

# 机器学习

## 第5周



**【声明】** 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

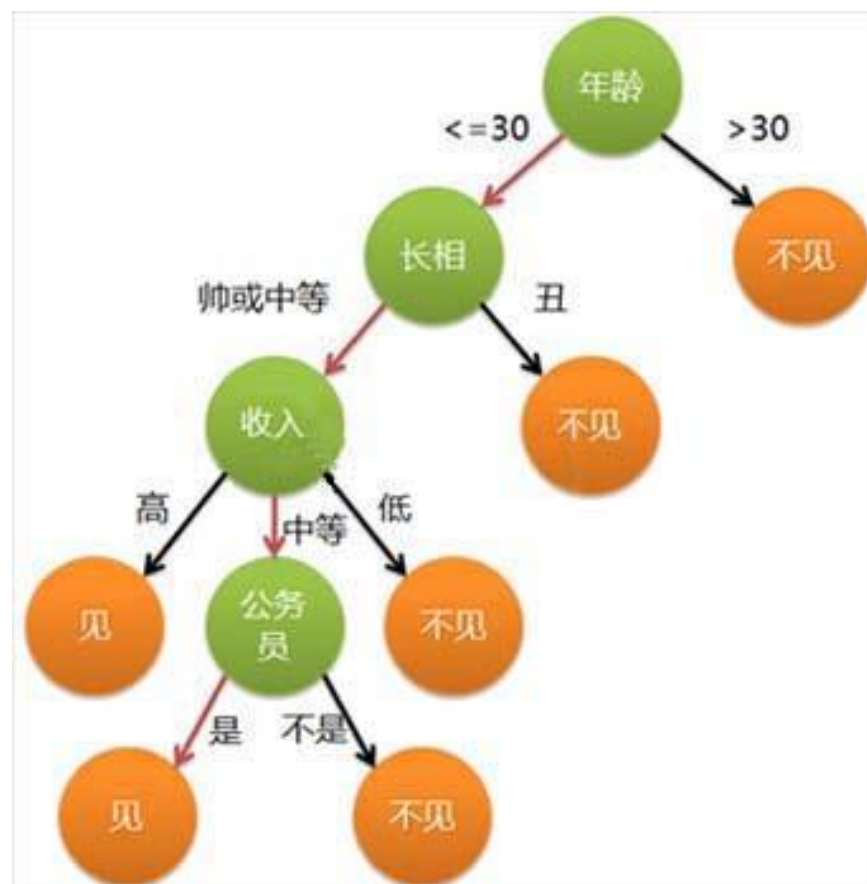
课程详情访问炼数成金培训网站

<http://edu.dataguru.cn>

# 分类：分类的意义

- 传统意义下的分类：生物物种
- 预测：天气预报
- 决策：yes or no
- 分类的传统模型
- 分类（判别分析）与聚类有什么差别？
- 有监督学习，无监督学习，半监督学习

- 线性判别法
- 距离判别法
- 贝叶斯分类器
- 决策树
- 支持向量机(SVM)
- 神经网络



- 网页自动分类
- 垃圾邮件判断
- 评论自动分析
- 通过用户访问内容判别用户喜好

- 自动化门户系统（百度新闻，谷歌新闻等）
- 搜索引擎根据用户标签类型推送不同类别的搜索结果

### 焦点新闻

#### 日本砸22亿元应对钓鱼岛局势 将建专属部队

人活每一天 马英九外甥吓退绑匪 王岳伦死磕到底  
张安薇大哥张大公谈救妹过程 特别感谢余靖

- 中国军方高度评价AK-47之父 美媒：人类悲剧 08:11
- 美政界盘点奥巴马政绩：内外交困 被中俄夺主动权 08:33
- 澳媒：中国人即将给亲日的澳领导人一个教训 08:27
- 罗德曼访问朝鲜 金正恩竟为他安排色情服务 12-24 08:17
- 朝鲜第一夫人为张成泽提供性服务被朝鲜证实 12-12 11:30
- 盘点2013中国军队出国十大事件 东海识别区上榜 10:20
- 甲骨文记载：巨人帮助古代中国人大战外星人 12-12 11:41
- 嫦娥之父：美国人去过月球 中国人也一定要去 08:13
- 外媒：中国连射洲际导弹意义重大 令美国不安 08:35
- 安理会通过向南苏丹大规模增派维和部队决议 16:16
- 空军上将：围绕强军目标学习研究毛泽东军事思想 08:59
- 叙利亚称化武储藏点遭反对派袭击 11:21
- 俄媒：直20先用直10发动机 量产型动力含外国技术 15:57
- 解放军四总部党的群众路线教育实践活动取得成效 15:04
- 共和国“第一号烈士”段德昌：被冤杀的未来元帅 13:10
- 一专多能的女兵台长 05:52
- 航空军工行业：大军工时代的到来 15:32



媒体称中国十天连射两洲际导弹 核打击能力增强？



南苏丹冲突至少8万人流离失所 联合国关切(图)



2013，这些事件峰回路转？

## 军事评论

- 毛泽东军事思想的伟大建树
- 美驱日制华“鹰犬战略”很危
- 华报：应重视俄罗斯对中日关
- “AK-47之父”曾称其枪支发
- 陈政雄：认识自我核心能力
- 中国罕见海战利器！“潜水战
- 解放军接连展示两大战略神器
- 面对朝鲜变局，韩国有必要紧

## 图片报道



国际晚班车：《时代》称奥巴马成2013



中英美印四国航空母舰“正脸”照大比拼



[酒店详情](#)[酒店点评 \(3027\)](#)[立即预订](#)

luya\*\*\*\*  
2013-12-23

总评:  5.0 卫生: 5 服务: 5 设施: 5 位置: 5


价格便宜 性价比高 交通便捷 靠近市区 服务不错。[详情]

豪华房

有用(0)



luya\*\*\*\*  
2013-12-23

总评:  5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格公道 性价比高 交通便捷 酒店餐厅很好吃 服务也很到位。[详情]

高级房

有用(0)



luya\*\*\*\*  
2013-12-23

总评:  5.0 卫生: 5 服务: 5 设施: 5 位置: 5


五星级酒店而言 价格便宜 性价比高 交通便捷 服务到位。[详情]

豪华房

有用(0)



1100\*\*\*\*  
2013-12-23

总评:  5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格合理, 出行方便[详情]

高级房


有用(0)

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



300720\*\*\*\*  
2013-12-23

总评:  3.8 卫生: 5 服务: 5 设施: 3 位置: 2

在携程订购的话给的房间都是最小的。别的还行[详情]

高级单人房

有用(0)


来自: 手机用户

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



109216\*\*\*\*  
2013-12-23

总评:  5.0 卫生: 5 服务: 5 设施: 5 位置: 5

还不错。[详情]

高级单人房

有用(0)

## ■ 例子：天气预报数据

```
G=c(1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2)
```

```
x1=c(-1.9,-6.9,5.2,5.0,7.3,6.8,0.9,-12.5,1.5,3.8,0.2,-0.1,0.4,2.7,2.1,-4.6,-1.7,-2.6,2.6,-  
2.8)
```

```
x2=c(3.2,0.4,2.0,2.5,0.0,12.7,-5.4,-  
2.5,1.3,6.8,6.2,7.5,14.6,8.3,0.8,4.3,10.9,13.1,12.8,10.0)
```

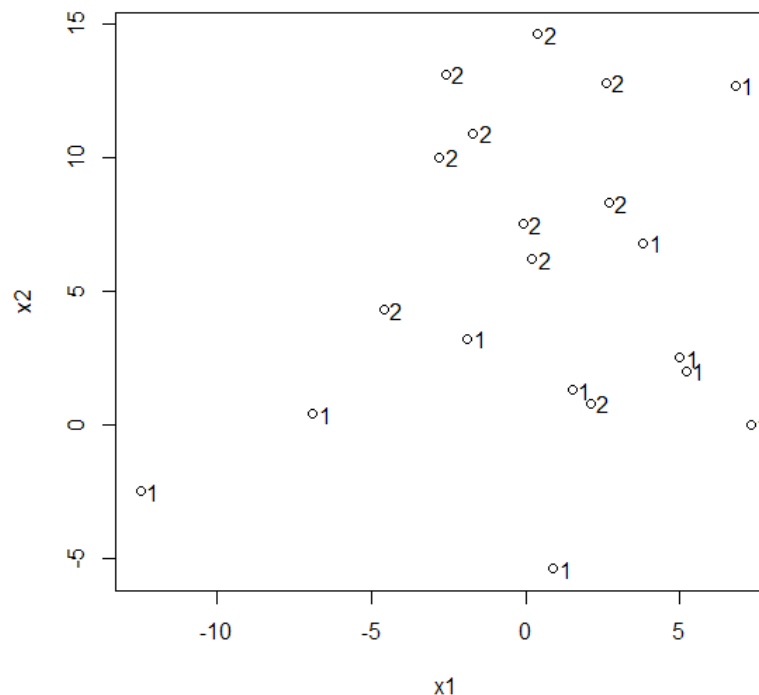
```
a=data.frame(G,x1,x2)
```

```
plot(x1,x2)
```

```
text(x1,x2,G,adj=-0.5)
```



- 用一条直线来划分学习集（这条直线一定存在吗？）
- 然后根据待测点在直线的哪一边决定它的分类



# MASS包与线性判别函数lda( )

```
library(MASS)
```

```
ld=lda(G~x1+x2)
```

```
ld
```

```
> ld
```

```
Call:
```

```
lda(G ~ x1 + x2)
```

```
Prior probabilities of groups:
```

```
  1    2  
0.5 0.5
```

```
Group means:
```

```
      x1    x2  
1  0.92  2.10  
2 -0.38  8.85
```

```
Coefficients of linear discriminants:
```

```
      LD1  
x1 -0.1035305  
x2  0.2247957  
█
```

```
z=predict(ld)
newG=z$class
newG
[1] 1 1 1 1 1 2 1 1 1 1 1 2 2 2 2 1 2 2 2 2 2
Levels: 1 2
cbind=(G,z$x,newG)
y=cbind(G,z$x,newG)
y
```

|    | G | LD1         | newG |
|----|---|-------------|------|
| 1  | 1 | -0.28674901 | 1    |
| 2  | 1 | -0.39852439 | 1    |
| 3  | 1 | -1.29157053 | 1    |
| 4  | 1 | -1.15846657 | 1    |
| 5  | 1 | -1.95857603 | 1    |
| 6  | 1 | 0.94809469  | 2    |
| 7  | 1 | -2.50987753 | 1    |
| 8  | 1 | -0.47066104 | 1    |
| 9  | 1 | -1.06586461 | 1    |
| 10 | 1 | -0.06760842 | 1    |
| 11 | 2 | 0.17022402  | 2    |
| 12 | 2 | 0.49351760  | 2    |
| 13 | 2 | 2.03780185  | 2    |
| 14 | 2 | 0.38346871  | 2    |
| 15 | 2 | -1.24038077 | 1    |
| 16 | 2 | 0.24005867  | 2    |
| 17 | 2 | 1.42347182  | 2    |
| 18 | 2 | 2.01119984  | 2    |
| 19 | 2 | 1.40540244  | 2    |
| 20 | 2 | 1.33503926  | 2    |

- 原理：计算待测点与各类的**距离**，取最短者为其所属分类
- 马氏距离（薛毅书p445，为什么不用欧氏距离？），计算函数mahalanobis()

**定义 8.1** 设  $x, y$  是服从均值为  $\mu$ ，协方差阵为  $\Sigma$  的总体  $X$  中抽取的样本，则总体  $X$  内两点  $x$  与  $y$  的 *Mahalanobis* 距离（简称马氏距离）定义为

$$d(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}. \quad (8.1)$$

定义样本  $x$  与总体  $X$  的 *Mahalanobis* 距离为

$$d(x, X) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}. \quad (8.2)$$

## ■ 情形一（薛毅书p445）

首先考虑两个总体  $X_1$  和  $X_2$  的协方差相同的情况，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 = \Sigma_2 = \Sigma.$$

要判断  $x$  是属于哪一个总体，需要计算  $x$  到总体  $X_1$  和  $X_2$  的 Mahalanobis 距离的平方  $d^2(x, X_1)$  和  $d^2(x, X_2)$ ，然后进行比较，若  $d^2(x, X_1) \leq d^2(x, X_2)$ ，则判定  $x$  属于  $X_1$ ；否则判定  $x$  来自  $X_2$ 。由此得到如下判别准则：

$$R_1 = \{x \mid d^2(x, X_1) \leq d^2(x, X_2)\}, \quad R_2 = \{x \mid d^2(x, X_1) > d^2(x, X_2)\}. \quad (8.3)$$

令

$$w(x) = (x - \bar{\mu})^T \Sigma^{-1} (\mu_1 - \mu_2), \quad (8.5)$$

称  $w(x)$  为两总体距离的判别函数，因此判别准则 (8.3) 变为

$$R_1 = \{x \mid w(x) \geq 0\}, \quad R_2 = \{x \mid w(x) < 0\}. \quad (8.6)$$

- 情形二 ( 薛毅书p447 )
- 例子 ( 薛毅书p449 )

对于样本  $x$ , 在协方差阵不同的情况下, 判别函数为

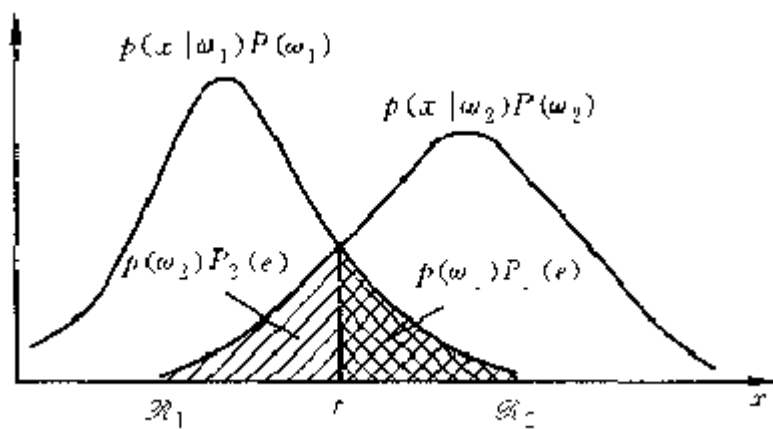
$$w(x) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1). \quad (8.12)$$

## ■ 算法主要思想：

- 1 选取**k**个和待分类点**距离**最近的样本点
- 2 看1中的样本点的分类情况，**投票**决定待分类点所属的类



## ■ 原理 (薛毅书p455)



$$R_1 = \left\{ x \mid \frac{f_1(x)}{f_2(x)} \geq \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1} \right\}, \quad R_2 = \left\{ x \mid \frac{f_1(x)}{f_2(x)} < \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1} \right\}.$$

- 对于总体协方差矩阵相同的情形

$$R_1 = \left\{ x \mid W(x) \geq \beta \right\}, \quad R_2 = \left\{ x \mid W(x) < \beta \right\}, \quad (8.26)$$

其中

$$\begin{aligned} W(x) &= \frac{1}{2}(x - \mu_2)^T \Sigma^{-1}(x - \mu_2) - \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1) \\ &= \left[ x - \frac{1}{2}(\mu_1 + \mu_2) \right]^T \Sigma^{-1}(\mu_1 - \mu_2), \end{aligned} \quad (8.27)$$

$$\beta = \ln \frac{L(1|2) \cdot p_2}{L(2|1) \cdot p_1}. \quad (8.28)$$

- 对于总体协方差矩阵不同的情形

$$R_1 = \left\{ x \mid W(x) \geq \beta \right\}, \quad R_2 = \left\{ x \mid W(x) < \beta \right\}, \quad (8.29)$$

其中

$$W(x) = \frac{1}{2}(x - \mu_2)^T \Sigma_2^{-1}(x - \mu_2) - \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1}(x - \mu_1), \quad (8.30)$$

$$\beta = \ln \frac{L(1|2) \cdot p_2}{L(2|1) \cdot p_1} + \frac{1}{2} \ln \left( \frac{|\Sigma_1|}{|\Sigma_2|} \right). \quad (8.31)$$

- 薛毅书P457
- 利用贝叶斯分类器判断垃圾邮件

- 多分类下的距离判别法（薛毅书p452）
- 多分类下的贝叶斯（薛毅书p460）

## 背景知识：朴素贝叶斯文本分类器原理

- 一个文档代表一个样本，一个词代表一个特征，类别就是目标变量，假设：
  - 文档集为  $D = \{d_1, d_1, \dots, d_n\}$ ；
  - 目标类别集为  $C = \{c_1, c_1, \dots, c_m\}$ ；
  - 特征集为  $X = \{x_1, x_1, \dots, x_k\}$ ；
- 朴素假设：给定的文档集中，文档的特征是相互独立的
- 贝叶斯原理：利用贝叶斯条件概率公式，计算出已知文档属于不同文档类别的后验概率，然后根据最大后验假设将该文档归结为具有最大后验概率的那一类
  - 类内条件概率：文档 $d$ 在类 $c_i$ 中出现的概率，记为 $P(d|c_i)$
  - 先验概率：类 $c_i$ 出现的概率，记为 $P(c_i)$
  - 后验概率： $P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)}$
  - 类别判断公式： $L(d) = \underset{i}{\operatorname{argmax}} P(c_i|d)$ ，可以简化为 $L(d) = \underset{i}{\operatorname{argmax}} P(d|c_i)P(c_i)$

# 背景知识：朴素贝叶斯文本分类器算法

- 根据对类条件概率 $P(d|c_i)$ 的计算方式不同，朴素贝叶斯算法可以分为多变量伯努利模型、多项式模型、泊松模型，等。

## ■ 多项式模型

- 设 $t_j$ 为特征 $x_j$ 在文档 $d$ 中出现的次数， $P(x_j|c_i)$ 为特征 $x_j$ 在类 $c_i$ 的文档中出现一次的概率

- 文档 $d$ 在类 $c_i$ 中出现的概率：
$$P(d|c_i) = \frac{(\sum_j t_j)!}{\prod_j t_j!} \prod_j P(x_j|c_i)^{t_j}$$

- 类别判断公式：
$$L(d) = \operatorname{argmax}_i \left[ \log P(c_i) + \sum_j t_j P(x_j|c_i) \right]$$

- $P(x_j|c_i)$  计算公式：
$$P(x_j|c_i) = \frac{x_j \text{在类 } c_i \text{ 中出现的次数} + \alpha}{c_i \text{ 中所有特征出现的总次数} + k}$$
，其中 $\alpha$ 是平滑因子， $k$ 是特征集的大小



A B C  
B  
12 B 14

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mtttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

研究表明，汉字的序顺并不定一能影阅响读，比如当你看完这句话后，才发这现里的字全是乱的。

Prof. Daniel Kahneman的研究

# 垃圾邮件判断

Foxmail [stswzh@sysu.edu.cn]

文件(E) 查看(V) 邮箱(B) 邮件(M) 工具(T) 帮助(H)

收取 发送 撰写 回复 全部回复 转发 删除 邮件提醒 地址簿 远程管理 中转站

反垃圾邮件设置

常规 规则过滤 贝叶斯过滤 黑名单 白名单

在学习邮件前需要整理您的邮件夹，以避免把垃圾邮件作为非垃圾邮件学习或把非垃圾邮件作为垃圾邮件学习。

☐ 使用贝叶斯概率模型判定接收的邮件是否垃圾邮件(U)

已学习信息

|        |                        |       |        |
|--------|------------------------|-------|--------|
| 非垃圾邮件: | 2558                   | 垃圾邮件: | 4104   |
| 非垃圾词:  | 1079931                | 垃圾词:  | 786541 |
| 更新时间:  | 2013-12-25 下午 01:32:11 |       |        |

学习(L)... 高级(A)...

过滤强度

移动下面的标记设定过滤的强度。

低 中 高

过滤强度设定越高，邮件被判定为垃圾邮件的可能性越

☒ 自动删除垃圾邮件箱中以下天数之前的旧邮件

30 天之前

☐ "设定为非垃圾邮件"时不显示提示窗(D)

导入... 导出...

确定 取消

垃圾邮件箱

尊敬的老师:

| 主题                            | 日期          |
|-------------------------------|-------------|
| 高校特色专业建设与创新人才培养               | 2013年11月29日 |
| SciencePG: Fast Publication   | 2013年11月29日 |
| 韩编辑                           | 2013年11月29日 |
| stswzh                        | 2013年11月29日 |
| 真人外教一对一,一节课15元,仅限本周!          | 2013年11月29日 |
| 论文翻译: stswzh@mail.sysu.edu.cn | 2013年11月28日 |
| 教授导师,您好!麻烦看一下!                | 2013年11月28日 |
| 真人外教一对一,一节课15元,仅限本周!          | 2013年11月28日 |
| 经销商管理终端销量提升                   | 2013年11月27日 |
| 与客户打交道的9个基本原则                 | 2013年11月26日 |
| 在线企业网络安全线上研讨会 赢千元健康手环...      | 2013年11月26日 |
| 在线申请平安小额借,15万一天到手(ad)         | 2013年11月25日 |

- 分词
- 贝叶斯公式与贝叶斯分类器

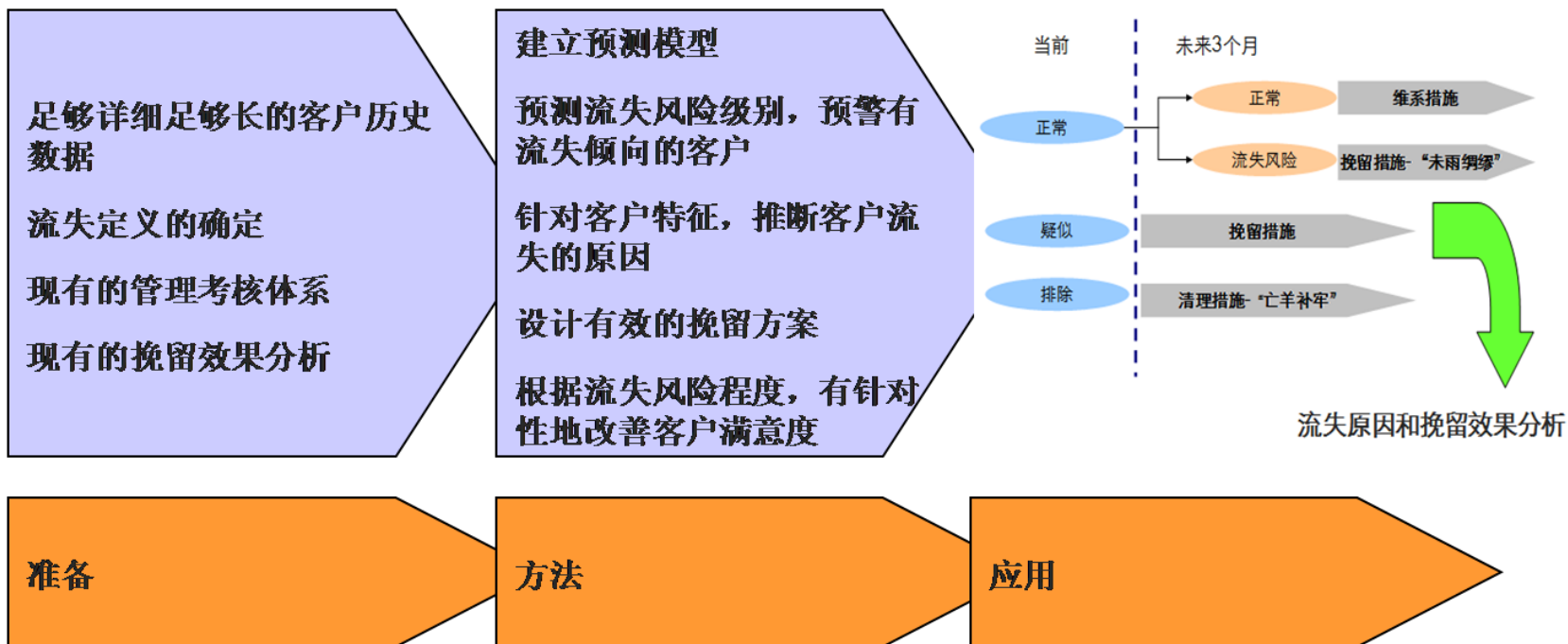
若  $B_1, B_2, \dots$  为一系列互不相容的事件，且

$$\bigcup_{i=1}^{\infty} B_i = \Omega, \quad P(B_i) > 0, i = 1, 2, \dots$$

则对任一事件  $A$ ，有

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{k=1}^{\infty} P(B_k)P(A|B_k)}, \quad i = 1, 2, \dots$$





# 用户标签系统



- Bayes Belief Network , 简称BBN
- 朴素贝叶斯分类器需要特征之间互相独立的强条件 , 制约了模型的适用
- 用有向无环图表达变量之间的依赖关系 , 变量用节点表示 , 依赖关系用边表示
- 祖先 , 父母和后代节点。贝叶斯网络中的一个节点 , 如果它的父母节点已知 , 则它条件独立于它的所有非后代节点
- 每个节点附带一个条件概率表 ( CPT ) , 表示该节点和父母节点的联系概率



- 创建网络结构（专业人员知识）
- 计算CPT（通过学习数据）
- 如果数据不完备，则需要训练计算（类似神经网络，采用梯度下降法）

- 如果节点X没有父母节点，则它的CPT之包含先验概率 $P(X)$
- 如果节点X只有一个父母节点Y，则CPT中包含条件概率 $P(X|Y)$
- 如果节点X有多个父母节点 $Y_1, Y_2, \dots, Y_k$ ，则CPT中包含条件概率 $P(X|Y_1, Y_2, \dots, Y_k)$

## ■ 韩家炜书第256页

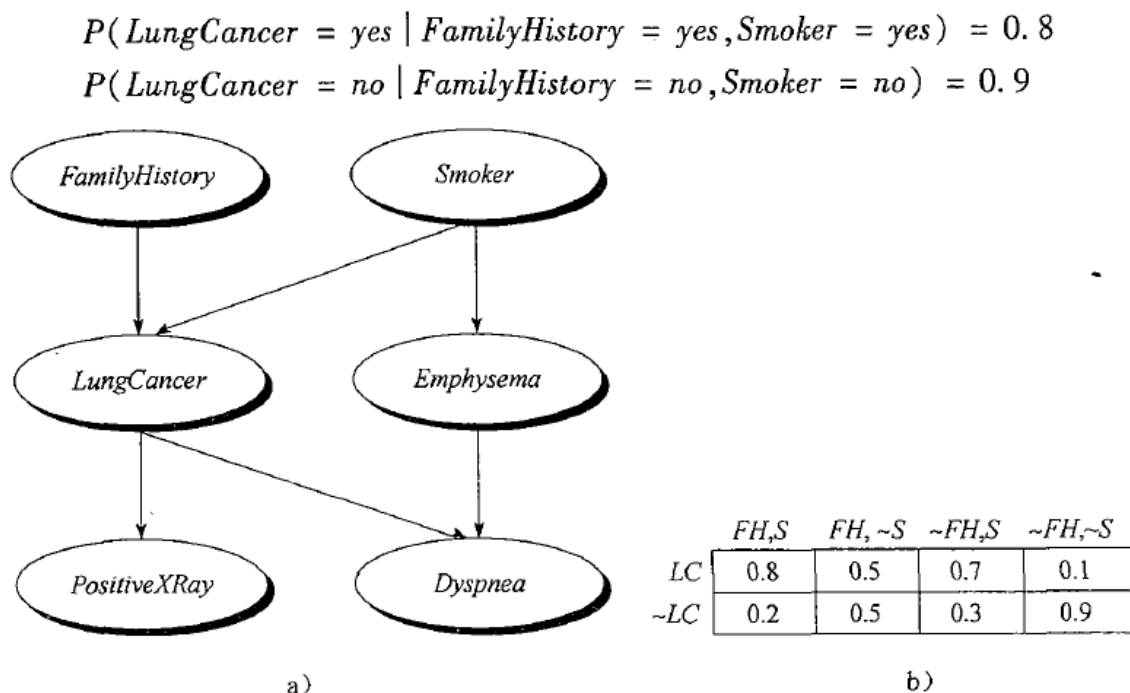


图 9.1 一个简单的贝叶斯信念网络：a) 一个提议的因果模型，用有向无环图表示；b) 变量 *LungCance*(*LC*) 的条件概率表，给出其双亲节点 *FamilyHistory* 和 *Smoke* 的每个可能值组合的条件概率。取自 Russell、Binder、Koller 和 Kanazawa[RBKK95]

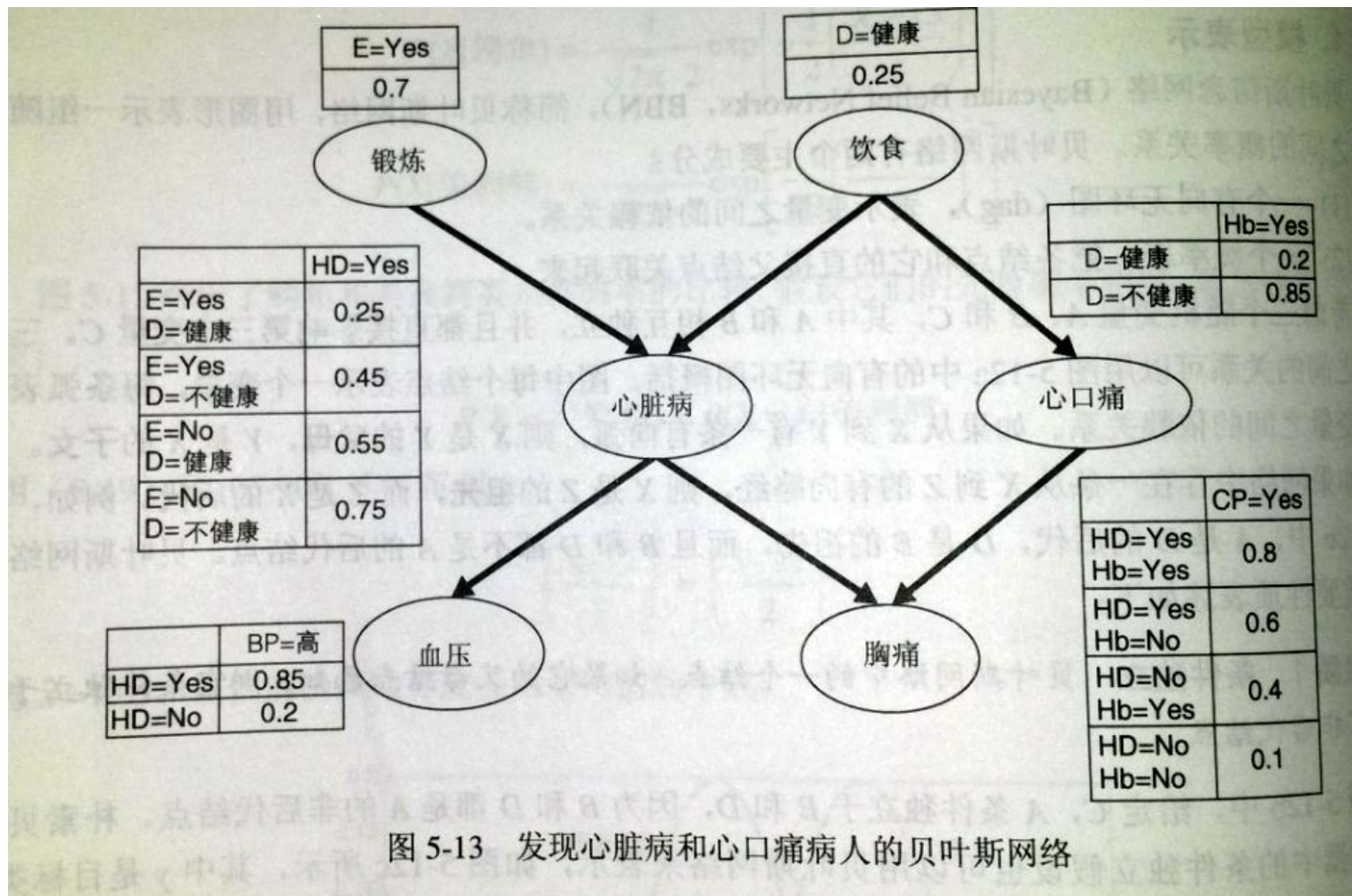


图 5-13 发现心脏病和心口痛病人的贝叶斯网络

- 从CPT中基于父母节点的条件概率推出某节点（变量）的概率

$$\begin{aligned} P(\text{HD}=\text{Yes}) &= \sum_{\alpha} \sum_{\beta} P(\text{HD} = \text{Yes} \mid E = \alpha, D = \beta) P(E = \alpha, D = \beta) \\ &= \sum_{\alpha} \sum_{\beta} P(\text{HD} = \text{Yes} \mid E = \alpha, D = \beta) P(E = \alpha) P(D = \beta) \\ &= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 + 0.75 \times 0.3 \times 0.75 \\ &= 0.49 \end{aligned}$$

- 计算某节点基于后代节点的条件概率

$$\begin{aligned} P(\text{BP} = \text{高}) &= \sum_{\gamma} P(\text{BP} = \text{高} \mid \text{HD} = \gamma) P(\text{HD} = \gamma) \\ &= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185 \end{aligned}$$

- 计算某节点基于父母节点，后代节点的条件概率

$$\begin{aligned} & P(HD=Yes|BP=高, D=健康, E=Yes) \\ &= \left[ \frac{P(BP=高|HD=Yes, D=健康, E=Yes)}{P(BP=高|D=健康, E=Yes)} \right] \times P(HD=Yes|D=健康, E=Yes) \\ &= \frac{P(BP=高|HD=Yes)P(HD=Yes|D=健康, E=Yes)}{\sum_{\gamma} P(BP=高|HD=\gamma)P(HD=\gamma|D=健康, E=Yes)} \\ &= \frac{0.85 \times 0.25}{0.85 \times 0.25 + 0.2 \times 0.75} \\ &= 0.5862 \end{aligned}$$



- 其它非父母，非后代节点与该节点本身是条件独立的

- 韩家炜书第257页
- 什么时候需要训练？

- **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



# Thanks

## FAQ时间