数据挖掘原理

主讲教师：李志勇（博士）

硕士生导师

讲师

数据科学系

数字农业工程技术研究中心

移动：13882213811 电邮: lzy@sicau.edu.cn

第七章：贝叶斯

——朴素是一种美德

主讲教师：李志勇

主要介绍内容

- 7.1 贝叶斯奇幻之旅
- 7.2 朴素贝叶斯
- 7.3 贝叶斯的预测算法
- 7.4 贝叶斯的诊断算法
- 7.5 贝叶斯的预测和诊断综合算法

7.1 贝叶斯奇幻之旅

- 先验概率：根据历史的资料或主观判断所确定的各种时间发生的概率

- 后验概率：通过贝叶斯公式，结合调查等方式获取了新的附加信息，对先验概率修正后得到的更符合实际的概率

- 条件概率：某事件发生后该事件的发生概率

$$\mathrm{P(A|B)} = \frac{\mathrm{P}(A,B)}{\mathrm{P(B)}} \qquad \mathrm{P(B|A)} = \frac{\mathrm{P}(A,B)}{\mathrm{P(A)}}$$

7.1 贝叶斯奇幻之旅

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

- 基本事件的互斥性 $B_iB_j=\phi, i\neq j, i,j=1,2,\ldots\ldots,n$
- 基本事件的完备性 $B_1\cup B_2\cup\ldots\ldots\cup B_n=\Omega$

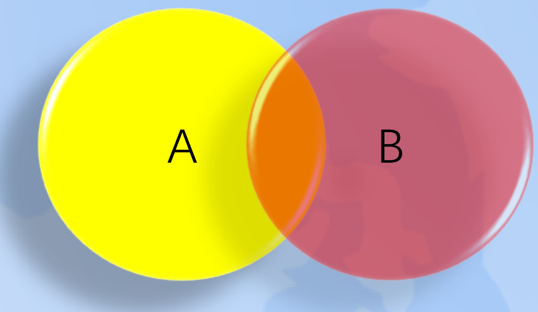
贝叶斯公式 $$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)}$$

- 独立互斥且完备的先验事件概率可以由后验事件的概率和相应条件概率决定

7.1 贝叶斯奇幻之旅

$$P(A\cup B)=P(A)+P(B)-P(A\cap B)$$

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



Likelihood of evidence B if A is true

Prior probability

$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)}$$

Posterior probability of A given the evidence B

Prior probability that evidence B is true

7.1 贝叶斯奇幻之旅

- **Salmon vs. Tuna**
- 随机抓一条鱼.
- $P(\omega_1)=P(\omega_2)$
- $P(\omega_1)>P(\omega_2)$
- 需要额外信息

$$P(\omega_i \mid x)=\frac{P(x \mid \omega_i)P(\omega_i)}{P(x)}$$

7.1 贝叶斯奇幻之旅

- **Probability of Kill**
  - **P(A): 0.6**
  - **P(B): 0.5**

- **The target is killed with:**
  - **One shoot from A**
  - **One shoot from B**

- **What is the probability that it is shot down by A?**
  - **C: The target is killed.**

$$P(A \mid C) = \frac{P(C \mid A)P(A)}{P(C)} = \frac{1 \times 0.6}{0.6 \times 0.5 + 0.4 \times 0.5 + 0.6 \times 0.5} = \frac{3}{4}$$

7.1 贝叶斯奇幻之旅

- ω_1: 癌症;　ω_2: 正常
- $P(\omega_1)=0.008$; $P(\omega_2)=0.992$
- 实验室测试结果: + vs. –
- $P(+|\omega_1)=0.98$; $P(-|\omega_1)=0.02$
- $P(+|\omega_2)=0.03$; $P(-|\omega_2)=0.97$
- 如果某人检验呈阳性...
- Is he/she doomed?

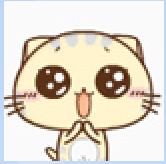
7.1 贝叶斯奇幻之旅

$$P(\omega_1 \mid +) \propto P(+ \mid \omega_1) P(\omega_1) = 0.98 \times 0.008 = 0.0078$$

$$P(\omega_2 \mid +) \propto P(+ \mid \omega_2) P(\omega_2) = 0.03 \times 0.992 = 0.0298$$

$$P(\omega_1 \mid +) < P(\omega_2 \mid +)$$

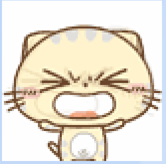
$$P(\omega_1 \mid +) = \frac{0.0078}{0.0078 + 0.0298} = 0.21 >> P(\omega_1)$$



7.1 贝叶斯奇幻之旅

- H="头痛"
- F="流感"
- P(H)=1/10; P(F)=1/40; P(H|F)=1/2

- 某一天你觉得头痛...
- 由于得了流感的人有 50% 会觉得头痛 ...
- 我有50%的可能性得了流感!



7.1 贝叶斯奇幻之旅

The truth is ···

$$P(F \mid H) = \frac{P(H \mid F)P(F)}{P(H)} = \frac{1/2 \times 1/40}{1/10} = \frac{1}{8}$$

7.2 朴素贝叶斯

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i \mid a_1, a_2, \ldots, a_n)$$

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \frac{P(a_1, a_2, \ldots, a_n \mid \omega_i) P(\omega_i)}{P(a_1, a_2, \ldots, a_n)}$$

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} \underline{P(a_1, a_2, \ldots, a_n \mid \omega_i)} P(\omega_i)$$

Conditionally Independent

$$\omega_{MAP} = \arg\max_{\omega_i \in \omega} P(\omega_i) \prod_j P(a_j \mid \omega_i)$$

MAP: Maximum A Posterior

7.2 朴素贝叶斯

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

$$\Downarrow$$

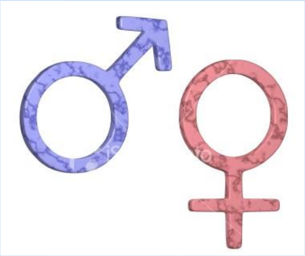
$$P(A \cap B) = P(A)P(B)$$

Conditionally Independent

$$P(A,B \mid G) = P(A \mid G)P(B \mid G) \Leftrightarrow \underline{P(A \mid G,B)} = P(A \mid G)$$

$$P(A, B \mid G) = P(A, B, G) / P(G) = P(A \mid B, G) \times P(B, G) / P(G)$$
$$= \underline{P(A \mid B, G)} \times P(B \mid G)$$

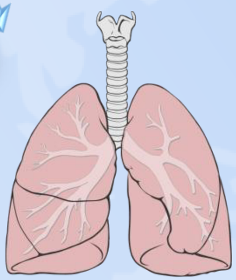
7.2 朴素贝叶斯



$$P(Cancer|Male) = 65/100,000$$
$$P(Cancer|Female) = 48/100,000$$

- Male/Female 与 Cancer是否独立?

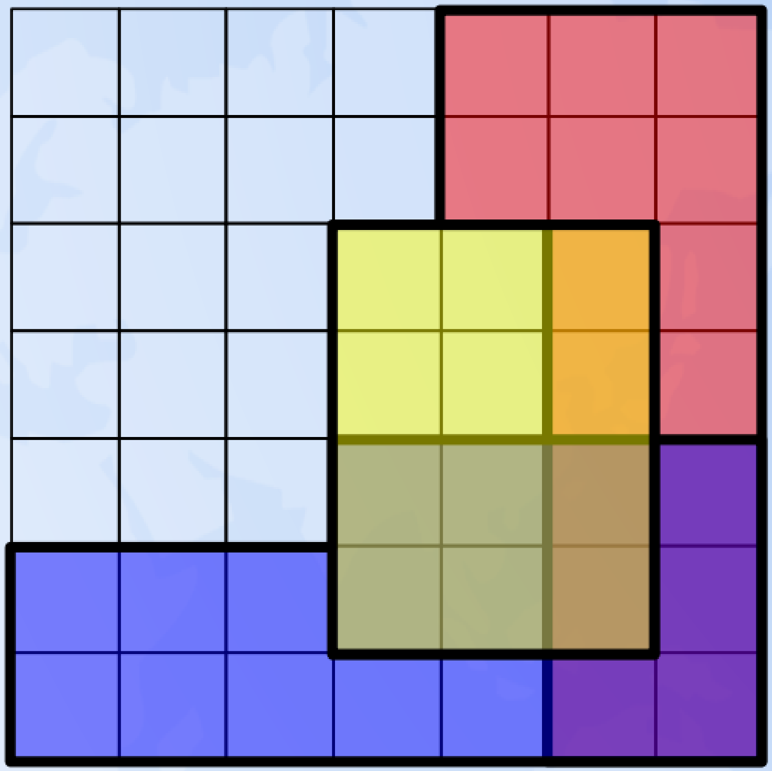
- 假设吸烟室导致肺癌的唯一因素.



Conditionally Independent

$$P(Cancer|Male, Smoking) = P(Cancer|Smoking)$$

7.2 朴素贝叶斯



$$P(R \cap B) = 6/49$$
$$P(R) = 16/49$$
$$P(B) = 18/49$$

$$P(R \cap B) \neq P(R)P(B)$$

Not Independent

$$P(R \cap B|Y) = 1/6$$
$$P(R|Y) = 1/3$$
$$P(B|Y) = 1/2$$

$$P(R \cap B|Y) = P(R|Y)P(B|Y)$$

Conditionally Independent

7.2 朴素贝叶斯

- 两枚硬币: 正常 vs. 异常 (two-headed)

- 随机选择一枚硬币抛两次.
- A: 第一次抛硬币头朝上.

- B: 第二次抛硬币头朝上.
- C: 你选择的是正常的硬币.

$$P(A) = P(B) = 0.5 \times 0.5 + 0.5 \times 1.0 = 0.75$$

$$P(C|A) = \frac{P(A|C)P(C)}{P(A|C)P(C)+P(A|\neg C)P(\neg C)} = \frac{0.5\times0.5}{0.5\times0.5+1\times0.5} = \frac{1}{3}$$

$$P(B|A) = \frac{1}{3} \times 0.5 + \frac{2}{3} \times 1.0 = \frac{5}{6} \neq P(B)$$ **Not Independent**

$$P(B|A,C) = P(B|C) = 0.5$$ **Conditionally Independent**

7.2 朴素贝叶斯

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E\big((X-\mu_X)(Y-\mu_Y)\big)}{\sigma_X \sigma_Y}$$

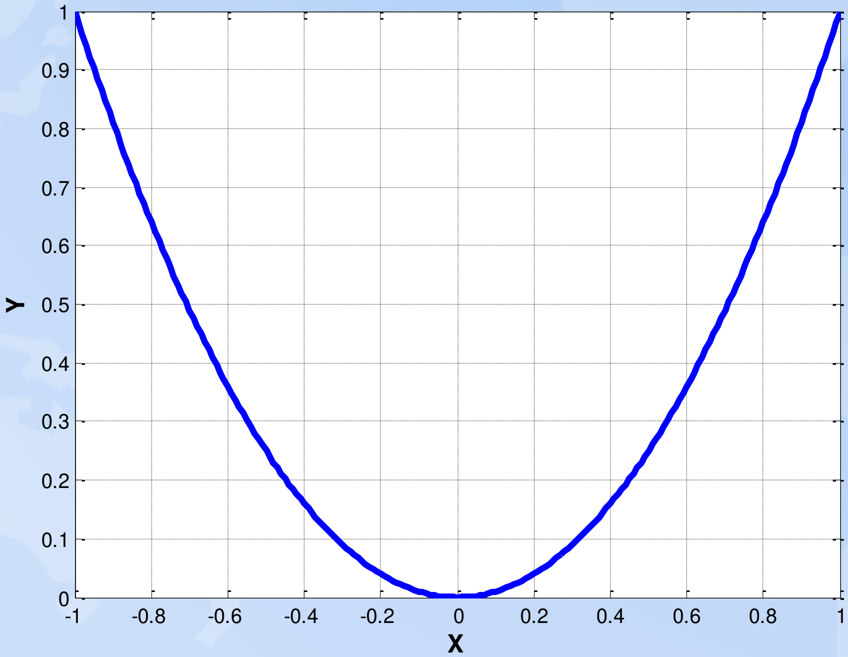
$X \in [-1,\ 1]$

$Y = X^2$

Cov (X,Y)=0 → X and Y are uncorrelated.

However, Y is completely determined by X.

X	Y
1	1
0.5	0.25
0.2	0.04
0	0
-0.2	0.04
-0.5	0.25
-1	1



7.2 朴素贝叶斯

α1	α2	α3	ω
	+		ω_1
			ω_2
	-		ω_1
	+		ω_1
			ω_2

$$P(\omega_1)=3/5; \qquad P(\omega_2)=2/5$$

$$P(a_2='+'|\omega_1)=2/3$$

$$P(a_2='-'|\omega_1)=1/3$$

Laplace Smoothing $$P(a_{jk}|\omega_i)=\frac{|a_j=a_{jk}\wedge\omega=\omega_i|+1}{|\omega=\omega_i|+|a_j|}$$

How about continuous variables?

7.2 朴素贝叶斯

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

7.2 朴素贝叶斯

$\text{Given}:$

$< \text{Outlook} = sunny, \text{Temperature} = cool, \text{Humidity} = high, \text{Wind} = strong >$

$\text{Predict}:$

$PlayTennis\,(yes\ or\ no)$

$Bayes\ Solution:$

$$P\left(PlayTennis = yes\right) = 9/14$$

$$P\left(PlayTennis = no\right) = 5/14$$

$$P\left(Wind = strong \mid PlayTennis = yes\right) = 3/9$$

$$P\left(Wind = strong \mid PlayTennis = no\right) = 3/5$$

...

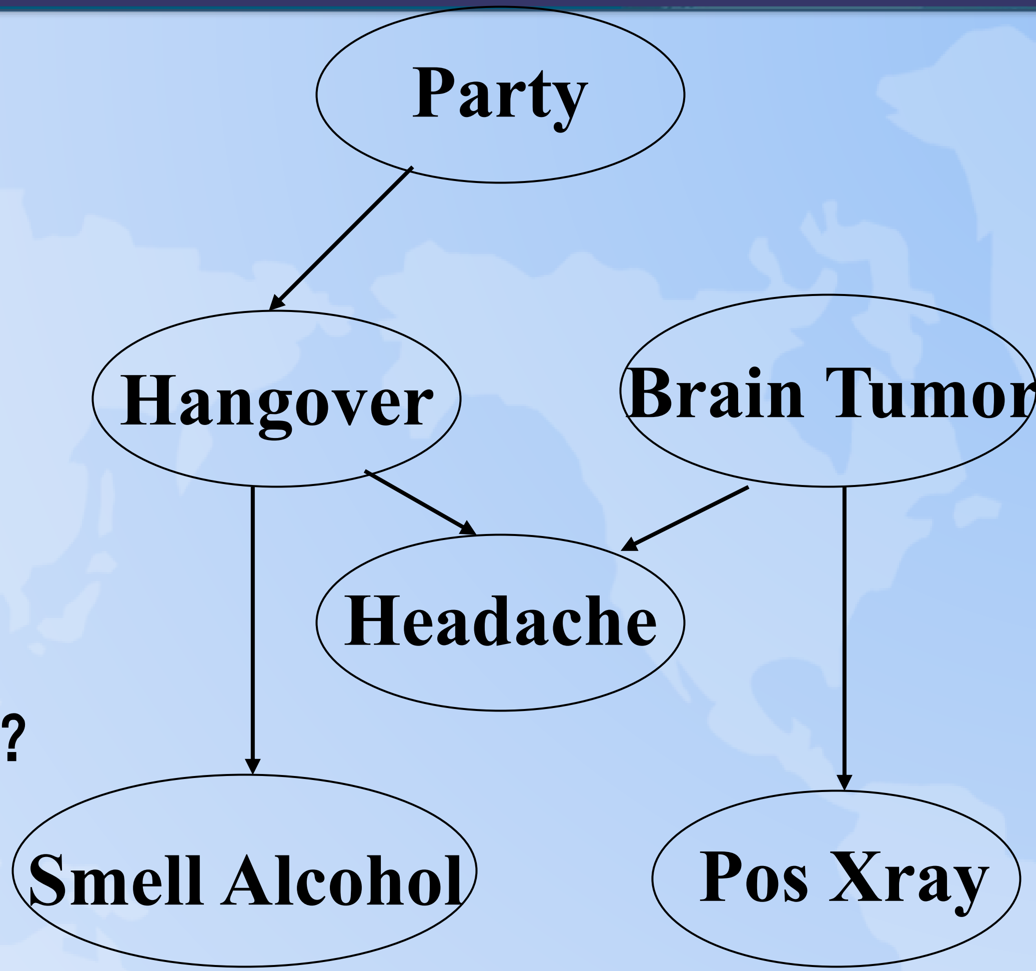
$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0053$$

$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = 0.0206$$

$$\text{The conclusion is not to play tennis with probability}: \frac{0.0206}{0.0206 + 0.0053} = 0.795$$

7.3 贝叶斯的预测算法

- 参加晚会后，第二天早晨呼吸中有酒精味的可能性有多大？
- 如果头疼，患脑瘤的概率有多大？
- 如果参加了晚会，并且头疼，那么患脑瘤的概率有多大？



- 为了方便表示，约定：对于一个一点point，P(+point)表示point发生的概率，P(-point)表示不发生的概率

7.3 贝叶斯的预测算法

已知条件节点发生与否，
推断结果结点发生的概率

例7.4 计算节点HA的概率。

目标：$P(+HA)=P(+HA|+BT,+HO)P(+BT,+HO)$
$+P(+HA|-BT,-HO)P(-BT,-HO)$
$+P(+HA|-BT,+HO)P(-BT,+HO)$
$+P(+HA|+BT,-HO)P(+BT,-HO)$
$=P(+HA|+BT,+HO)P(+BT)P(+HO)$
$+P(+HA|-BT,-HO)P(-BT)P(-HO)$
$+P(+HA|-BT,+HO)P(-BT)P(+HO)$
$+P(+HA|+BT,-HO)P(+BT)P(-HO)$

- $P(+HO)=P(+PT)P(+HO|+PT)+ P(-PT)P(+HO|-PT)$
  $=0.2*0.7+0.8*0=0.14$
  $P(-HO)=1 - P(+HO)=0.86$
- $P(+HA)=0.99*0.001*0.14+0.02*0.999*0.86$
  $+0.7*0.999*0.14+0.9*0.001*0.86$
  $=0.1159974$

7.3 贝叶斯的预测算法

例7.5 计算已知参加晚会的情况下，第二天早上呼吸有酒精味的概率。

已知条件节点发生与否，推断结果结点发生的概率

目标：P(+SA)=P(+SA|+HO)P(+HO)+P(+SA|-HO)P(-HO)

- P(+HO) =P(+PT)P(+HO|+PT)=1*0.7
  P(-HO)=1 - P(+HO)=0.3

- P(+SA)=P(+SA|+HO)P(+HO)+P(+SA|-HO)P(-HO)
  =0.7*0.8+0.3*0.1=0.59

7.3 贝叶斯的预测算法

例7.6 计算已知参加晚会的情况下，头痛发生的概率。

已知条件节点发生与否，推断结果结点发生的概率

目标：$P(+HA)= P(+HA|+BT,+HO)P(+BT)P(+HO)$
$+P(+HA|-BT,-HO)P(-BT)P(-HO)$
$+P(+HA|-BT,+HO)P(-BT)P(+HO)$
$+P(+HA|+BT,-HO)P(+BT)P(-HO)$

- $P(+HO)=P(+PT)P(+HO|+PT)=1*0.7$
  $P(-HO)=1 - P(+HO) =0.3$

- $P(+HA)=0.99*0.001*0.7+0.02*0.999*0.3$
  $+0.7*0.999*0.7+0.9*0.001*0.3$
  $=0.496467$

7.4 贝叶斯的诊断算法

例7.7 计算已知X光检查呈阳性的情况下，患脑瘤的概率。

已知结果结点发生与否，推断条件节点发生的概率

目标：P(+BT|+PX)={P (+BT)P(+PX|+BT)} / P(+PX)

- P(+PX)= P (+BT)P(+PX|+BT)+ P (-BT)P(+PX|-BT)
  =0.001*0.98+0.999*0.01=0.01097=0.011

- P(+BT|+PX)={P (+BT)P(+PX|+BT)} / P(+PX)
  =0.001*0.98/0.011=0.0890909

7.4 贝叶斯的诊断算法

例7.8 计算已知头痛的情况下，患脑瘤的概率。

已知结果结点发生与否，推断条件节点发生的概率

目标：P(+BT|+HA)={P(+BT)P(+HA|+BT)}/{P(HA)}
={0.001*0.9123}/{0.016}=0.007867

- P(+HA|+BT)=P(+HA|+BT,+HO)*P(+HO)
  + P(+HA|+BT,-HO)*P(-HO)
  =0.99*0.14+0.9*0.86=0.912

其中 P(+HO)=P(+PX)P(+HO|+PX)+ P(-PX)P(+HO|-PX)=0.14
P(-HO)=0.86

7.5 贝叶斯的预测和诊断综合算法

例7.9 计算参加晚上并且第二天早上呼吸有酒精味的情况下，宿醉发生的概率。

目标：P(+HO|+SA)={P(+HO)P(+SA|+HO)}/{P(+SA)}

- P(+HO)=P(+PX)P(+HO|+PT)+ P(-PX)P(+HO|-PT) =0.7
  P(-HO)=1-+0.7=0.3
- P(+SA)=P(+HO)P(+SA|+HO)+P(-HO)P(+SA|-HO)
  =0.7*0.8+0.3*0.1=0.59

- P(+HO||+SA)={P(+HO)P(+SA|+HO)}/{P(+SA)}
  =0.7*0.8/0.59=0.94915

7.5 贝叶斯的预测和诊断综合算法

例7.10 计算已知有酒精味、头痛的情况下，患脑瘤的概率。

目标：P(+BT|+HA)={P(+BT)P(+HA|+BT)}/{P(+HA)}

- P(+HO|+SA)={P(+HO)P(+SA|+HO)}/{P(+SA)}
  ={0.14*0.8}/{0.14*0.8+0.86*0.1}=0.5657
  P(+HO)=P(+HO|+SA)=0.4343
- P(+HA)=P(+HA|+BT,+HO)P(+BT)P(+HO)
  +P(+HA|-BT,-HO)P(-BT) P(-HO)
  +P(+HA|-BT,+HO)P(-BT)P(+HO)
  +P(+HA|+BT,-HO)P(+BT)P(-HO)
  =0.99*0.001*0.5657+0.02*0.999*0.4343
  +0.7*0999*0.5657+0.9*0.001*0.4343=0.4052
- P(+HA|+BT)=P(+HA|+BT,+HO)P(+HO)
  + P(+HA|+BT,-HO)P(-HO)
  =0.99* 0.5657 +0.9*0.4343=0.950913

- P(+BT|+HA)= ｛0.001*0.950913｝/｛ 0.4052 ｝=0.0023467