

分类:NB-丁兆云

分类:NB-丁兆云

学习目标

- ◆ 描述分类的一般过程
- ◆ 掌握朴素贝叶斯分类原理

主要内容

- ◆ 1. 分类概念及一般方法
- ◆ 2. 朴素贝叶斯

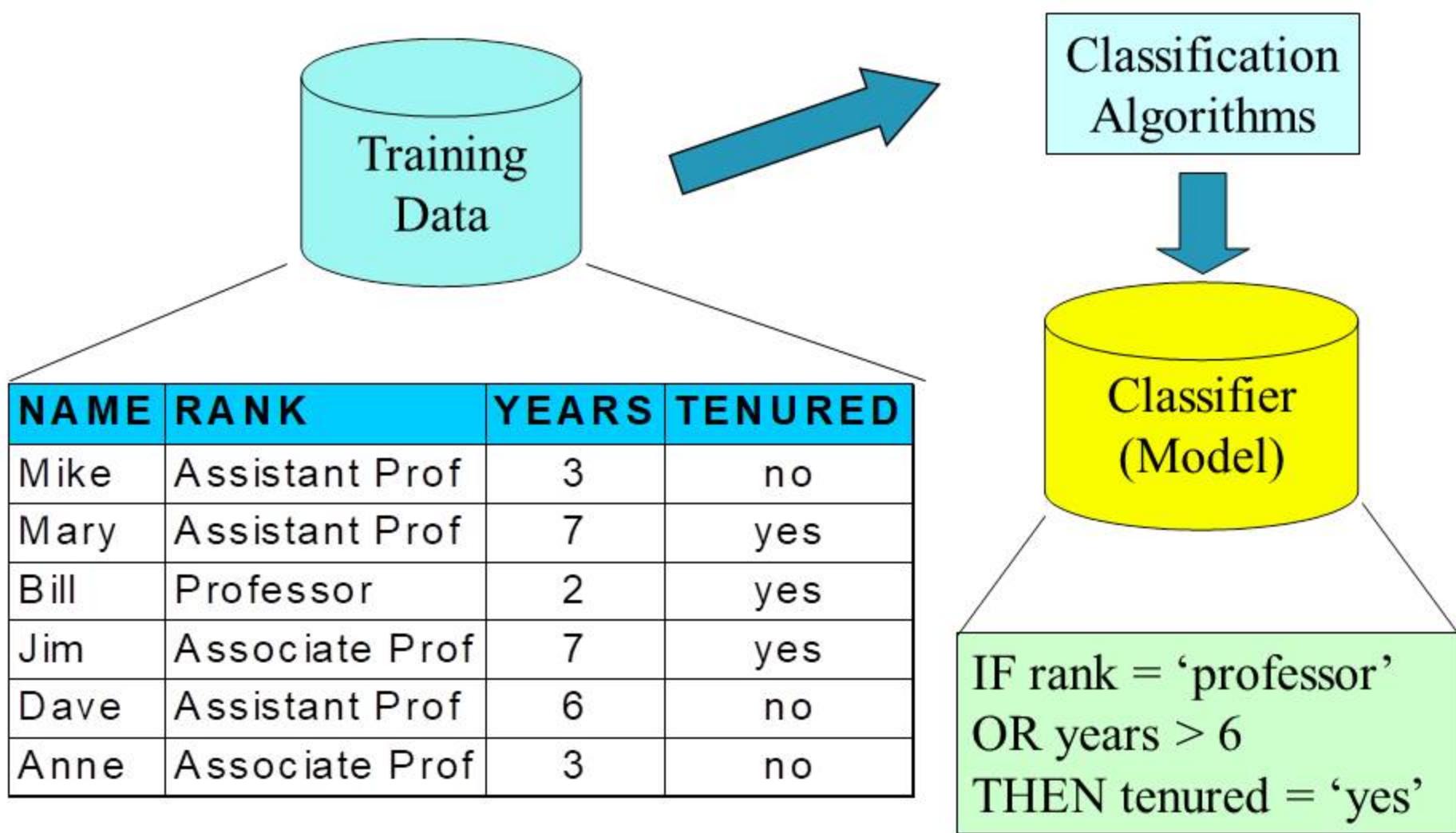
1. 分类概念

- ◆ 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

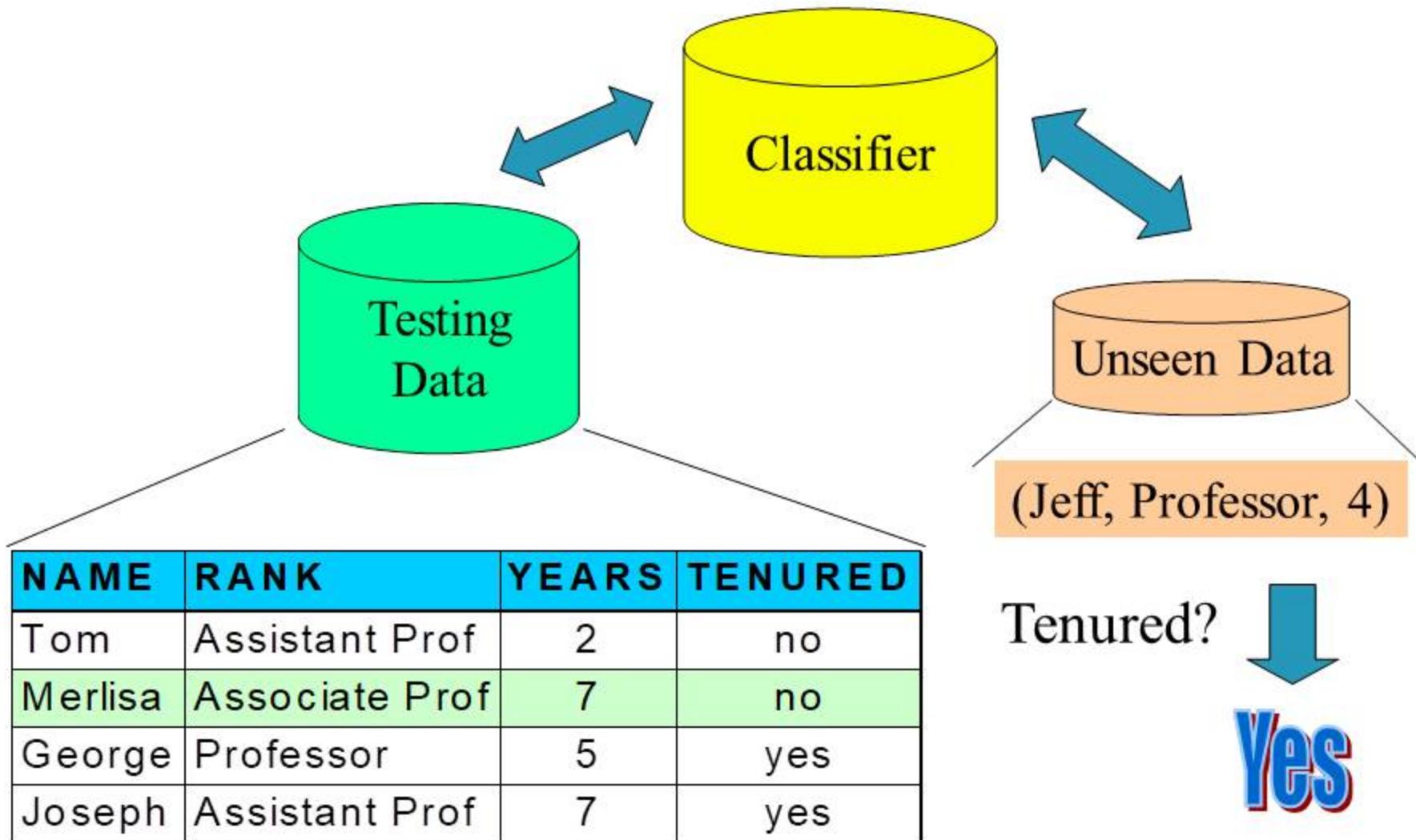
1. 分类概念

- ◆ 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
 - 分类与预测
 - 分类是预测分类（离散、无序）标号；
 - 预测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



示例：分类阶段



2. 朴素贝叶斯分类

■ 介绍

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



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

一个例子

- ◆ 描述

- 一所学校里面有 60% 的男生(boy), 40% 的女生(girl)。男生总是穿长裤(pants), 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?

◆ 描述

- 一所学校里面有 60% 的男生(boy), 40% 的女生(girl)。男生总是穿长裤(pants), 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 $P(\text{Boy}) = \text{[填空1]}$, $P(\text{Girl}) = \text{[填空2]}$, $P(\text{Pants} | \text{Girl}) = \text{[填空3]}$,
 $P(\text{Pants} | \text{Boy}) = \text{[填空4]}$
- 求: $P(\text{Girl} | \text{Pants})$

正常使用填空题需3.0以上版本雨课堂

作答

一个例子

- ◆ 描述
 - 一所学校里面有 60% 的男生(boy), 40% 的女生(girl)。男生总是穿长裤(pants), 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)是女生的概率有多大?
- ◆ 形式化
 - 已知 $P(\text{Boy})=60\%$, $P(\text{Girl})=40\%$, $P(\text{Pants} \mid \text{Girl})=50\%$, $P(\text{Pants} \mid \text{Boy})=100\%$
 - 求: $P(\text{Girl} \mid \text{Pants})$

一个例子

- ◆ 描述
 - 一所学校里面有 60% 的男生(boy), 40% 的女生(girl)。男生总是穿长裤(pants), 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 $P(\text{Boy})=60\%$, $P(\text{Girl})=40\%$, $P(\text{Pants} \mid \text{Girl})=50\%$, $P(\text{Pants} \mid \text{Boy})=100\%$
 - 求: $P(\text{Girl} \mid \text{Pants})$
- ◆ 解答
 - $$P(\text{Girl} \mid \text{Pants}) = \frac{P(\text{Girl})P(\text{Pants} \mid \text{Girl})}{P(\text{Boy})P(\text{Pants} \mid \text{Boy}) + P(\text{Girl})P(\text{Pants} \mid \text{Girl})} = \frac{P(\text{Girl})P(\text{Pants} \mid \text{Girl})}{P(\text{Pants})}$$
- ◆ 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。

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？

2.2 定义

$$P(Girl|Pants) = \frac{P(Pants|Girl)P(Girl)}{P(Pants)}$$

D: 待测试数据
h: 假设类别

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

问题

- ◆ 观察知识：一所学校里面有 60% 的男生(boy), 40% 的女生(girl)。男生总是穿长裤(pants)，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，往往会有多个假设。而对于不确定的事物，往往会有多个假设。

D: 待测试数据

h: 假设类别

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



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

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

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

- 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

- 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。

D: 待测试数据

h: 假设类别

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



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

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

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

概率分别多大？

- 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

- 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。

D: 待测试数据

h: 假设类别

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



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

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

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

概率分别多大？

- 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？

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

$$h_{MAP} = \max_{h \in H} P(h|D)$$

$$= \max_{h \in H} P(D|h)P(h)/P(D)$$

$$= \max_{h \in H} P(D|h)P(h)$$

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

D: 待测试数据
h: 假设类别

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

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？

D: 待测试数据
h: 假设类别

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

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

D待测试数据到底是什么呢？

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？

D: 待测试数据
h: 假设类别

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

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

D待测试数据到底是什么呢？

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
 - $D = \langle a_1, a_2, \dots, a_n \rangle$ 一个收入中等、信用度良好的青年爱好游戏顾客。
- ◆ 目标
$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h|D) \\ &= \max_{h \in H} P(D|h)P(h)/P(D) \\ &= \max_{h \in H} P(D|h)P(h) \end{aligned}$$

↓

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

D: 待测试数据
h: 假设类别

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

- 解决方法

- 假设D的属性 a_i 之间相互独立

- $P(< a_1, a_2, \dots, a_n > | h) = \prod_i P(a_i | h)$

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

- $= \max_{h \in H} P(< a_1, a_2, \dots, a_n > | h)P(h)$

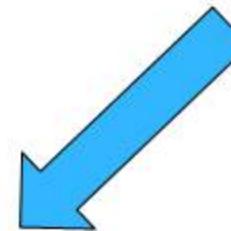
- $= \max_{h \in H} \prod_i P(a_i | h) P(h)$

- 优点

- 获得估计的 $P(a_i | h)$ 比 $P(< a_1, a_2, \dots, a_n > | h)$ 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

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



2.3 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？

D: 待测试数据
h: 假设类别

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

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

$$\begin{aligned} h_{MAP} &= \max_{h \in H} P(h | a_1, a_2, \dots, a_n) \\ &= \max_{h \in H} P(a_1, a_2, \dots, a_n | h) P(h) \\ &= \max_{h \in H} \prod_{i=1}^n P(a_i | h) P(h) \end{aligned}$$



一个收入中等、信用度良好的青年爱好游戏顾客。(答案保留小数点后三位)

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

$$P(\text{青年} \mid \text{购买}) = [\text{填空1}]$$

$$P(\text{收入中等} \mid \text{购买}) = [\text{填空2}]$$

$$P(\text{爱好} \mid \text{购买}) = [\text{填空3}]$$

$$P(\text{信用中} \mid \text{购买}) = [\text{填空4}]$$

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

正常使用填空题需3.0以上版本雨课堂

$$P(\mathbf{X} \mid \text{购买}) = [\text{填空5}]$$

作答

2.3 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

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

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

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

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

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

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



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

$$P(\text{青年} \mid \text{不买}) = [\text{填空1}]$$

$$P(\text{收入中等} \mid \text{不买}) = [\text{填空2}]$$

$$P(\text{爱好} \mid \text{不买}) = [\text{填空3}]$$

$$P(\text{信用中} \mid \text{不买}) = [\text{填空4}]$$

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

$$P(\mathbf{X} \mid \text{不买}) = [\text{填空5}]$$

正常使用填空题需3.0以上版本雨课堂

作答

2.3 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	P(青年 不买) = 3/5 = 0.6
1	青	高	否	中	否	P(收入中等 不买) = 2/5 = 0.4
2	青	高	否	优	否	P(爱好 不买) = 1/5 = 0.2
6	老	低	是	优	否	P(信用中 不买) = 2/5 = 0.4
8	青	中	否	中	否	
14	老	中	否	优	否	

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

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

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	正常使	是
13	中	高	是	中	是
14	老	中	否	优	否

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

$$= \max_{h \in H} P(< a_1, a_2, \dots, a_n > | h) P(h)$$

$$= \max_{h \in H} \prod_i P(a_i | h) P(h)$$

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

$$P(C_{\text{买}}) = [\text{填空1}]$$

$$P(C_{\text{不买}}) = [\text{填空2}]$$

$$P(\text{购买} | \mathbf{X}) = [\text{填空3}]$$

$$P(\text{不买} | \mathbf{X}) = [\text{填空4}]$$

作答

3.0以上版本雨课堂

2.3 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

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

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

$$P(C_{\text{买}}) = 9/14 = 0.643$$

$$P(C_{\text{不买}}) = 5/14 = 0.357$$

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

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

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



一个例子

- 求解 $P(h+ | D) = P(h+) * P(D|h+)/P(D)$
 - $P(h+)$
 - 即计算已有训练集合中垃圾邮件的比例
 - $P(D|h+) = P(< a_1, a_2, \dots, a_n > | h+)$
 - 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【思考：为什么？】。
 - 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 $P(h+ | D) = P(h+) * P(D|h+)/P(D)$
 - $P(h+)$
 - 即计算已有训练集合中垃圾邮件的比例。
 - $P(D|h+) = P(< a_1, a_2, \dots, a_n > | h+)$
 - 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 在垃圾邮件训练集合中出现的频率。
 - $P(D)$ 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 $P(h+|D) = P(h+) * P(D|h+)/P(D)$
 - $P(h+)$
 - 即计算已有训练集合中垃圾邮件的比例
 - $P(D|h+) = P(< a_1, a_2, \dots, a_n > |h+)$
 - 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

- ◆ 已知 ■ $P(h+|D) = P(h+) * P(D|h+)/P(D)$
 - 训练集合中垃圾邮件的比例为 $P(h+) = 0.2$
 - 训练集合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 $D = \langle \text{“免费”}, \text{“奖励”}, \text{“网站”} \rangle$ 是否是垃圾邮件

◆ 已知

- 训练集合中垃圾邮件的比例为 $P(h+) = 0.2$
- 训练集合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

$$P(h+|D) = P(h+) * \frac{P(D|h+)}{P(D)} = [\text{填空1}] \quad \text{假设 } p(D)=1$$

正常使用填空题需3.0以上版本雨课堂

$$P(h-|D) = P(h-) * \frac{P(D|h-)}{P(D)} = [\text{填空2}]$$

作答

一个例子

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

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

2.4 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

2.4 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

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

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

利用参数估计两组正态分布期望和方差

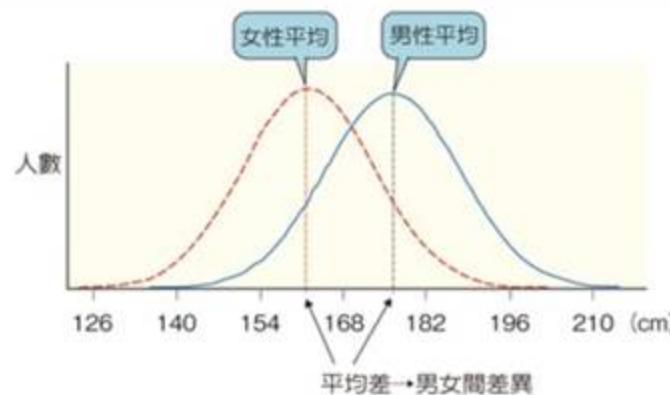
2.4 朴素贝叶斯分类-连续数据如何求概率

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

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

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



利用参数估计两组正态分布期望和方差

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

$$= 0.0072$$

2.5 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

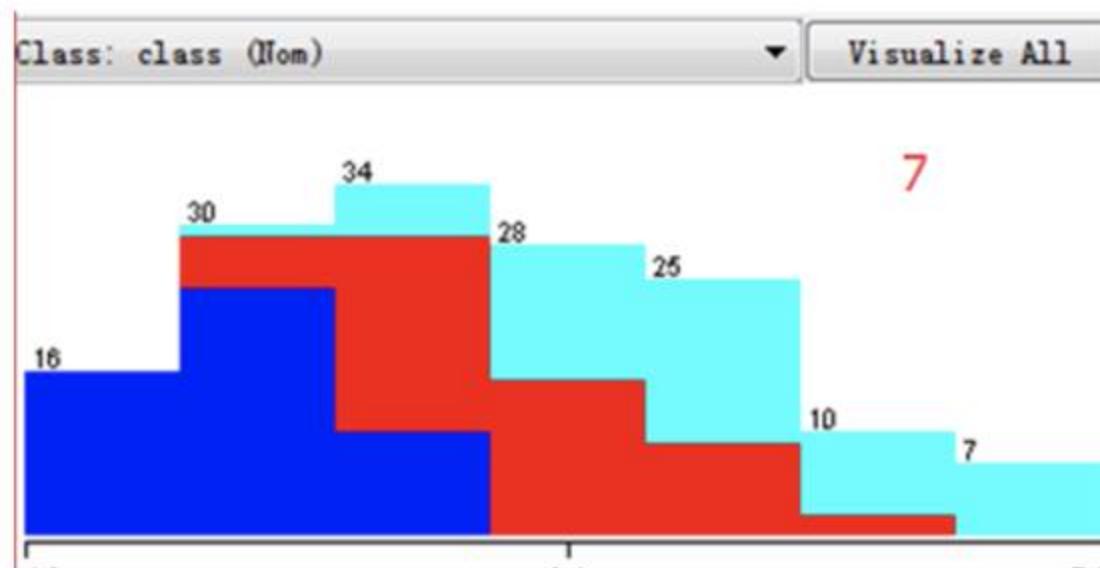
Iris数据集中每个属性在每个类别上的分布如下，
请同学们预估下，贝叶斯分类器是否适合iris数据集

A

是

B

否



提交

2.6参考文献

- ◆ 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

第8次课后作业

- ◆ 第八次课后作业-在educoder平台上完成作业
- ◆ <https://www.educoder.net/shixuns/uyl5pk2q/challenges>
- ◆ <https://www.educoder.net/shixuns/fg8nkf9y/challenges>

提交作业截至时间：2020年3月13日

◆ 问题？

历史最佳排名：10/1060 (8分)

历史最佳排名：10/1060 (8分)

www.pkbigdata.com/common/cmpt/汽车目的地智能预测大赛_排行榜.html

参赛队伍：958
参赛人数：1060
作品提交数：2003

排行榜

排名	排名变化	团队logo	队名	最高得分	提交次数	最后提交时间
10	-		2018数据挖掘丁兆云	0.42730	1	2018-11-
25	-		小虎牛牛无奇	0.38681	25	2018-11-

关注微信公众号

TOP

9:00
2018/11/10



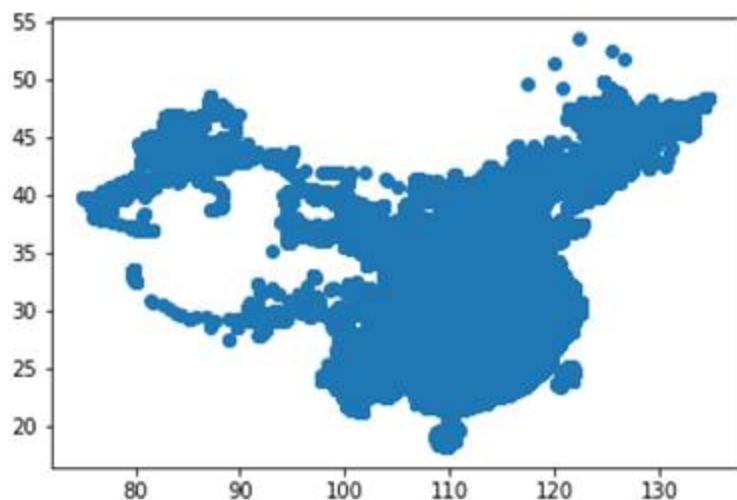
训练集 (1.5亿条数据) (6000辆车8个月数据)	测试集 (5.8万条数据) (最后一个月部分数据)
记录编号	记录编号
汽车编号	汽车编号
出发时间 (年月日时分秒)	出发时间 (年月日时分秒)
出发经度	出发经度
出发维度	出发维度
到达时间	
到达经度	预测值
到达维度	预测值

数据观察

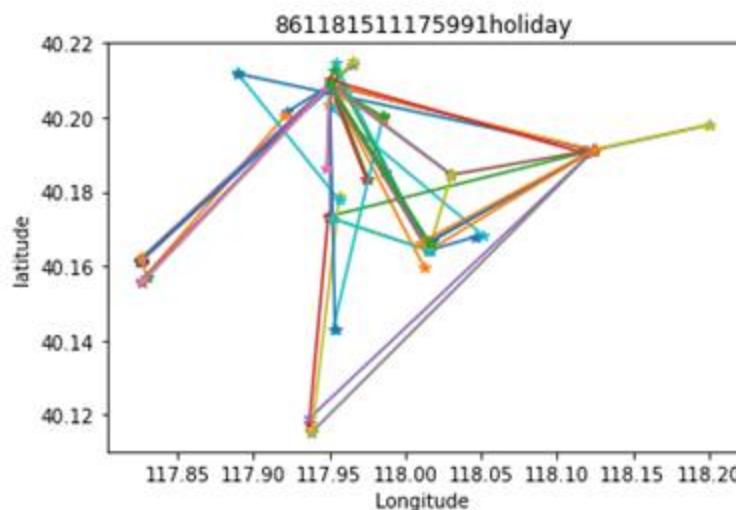
1.5亿条训练数据，5.8万条测试数据

5033辆车需要进行目的地预测

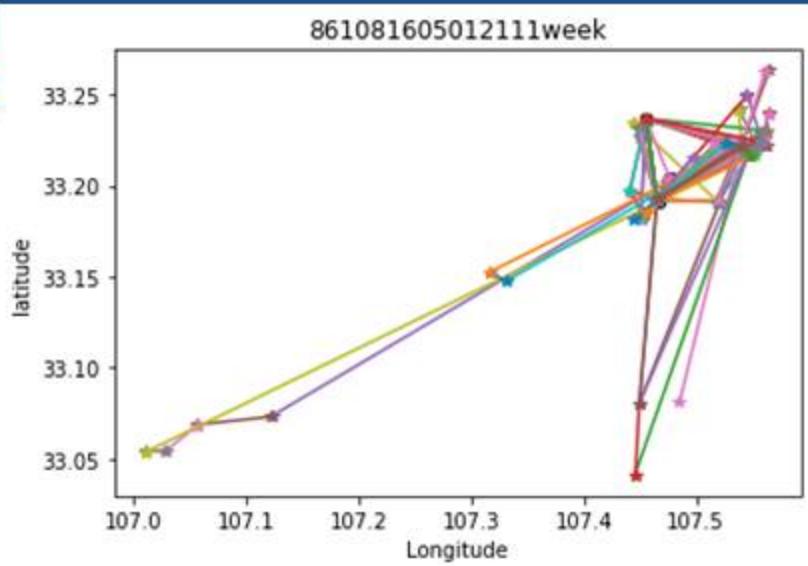
对于每一辆车，训练数据100-500条不等，测试数据1-20条不等



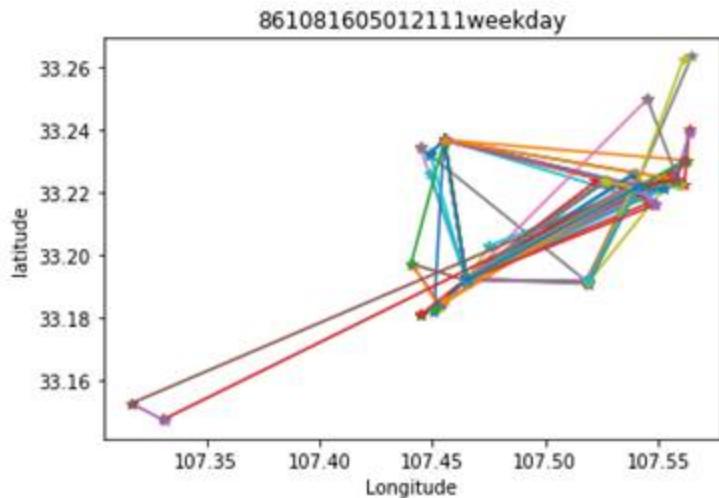
6000多辆车轨迹遍布全国



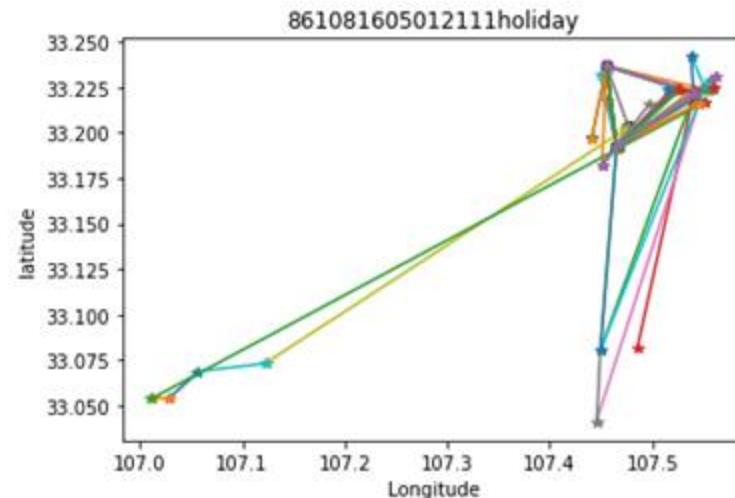
单辆车轨迹有规律



一辆车的日常



一辆车的工作日



一辆车的节假日

假设

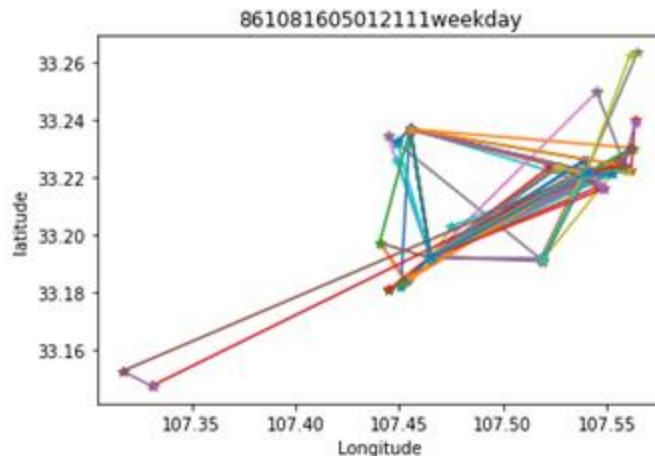
研究表明，人类93%的行为都是可以预测的

假设同一个用户在同样的时间段同一出发点会更倾向于去同一个地方。

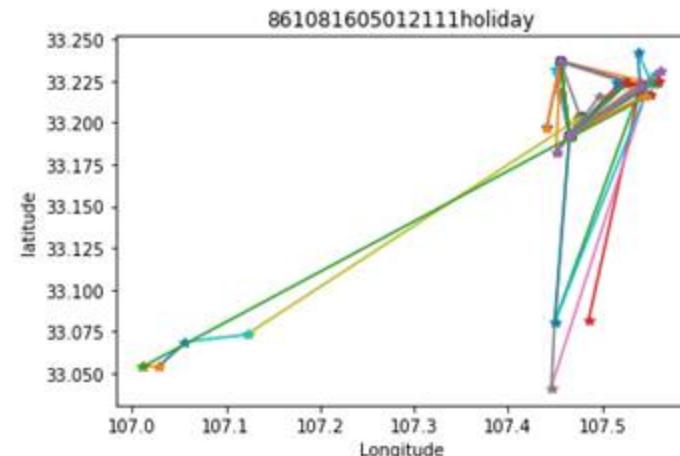
工作日：家和公司之间

周末：会有几个常去的地方商场

除了紧急出差、旅游、生病之类的没有办法预测，因此允许误差



一辆车的工作日



一辆车的节假日

模型——官网baseline

找到每个星期几用户最常去的七个地方作为备选目的地，预测去这七个地方的概率（训练时的Y为用户是否去了这七个地方）

出发经度

基于历史数据，得到预测的模型

出发维度

星期几



到七个最常去地方的概率

与七个最常去地方的距离

七个最常去地方的频数

特征工程

出发时间

年/月/日 时/分/秒



- 星期几
- 是否为节假日(周末和法定节假日如端午十一均属于节假日)
- 分成8个时间段 (0:00-7:00, 7:00-9:00, 9:00-11:00, 11:00-13:00, 13:00-15:00, 15:00-17:00, 17:00-19:00, 19:00-22:00, 12:00-24:00) **不均分**

特征工程

出发地点

地理位置的经纬度

目的地



离散化: 进行**密度聚类**, 按其所属的类别进行编号
Geohash对地理位置进行分区编码

模型

出发时间
(年月日 时分秒)

出发地点
(经纬度)

基于历史数据,
预测

目的地
(经纬度)

连续数据的
回归问题

用户ID

出发地点ID

星期几

是否节假日

时段

基于历史数据,
预测

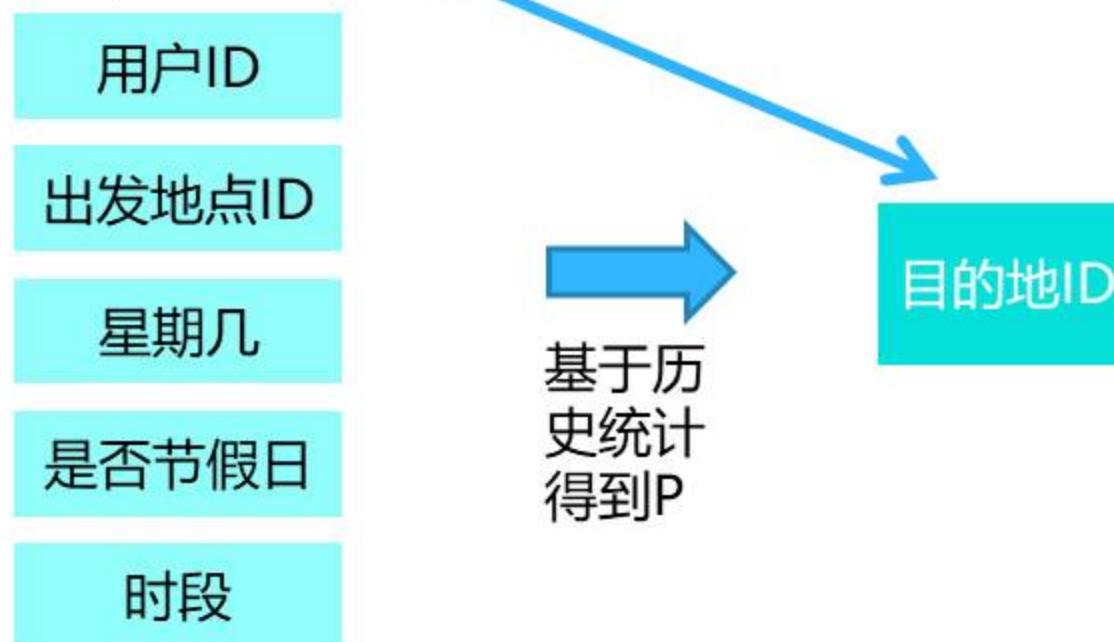
目的地ID

离散数据的
分类问题

模型——朴素贝叶斯分类器

25号车9月20日 (星期四) 7:30 (时段编号1) 从 (区域编号为10)
的地方出发的目的地——条件概率

$$p(c|X) = \frac{P(c)p(c|X)}{P(X)} = \frac{P(c)}{P(X)} \prod_{i=1}^d p(x_i|c)$$



模型——朴素贝叶斯分类器

25号车9月20日 (星期四) 7:30 (时段编号1) 从 (区域编号为10)
的地方出发的目的地——条件概率

$$p(c|X) = \frac{P(c)p(c|X)}{P(X)} = \frac{P(c)}{P(X)} \prod_{i=1}^d p(x_i|c)$$

用户ID

P (ID=101025 |目的地=3)

出发地点ID

P (出发地=10|目的地=3)

P (目的地=3)

星期几

P (周四|目的地=3)

$P(X)$ 做归一化

是否节假日

P (非节假日|目的地=3)

$p(c = 3|X)$

时段

P (时段=1|目的地=3)

同理推广大各个可能的目的地

题目来源&内容

DataCastle平台

“神策杯” 2018高校算法大师赛

神策杯
2018高校算法大师赛

MacBook Pro等你来拿，玩法 aPKi 等你来拿

赛题背景：

神策数据推荐系统是基于神策分析平台的智能推荐系统。本次竞赛是模拟业务场景，以新闻文本的核心词提取为目的，最终结果达到提升推荐和用户画像的效果。

赛题内容：

以已标注关键词的1000篇文档为训练集，训练出一个“关键词提取”的模型，来提取10万篇文档的关键词。

评分原则：

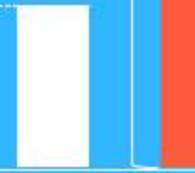
选手上交对10万篇文档的标注结果，每篇文档标注两个关键词，官方在10万篇选取1000篇作为评分依据，每篇命中一个关键词记0.5分，命中两个记1分。

训练集



1000条

所有文档集



10万条

- all_docs.txt, 108295篇资讯文章数据，数据格式为：ID 文章标题 文章正文，中间由\001分割。
- train_docs_keywords.txt, 1000篇文章的关键词标注结果，数据格式为：ID 关键词列表，中间由\t分割。
- 所标注的文档和评分文档关键词数量大于1小于5。

all_docs.txt		
ID	标题	正文
D083417	LOL: faker和恩静的前世今生恩静要结婚了，那飞科变捞的原因？	近来李哥仿佛又开始替补了，的确今年锻练的锅真的很大，算了不说了，我们聊点开心的。Faker和恩静…
D026238	可爱担当吴芊盈甜美笑容感染全场蓄力绽放非凡魅力惹人爱	近日，SDT娱乐练习生吴芊盈在新一期的《创造101》中，表现优异展现了不凡实力，成功晋级…
D066225	生一个孩子和生两个孩子有哪些区别？	虽然二胎开放了，但是有些家庭却坚持一个好，而有些家庭政策积极响应，生了二胎。一胎家庭和…
D000212	复仇者联盟3：无限战争结局，如何影响漫威影集神盾局特工	《甄嬛传》想必很多人已经二刷三刷，剧中5位小主的命运差异好大，剧…
D011909	【NCT127成员介绍】谁是认证的拥有克	13日，TOWER_官方推特公开秦容相关宣传

ID	label
D083417	LOL, faker, 恩静
D026238	吴芊盈, 创造101
D066225	一胎, 二胎
D000212	复仇者联盟3, 无限战争, 漫威, 神盾局特工
D011909	NCT127, 泰荣君
ID	类别
1-40000	娱乐新闻
40001-44060	体育新闻
44061-54060	健康新闻
54061-64060	军事新闻
64061-74060	正文本
74061-	教育新闻

求解思路

简单
规则

根据训练集的
观察结果，制
定简单规则，
预测所有文档
级。 (无监督)

二分类

将关键词提取
问题转化为二
分类的问题，
对每个词判断
其是否为关键
词的概率。

word2ve
c+神经网
络

将文档和对应
关键词表示为
向量，利用神
经网络。进行
预测。

详细过程

外部依赖

pandas、numpy、jieba、jieba.analyse、
re、math、sklearn、lightgbm、
collections、tqdm



数据清洗和预处理



TF-IDF的
baseline



特征工程



验证结果



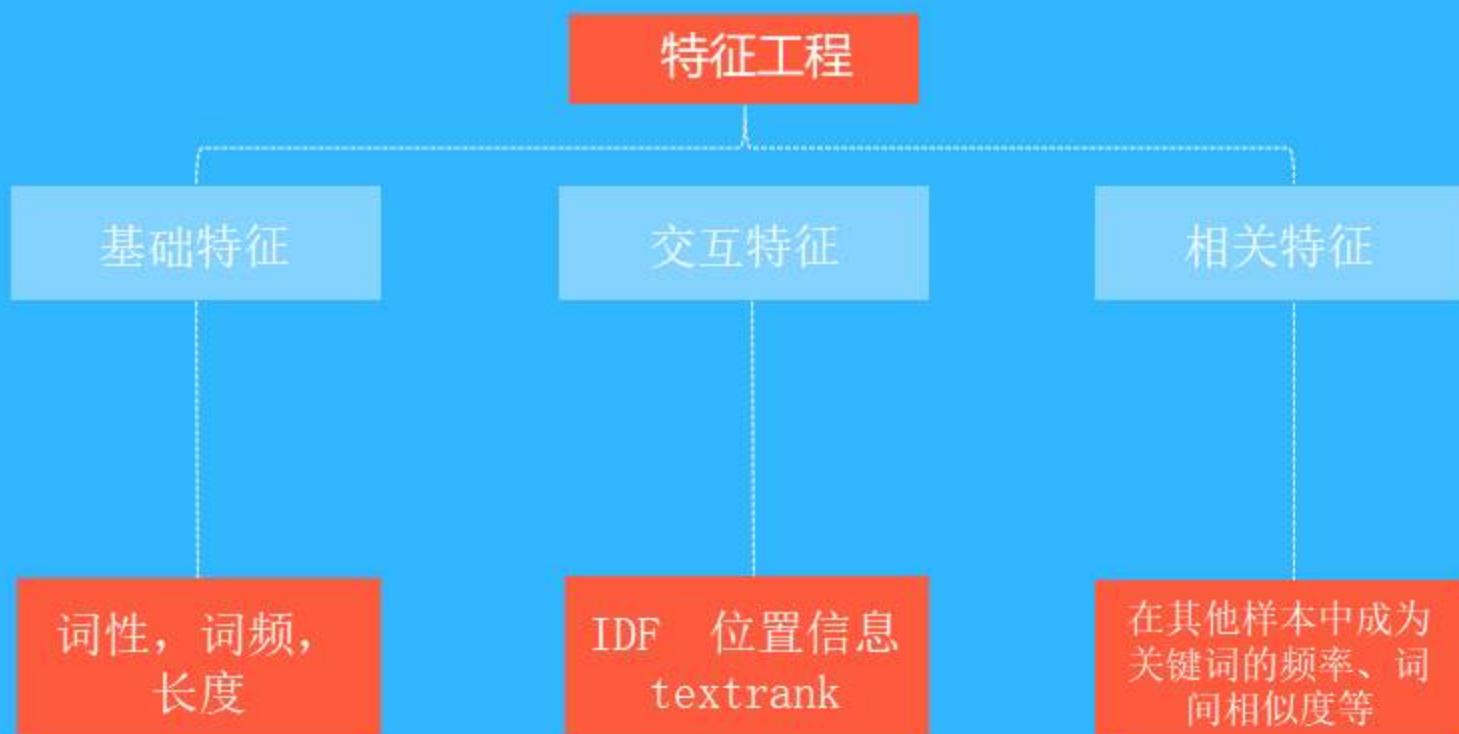
训练

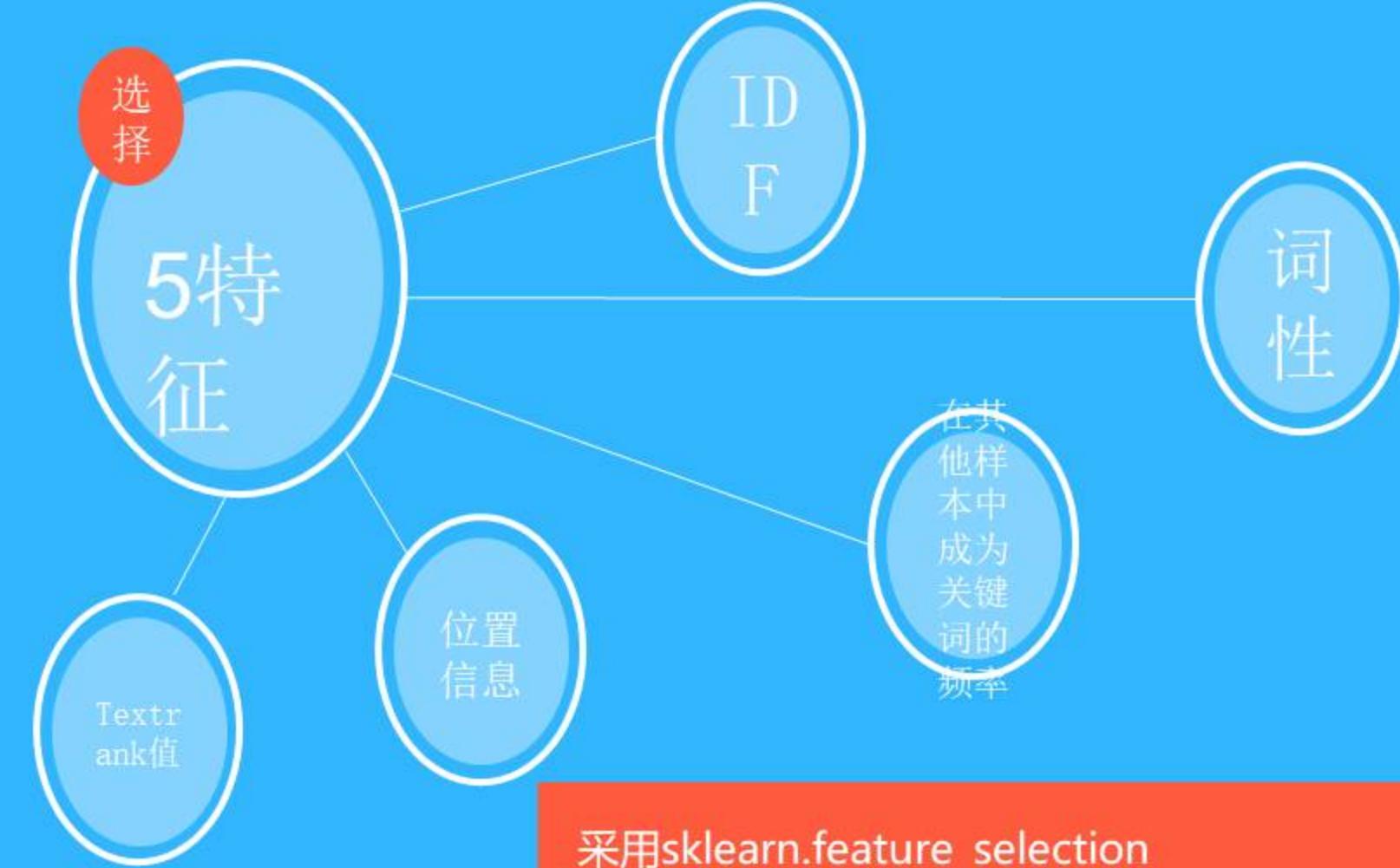
清洗及预处理



特征工程

心得：特征工程的好坏程度直接决定了结果的上限





特征选择

采用sklearn.feature_selection
SelectKBest、
ExtraTreesClassifier
决策树分析
其他：L1、L2正则化、循环提取

结果

