



# 贝叶斯分类器

---

杜逆索



# 贝叶斯分类算法

---

## 1. 基本概念

### 1.1 主观概率

贝叶斯方法是一种研究不确定性的推理方法，不确定性常用贝叶斯概率表示，它是一种主观概率，是人的认识，是个人主观的估计，随个人的主观认识的变化而变化。对它的估计取决于先验知识的正确和后验知识的丰富和准确，因此贝叶斯概率常常可能随个人掌握信息的不同而发生变化，基于后验知识的一种判断，取决于对各种信息的掌握。



# 贝叶斯分类算法

---

## 1. 基本概念

### 1.2 贝叶斯定理

#### 1. 基础知识

(1) 已知事件A发生的条件下，事件B发生的概率，叫做事件B在事件A发生下的条件概率，记为  $P(B|A)$ ，其中  $P(A)$  叫做先验概率， $P(B|A)$  叫做后验概率，计算条件概率的公式为：

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

条件概率公式通过变形得到乘法公式：

$$P(A \cap B) = P(B|A)P(A)$$



# 贝叶斯分类算法

---

## 1. 基本概念

### 1.2 贝叶斯定理

#### 1. 基础知识

(2) 设A,B为两个随机事件，如果有 $P(AB) = P(A)P(B)$ 成立，则称事件A和B相互独立。此时有 $P(A|B) = P(A)$ ， $P(AB) = P(A)P(B)$ 成立。



# 贝叶斯分类算法

## 1. 基本概念

### 1.2 贝叶斯定理

#### 1. 基础知识

(3) 设  $B_1, B_2, \dots, B_n$  为互不相容事件,  $P(B_i) > 0, i = 1, 2, \dots, n$ , 且  $\bigcup_{i=1}^n B_i = \Omega$ , 对任意的事件  $A \subset \bigcup_{i=1}^n B_i$ , 计算事件A概率的公式为:

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i)$$

设  $P(A) > 0$ , 则在事件A发生的条件下, 事件  $B_i$  发生的概率为:

$$P(B_i|A) = \frac{P(B_i A)}{P(A)} = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^n P(B_i)P(A|B_i)}$$

称该公式为贝叶斯公式。



# 贝叶斯分类算法

---

假设一个学校里有60%男生和40%女生。女生穿裤子的人数和穿裙子的人数相等，所有男生穿裤子。一个人在远处随机看到了一个穿裤子的学生。那么这个学生是女生的概率是多少？



# 贝叶斯分类算法

假设一个学校里有60%男生和40%女生。女生穿裤子的人数和穿裙子的人数相等，所有男生穿裤子。一个人在远处随机看到了一个穿裤子的学生。那么这个学生是女生的概率是多少？

$P(A)$ 是忽略其它因素，看到女生的概率，在这里是40%

$P(A')$ 是忽略其它因素，看到不是女生（即看到男生）的概率，在这里是60%

$P(B|A)$ 是女生穿裤子的概率，在这里是50%

$P(B|A')$ 是男生穿裤子的概率，在这里是100%

$P(B)$ 是忽略其它因素，学生穿裤子的概率，

$P(B) = P(B|A)P(A) + P(B|A')P(A')$ ，在这里是 $0.5 \times 0.4 + 1 \times 0.6 = 0.8$ 。

根据贝叶斯定理，我们计算出后验概率 $P(A|B)$



# 贝叶斯分类算法

假设一个学校里有60%男生和40%女生。女生穿裤子的人数和穿裙子的人数相等，所有男生穿裤子。一个人在远处随机看到了一个穿裤子的学生。那么这个学生是女生的概率是多少？

$P(A)$ 是忽略其它因素，看到女生的概率，在这里是40%

$P(A')$ 是忽略其它因素，看到不是女生（即看到男生）的概率，在这里是60%

$P(B|A)$ 是女生穿裤子的概率，在这里是50%

$P(B|A')$ 是男生穿裤子的概率，在这里是100%

$P(B)$ 是忽略其它因素，学生穿裤子的概率，

$P(B) = P(B|A)P(A) + P(B|A')P(A')$ ，在这里是 $0.5 \times 0.4 + 1 \times 0.6 = 0.8$ 。

根据贝叶斯定理，我们计算出后验概率 $P(A|B)$

$P(A|B) = P(B|A) * P(A) / P(B) = 0.25$





# 贝叶斯分类算法

---

## 1. 基本概念

### 1.2 贝叶斯定理

#### 2. 贝叶斯决策准则

如果对于任意  $i \neq j$ ，都有  $P(C_i|X) > P(C_j|X)$  成立，则样本模式  $X$  被判定为类别  $C_i$ 。



# 贝叶斯分类算法

---

## 1. 基本概念

### 1.2 贝叶斯定理

#### 3. 极大后验假设

根据贝叶斯公式可得到一种计算后验概率的方法：在一定假设的条件下，根据先验概率和统计样本数据得到的概率，可以得到后验概率。

令 $P(c)$ 是假设 $c$ 的先验概率，它表示 $c$ 是正确假设的概率， $P(X)$ 表示的是训练样本 $X$ 的先验概率， $P(X|c)$ 表示在假设 $c$ 正确的条件下样本 $X$ 发生或出现的概率，根据贝叶斯公式可以得到后验概率的计算公式：

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



# 贝叶斯分类算法

---

## 1. 基本概念

### 1.2 贝叶斯定理

#### 3. 极大后验假设

设  $C$  为类别集合也就是待选假设集合，在给定未知类别标号样本  $X$  时，通过计算找到可能性最大的假设  $c \in C$ ，具有最大可能性的假设或类别被称为极大后验假设(maximum a posteriori)，记作：

$$c_{map} = \arg \max_{c \in C} P(c|X) = \arg \max_{c \in C} \frac{P(X|c)P(c)}{P(X)}$$



# 贝叶斯分类算法

---

## 2. 贝叶斯分类算法原理

### 2.1 朴素贝叶斯分类模型

朴素贝叶斯分类模型：算法逻辑简单、运算速度快、分类耗时短、精度高。

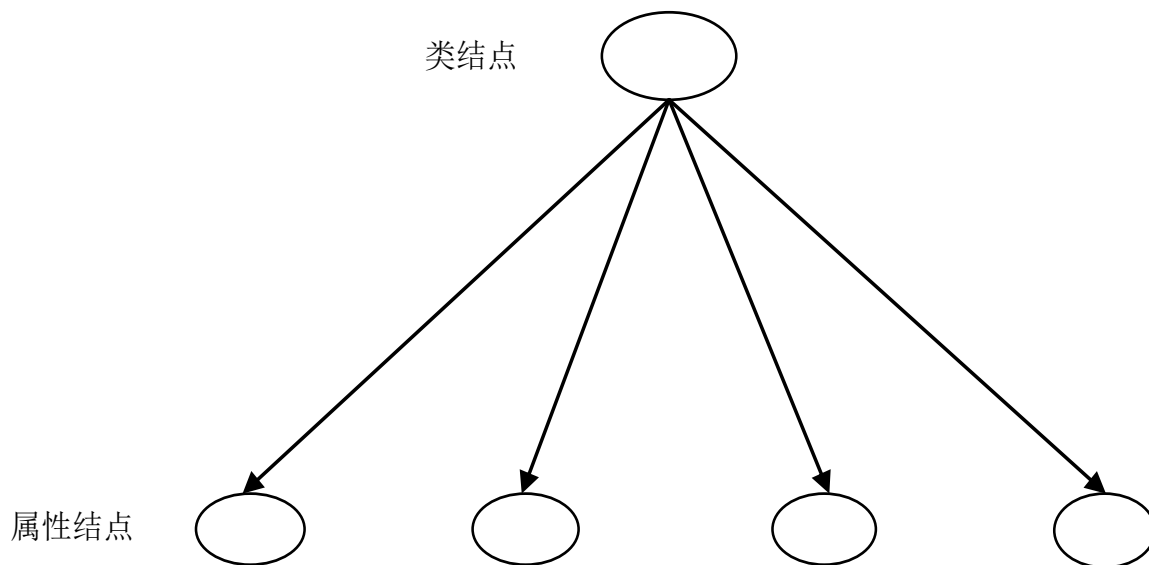
以属性的类条件独立性假设为前提，即在给定类别状态条件下，属性之间是相互独立的。

# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

### 2.1 朴素贝叶斯分类模型

朴素贝叶斯分类器的结构示意图如下图所示：





# 贝叶斯分类算法

---

## 2. 贝叶斯分类算法原理

### 2.1 朴素贝叶斯分类模型

朴素贝叶斯分类模型的算法描述如下：

- (1) 对训练样本数据集和测试样本数据集进行离散化处理和缺失值处理；
- (2) 扫描训练样本数据集，分别统计训练集中类别  $C_i$  的个数  $d_i$  和属于类别  $C_i$  的样本中属性  $A_k$  取值为  $x_k$  的实例样本个数  $d_{ik}$ ，构成统计表；
- (3) 计算先验概率和条件概率  $P(A_k = x_k | C_i) = d_{ik} / d_i$ ，构成概率表；
- (4) 构建分类模型  $V(X) = \arg \max_i P(C_i) P(X | C_i)$
- (5) 扫描待分类的样本数据集，调用已得到的统计表、概率表以及构建好的分类准则，得出分类结果。

# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

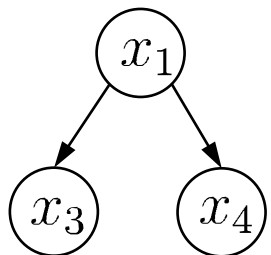
### 2.2 贝叶斯信念网

#### 1. 模型表示

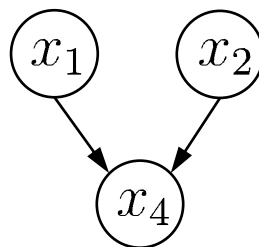
贝叶斯信念网络,简称贝叶斯网络, 用图形表示一组随机变量之间的概率关系。

贝叶斯网络有两个主要成分:

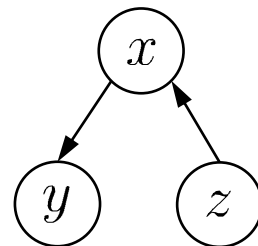
- (1) 一个有向无环图, 表示变量之间的依赖关系。
- (2) 一个概率表, 把各结点和它的直接父结点关联起来。



同父结构



V型结构



顺序结构

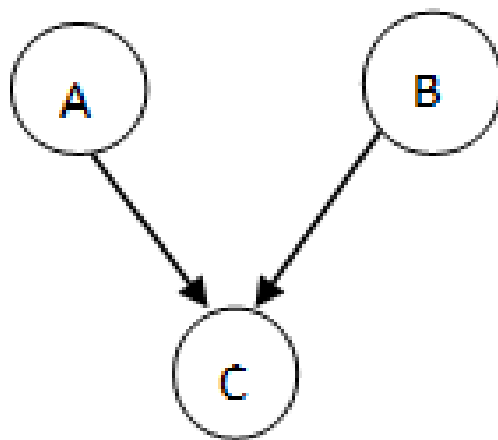
# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型表示

考虑三个随机变量A、B和C，A和B相互独立，都直接影响变量C。三个之间的关系可以用图中的有向无环图概括。图中每个结点表示一个变量，每条弧表示变量之间的依赖关系。





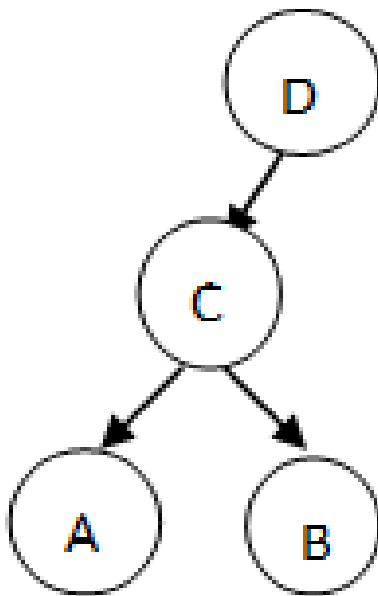
# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型表示

A是D的后代，D是B的祖先。给定变量C,A条件独立于B和D,因为B和D都是A的非后代结点。



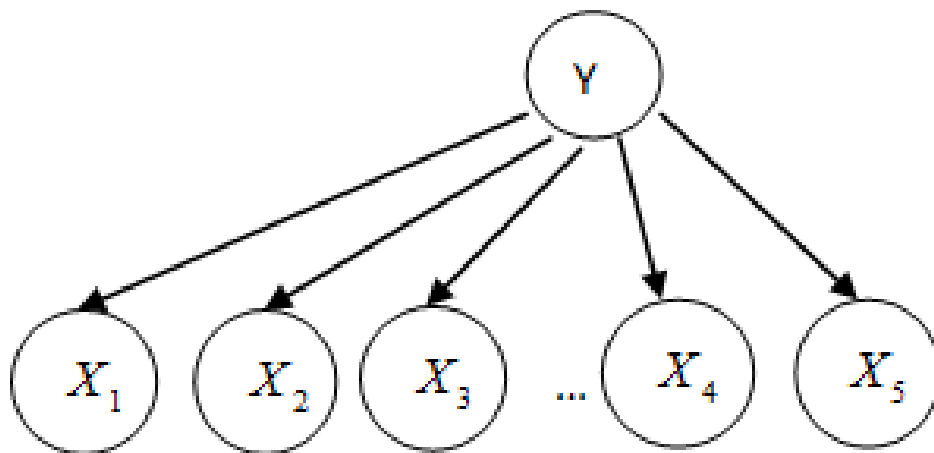
# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型表示

也可以用贝叶斯网络来表示，如图所示，其中 $y$ 是目标类， $\{X_1, X_2, \dots, X_d\}$ 是属性集。





# 贝叶斯分类算法

---

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型表示

在贝叶斯信念网中，每个结点还关联一个概率表。如果结点X没有父母结点，则表中只包含先验概率 $P(X)$ ，如果结点X只有一个父母结点Y，则表中包含条件概率 $P(X|Y)$ ，如果结点X有多个父母结点 $\{Y_1, Y_2, \dots, Y_k\}$ ，则表中包含条件概率 $P(X|Y_1, Y_2, \dots, Y_k)$ 。

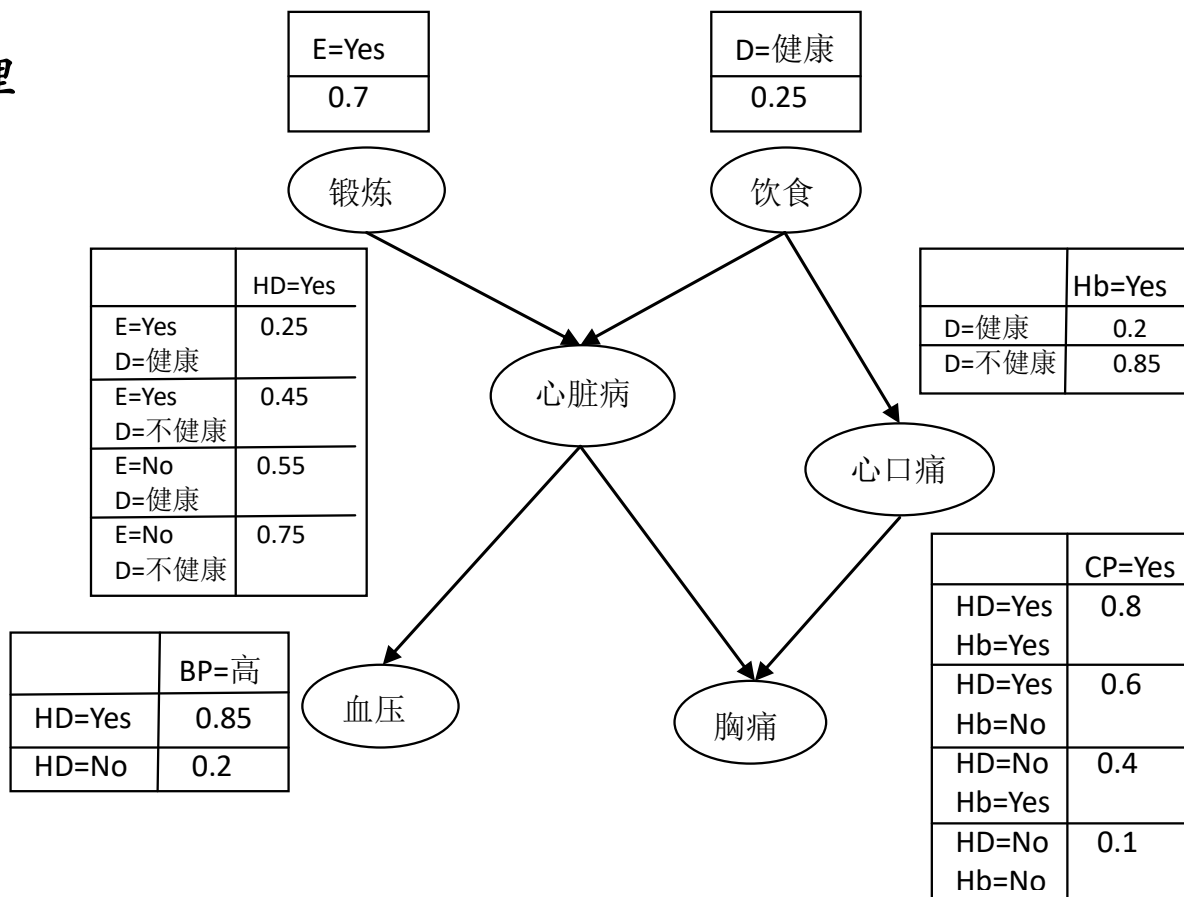
# 贝叶斯分类算法

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型表示

概率表:





# 贝叶斯分类算法

---

## 2. 贝叶斯分类算法原理

### 2.2 贝叶斯信念网

#### 1. 模型建立

贝叶斯网络的建模包括两个步骤：

- (1) 创建网络结构
- (2) 估计每一个结点在概率表中的概率值



# 贝叶斯分类算法

---

## 3. 贝叶斯算法特点及应用

### 3.1 朴素贝叶斯分类算法

#### 1. 朴素贝叶斯算法特点

优点：（1）逻辑简单、易于实现、分类过程中算法的时间空间开销比较小；  
（2）算法比较稳定、具有比较好的健壮性

缺点：有属性间类条件独立的这个假定，而很多实际问题中这个独立性假设并不成立，如果在属性间存在相关性的实际问题中忽视这一点，会导致分类效果下降。



# 贝叶斯分类算法

---

## 3. 贝叶斯算法特点及应用

### 3.1 朴素贝叶斯分类算法

#### 2. 朴素贝叶斯算法应用

- (1) 贝叶斯方法在中医证候和症状描述中的应用。
- (2) 贝叶斯方法在玉米叶部病害图像识别中的应用。



# 贝叶斯分类算法

---

## 3. 贝叶斯算法特点及应用

### 3.2 贝叶斯信念网

#### 1. 贝叶斯信念网特点

- (1) BBN提供了一种用图形模型来捕获特定领域的先验知识的方法。
- (2) 网络结构确定，添加新变量就十分容易。
- (3) 贝叶斯网络很适合处理不完整的数据。
- (4) 对模型的过分拟合问题是非常鲁棒的。





# 贝叶斯分类算法

---

## 3. 贝叶斯算法特点及应用

### 3.2 贝叶斯信念网

#### 2. 贝叶斯信念网应用

可以用于网络流量分类与识别研究