人工智能之机器学习

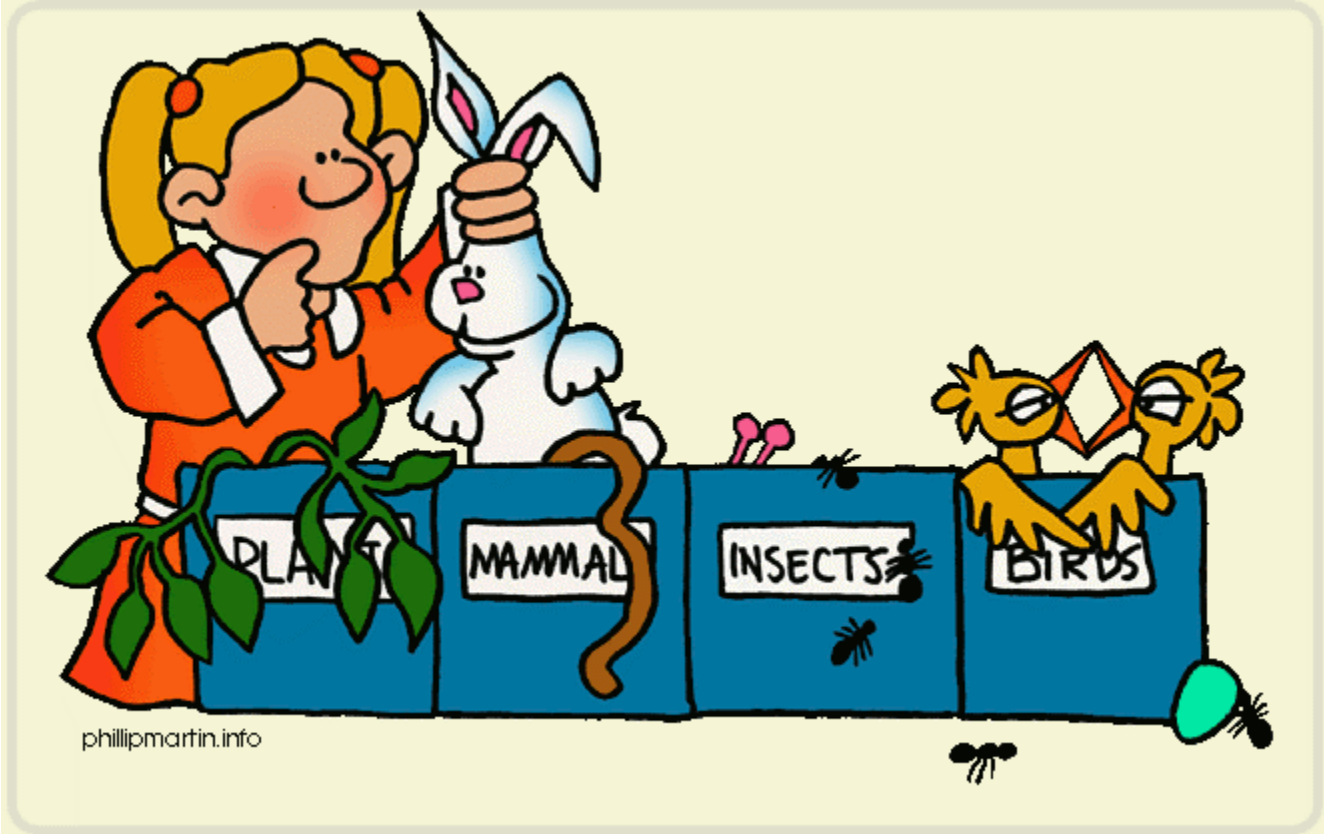
朴素贝叶斯

产品研发中心 -- 李军

凯通科技
www.ctt-net.com

用软件重新定义世界，让世界更加智能互联

朴素贝叶斯简介



样本数据

	F1	F2	F3	…	Fn	K
X1	1	2	3			A
X2	2	1	3			A
X3	3	2	1			B
.						
.						
.						
Xn						

- 朴素贝叶斯是使用概率论来分类的算法。其中
  - ✓ 朴素：各特征条件独立；
  - ✓ 贝叶斯：根据贝叶斯定理。

$$P(C \mid F_1) = \frac{P(CF_1)}{P(F_1)} = \frac{P(C) \cdot P(F_1 \mid C)}{P(F_1)}$$

条件概率分布

先验概率（Prior）：P(C)是C的先验概率，可以从已有的训练集中计算分为C类的样本占所有样本的比重得出。

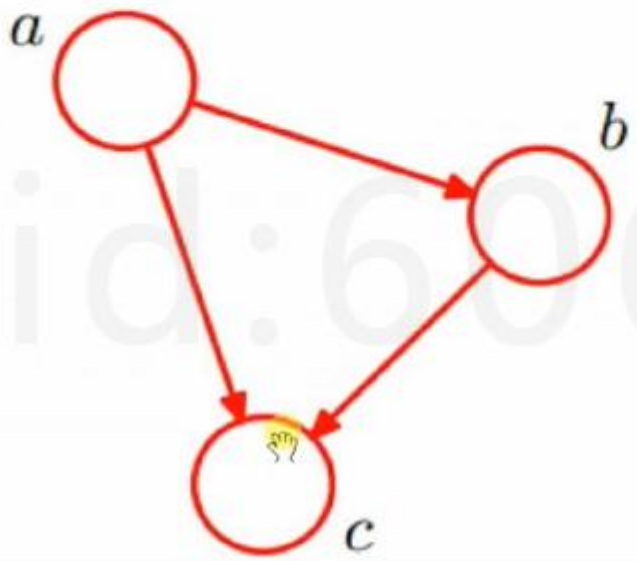
证据（Evidence）：即上式P(F1)，表示对于某测试样本，特征F1出现的概率。同样可以从训练集中F1特征对应样本所占总样本的比例得出。

似然（likelihood）：即上式P(F1|C)，表示如果知道一个样本分为C类，那么他的特征为F1的概率是多少。

一个简单的贝叶斯网络

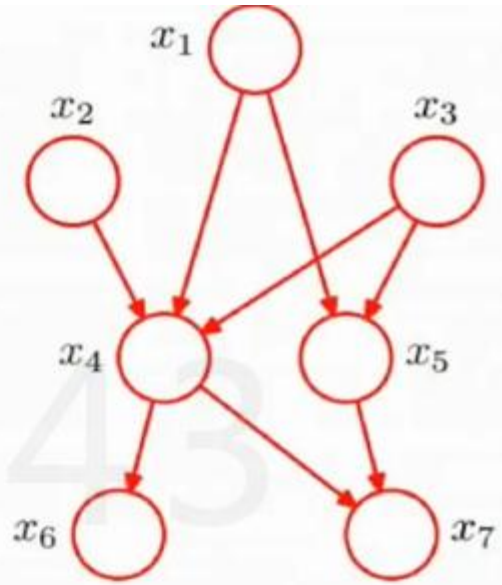
有向图（联合分布概率）

$$p(a,b,c) = p(c|a,b)p(b|a)p(a)$$



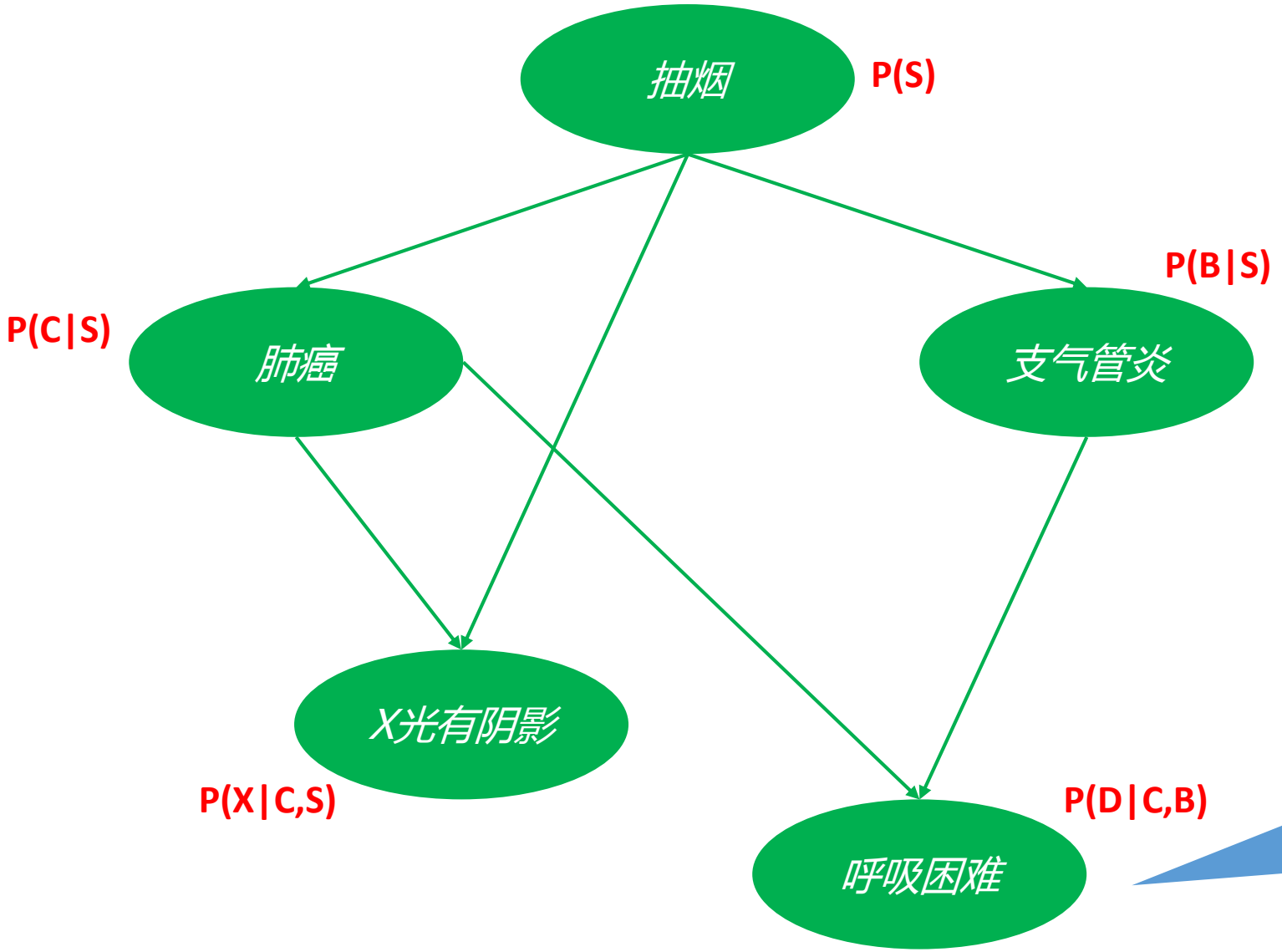
一个“正常”的贝叶斯网络

- 有些边缺失
- 直观上：
  - x1和x2独立
  - x6和x7在x4给定的条件下独立
- x1,x2,...x7的联合分布：



$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

朴素贝叶斯网络分布



1 + 2 + 2 + 4 + 4 = 13 VS 2^5

贝叶斯网络可以很好的降维
（防止过拟合）

CPD（条件概率表）：

C	B	D=0	D=1
0	0	0.9	0.1
0	1	0.3	0.7
1	0	0.2	0.8
1	1	0.1	0.9

三种朴素贝叶斯模型

高斯分布朴素贝叶斯（GaussianNB）

$$P(X_j = x_j|Y = C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} exp\left(-\frac{(x_j - \mu_k)^2}{2\sigma_k^2}\right)$$

- 样本特征的分布大部分是连续值，使用GaussianNB会比较好。

多项式分布朴素贝叶斯（MultinomialNB）

$$P(X_j = x_{jl}|Y = C_k) = \frac{x_{jl} + \lambda}{m_k + n\lambda}$$

- 如果样本特征的大部分是多元离散值，则使用MultinomialNB比较合适。

伯努利分布朴素贝叶斯（BernoulliNB）

$$P(X_j = x_{jl}|Y = C_k) = P(j|Y = C_k)x_{jl} + (1 - P(j|Y = C_k)(1 - x_{jl})$$

- 如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用BernoulliNB。

案例分析 – 病人诊断预测

某个医院早上收了六个门诊病人，如下表，现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

症状	职业	疾病
打喷嚏	护士	感冒
打喷嚏	农夫	过敏
头痛	建筑工人	脑震荡
头痛	建筑工人	感冒
打喷嚏	教师	感冒
头痛	教师	脑震荡

根据贝叶斯定理：P(A|B) = P(B|A) P(A) / P(B) 可得：

P(感冒|打喷嚏x建筑工人) = P(打喷嚏x建筑工人|感冒) x P(感冒) / P(打喷嚏x建筑工人)

假定"打喷嚏"和"建筑工人"这两个特征是独立的，因此，上面的等式就变成了

P(感冒|打喷嚏x建筑工人) = P(打喷嚏|感冒) x P(建筑工人|感冒) x P(感冒) / P(打喷嚏) x P(建筑工人)

P(感冒|打喷嚏x建筑工人) = 0.66 x 0.33 x 0.5 / 0.5 x 0.33 = 0.66

注意事项：

P(感冒)=3条感冒记录/总共6条记录=0.5

P(打喷嚏)=3条打喷嚏记录/总共6条记录=0.5

P(建筑工人)=2条建筑工人/总共6条记录=0.33

P(打喷嚏|感冒)=2条打喷嚏/总感冒记录3=0.66（即：总共3条感冒记录里面有2条是因为打喷嚏）

P(建筑工人|感冒)=1条建筑工人/总感冒记录3=0.33（即：总共3条感冒记录里面有1条是因为建筑工人）

因此，这个打喷嚏的建筑工人，有66%的概率是得了感冒。同理，可以计算这个病人患上过敏或脑震荡的概率。比较这几个概率，就可以知道他最可能得什么病。

案例分析 – 账号真实性预测

根据某社区网站的抽样统计，该站10000个账号中有89%为真实账号（设为C0），11%为虚假账号（设为C1）。假定某一个账号有以下三个特征：日志密度F1=s，好友密度为F2=s，是否真实头像为F3=yes，预测是否真实账号？

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

P(C0|F1F2F3) = P(F1|C0) P(F2|C0) P(F3|C0) P(C0) / P(F1) * P(F2) * P(F3)

P(C0)=7/10=0.7

P(F1)=3/10=0.3

P(F2)=4/10=0.4

P(F3)=5/10=0.5

□ 真实账号概率

P(F1|C0)=1/7=0.14

P(F2|C0)=1/7=0.14

P(F3|C0)=4/7=0.57

P(C0|F1F2F3) = 0.14 * 0.14 * 0.57 * 0.7 / 0.3 * 0.4 * 0.5 = 0.13034

P(C1|F1F2F3) = P(F1|C1) P(F2|C1) P(F3|C1) P(C1) / P(F1) * P(F2) * P(F3)

P(C1)=3/10=0.3

P(F1)=3/10=0.3

P(F2)=4/10=0.4

P(F3)=5/10=0.5

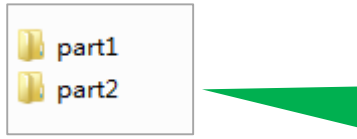
□ 虚假账号概率

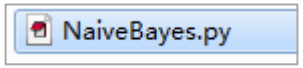
P(F1|C1)=2/3=0.66

P(F2|C1)=3/3=1

P(F3|C1)=1/3=0.33

P(C1|F1F2F3) = 0.66 * 1 * 0.33 * 0.3 / 0.3 * 0.4 * 0.5 = 1.089





案例实践 – 垃圾邮件过滤

☐ 设计思路

1）收集数据：将文本文件解析成词条向量，并取top100个关键词作为邮件特征

2）分析数据：去掉停用词、标点符号、语气词等噪声数据

3）规整数据：构建（n行 * m维）的文本特征向量矩阵

4）训练算法：使用朴素贝叶斯模型进行训练数据

5）测试算法：使用测试样本验证准确性

part1是训练数据集
part2是测试数据集

源代码程序

```
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.]
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  1.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.  1.
  1.  1.  1.  1.  1.  1.  1.  1.  0.  1.  1.  1.  0.  1.  0.  1.  1.  1.
  1.]
测试样本准确度：97.58%
```

测试分类结果

预测分类结果

朴素贝叶斯总结

优点：

- 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练；
- 对缺失数据不太敏感，算法也比较简单，常用于文本分类。

缺点：

- 分类决策存在错误率，依赖数据准确性，容易欠拟合；
- 对输入数据的表达形式很敏感。

感谢您的聆听！

Thank you for your time!

凯通科技
www.ctt-net.com

用软件重新定义世界，让世界更加智能互联