计算机应用数学

Applied Mathematics for Computer Science

主讲教师：宋骐

https://songqi1990.github.io/

中国科学技术大学 University of Science and Technology of China

1

1.4 贝叶斯分类器

24 October 2025

2

贝叶斯分类器

- 每个记录用一个d维特征向量$\mathbf{X} = (x_1, x_2, \ldots, x_d)$表示
- 假定有k个类$y_1, y_2, \ldots, y_k$
- 给定X，X属于y_j类的后验概率$\mathbf{P}(y_j|\mathbf{X})$ 满足贝叶斯(Bayes)定理：

$$P(y_j|X) = \frac{P(X|y_j)P(y_j)}{P(X)}$$

3

贝叶斯分类器

$$P(y_j|X)=\frac{P(X|y_j)P(y_j)}{P(X)}$$

- MAP (Maximum Posteriori Hypothesis, 最大后验假设)
  - 将X指派到具有最大后验概率$P(y_j|X)$的类y_j
    - 即将X指派到$P(X|y_j)P(y_j)$ 最大的类y_j

4

朴素贝叶斯分类
Naïve Bayes Classifier

- 给定一个未知的数据样本X, 分类法将预测X属于具有最高后验概率的类。即未知的样本分配给类y_j, 当且仅当

$$P(y_j|\mathbf{X})>P(y_i|\mathbf{X}),\quad 1\le i\le k,\quad i\ne j$$

根据贝叶斯定理, 我们有

$$P(y_j|\mathbf{X})=\frac{P(\mathbf{X}|y_j)P(y_j)}{P(\mathbf{X})}$$

- 由于P(X) 对于所有类为常数, 只需要最大化 $P(X|y_j)P(y_j)$即可.

5

估计$P(X|y_j)P(y_j)$

- 估计$P(y_j)$
  - 类别y_j的先验概率可以用下式估计

    $P(y_j)=n_j/n$
  - 其中, n_j是类y_j中的训练样本数,而n是训练样本总数
- 估计$P(X|y_j)$
  - 为便于估计$P(X|y_j)$,假定类条件独立,即给定样本的类标号, 假定属性值条件地相互独立
  - 于是, $P(X|y_j)$可以用下式估计

$$P(\mathbf{X}|y_j)=\prod_{i=1}^{d}P(x_i|y_j)$$

  - 其中, $P(x_i|y_j)$可以由训练样本估值

6

估计$P(x_i|y_j)$

- 设第i个属性A_i是分类属性, 则

$$P(x_i|y_j) = n_{ij}/n_j$$

其中n_{ij}是在属性A_i上具有值x_i的y_j类的训练样本数, 而n_j是y_j类的训练样本数

- 设第i个属性A_i是连续值属性
  - 把A_i离散化
  - 假定A_i服从正态分布

$$P(x_i \mid y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - 其中, μ_{ij}, σ_{ij}分别为给定y_j类的训练样本在属性A_i上的均值和标准差

- 朴素贝叶斯分类器所需要的信息
  - 计算每个类的先验概率$P(y_j)$

$$P(y_j)=n_j/n$$

其中, n_j是y_j类的训练样本数,而n是训练样本总数

  - 对于离散属性A_i，设的不同值为a_{i1}, a_{i2}, …,a_{il}，
    - 对于每个类y_j，计算后验概率$P(a_{ik}|y_j)$, $1 \le k \le l$

$$P(a_{ik}|y_j)= n_{ikj}/n_j$$

其中n_{ikj}是在属性A_i上具有值a_{ik}的y_j类的训练样本数, 而n_j是y_j类的训练样本数

  - 对于连续属性A_i和每个类y_j, 计算y_j类样本的均值μ_{ij}, 标准差σ_{ij}

- X=（有房=否，婚姻状况=已婚，年收入=120K），求X的分类

Tid	有房	婚姻状况	年收入	拖欠贷款
1	是	单身	125K	No
2	否	已婚	100K	No
3	否	单身	70K	No
4	是	已婚	120K	No
5	否	离婚	95K	Yes
6	否	已婚	60K	No
7	是	离婚	220K	No
8	否	单身	85K	Yes
9	否	已婚	75K	No
10	否	单身	90K	Yes

P(Yes)=3/10
P(No)=7/10
P(有房=是|No) =3/7
P(有房=否|No) =4/7
P(有房=是|Yes) =0
P(有房=否|Yes) =1
P(婚姻状况=单身|No) =2/7
P(婚姻状况=离婚|No) =1/7
P(婚姻状况=已婚|No) =4/7
P(婚姻状况=单身|Yes) =2/3
P(婚姻状况=离婚|Yes) =1/3
P(婚姻状况=已婚|Yes) =0

年收入：
类=No：样本均值=110
样本方差=2975
类=Yes：样本均值=90
样本方差=25

$$P(x_i \mid y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

计算P(X| No)P(No)和P(X| Yes)P(Yes)

第1.4节 贝叶斯分类器 第10页

- X=（有房=否，婚姻状况=已婚，年收入=120K），求X的分类
- 计算$P(X|No)$和$P(X|Yes)$

$P(X|No) = P(有房=否|No) \times P(婚姻状况=已婚|No) \times P(年收入=\$120K|No)$

$= 4/7 \times 4/7 \times \frac{1}{\sqrt{2\pi}\times 2975} e^{-\frac{(120-110)^2}{2\times 2975}} = 4/7 \times 4/7 \times 0.0072 = 0.0024$

$P(X|Yes) = P(有房=否|Yes) \times P(婚姻状况=已婚|Yes) \times P(年收入=\$120K|Yes)$

$= 1 \times 0 \times 1.2 \times 10^{-9} = 0$

- 计算$P(X|No)P(No)$和$P(X|Yes)P(Yes)$

$P(X|No)P(No) = 0.0024 \times 0.7 = 0.00168$

$P(X|Yes)P(Yes) = 0 \times 0.3 = 0$

- 因为$P(X|No)P(No) > P(X|Yes)P(Yes)$, 所以X分类为No

P(Yes)=3/10
P(No)=7/10
P(有房=是|No) =3/7
P(有房=否|No) =4/7
P(有房=是|Yes) =0
P(有房=否|Yes) =1
P(婚姻状况=单身|No) =2/7
P(婚姻状况=离婚|No) =1/7
P(婚姻状况=已婚|No) =4/7
P(婚姻状况=单身|Yes) =2/3
P(婚姻状况=离婚|Yes) =1/3
P(婚姻状况=已婚|Yes) =0

年收入：
类=No：样本均值=110
样本方差=2975
类=Yes：样本均值=90
样本方差=25

10

第1.4节 贝叶斯分类器 第11页

- 存在问题
  - 如果诸条件概率$P(X_i=x_i|Y=y_j)$ 中的一个为0，则它们的乘积（计算$P(X|Y=y_j)$的表达式）为0
  - 很可能每个$P(X|Y=y_j)$都为0
- 解决方法
  - 使用Laplace 估计:

原估计: $P(X_i=x_i|Y=y_j) = n_{ij}/n_j$

$$\text{Laplace}: P(X_i = x_i | Y = y_i) = \frac{n_{ij}+1}{n_j + k}$$

11

第1.4节 贝叶斯分类器 第12页

m-估计

1、当n_{ij}过小n_{ij}/n_j产生了一个有偏的过低估计概率。
2、当此概率估计为0时，将来的查询此概率项将会在贝叶斯分类器中占统治地位，很可能每个P(X|Y=yj)都为0。

为了避免此问题，所以需要采用一种估计概率，即如下定义的m-估计

$$\frac{n_{ij}+mp}{n_j+m}$$

其中n_{ij}是具有属性y_j类的训练样本数，而n_j是y_j类的训练样本数
p为将要确定的概率的先验估计，m为等效样本大小的常量

12

m-估计

$$\frac{n_{ij}+mp}{n_j+m}$$

$$\begin{aligned}
&P(x_i|y_j) = n_{ij}/n_j \\
&= n_{ij}(n_j+m)/n_j(n_j+m) \\
&= (n_{ij}n_j + n_{ij}m)/n_j(n_j+m) \\
&= (n_{ij} + m*(n_{ij}/n_j))/(n_j+m) \\
&\approx (n_{ij} + mp)/(n_j+m) \qquad \text{用 } p \approx n_{ij}/n_j \text{代入} \\
&= (n_{ij} + mp)/(n_j + m)
\end{aligned}$$

13

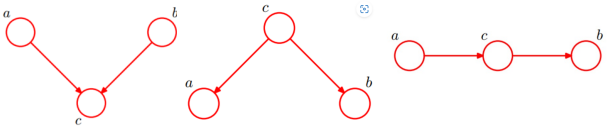
贝叶斯分类特点

- 对孤立的噪声点的鲁棒性
  - 个别点对概率估计的影响很小
- 容易处理缺失值
  - 在估计概率时忽略缺失值的训练实例
- 对不相关属性的鲁棒性
  - 各类在不相关属性上具有类似分布
- 类条件独立假设可能不成立
  - 使用其他技术，如贝叶斯信念网络（Bayesian Belief Networks）

14

贝叶斯网络结构

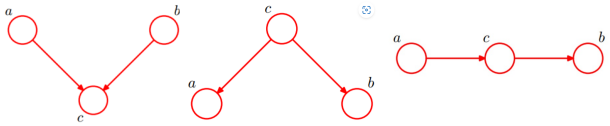
- $P(a,b,c) = P(c|a,b)P(b)P(a)$，可得：$P(a,b) = P(a)*P(b)$，在$c$未知的条件下，$a,b$被阻断(blocked)，是独立的，称之为head-to-head条件独立。
- $P(a,b,c) = P(b|c)P(a|c)P(c)$，可得：$P(a,b|c) = P(a|c)*P(b|c)$，在$c$给定的条件下，$a,b$被阻断(blocked)，是独立的，称之为tail-to-tail条件独立。



15

贝叶斯网络结构

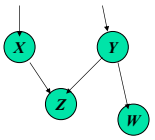
- head-to-head条件独立：$P(a,b) = P(a) * P(b)$。
- tail-to-tail条件独立：$P(a,b|c) = P(a|c) * P(b|c)$
- $P(a,b,c) = P(b|c)P(c|a)P(a)$，化简可得：$P(a,b,c) = P(a) * P(c|a) * P(b|c)$，在$c$给定的条件下，$a,b$被阻断(blocked)，是独立的，称之为head-to-tail条件独立



16

贝叶斯信念网络

- BBN允许在变量的子集间定义类条件独立性
- 因果关系图模型
  - 表示变量之间的依赖
  - 给出联合概率分布的说明
- 图示
  - 节点: 随机变量
  - 边: 依赖
  - X, Y 是Z的父节点/前驱，并且Y 是W的父节点/前驱
  - Z 和W之间没有依赖关系
  - 图中没有环



17

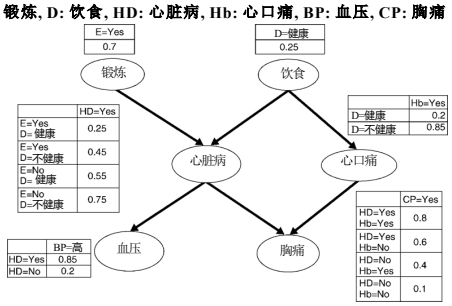
贝叶斯信念网络训练

- 若干情况
  - 给定网络结构和所有可观测变量
    - 只需要学习CPT
  - 网络结构未知, 所有的变量可以观测
    - 搜索模型空间, 构造网络拓扑结构
  - 网络结构已知,而某些变量是隐藏的
    - 使用梯度下降法或类似于神经网络的方法训练信念网络
  - 网络结构未知, 所有变量是隐藏的
    - 没有已知的好算法

18

贝叶斯信念网络推理

- E: 锻炼, D: 饮食, HD: 心脏病, Hb: 心口痛, BP: 血压, CP: 胸痛



19

Case 1: 没有先验信息

- 通过计算先验概率$P(HD=Yes)$和$P(HD=No)$来确定一个人是否可能患心脏病
- 设$\alpha \in \{Yes, No\}$表示锻炼的两个值，$\beta \in \{健康, 不健康\}$表示饮食的两个值，由全概率公式

$$P(HD=Yes) = \sum_{\alpha}\sum_{\beta} P(HD=Yes \mid E=\alpha, D=\beta)P(E=\alpha, D=\beta)$$
$$= \sum_{\alpha}\sum_{\beta} P(HD=Yes \mid E=\alpha, D=\beta)P(E=\alpha)P(D=\beta)$$
$$= 0.25\times0.7\times0.25 + 0.45\times0.7\times0.75 + 0.55\times0.3\times0.25 + 0.75\times0.3\times0.75$$
$$= 0.49$$

- 因为$P(HD=No) = 1 - P(HD=Yes) = 0.51$，所以，此人不得心脏病的机率略微大一点

20

Case 2: 高血压

- 如果一个人有高血压，可以通过比较后验概率$P(HD=Yes|BP=高)$和$P(HD=No|BP=高)$来诊断他是否患有心脏病
- 先用全概率公式，计算$P(BP=高)$

$$P(BP=高) = \sum_{g} P(BP=高 \mid HD=g)P(HD=g)$$
$$= 0.85\times0.49 + 0.2\times0.51 = 0.5185$$

  - 其中$\gamma \in \{Yes, No\}$
- 用贝叶斯公式计算此人患心脏病的后验概率

$$P(HD=Yes \mid BP=高) = \frac{P(BP=高 \mid HD=Yes)P(HD=Yes)}{P(BP=高)} = \frac{0.85\times0.49}{0.5185} = 0.8033$$

21