

2025-2026学年 第1学期(秋)



数据挖掘

第4章 朴素贝叶斯分类

2025 年 10 月

学习目标

- **掌握分类的概念和基本过程**
- **掌握朴素贝叶斯分类原理**



南京大學
NANJING UNIVERSITY

目录

01

分类概念及一般方法

02

朴素贝叶斯

分类概念

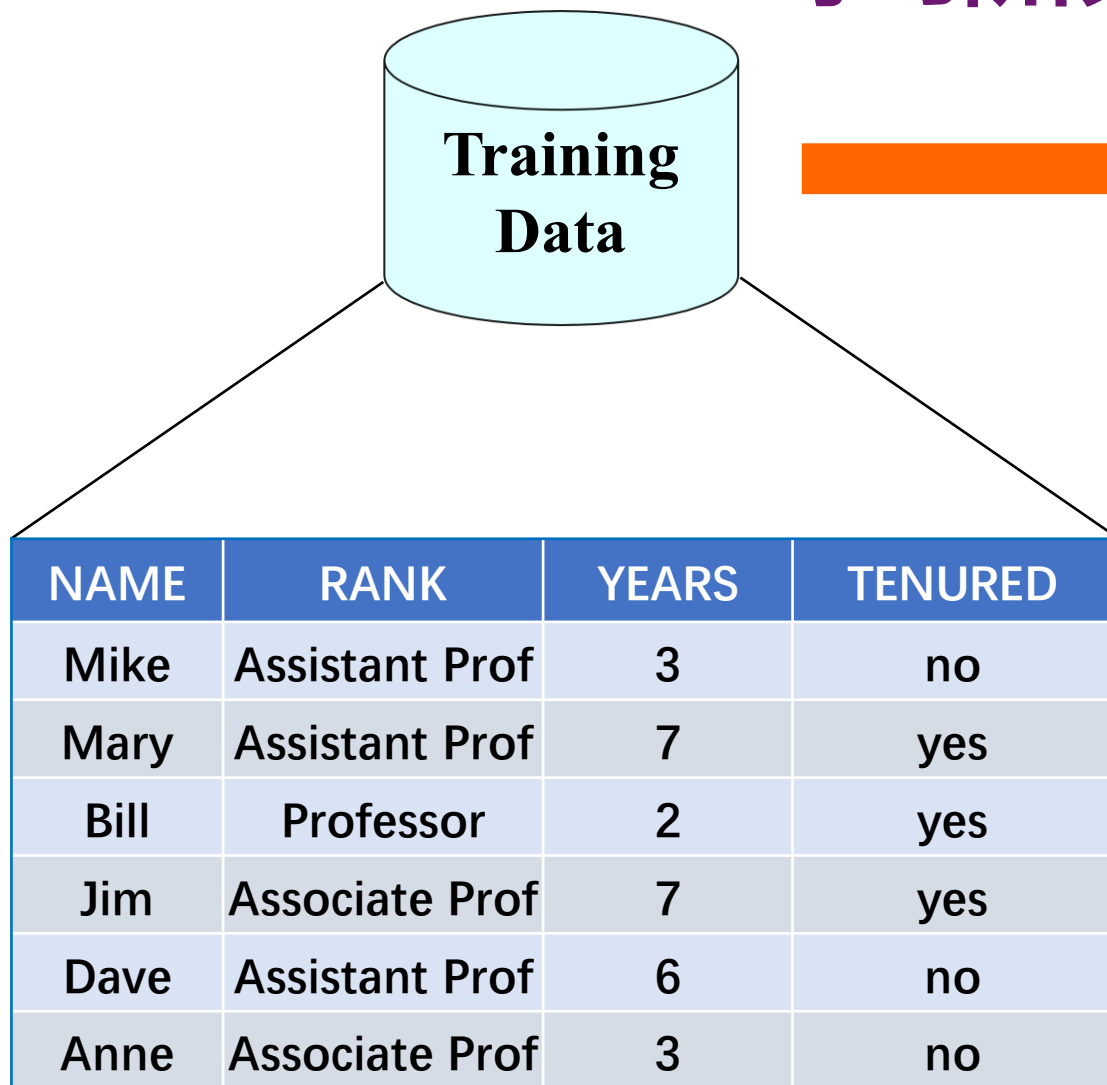
- 什么是分类?
 - 找出描述和区分数据类或概念的模型，以便能够使用模型预测类标号未知的对象的类标号
- 一般过程
 - 学习阶段
 - 建立描述预先定义的数据类或概念集的分类器
 - 训练集提供了每个训练元组的类标号，分类的学习过程也称为监督学习 (supervised learning)
 - 分类阶段
 - 使用定义好的分类器进行分类的过程

分类概念

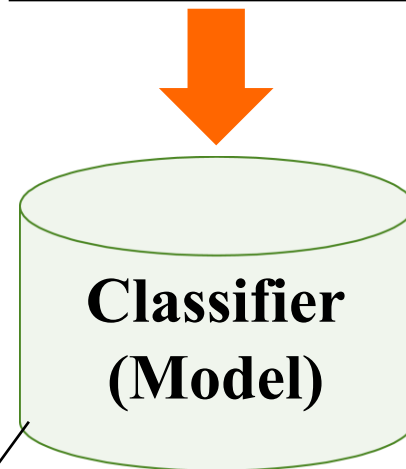
- 什么是分类?
 - 找出描述和区分数据类或概念的模型，以便能够使用模型预测类标号未知的对象的类标号
- 概念区分
 - 分类与回归
 - 分类是预测分类（离散、无序）标号
 - 回归建立连续值函数模型
 - 分类与聚类
 - 分类是有监督学习，提供了训练元组的类标号
 - 聚类是无监督学习，不依赖有类标号的训练实例

示例

学习阶段

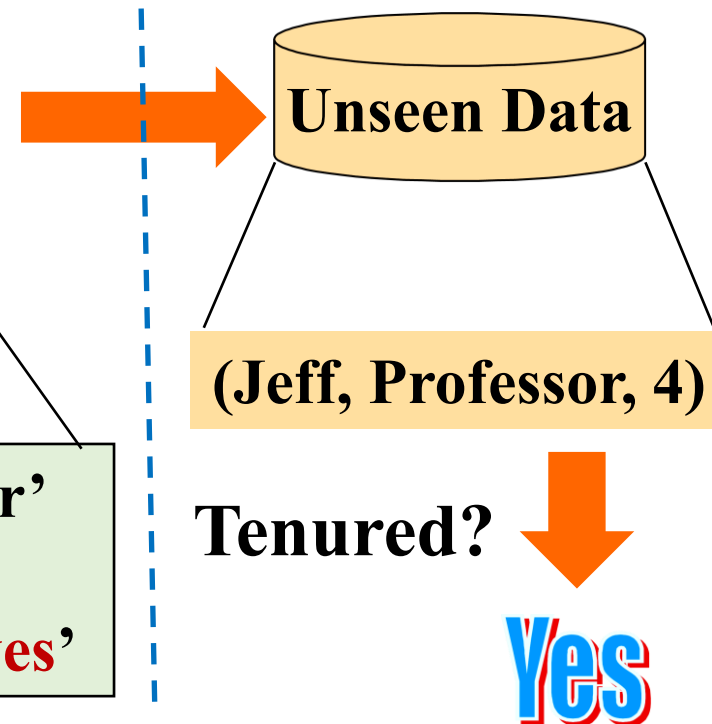


Classification Algorithms



IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

分类阶段





南京大學
NANJING UNIVERSITY

目录

01

分类概念及一般方法

02

朴素贝叶斯

朴素贝叶斯分类

- 托马斯·贝叶斯 Thomas Bayes (1701-1761)
- An essay towards solving a problem in the doctrine of chances, 1763

$$\overset{\text{后验}}{\boxed{P(h|D)}} = \frac{\overset{\text{似然}}{\boxed{P(D|h)}} \overset{\text{先验}}{\boxed{P(h)}}}{P(D)}$$



一个例子

- 描述

一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。随机选取一个穿长裤的学生，他（她）是女生的概率是多大？

一个例子

- 描述

一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。随机选取一个穿长裤的学生，他（她）是女生的概率是多大？

- 形式化

已知 $P(\text{男生}) = 60\%$, $P(\text{女生}) = 40\%$, $P(\text{长裤}|\text{女生}) = 50\%$, $P(\text{长裤}|\text{男生}) = 100\%$

求: $P(\text{女生}|\text{长裤})$

一个例子

- 形式化

已知 $P(\text{男生}) = 60\%$, $P(\text{女生}) = 40\%$, $P(\text{长裤}|\text{女生}) = 50\%$, $P(\text{长裤}|\text{男生}) = 100\%$

求: $P(\text{女生}|\text{长裤})$

- 解答

$$P(\text{女生}|\text{长裤}) = \frac{P(\text{女生})P(\text{长裤}|\text{女生})}{P(\text{男生})P(\text{长裤}|\text{男生}) + P(\text{女生})P(\text{长裤}|\text{女生})} = \frac{P(\text{女生})P(\text{长裤}|\text{女生})}{P(\text{长裤})}$$

- 直观理解

算出学校里面有多少穿长裤的，然后在这些人里面再算出有多少女生。

定义

$$P(\text{女生}|\text{长裤}) = \frac{P(\text{长裤}|\text{女生})P(\text{女生})}{P(\text{长裤})}$$

h的似然概率

h的先验概率

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

h的后验概率

D的先验概率

D: 待测试数据
h: 假设类别

问题

- 观察知识：

一所学校里面有 60% 的男生，40% 的女生。男生总是穿长裤，女生则一半穿长裤一半穿裙子。

- 不能够直接观察：

随机选取一个穿长裤的学生，你倾向于认为学生是男生还是女生？

提出假设

不能够直接观察：随机选取一个穿长裤的学生，你倾向于认为学生是男生还是女生？

- 对于不能直接观察到的部分，往往会提出假设。而对于不确定的事物，往往会有多个假设。

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

D: 待测试数据
h: 假设类别



$$P(h_1|D) = \frac{P(D|h_1)P(h_1)}{P(D)}$$

$$P(h_2|D) = \frac{P(D|h_2)P(h_2)}{P(D)}$$

$$P(h_n|D) = \frac{P(D|h_n)P(h_n)}{P(D)}$$

- 对这些假设，往往涉及两个问题：
 1. 不同假设的可能性大小？
 2. 最合理的假设是什么？

提出假设

不能够直接观察：随机选取一个穿长裤的学生，你倾向于认为学生是男生还是女生？

- 对于不能直接观察到的部分，往往会提出假设。而对于不确定的事物，往往会有多个假设。

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

D: 待测试数据
h: 假设类别



$$\begin{aligned}P(h_1|D) &= \frac{P(D|h_1)P(h_1)}{P(D)} \\P(h_2|D) &= \frac{P(D|h_2)P(h_2)}{P(D)} \\P(h_n|D) &= \frac{P(D|h_n)P(h_n)}{P(D)}\end{aligned}$$

概率分别多大？

- 对这些假设，往往涉及两个问题：
 - 不同假设的可能性大小？
 - 最合理的假设是什么？

提出假设

不能够直接观察：随机选取一个穿长裤的学生，你倾向于认为学生是男生还是女生？

- 对于不能直接观察到的部分，往往会提出假设。而对于不确定的事物，往往会有多个假设。

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

D: 待测试数据
h: 假设类别



$$\begin{aligned}P(h_1|D) &= \frac{P(D|h_1)P(h_1)}{P(D)} \\P(h_2|D) &= \frac{P(D|h_2)P(h_2)}{P(D)} \\P(h_n|D) &= \frac{P(D|h_n)P(h_n)}{P(D)}\end{aligned}$$

概率分别多大？

- 对这些假设，往往涉及两个问题：
 - 不同假设的可能性大小？
 - 最合理的假设是什么？

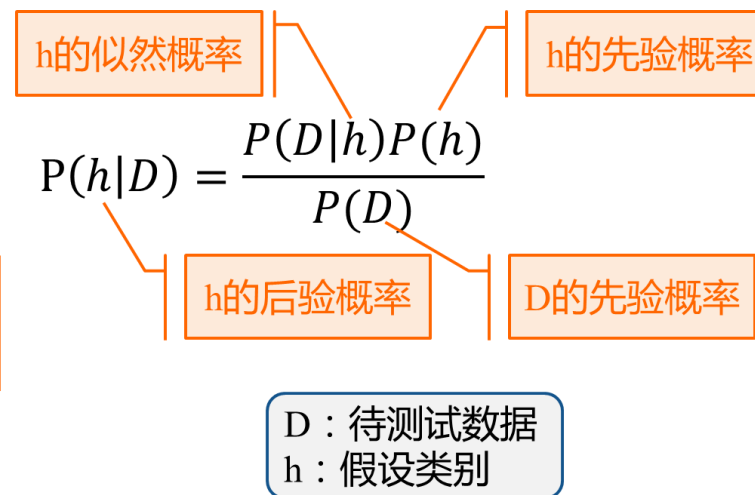
哪个概率更大，则认为
D属于哪种类别更合理

极大后验假设

- 极大后验假设定义
 - 学习器在候选假设集合H中寻找给定数据D时可能性最大的假设h，h被称为极大后验假设 (Maximum a posteriori: MAP)
 - 确定MAP的方法是用贝叶斯公式计算每个候选假设的后验概率，计算式如下

$$\begin{aligned}h_{MAP} &= \max_{h \in H} P(h|D) \\&= \max_{h \in H} P(D|h)P(h)/P(D) \\&= \max_{h \in H} P(D|h)P(h)\end{aligned}$$

$$\begin{aligned}P(h_1|D) &= \frac{P(D|h_1)P(h_1)}{P(D)} \\P(h_2|D) &= \frac{P(D|h_2)P(h_2)}{P(D)} \\P(h_n|D) &= \frac{P(D|h_n)P(h_n)}{P(D)}\end{aligned}$$



分类中的训练集与测试集

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年
爱好游戏顾客。是否会购买电脑呢？

分类中的训练集与测试集

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年爱好游戏顾客。
是否会购买电脑呢？

h的似然概率

h的先验概率

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

D：待测试数据
h：假设类别

h的后验概率

D的先验概率

D待测试数据到底是什么呢？

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h|D) \\ &= \max_{h \in H} P(D|h)P(h)/P(D) \\ &= \max_{h \in H} P(D|h)P(h) \end{aligned}$$

对象是一个多维向量

- 已知：对象D是由多个属性组成的向量

- $D = \langle a_1, a_2, \dots, a_n \rangle$

- 目标

一个收入中等、信用度良好的青年爱好游戏顾客。
是否会购买电脑呢？

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h|D) \\ &= \max_{h \in H} P(D|h)P(h)/P(D) \\ &= \max_{h \in H} P(D|h)P(h) \end{aligned}$$



$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle) \\ &= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h)P(h) \end{aligned}$$

- 问题

- 计算 $P(\langle a_1, a_2, \dots, a_n \rangle | h)$ 时，当维度过高时，可用数据变得很稀疏，难以获得结果。

独立性假设

The diagram illustrates the relationship between different probability components in a Bayesian model. On the left, the formula for the posterior probability is shown: $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$. Four orange boxes with lines pointing to the formula identify its parts: 'h的似然概率' (Likelihood of h) points to $P(D|h)$, 'h的先验概率' (Prior of h) points to $P(h)$, 'h的后验概率' (Posterior of h) points to $P(h|D)$, and 'D的先验概率' (Prior of D) points to $P(D)$. A blue box next to the formula defines 'D: 待测试数据' (Test data) and 'h: 假设类别' (Hypothesis class). A large orange arrow points from this formula to the MAP hypothesis formula on the right: $h_{MAP} = \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle) = \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h)$. A second large orange arrow points from the MAP formula down towards the '解决方法' (Solution method) section.

$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$

h的似然概率 h的先验概率

D: 待测试数据
h: 假设类别

h的后验概率 D的先验概率

$h_{MAP} = \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle)$
 $= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h)$

- 解决方法

- 假设D的属性 a_i 之间相互独立
- $P(\langle a_1, a_2, \dots, a_n \rangle | h) = \prod_i P(a_i | h)$
- $h_{MAP} = \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle)$
 $= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h)$
 $= \max_{h \in H} \prod_i P(a_i | h) P(h)$

- 优点

- 获得估计的 $P(a_i | h)$ 比 $P(\langle a_1, a_2, \dots, a_n \rangle | h)$ 容易很多
- 如果D的属性之间不满足相互独立，朴素贝叶斯分类的结果是贝叶斯分类的近似

朴素贝叶斯分类案例

训练集

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

测试集

一个收入中等、信用度良好的青年爱好游戏顾客。
是否会购买电脑呢？

h的似然概率

h的先验概率

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

D：待测试数据
h：假设类别

h的后验概率

D的先验概率

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle) \\ &= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h) \end{aligned}$$

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle) \\ &= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h) \\ &= \max_{h \in H} \prod_i P(a_i | h) P(h) \end{aligned}$$

朴素贝叶斯分类案例

一个收入中等、信用度良好的青年爱好游戏顾客。

id	年龄段	收入状况	爱好	信用度	购买电脑
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
7	中	低	是	优	是
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是

$$P(\text{青年} | \text{购买}) = 2/9 = 0.222$$

$$P(\text{收入中等} | \text{购买}) = 4/9 = 0.444$$

$$P(\text{爱好} | \text{购买}) = 6/9 = 0.667$$

$$P(\text{信用中} | \text{购买}) = 6/9 = 0.667$$

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

$$P(\mathbf{X} | \text{购买}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$$

朴素贝叶斯分类案例

一个收入中等、信用度良好的青年爱好游戏顾客。

id	年龄段	收入状况	爱好	信用度	购买电脑
1	青	高	否	中	否
2	青	高	否	优	否
6	老	低	是	优	否
8	青	中	否	中	否
14	老	中	否	优	否

$$P(\text{青年} | \text{不买}) = 3/5 = 0.6$$

$$P(\text{收入中等} | \text{不买}) = 2/5 = 0.4$$

$$P(\text{爱好} | \text{不买}) = 1/5 = 0.2$$

$$P(\text{信用中} | \text{不买}) = 2/5 = 0.4$$

$$P(\mathbf{X} | C_i) = \prod_{k=1}^n P(x_k | C_i) = P(x_1 | C_i) \times P(x_2 | C_i) \times \dots \times P(x_n | C_i)$$

$$P(\mathbf{X} | \text{不买}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$$

朴素贝叶斯分类案例

一个收入中等、信用度良好的青年爱好游戏顾客。

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	高	否	中	是
4	老	中	否	中	是
5	老	低	是	中	是
6	老	低	是	优	否
7	中	低	是	优	是
8	青	中	否	中	否
9	青	低	是	中	是
10	老	中	是	中	是
11	青	中	是	优	是
12	中	中	否	优	是
13	中	高	是	中	是
14	老	中	否	优	否

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h | \langle a_1, a_2, \dots, a_n \rangle) \\ &= \max_{h \in H} P(\langle a_1, a_2, \dots, a_n \rangle | h) P(h) \\ &= \max_{h \in H} \prod_i P(a_i | h) P(h) \end{aligned}$$

$$P(\mathbf{X} | C_i) P(C_i)$$

$$P(C_{\text{购买}}) = 9/14 = 0.643$$

$$P(C_{\text{不买}}) = 5/14 = 0.357$$

$$P(\text{购买} | \mathbf{X}) = 0.044 \times 0.643 = 0.028$$

$$P(\text{不买} | \mathbf{X}) = 0.019 \times 0.357 = 0.007$$

一个例子

● 问题

- 给定一封邮件，判定它是否属于垃圾邮件。按照先例，用 D 来表示邮件（注意 D 由 n 个单词的属性合取 $\langle a_1, a_2, \dots, a_n \rangle$ 组成）。用 $h+$ 来表示垃圾邮件， $h-$ 表示正常邮件，即目标空间 $H = \langle h+, h- \rangle$ 。

● 形式化描述：

- $P(h+|D) = P(h+) * P(D|h+)/P(D)$
- $P(h-|D) = P(h-) * P(D|h-)/P(D)$



一个例子

- 求解 $P(h + | D) = P(h +) * P(D|h+)/P(D)$
- $P(h +)$
 - 即计算已有训练集中垃圾邮件的比例



一个例子

- 求解 $P(h+|D) = P(h+) * P(D|h+)/P(D)$
- $P(h+)$
 - 即计算已有训练集中垃圾邮件的比例
- $P(D|h+) = P(< a_1, a_2, \dots, a_n > |h+)$
 - 即计算垃圾邮件中完全包含 a_1, a_2, \dots, a_n 这 n 个单词的邮件比例。当 n 很大时，这几乎不可能。
 - 利用朴素贝叶斯 $P(< a_1, a_2, \dots, a_n > |h+) = \prod_i P(a_i|h+)$ ，对于每个 $P(a_i|h+)$ ，就是要求解单词 a_i 在垃圾邮件训练集中出现的频率。



一个例子

- 求解 $P(h+|D) = P(h+) * P(D|h+)/P(D)$
- $P(h+)$
 - 即计算已有训练集中垃圾邮件的比例
- $P(D|h+) = P(< a_1, a_2, \dots, a_n > |h+)$
 - 即计算垃圾邮件中完全包含 a_1, a_2, \dots, a_n 这 n 个单词的邮件比例。当 n 很大时，这几乎不可能。
 - 利用朴素贝叶斯 $P(< a_1, a_2, \dots, a_n > |h+) = \prod_i P(a_i|h+)$ ，对于每个 $P(a_i|h+)$ ，就是要求解单词 a_i 在垃圾邮件训练集中出现的频率。
- $P(D)$ 即单词 a_1, a_2, \dots, a_n 同时出现在一封邮件中的概率，可假设为常量。



一个例子

- 求解 $P(h+|D) = P(h+) * P(D|h+)/P(D)$
- $P(h+)$
 - 即计算已有训练集中垃圾邮件的比例
- $P(D|h+) = P(< a_1, a_2, \dots, a_n > |h+)$
 - 即计算垃圾邮件中完全包含 a_1, a_2, \dots, a_n 这n个单词的邮件比例。当n很大时，这几乎不可能。
 - 利用朴素贝叶斯 $P(< a_1, a_2, \dots, a_n > |h+) = \prod_i P(a_i|h+)$ ，对于每个 $P(a_i|h+)$ ，就是要求解单词 a_i 在垃圾邮件训练集中出现的频率。
- $P(D)$ 即单词 a_1, a_2, \dots, a_n 同时出现在一封邮件中的概率，可假设为常量。
- 同理求解 $P(h-|D) = P(h-) * P(D|h-)/P(D)$
- 比较 $P(h+|D)$ 和 $P(h-|D)$ 的大小



一个例子

$$P(h + |D) = P(h +) * P(D|h+)/P(D)$$

- 已知

- 训练集中垃圾邮件的比例为 $P(h +) = 0.2$
- 训练集中正常邮件的比例为 $P(h -) = 0.8$
- 单词出现频率表

分词	在垃圾邮件中出现的比例	在正常邮件中出现的比例
免费	0.3	0.01
奖励	0.2	0.01
网站	0.2	0.2

- 求解

- 判断一封邮件 $D = \langle \text{“免费”, “奖励”, “网站”} \rangle$ 是否是垃圾邮件

一个例子

$$P(h + |D) = P(h +) * P(D|h+)/P(D)$$

$$\begin{aligned} P(h + |D) &= P(h +) * \frac{P(D|h+)}{P(D)} \\ &= 0.2 * \frac{(0.3*0.2*0.2)}{P(D)} = 0.0096/P(D) \\ P(h - |D) &= P(h -) * \frac{P(D|h-)}{P(D)} \\ &= 0.8 * \frac{(0.01*0.01*0.2)}{P(D)} = 0.000016/P(D) \end{aligned}$$

$$P(h + |D) > P(h - |D)$$

朴素贝叶斯分类 —— 连续数据如何求概率

id	年龄	收入	爱好	信用	购买
1	青	高	否	中	否
2	青	高	否	优	否
3	中	低	否	中	否
4	老	高	否	中	否
5	老	中	是	中	是
6	老	低	是	优	否
7	中	高	是	优	否
8	青	中	否	中	是
9	青	低	是	中	否
10	老	中	是	中	是

id	年龄	收入	爱好	信用	购买
1	青	125	否	中	否
2	青	100	否	优	否
3	中	70	否	中	否
4	老	120	否	中	否
5	老	95	是	中	是
6	老	60	是	优	否
7	中	220	是	优	否
8	青	85	否	中	是
9	青	75	是	中	否
10	老	90	是	中	是

预测收入为121，无游戏爱好、信用良好的中年人，是否购买

朴素贝叶斯分类 —— 连续数据如何求概率

id	收入	购买
1	125	否
2	100	否
3	70	否
4	120	否
5	95	是
6	60	否
7	220	否
8	85	是
9	75	否
10	90	是

假设不同类别收入分别服从不同正态分布

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

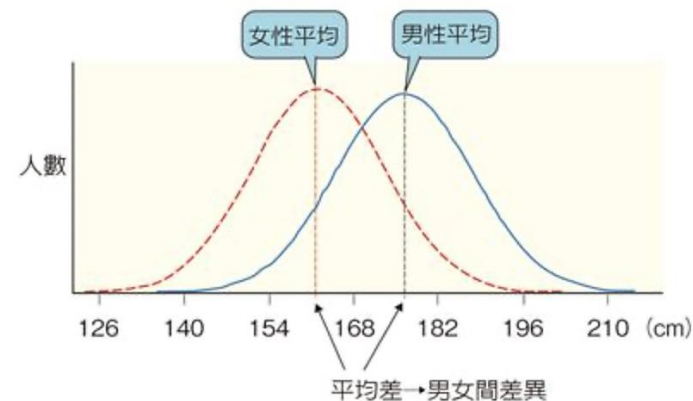
利用参数估计两组正态分布期望和方差

朴素贝叶斯分类 —— 连续数据如何求概率

id	收入	购买
1	125	否
2	100	否
3	70	否
4	120	否
5	95	是
6	60	否
7	220	否
8	85	是
9	75	否
10	90	是

假设不同类别收入分别服从不同正态分布

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



利用参数估计两组正态分布期望和方差

$$P(\text{收入} = 121 | No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(121-110)^2}{2(2975)}}$$
$$= 0.0072$$

贝叶斯分类器总结

- 本质上是同时考虑了先验概率和似然概率的重要性
- 特点
 - 属性可以离散、也可以连续
 - 数学基础坚实、分类效率稳定
 - 对缺失和噪声数据不太敏感
 - 属性如果不相关，分类效果很好

参考文献

- **数学之美番外篇：平凡而又神奇的贝叶斯方法. 网络文章.**
- **贝叶斯学派与频率学派有何不同？**

<http://www.zhihu.com/question/20587681/answer/16023547>

贝叶斯分类编程实践

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html#sklearn.naive_bayes.GaussianNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html#sklearn.naive_bayes.MultinomialNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html#sklearn.naive_bayes.ComplementNB

https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB

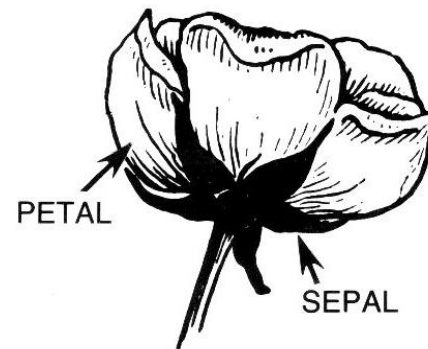
https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes.BernoulliNB

同学们可以尝试利用python读入本地iris数据集，来完成贝叶斯分类，分析其分类效果

贝叶斯分类编程实践

- Iris数据集中包含了: 山鸢尾(Setosa), 杂色鸢尾(Versicolour), 维吉尼亚鸢尾(Virginica), 每种50个数据, 共含150个数据。
- 在每个数据包含四个属性:
 - ✓ 花萼长度(sepal length in cm),
 - ✓ 花萼宽度(sepal width in cm),
 - ✓ 花瓣长度(petal length in cm),
 - ✓ 花瓣宽度(petal width in cm)
- 可通过这四个属性预测鸢尾花卉属于 (山鸢尾, 杂色鸢尾, 维吉尼亚鸢尾) 哪一类。

数据集中部分数据如右图所示:



IRIS (sepal:萼片, petal:花瓣)

```
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica
5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
```

参考资料: <https://www.kaggle.com/datasets/uciml/iris>

贝叶斯分类编程实践

具体要求

- 建议使用scikit-learn中的load_iris函数载入数据集，如使用本地数据集，文件路径设置为“./iris.csv”（即当前代码文件路径下）。
- 对数据进行预处理，如缺失值处理、数据标准化等。
- 利用Python实现朴素贝叶斯分类器。推荐使用scikit-learn库中的相关函数，但鼓励自己编写核心逻辑以加深理解。
- 将数据集分为训练集和测试集，比例自定，至少保持70%的数据用于训练。
- 训练模型并对测试集进行预测。
- 计算模型的准确率（需不低于95%）等评价指标。
- 可选：尝试调整参数或采用不同的特征选择方法来提高模型性能。

贝叶斯分类编程实践

上交形式

- 提交完整的代码文件（使用Jupyter Notebook格式）。
- 文件命名为 “**学号-姓名-准确率.ipynb**” （如602024710021-张三-97.ipynb）。

注意事项

- 确保代码具有良好的可读性，适当添加注释说明。
- 遵守学术诚信原则，严禁抄袭他人作品。