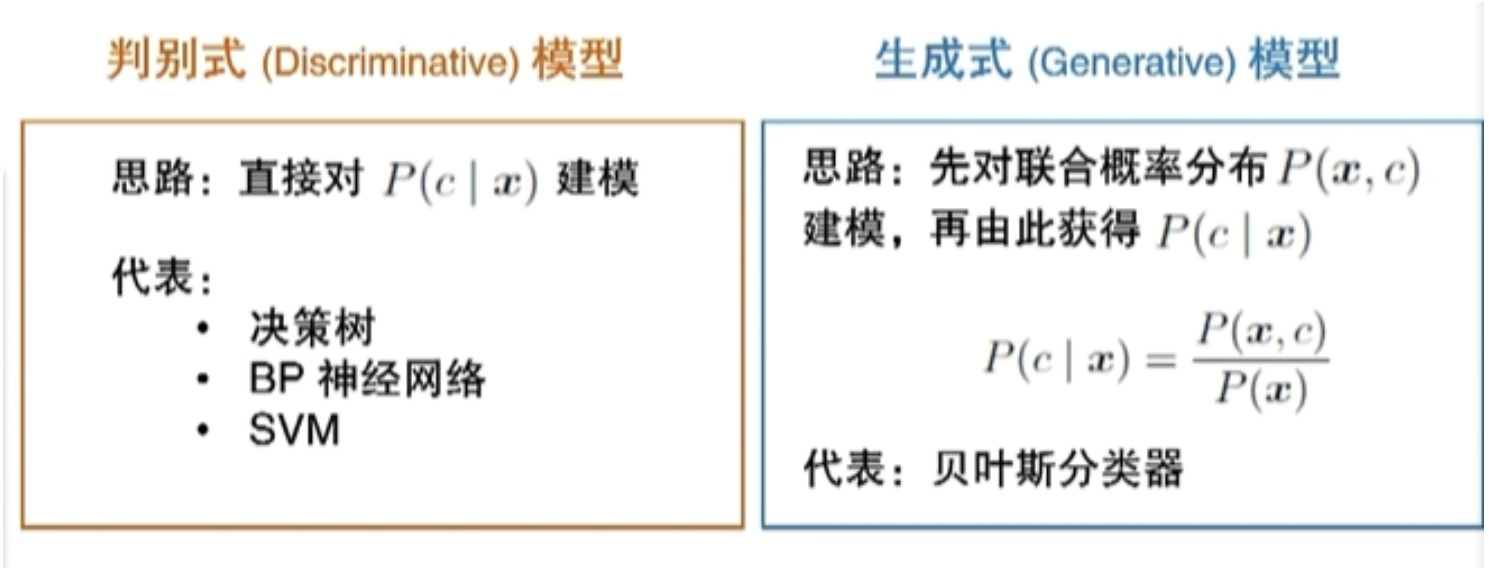
- h(x)称为贝叶斯最优分类器，其总体风险称为贝叶斯风险

一、贝叶斯决策论

判别式 VS. 生成式

p（c | x）后验概率在现实中通常难以直接获得
从这个角度来看，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

两种基本策略：



一、贝叶斯决策论

基于贝叶斯定理，$P(c \mid \boldsymbol{x})$可写为

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

先验概率p(c)：表达了样本空间中各类样本所占的比例，当训练集包含充足的独立同分布样本时，p(c)可通过各类样本出现的频率来进行估计。

似然概率p(x|c)：是样本x相对于类标记c的类条件概率，也称为似然概率。

证据 p(x)：用于归一化的“证据”因子。

二、最大似然估计

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(x | c)具有确定的概率分布形式，且被参数θ唯一确定，则任务就是利用训练集D来估计参数θ。极大似然估计本质是得到类条件概率p(x|c)，

θ对于训练集D中第c类样本组成的集合D的似然为

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

连乘易造成下溢因此通常使用对数似然

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

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

三、朴素贝叶斯分类器

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

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
组合爆炸；样本稀疏

基本思路：假定属性相互独立

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

d为属性数，Xi为X在第i个属性上的取值，p(x)对所有类别相同，于是

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

三、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Dc中在第i个属性上取值为Xi的样本组成的集合，则

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

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

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

三、贝叶斯分类器

举例:

下面我们用西瓜数据集 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	?

首先估计类先验概率 $P(c)$, 显然有

$$P(\text{好瓜}=\text{是}) = \frac{8}{17} \approx 0.471 ,$$
$$P(\text{好瓜}=\text{否}) = \frac{9}{17} \approx 0.529 .$$

结果:

于是, 有

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038 ,$$
$$P(\text{好瓜}=\text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5} .$$

由于 $0.038 > 6.80 \times 10^{-5}$, 因此, 朴素贝叶斯分类器将测试样本“测 1”判别为“好瓜”.

然后, 为每个属性估计条件概率 $P(x_i \mid c)$:

$$P_{\text{青绿}|\text{是}} = P(\text{色泽}=\text{青绿} \mid \text{好瓜}=\text{是}) = \frac{3}{8} = 0.375 ,$$
$$P_{\text{青绿}|\text{否}} = P(\text{色泽}=\text{青绿} \mid \text{好瓜}=\text{否}) = \frac{3}{9} \approx 0.333 ,$$
$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂}=\text{蜷缩} \mid \text{好瓜}=\text{是}) = \frac{5}{8} = 0.375 ,$$
$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂}=\text{蜷缩} \mid \text{好瓜}=\text{否}) = \frac{3}{9} \approx 0.333 ,$$
$$P_{\text{浊响}|\text{是}} = P(\text{敲声}=\text{浊响} \mid \text{好瓜}=\text{是}) = \frac{6}{8} = 0.750 ,$$
$$P_{\text{浊响}|\text{否}} = P(\text{敲声}=\text{浊响} \mid \text{好瓜}=\text{否}) = \frac{4}{9} \approx 0.444 ,$$
$$P_{\text{清晰}|\text{是}} = P(\text{纹理}=\text{清晰} \mid \text{好瓜}=\text{是}) = \frac{7}{8} = 0.875 ,$$
$$P_{\text{清晰}|\text{否}} = P(\text{纹理}=\text{清晰} \mid \text{好瓜}=\text{否}) = \frac{2}{9} \approx 0.222 ,$$
$$P_{\text{凹陷}|\text{是}} = P(\text{脐部}=\text{凹陷} \mid \text{好瓜}=\text{是}) = \frac{6}{8} = 0.750 ,$$

$$p_{\text{密度: 0.697}|\text{是}} = p(\text{密度}=0.697 \mid \text{好瓜}=\text{是}) = \frac{1}{\sqrt{2\pi}\cdot 0.129} \exp\left(-\frac{(0.697-0.574)^2}{2\cdot 0.129^2}\right) \approx 1.959 ,$$
$$p_{\text{密度: 0.697}|\text{否}} = p(\text{密度}=0.697 \mid \text{好瓜}=\text{否}) = \frac{1}{\sqrt{2\pi}\cdot 0.195} \exp\left(-\frac{(0.697-0.496)^2}{2\cdot 0.195^2}\right) \approx 1.203 ,$$

四、EM算法

未观测变量称为“隐变量”，令X表示已观测变量集，Z表示隐变量集，θ表示模型参数，若对θ做极大似然估计，则最大化对数似然：

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

通过对Z计算期望，来最大化已观测数据的对数“边际似然”

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

三、EM算法

基本思想：

- 若参数θ已知，则可以根据训练数据推断出最优隐变量Z的值.（E步）

$$Q(\Theta \mid \Theta^t) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^t}\, LL(\Theta \mid \mathbf{X}, \mathbf{Z}) \ .$$

- 若Z的值已知，则可方便地去参数θ做极大似然估计(M步)

$$\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta \mid \Theta^t) \ .$$

于是, 以初始值 Θ^0 为起点, 对式(7.35), 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.

五、课后习题

- 试用极大似然法估算西瓜集3.0中前3个属性的类条件概率。

好瓜\色泽	乌黑	浅白	青绿
否	2	4	3
是	4	1	3

$p_1 = P_{乌黑|是}$　$p_2 = P_{浅白|是}$　$p_3 = P_{青绿|是}$　$p_4 = 1-p_1-p_2$

对色泽各性别的似然概率 $L(p) = P(X_{色泽}|下=是)$

$$= \prod_{x \in 是} P(x) = p_1^4 p_2^1 (1-p_1-p_2)^3$$

对数似然为 $LL(p) = \ln(p_1) + \ln(p_2) + 3\ln(1-p_1-p_2)$

分别对 p_1、p_2 求偏导并使其为 0，得到 p_1, p_2 的极大似然估计

$\hat{p}_1 = \frac{1}{2}$　$\hat{p}_2 = \frac{1}{8}$　$\hat{p}_3 = 1-\hat{p}_1-\hat{p}_2 = \frac{3}{8}$

同理

$P_{青|否} = \frac{3}{5}$　$P_{乌|否} = \frac{2}{5}$　$P_{白|否} = \frac{4}{5}$

$P_{蜷|是} = \frac{5}{8}$　$P_{蜷|是} = \frac{3}{8}$　$P_{硬|是} = 0$

$P_{蜷|是} = \frac{3}{5}$　$P_{蜷|否} = \frac{1}{5}$　$P_{硬|否} = \frac{1}{5}$

$P_{沉|是} = \frac{5}{8}$　$P_{沉|是} = \frac{1}{8}$　$P_{响|是} = \frac{0}{8}$

$P_{沉|否} = \frac{4}{5}$　$P_{沉|否} = \frac{3}{5}$　$P_{响|否} = \frac{2}{5}$

五 、课后习题

- 试证明：条件独立性假设不成立时，朴素贝叶斯分类器任有可能产生最优分类器。

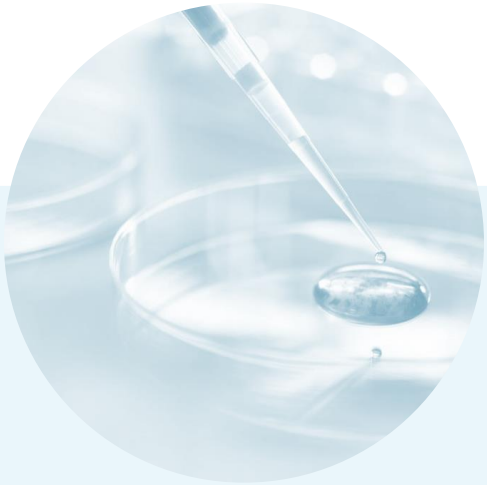
  朴素贝叶斯分类器就是建立在条件独立性假设上的。当有不独立的属性时，假如所有样本不独立的属性取值相同时分类也是相同的，那么此时朴素贝叶斯分类器也将产"生最优分类器。

- **实践中使用式(7.15)决定分类类别时，若数据的维数非常高，则概率连乘的结果通常会非常接近于0从而导致下溢。试述防止下溢的可能方案**

  这在p153中已经给出答案。即取对数将连乘转化为"连加"防止下溢。

  连乘易造成下溢因此通常使用对数似然

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



谢谢