数据挖掘与机器学习

1

潘斌
panbin@nankai.edu.cn
范孙楼227

实验课安排

- 本周四（4.22）上午1、2节，二主楼B403

上节回顾

- 贝叶斯分类器
  - 后验概率
  - 决策规则
- 最小错误率
- 最小风险
- 正态决策面

4.5.2 正态分布决策面

$$p(\boldsymbol{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu)^T\Sigma^{-1}(\boldsymbol{x}-\mu)\right\}$$

- 正态分布时分类器的决策面方程
  - 后验概率函数

③ $g_i(\boldsymbol{x})=\ln p(\boldsymbol{x}|\omega_i)+\ln P(\omega_i)$

$$g_i(\boldsymbol{x})=-\frac{1}{2}(\boldsymbol{x}-\mu_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\mu_i)-\frac{d}{2}\ln 2\pi-\frac{1}{2}\ln|\Sigma_i|+\ln P(\omega_i)$$

  - 决策面方程

$$g_i(\boldsymbol{x})=g_j(\boldsymbol{x})$$

$$-\frac{1}{2}[(\boldsymbol{x}-\mu_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\mu_i)-(\boldsymbol{x}-\mu_j)^T\Sigma_j^{-1}(\boldsymbol{x}-\mu_j)]-\frac{1}{2}\ln\frac{|\Sigma_i|}{|\Sigma_j|}+\ln\frac{P(\omega_i)}{P(\omega_j)}=0$$

4.5.2 正态分布决策面

- **正态分布时分类器的决策面方程**
  - 二次型判别函数

展开

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

$$= \boldsymbol{x}^T W_i \boldsymbol{x} + \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

$$W_i = -\frac{1}{2}\Sigma_i^{-1} \quad (d \times d \text{ 矩阵})$$

$$\boldsymbol{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i \quad (d \text{ 维列向量})$$

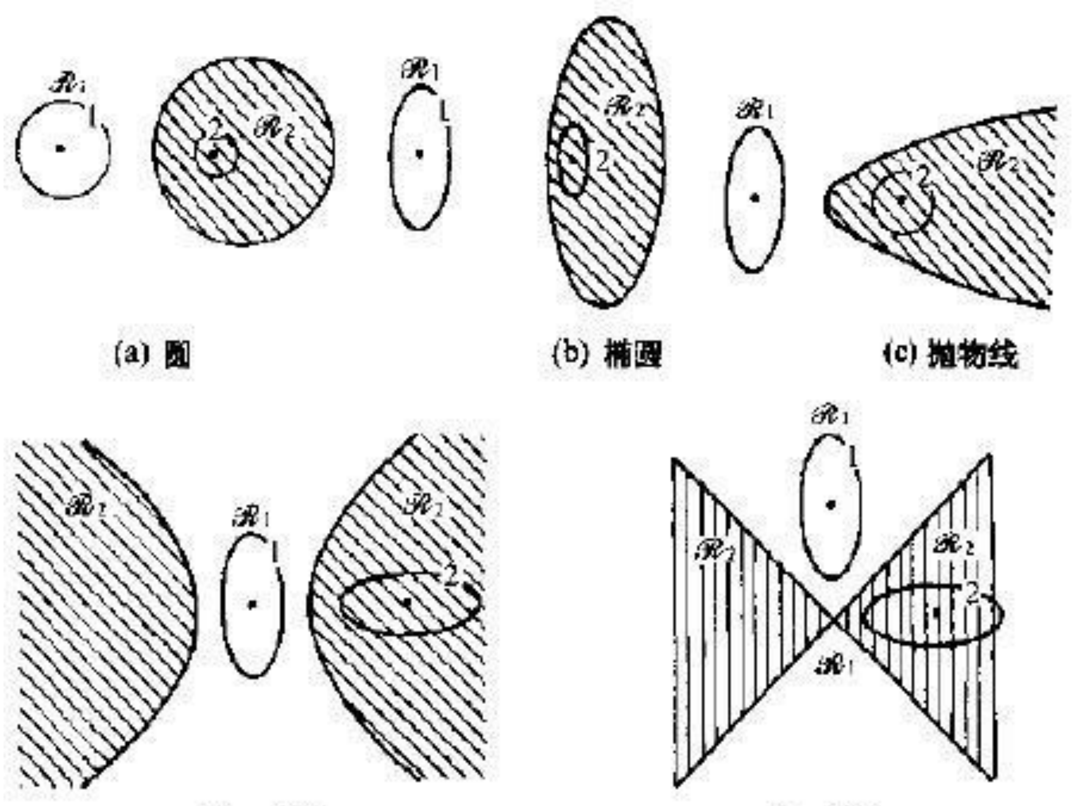
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

  - 决策面方程

相减

4.5.2 正态分布决策面

- 决策面示例



(a) 圆　(b) 椭圆　(c) 抛物线

(d) 双曲线　(e) 直线

4.5.2 正态分布决策面

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$
$$= x^T W_i x + w_i^T x + w_{i0}$$

- 决策面特例一
  - $\Sigma_i = \sigma^2 I \qquad i = 1, 2, \ldots, C$

$$\Sigma_i = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}$$

朴素贝叶斯
分类器

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \ln P(\omega_i)$$

准垂直平分

$$w^T(x-x_0) = 0$$

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i+\mu_j) - \frac{\sigma^2}{\|\mu_i-\mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i-\mu_j)$$

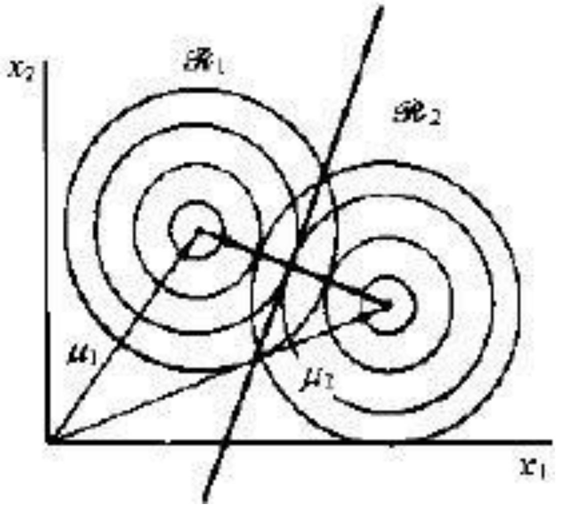
4.5.2 正态分布决策面

- 决策面特例一

$$w^T(x-x_0)=0$$

$$w=\mu_i-\mu_j$$

$$x_0=\frac{1}{2}(\mu_i+\mu_j)-\frac{\sigma^2}{\|\mu_i-\mu_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\mu_i-\mu_j)$$



4.5.2 正态分布决策面

- 决策面特例二
  - $\Sigma_i = \Sigma$　　　　　　　i = 1, 2, …, C

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0) = 0$$

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\dfrac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

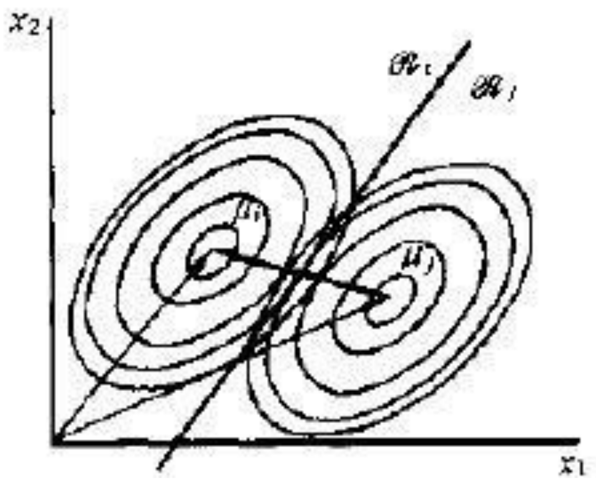
4.5.2 正态分布决策面

- 决策面特例二

$$w^T(x-x_0)=0$$

$$w=\Sigma^{-1}(\mu_i-\mu_j)$$

$$x_0=\frac{1}{2}(\mu_i+\mu_j)-\frac{\ln\frac{P(\omega_i)}{P(\omega_j)}}{(\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i-\mu_j)}(\mu_i-\mu_j)$$



4.5.3 错误率估计

- **Bayes分类器错误率的估计方法**
  - 按理论公式求解计算*
  - 计算错误率上界
  - 实验估计错误率

  - *注：直接计算十分困难，只能在特例情况下进行。

4.5.3 错误率估计

- **错误率的计算**
  - 两类正态分布等协方差阵情况下

$$P(e)=\int_{\Re_1}P(\omega_2)\,p(x\mid\omega_2)dx+\int_{\Re_2}P(\omega_1)\,p(x\mid\omega_1)dx$$

$$P(e_{12})=\int_{\Re_1}P(\omega_2)\,p(x\mid\omega_2)dx=P(\omega_2)\int_{\Re_1}p(x\mid\omega_2)dx=P(\omega_2)P_2(e)$$

$$P(e_{21})=\int_{\Re_2}P(\omega_1)\,p(x\mid\omega_1)dx=P(\omega_1)\int_{\Re_2}p(x\mid\omega_1)dx=P(\omega_1)P_1(e)$$

4.5.3 错误率估计

- 错误率的计算

回顾最小错误率贝叶斯决策规则的负对数似然比形式。

$$h(\mathbf{x}) = -\ln l(\mathbf{x}) = -\ln p(\mathbf{x}|\omega_1) + \ln p(\mathbf{x}|\omega_2) \lessgtr \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right] \to \mathbf{x} \in \begin{cases}\omega_1 \\ \omega_2\end{cases}$$

$h(\mathbf{x})$是$\mathbf{x}$的函数，$\mathbf{x}$是随机向量，因此$h(\mathbf{x})$是随机变量。我们记它的分布密度函数为$p(h|\omega_i)$。由于它是一维密度函数，因此易于积分，所以用它计算错误率有时较为方便。这样式(2-113)可表示为

$$P_1(e) = \int_{\mathscr{R}_2} p(\mathbf{x}|\omega_1)\mathrm{d}\mathbf{x} = \int_t^{\infty} p(h|\omega_1)\mathrm{d}h \qquad (2\text{-}114)$$

$$P_2(e) = \int_{\mathscr{R}_1} p(\mathbf{x}|\omega_2)\mathrm{d}\mathbf{x} = \int_{-\infty}^{t} p(h|\omega_2)\mathrm{d}h \qquad (2\text{-}115)$$

其中 $t = \ln[P(\omega_1)/P(\omega_2)]$ (2-116)

4.5.3 错误率估计

$$p(x)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

$$\mu=E\{x\}$$

$$\Sigma=E\{(x-\mu)(x-\mu)^T\}$$

- 错误率的计算

$$\begin{aligned}
h(x) &= -\ln l(x) = -\ln p(x|\omega_1) + \ln P(x|\omega_2) \\
&= -\left[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_1|\right] \\
&\quad + \left[-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_2|\right] \\
&= \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2) + \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \\
&\lessgtr \ln\frac{P(\omega_1)}{P(\omega_2)} \rightarrow x \in \begin{cases}\omega_1 \\ \omega_2\end{cases}
\end{aligned}$$

4.5.3 错误率估计

- **错误率的计算**
  - 两类正态分布等协方差阵情况下

$$h(\mathbf{x}) = (\mu_2 - \mu_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)$$

$$\lessgtr \ln \frac{P(\omega_1)}{P(\omega_2)} \rightarrow \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

4.5.3 错误率估计

$$h(x)=(\mu_2-\mu_1)^T\Sigma^{-1}x+\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2)$$
$$\lessgtr\ln\frac{P(\omega_1)}{P(\omega_2)}\to x\in\begin{cases}\omega_1\\ \omega_2\end{cases}$$

- **错误率的计算**
  - 线性组合仍然服从正态分布：通过均值和方差来估计错误率

$$\eta_1=E[h(x)|\omega_1]=(\mu_2-\mu_1)^T\Sigma^{-1}\mu_1+\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1-\mu_2^T\Sigma^{-1}\mu_2)$$
$$=-\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)$$

$$\eta=\frac{1}{2}[(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)]$$

X是正态分布

$$\eta_1=-\eta$$
$$\sigma_1^2=E\{[h(x)-\eta]^2|\omega_1\}=(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)=2\eta$$

$$\eta_2=\frac{1}{2}(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)=\eta$$
$$\sigma_2^2=(\mu_1-\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2)=2\eta$$

4.5.3 错误率估计

- **错误率的计算**
  - 线性组合仍然服从正态分布

$$
\begin{aligned}
P_1(e) &= \int_t^{\infty} p(h|\omega_1)\mathrm{d}h \\
&= \int_t^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{h+\eta}{\sigma}\right)^2\right\}\mathrm{d}h \\
&= \int_t^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{h+\eta}{\sigma}\right)^2\right\}\mathrm{d}\left(\frac{h+\eta}{\sigma}\right) \\
&= \int_{\left(\frac{t+\eta}{\sigma}\right)}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\xi^2\right)\mathrm{d}\xi
\end{aligned}
$$

4.5.3 错误率估计

- **错误率的计算**
  - 线性组合仍然服从正态分布

$$P_2(e)=\int_{-\infty}^{t} p(h|\omega_2)\mathrm{d}h$$

$$=\int_{-\infty}^{t}(2\pi)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{h-\eta}{\sigma}\right)^2\right\}\mathrm{d}\left(\frac{h-\eta}{\sigma}\right)$$

$$=\int_{-\infty}^{\frac{t-\eta}{\sigma}}(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\xi^2\right)\mathrm{d}\xi$$

$$t=\ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right],\sigma=\sqrt{2\eta}$$

4.5.3 错误率上界

- **两类情况下最小错误率的上界**
  - Chernoff切尔诺夫上界

$$P(e_{12}) \le P(\omega_2)e^{[-\mu(s)+(1-s)t]}$$

$$P(e_{21}) \le P(\omega_1)e^{[-\mu(s)-st]}$$

  - Bhattacharyya巴氏系数上界

$$J_B = -\ln\left[\int \sqrt{p(x|\omega_1)p(x|\omega_2)}dx\right]$$

$$P(e) \le \sqrt{P(\omega_1)P(\omega_2)}\exp\{-J_B\}$$

4.5.3 错误率估计

- **错误率的实验估计**
  - 测试样本集（区别于训练样本集）
  - 测试结果（有限样本情况，混淆矩阵）
  - 错误率评价

4.5.4 讨论

- **Bayes决策的先决条件**
  - 类条件概率密度的估计
  - 例：密度函数估计的简化
    - 假设特征向量各分量相互独立

4.5.4 讨论

- **Bayes决策的最小错误概率**
  – 设计难度大（类条件概率密度估计）

  – 寻求高“性价比”的分类器
    - 错误概率逼近
    - 决策面逼近

朴素贝叶斯分类器（NAÏVE BAYES CLASSIFIER）

- 在给定目标值时，属性值之间相互条件独立，即

$$P(a_1, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$$

$$P(x_1, \ldots, x_n | w_j) = \prod_i P(x_i | w_j)$$

$$v_{NB} = \underset{v_j \in V}{\arg\max} P(v_j) \prod_i P(a_i | v_j)$$

Naive_Bayes_Learn($examples$)

For each target value v_j

$\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$

For each attribute value a_i of each attribute a

$\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

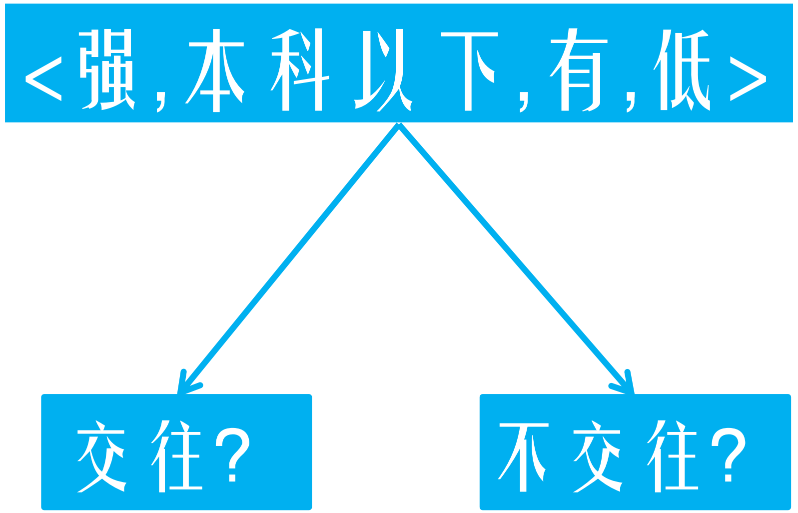
Classify_New_Instance(x)

$$v_{NB} = \operatorname*{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

对比 $\longrightarrow P(X|\theta) \times P(\theta)$

例：

上进心	学历	房产	年薪	交往
强	研究生	有	高	是
强	研究生	有	低	是
弱	研究生	有	高	否
一般	本科	有	高	否
一般	本科以下	无	高	否
一般	本科以下	无	低	是
弱	本科以下	无	低	否
强	本科	有	高	是
强	本科以下	无	高	否
一般	本科	无	高	否
强	本科	无	低	否
弱	本科	有	低	否
弱	研究生	无	高	否
一般	本科	有	低	是



- 例（续）$v_{NB} = \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$

$$= \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) P(强|v_j) P(本科以下|v_j) P(有|v_j) P(低|v_j)$$

  - 根据数据集，可以计算出上式需要的概率值
    - P(否)=9/14=0.64；P(是)=5/14=0.36；
    - P(强|否)=2/9=0.22；P(强|是)=3/5=0.6；
    - P(弱|否)=4/9=0.44；P(弱|是)=0/5=0；
    - P(一般|否)=3/9=0.33；P(一般|是)=2/5=0.4；
    - P(研究生|否)= 2/9=0.22；P(研究生|是)=2/5=0.4；
    - P(本科|否)= 4/9=0.44；P(本科|是)=2/5=0.4；
    - P(本科以下|否)=3/9=0.33；P(本科以下|是)=1/5=0.2；
    - P(无|否)=6/9=0.67；P(无|是)=1/5=0.2；
    - P(有|否)=3/9=0.33；P(有|是)=4/5=0.8；
    - P(高|否)=6/9=0.67；P(高|是)=2/5=0.4；
    - P(低|否)=3/9=0.33；P(低|是)=3/5=0.4

$$v_{NB} = \underset{v_j \in \{\text{是},\text{否}\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$
$$= \underset{v_j \in \{\text{是},\text{否}\}}{\operatorname{argmax}} P(v_j)P(\text{强}|v_j)P(\text{本科以下}|v_j)P(\text{有}|v_j)P(\text{低}|v_j)$$

- 求 v_{NB}
  - P(否)P(强|否)P(本科以下|否)P(有|否)P(低|否) =0.0053
  - P(是)P(强|是)P(本科以下|是)P(有|是)P(低|是) =0.0206
  - v_{NB}=是
- 通过将上述的量归一化，可计算给定观察值下目标值为“是”的条件概率为0.0206/(0.0206+0.0053)=0.795。

- 练习：

Name	Give Birth	Can Fly	ve in Wat	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mamma
salmon	no	no	yes	no	non-mamma
whale	yes	no	yes	no	mammals
frog	no	no	sometime	yes	non-mamma
komodo	no	no	no	yes	non-mamma
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mamma
cat	yes	no	no	yes	mammals
leopard shar	yes	no	yes	no	non-mamma
turtle	no	no	sometime	yes	non-mamma
penguin	no	no	sometime	yes	non-mamma
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mamma
salamander	no	no	sometime	yes	non-mamma
gila monster	no	no	no	yes	non-mamma
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mamma
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mamma

Give Birth	Can Fly	ive in Wate	Have Legs	Class
yes	no	yes	no	?

例

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$?

估计连续型属性的条件概率

- 将每一连续属性离散化，然后利用相应的离散区间替换连续属性值
- 把“Taxable Income”划分成两个区间，最佳的候选划分点为97K，对应区间为（0,97）和[97,10000）。通过计算不同类别中属性“Taxable Income”落入对应区间的比例来估计条件概率。

Tid	Refund	Marital Status	Taxable Income< 97K	Evade
1	Yes	Single	No	No
2	No	Married	No	No
3	No	Single	Yes	No
4	Yes	Married	No	No
5	No	Divorced	Yes	Yes
6	No	Married	Yes	No
7	Yes	Divorced	No	No
8	No	Single	Yes	Yes
9	No	Married	Yes	No
10	No	Single	Yes	Yes

- 用Bayes方法估计每个条件概率后，对之前新给出的样本可以进行判别。

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = \text{No})$$

Tid	Refund	Marital Status	Taxable Income< 97K	Evade
1	Yes	Single	No	No
2	No	Married	No	No
3	No	Single	Yes	No
4	Yes	Married	No	No
5	No	Divorced	Yes	Yes
6	No	Married	Yes	No
7	Yes	Divorced	No	No
8	No	Single	Yes	Yes
9	No	Married	Yes	No
10	No	Single	Yes	Yes

- P(X|Class=No) = P(Refund=No|Class=No)
  P(Married| Class=No)
  P(Income=No| Class=No)
  $= 4/7 \times 4/7 \times 4/7 = 0.1866$

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  P(Married| Class=Yes)
  P(Income=No| Class=Yes)
  $= 1 \times 0 \times 0 = 0$

- Since P(X|No)P(No) > P(X|Yes)P(Yes)

  Therefore P(No|X) > P(Yes|X)
  => Class = No

用概率分布来估计条件概率

- 假设连续型属性服从某种概率分布（通常假设服从正态分布），然后用训练数据估计出分布的参数，进而计算相应的条件概率。如上例中，假设“Taxable Income”属性为随机变量

$$X_3 \sim N(\mu, \sigma^2)$$

- 对于每个类 Ci，属性值x_j属于类Ci的概率为

$$P(x_j|C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp(-\frac{(x_j-\mu_{ij})^2}{2{\sigma_{ij}}^2})$$

  - μ_{ij}和σ_{ij}分别为类C_i中随机变量x_j的期望和方差，可分别用 C_i中x_j的观察值的样本均值和标准差估计

- 如上表数据中“Taxable Income”数据，分别属于两类，设类别 C1=“No”，C2=“Yes”，对应的观察值如下：

Taxable Income	125	100	70	120	95	60	220	85	75	90
Evade	No	No	No	No	Yes	No	No	Yes	No	Yes

- 类别 C1=“No”的两个参数估计如下：
  - $\bar{X}=\frac{1}{7}(125+100+70+120+60+220+75)=110$
  - $S^2=\frac{1}{6}\{(125-110)^2+(100-110)^2+(70-110)^2$ $+(120-110)^2+(60-110)^2+(220-110)^2+(75-110)^2\}=2975$
  - $(\mu,\sigma^2)=(110,54.54^2)$
- 同理，类别 C2=“Yes”的两个参数估计为：$(\mu,\sigma^2)=(90,5^2)$
  - $P(Income=120|No)=\frac{1}{\sqrt{2\pi}(54.54)}e^{-\frac{(120-110)^2}{2(2975)}}=0.0072$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

$$
\begin{aligned}
P(X|\text{Class=No}) &= P(\text{Refund=No}|\text{Class=No}) \\
&\quad \times P(\text{Married}|\ \text{Class=No}) \\
&\quad \times P(\text{Income=120K}|\ \text{Class=No}) \\
&= 4/7 \times 4/7 \times 0.0072 = 0.0024
\end{aligned}
$$

$$
\begin{aligned}
P(X|\text{Class=Yes}) &= P(\text{Refund=No}|\ \text{Class=Yes}) \\
&\quad \times P(\text{Married}|\ \text{Class=Yes}) \\
&\quad \times P(\text{Income=120K}|\ \text{Class=Yes}) \\
&= 1 \times 0 \times 1.2 \times 10^{-9} = 0
\end{aligned}
$$

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$

Therefore $P(No|X) > P(Yes|X)$

=> Class = No

问题

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No: sample mean=110
sample variance=2975

If class=Yes: sample mean=90
sample variance=25

- 利用训练样例估计后验概率的潜在问题
  - 若一属性的类条件概率为0，在整个类的后验概率为0
  - 一个极端的例子，训练样例不能覆盖那么多属性值时，可能无法分类某些测试记录
- 一种解决方案
  - 当样本很小时，采用平滑技术，m-估计$\frac{n_a+mp}{n_c+m}$
    - p是将要确定的概率的先验估计，而m是一称为等效样本大小的常量，可取属性个数
  - 在缺少其他信息时，选择p的一种典型的方法是均匀概率，比如某属性有k个可能值，那么p=1/k。

- 例（续）
- 前例中，P(Married| Class=Yes)=0
- 使用m-估计，m=3，p=1/3
- P(Married| Class=Yes)=(0+ 3× 1/3)/(3+3)= 1/6
- P(X|Class=No) = P(Refund=No|Class=No)P(Married| Class=No)P(Income=120K| Class=No) =6/10 × 6/10 ×0.0072= 0.0026
- P(X|Class=Yes) = P(Refund=No| Class=Yes) P(Married| Class=Yes) P(Income=120K| Class=Yes) =4/6 × 1/6 ×1.2 × 10^(-9)= 1.3 × 10^(-10)

- 朴素贝叶斯分类器假定了属性a1,…,an的值在给定目标值v下是条件独立的。
- 这一假定显著地减小了目标函数学习的计算复杂度。
- 当此条件成立时，朴素贝叶斯分类器可得到最优贝叶斯分类。
- 但在多数情况下，这一条件独立假定过于严厉了。

实验3：基于朴素贝叶斯的犯罪类型预测

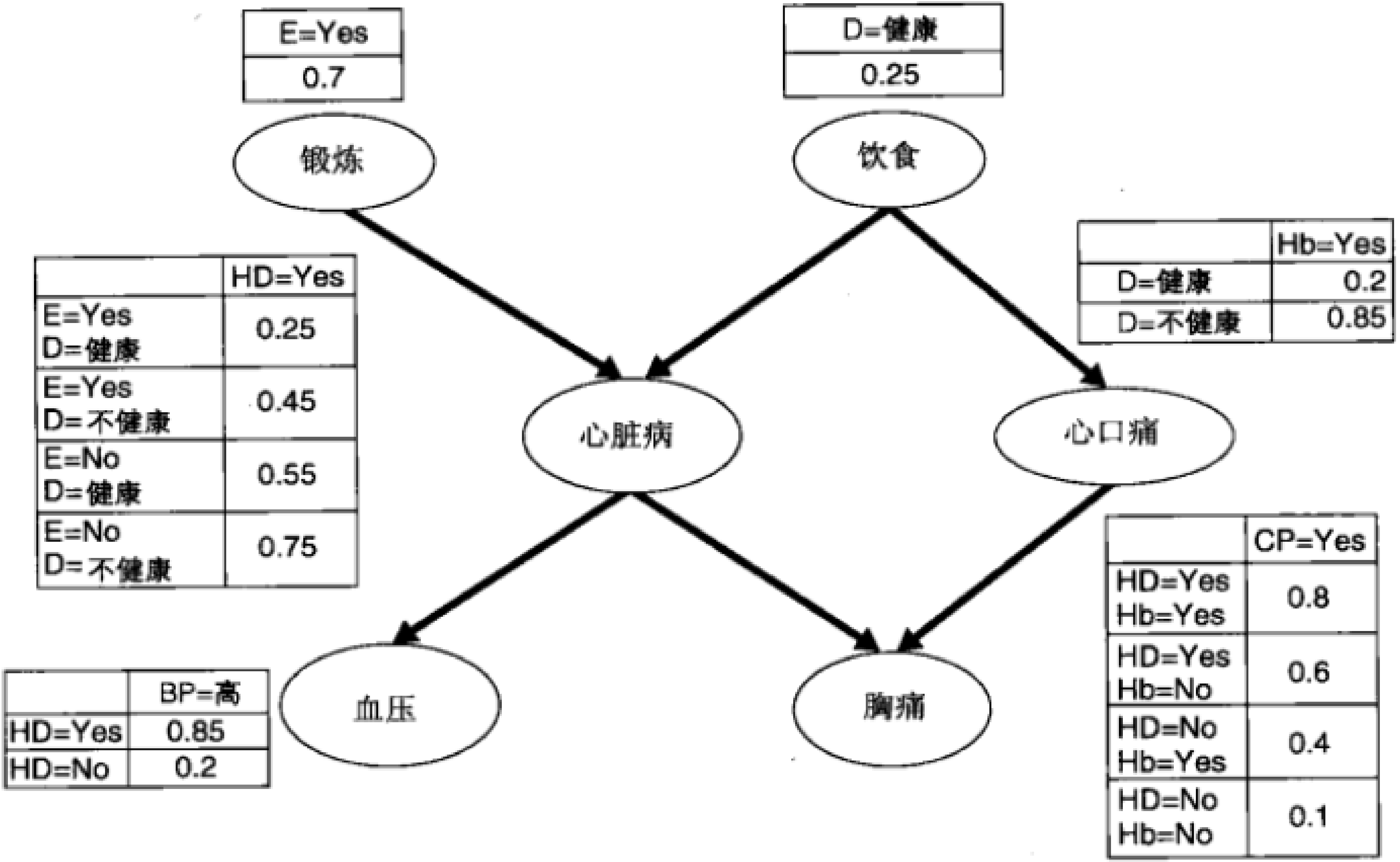
- 给定LA犯罪数据集
- 根据某案件的“案发时间”“所属警区”“案发地点”等属性，预测案件的类型，例如“抢劫”“偷窃”“杀人”等
- 属性均为离散的
- 共44948个训练样本，14983个测试样本
- Python编程实现

贝叶斯信念网络（BAYES BELIEF NETWORKS, BBN）

- 简称贝叶斯网络，表述变量的一个子集上的条件独立性假定。
- 贝叶斯网络中的一个节点，如果它的父母节点已知，则它条件独立与它的所有非后代节点。

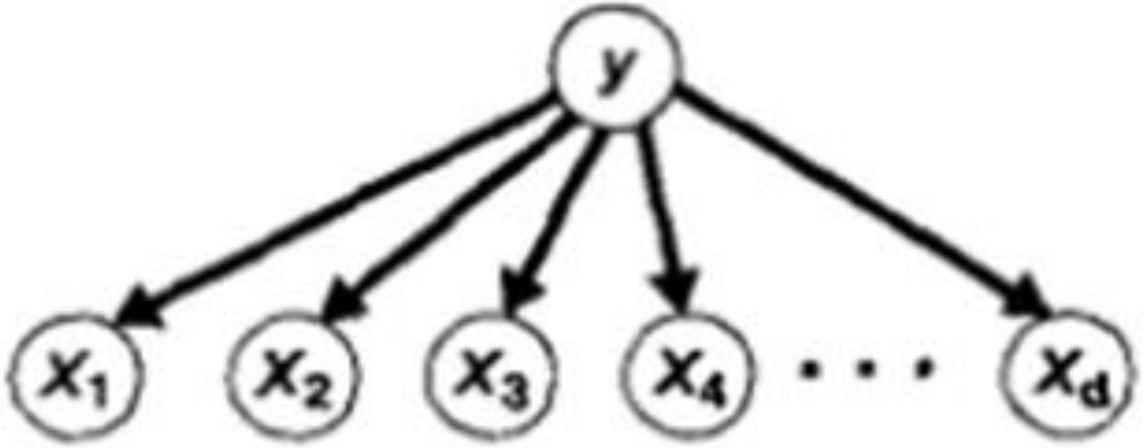
贝叶斯信念网络的表示

- 一个有向无环图(directed acyclic graph, or DAG)，指定一组条件独立性假定，表示变量间的依赖关系
  - 每个节点表示一个变量，每条弧表示两个变量间的依赖关系
  - 如从X到Y有一条有向弧，则X是Y的父母，Y是X的子女
  - 如网络中存在一条从X到Z的有向路径，则X是Z的祖先，Z是X的后代
- 一个概率表，即一组局部条件概率集合，把各节点和它的直接父节点关联起来
  - 如节点X没有父母节点，则表中只包含先验概率P(X)
  - 如节点X只有一个父母节点Y，则表中包含条件概率P(X|Y)
  - 如节点X有多个父母节点{Y1,…,YK}，则表中包含条件概率P(X| Y1,…,YK)

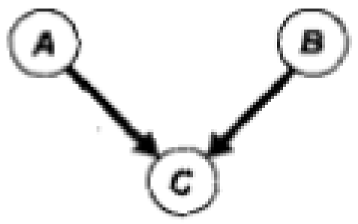


发现心脏病和心口痛病人的贝叶斯网络

朴素贝叶斯分类器中的条件独立假设也可以用贝叶斯网络表示。

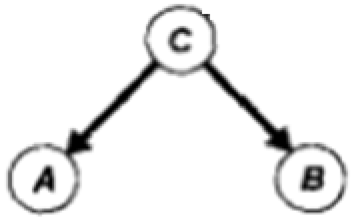


贝叶斯网络中三个变量间的典型依赖关系



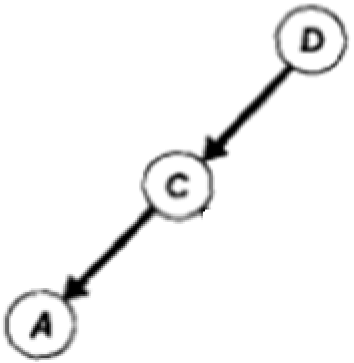
(a)

V型结构（V-Structure）、冲撞结构



(b)

“同父”结构（CommonParent）



(c)

顺序结构

节点

$B = \langle G, \Theta \rangle$

$\boxed{\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)}$

给定父结点集，贝叶斯网假设每个属性与它的非后裔属性独立，于是 $B = \langle G, \Theta \rangle$ 将属性 $x_1, x_2, \ldots, x_d$ 的联合概率分布定义为

$$P_B(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} .$$

随机变量，XI父节点的集合

- 学习贝叶斯信念网络
  - 从训练数据中学到贝叶斯信念网络，有多种讨论的框架：
    - 网络结构可以预先给出，或由训练数据中得到
    - 所有的网络变量可以直接从每个训练样例中观察到，或某些变量不能观察到
  - 如果网络结构已知且变量可以从训练样例中完全获得，那么得到条件概率表就比较简单
  - 如果网络结构已知，但只有一部分变量值能在数据中观察到，学习问题就困难多了。这类似于在人工神经网络中学习隐藏单元的权值
    - Russtll（1995）提出了一个简单的梯度上升过程以学习条件概率表中的项，相当于对表项搜索极大似然假设

- 学习贝叶斯网的结构
  - 如果贝叶斯网的结构未知，那么需要学习贝叶斯网的结构
  - 一种方法是“评分搜索”
    - 定义一个“评分函数（score function）”
    - 常基于信息论准则
    - 引入了归纳偏好

学习问题 ⟶ 数据压缩任务 ⟶ 找到能以最短编码长度描述训练数据的模型 { 模型自身 / 描述数据 }

选择综合编码长度最短的贝叶斯网——最小长度描述准则

计算编码贝叶斯网B所需的字节数

计算B所对应的概率分布描述D时所需的字节数

给定训练集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$, 贝叶斯网 $B=\langle G,\Theta\rangle$ 在 D 上的评分函数可写为

$$s(B \mid D)=f(\theta)|B|-LL(B \mid D) ,$$

其中, $|B|$ 是贝叶斯网的参数个数; $f(\theta)$ 表示描述每个参数 θ 所需的字节数;

$$LL(B \mid D)=\sum_{i=1}^{m} \log P_B\left(\boldsymbol{x}_i\right)$$

是贝叶斯网 B 的对数似然.

寻找一个贝叶斯网B使评分函数s(B|D)最小

$$s(B \mid D) = f(\theta)|B| - LL(B \mid D)$$

若 $f(\theta) = 1$, 即每个参数用 1 字节描述, 则得到 AIC (Akaike Information Criterion)评分函数

$$\mathrm{AIC}(B \mid D) = |B| - LL(B \mid D) \ .$$

若 $f(\theta) = \frac{1}{2}\log m$, 即每个参数用 $\frac{1}{2}\log m$ 字节描述, 则得到 BIC (Bayesian Information Criterion)评分函数

$$\mathrm{BIC}(B \mid D) = \frac{\log m}{2}|B| - LL(B \mid D) \ .$$

显然, 若 $f(\theta) = 0$, 即不计算对网络进行编码的长度, 则评分函数退化为负对数似然, $s(B \mid D) = -LL(B \mid D)$, 相应的, 学习任务退化为极大似然估计.

而参数 $\theta_{x_i|\pi_i}$ 能直接在训练数据 D 上通过经验估计获得, 即

$$\theta_{x_i|\pi_i} = \hat{P}_D(x_i \mid \pi_i) \ ,$$

其中 $\hat{P}_D(\cdot)$ 是 D 上的经验分布. 因此, 为了最小化评分函数 $s(B \mid D)$, 只需对网络结构进行搜索, 而候选结构的最优参数可直接在训练集上计算得到.

- 网络拓扑结构可以通过主观的领域专家知识编码获得。下面是归纳贝叶斯网络拓扑结构的一个系统的过程

算法 贝叶斯网络拓扑结构的生成算法

设 $T=(X_1, X_2, \cdots, X_d)$表示变量的全序

for j=1 to d **do**

  令 $X_{T(j)}$表示 T 中第 j 个次序最高的变量

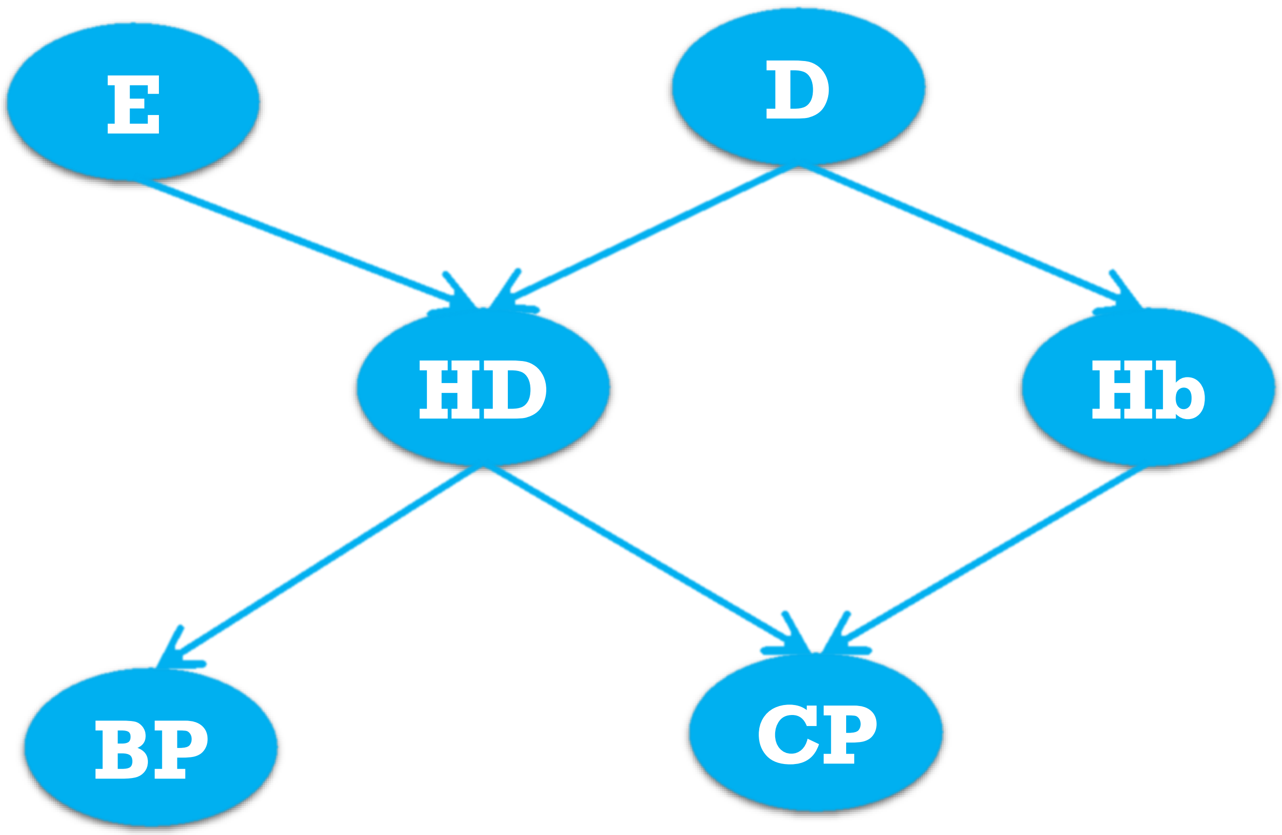
  令$\pi(X_{T(j)})=\{X_{T(1)}, X_{T(2)}, \cdots, X_{T(j-1)}\}$表示排在 $X_{T(j)}$前面的变量的集合

  从$\pi(X_{T(j)})$中去掉对 X_j 没有影响的变量（使用先验知识）

  在 $X_{T(j)}$和$\pi(X_{T(j)})$中剩余的变量之间画弧

end for

- 例：对心脏病或心口痛患者建模。假设每个变量都是二值的，心脏病节点（HD）的父母节点对应于影响该疾病的危险因素，如锻炼（E）和饮食（D）等，其子节点对应该病的症状，如胸痛（CP）和高血压（BP）等。心口痛（Hb）可能源于不健康饮食，同时有可能导致胸痛。
  - 执行步骤1后，设变量的次序为(E, D, HD, Hb, CP, BP)。经过步骤2到步骤7，得到如下条件概率
    - P(D|E)化简为P(D)
    - P(HD|E,D)不能化简
    - P(Hb|HD,E,D)化简为P(Hb|D)
    - P(CP|Hb,HD,E,D)化简为P(CP|Hb,HD)
    - P(BP|CP,Hb,HD,E,D)化简为P(BP|HD)
    - 基于以上条件概率，创建节点之间的弧(E,HD), (D,HD), (D,Hb), (HD,CP), (Hb,CP), (HD,BP)


E
D
HD
Hb
BP
CP

	S,B	$S,\neg B$	$\neg S,B$	$\neg S,\neg B$
C	0.4	0.1	0.8	0.2
$\neg C$	0.6	0.9	0.2	0.8

- 已知网络结构后:
- 可通过**梯度上升**来优化贝叶斯网络（最大化、概率）
  - 令w_{ijk}代表条件概率表的一个表项，即在给定父节点U_i取值u_{ik}时，网络变量Y_i值为y_{ij}的概率
  - 例如，w_{ijk}为最右上方的表项，那么Y_i为变量Campfire，U_i是其父节点的元组<Storm, BusTourGroup>，y_{ij}=True，且u_{ik}=<False, False>

- P(D|h)表示假设h成立的情形下观察到数据D的概率

	S,B	S,¬B	¬S,B	¬S,¬B
C	0.4	0.1	0.8	0.2
¬C	0.6	0.9	0.2	0.8

Campfire

- lnP(D|h)的梯度由对每个w_{ijk}求导数得到

$$\frac{\partial \ln P(D|h)}{\partial w_{ijk}} = \sum_{d \in D} \frac{P(Y_i = y_{ij}, U_i = u_{ik} | d)}{w_{ijk}}$$ ★

- 例如，为计算表右上方的表项的lnP(D|h)的导数，需要对D中每个训练样例d计算P(Campfire=True, Storm=False, BusTourGroup=False|d)
- 当训练样例中无法观察到这些变量时，这些概率可用标准的贝叶斯网络从d中观察到的变量中推理得到。

- 星式的推导
  - 用$P_h(D)$来表示$P(D|h)$
  - 假定在数据集D中的各样例d都是独立抽取的

$$
\begin{aligned}
\frac{\partial \ln P_h(D)}{\partial w_{ijk}} &= \frac{\partial}{\partial w_{ijk}} \ln \prod_{d \in D} P_h(d) \\
&= \sum_{d \in D} \frac{\partial \ln P_h(d)}{\partial w_{ijk}} \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial P_h(d)}{\partial w_{ijk}} \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial}{\partial w_{ijk}} \sum_{j',k'} P_h(d|y_{ij'}, u_{ik'}) P_h(y_{ij'}, u_{ik'}) \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial}{\partial w_{ijk}} \sum_{j',k'} P_h(d|y_{ij'}, u_{ik'}) P_h(y_{ij'}|u_{ik'})\, P_h(u_{ik'}) \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial}{\partial w_{ijk}} P_h(d|y_{ij}, u_{ik}) P_h(y_{ij}|u_{ik}) P_h(u_{ik}) \quad \leftarrow \text{仅对j,k求偏导} \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{\partial}{\partial w_{ijk}} P_h(d|y_{ij}, u_{ik}) w_{ijk} P_h(u_{ik}) \\
&= \sum_{d \in D} \frac{1}{P_h(d)} P_h(d|y_{ij}, u_{ik}) P_h(u_{ik}) \\
&= \sum_{d \in D} \frac{1}{P_h(d)} \frac{P_h(y_{ij}, u_{ik}|d) P_h(d) P_h(u_{ik})}{P_h(y_{ij}, u_{ik})} \\
&= \sum_{d \in D} \frac{P_h(y_{ij}, u_{ik}|d) P_h(u_{ik})}{P_h(y_{ij}, u_{ik})} \\
&= \sum_{d \in D} \frac{P_h(y_{ij}, u_{ik}|d)}{P_h(y_{ij}|u_{ik})} \\
&= \sum_{d \in D} \frac{P_h(y_{ij}, u_{ik}|d)}{w_{ijk}}
\end{aligned}
$$

- 更新权值

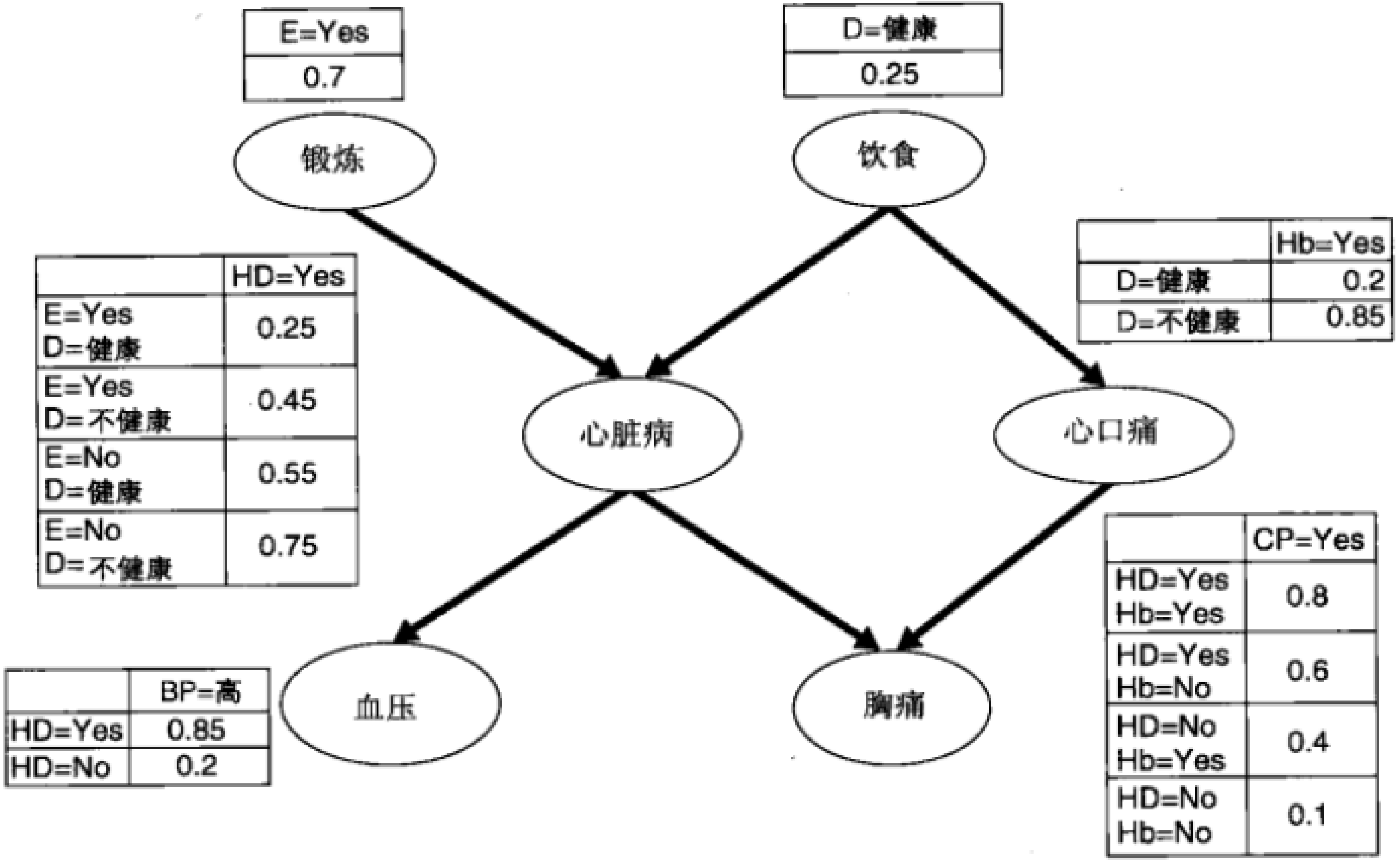
$$w_{ijk} \leftarrow w_{ijk} + \eta \sum_{d \in D} \frac{P_h(y_{ij}, u_{ik} | d)}{w_{ijk}}$$

- 归一化处理，保持在区间[0,1]之间，且$\sum_{j} \mathbf{w}_{\mathbf{ijk}}$对所有**i,k**保持为1

$$w_{ijk} \leftarrow \frac{w_{ijk}}{\sum_{j} w_{ijk}}$$

- 这个算法只保证找到局部最优解

- 例：利用下图的BBN诊断一个人是否有心脏病。



发现心脏病和心口痛病人的贝叶斯网络

情况1：没有先验信息。

- 可通过计算先验概率P(HD=Yes)和P(HD=No)来确定一个人是否可能患心脏病
- 为方便，设α∈{Yes, No}表示锻炼的两个值，β∈{健康，不健康}表示饮食的两个值

$$
\begin{aligned}
P(\mathrm{HD=Yes}) &= \sum_{\alpha}\sum_{\beta} P(\mathrm{HD=Yes} \mid E=\alpha, D=\beta)P(E=\alpha, D=\beta) \\
&= \sum_{\alpha}\sum_{\beta} P(\mathrm{HD=Yes} \mid E=\alpha, D=\beta)P(E=\alpha)P(D=\beta) \\
&= 0.25\times0.7\times0.25 + 0.45\times0.7\times0.75 + 0.55\times0.3\times0.25 + 0.75\times0.3\times0.75 \\
&= 0.49
\end{aligned}
$$

$$P(\mathrm{HD=No}) = 1 - P(\mathrm{HD=Yes}) = 0.51$$

此人不得心脏病的机率略微大一点。

情况2：已知该人有高血压。

- 此时可以通过比较后验概率P(HD=Yes|BP=高)和P(HD=No|BP=高)来诊断他是否患有心脏病
- 记$\gamma \in \{Yes, No\}$表示心脏病的两个值

$$P(\text{BP}=\text{高}) = \sum_{\gamma} P(\text{BP}=\text{高} \mid \text{HD}=\gamma) P(\text{HD}=\gamma)$$

$$= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185$$

因此，此人患心脏病的后验概率是：

$$P(\text{HD}=\text{Yes} \mid \text{BP}=\text{高}) = \frac{P(\text{BP}=\text{高} \mid \text{HD}=\text{Yes}) P(\text{HD}=\text{Yes})}{P(\text{BP}=\text{高})}$$

$$= \frac{0.85 \times 0.49}{0.5185} = 0.8033$$

$$P(\text{HD}=\text{No} \mid \text{BP}=\text{高}) = 1 - 0.8033 = 0.1967$$

因此，当一个人有高血压时，他患心脏病的危险就增加了。

情况3：又知此人常锻炼身体且饮食健康，这些新信息会对诊断造成何种影响？

- 加入新信息后，此人患心脏病的后验概率为

$$P(\text{HD=Yes}|\text{BP=高}, D=\text{健康}, E=\text{Yes})$$

$$= \left[\frac{P(\text{BP}=\text{高}\,|\,\text{HD}=\text{Yes}, D=\text{健康}, E=\text{Yes})}{P(\text{BP}=\text{高}\,|\,D=\text{健康}, E=\text{Yes})}\right] \times P(\text{HD}=\text{Yes}\,|\,D=\text{健康}, E=\text{Yes})$$

$$= \frac{P(\text{BP}=\text{高}\,|\,\text{HD}=\text{Yes})P(\text{HD}=\text{Yes}\,|\,D=\text{健康}, E=\text{Yes})}{\sum_\gamma P(\text{BP}=\text{高}\,|\,\text{HD}=\gamma)P(\text{HD}=\gamma\,|\,D=\text{健康}, E=\text{Yes})}$$

$$= \frac{0.85\times0.25}{0.85\times0.25+0.2\times0.75}$$

$$= 0.5862$$

而此人不患心脏病的概率是：

$$P(\text{HD=No}|\ \text{BP}=\text{高}, D=\text{健康}, E=\text{Yes}) = 1-0.5862 = 0.4138$$

因此模型暗示健康的饮食和有规律的体育锻炼可以降低患心脏病的危险。

5 非线性分类器

- 5.1 多类问题概述
- 5.2 最小距离分类器
- 5.3 分段线性分类器概述
- 5.4 决策树
- 5.5 近邻法分类器
- 5.6 Adaboost
- 5.7 SVM
- 5.8 人工神经网络

5.1 多类问题概述

- 5.1.1 多类问题
- 5.1.2 解决方案

5.1.1 多类问题

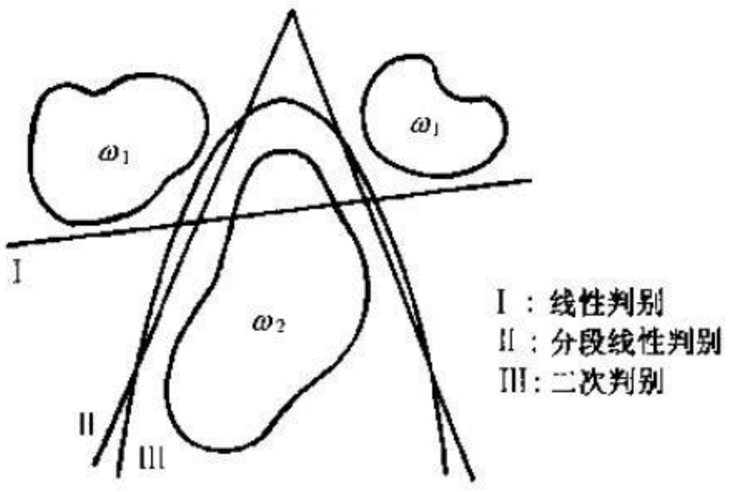
- **多类问题包括**
  - 两类情况（C = 2）
    - 样本集具有多峰分布

5.1.1 多类问题概述

- **多类问题包括**
  - 多类情况（类别数C > 2）
    - 单峰分布
    - 多峰分布

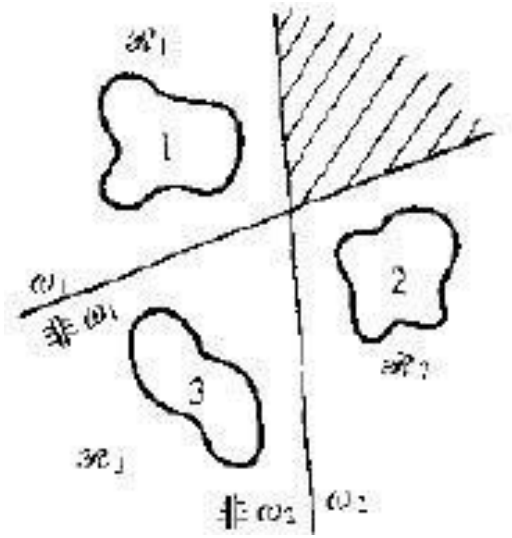
5.1.2 解决方案

- **如何解多类问题**
  - Bayes分类器
    - 二次型判别函数
  - 线性分类器
  - 其它分类器



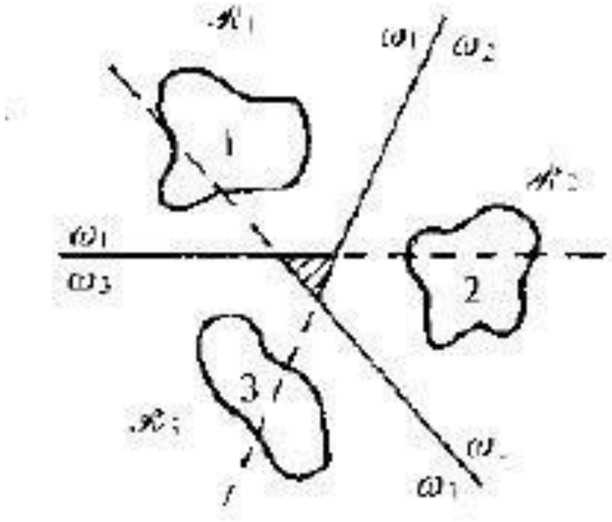
5.1.2 解决方案

- **多类问题是否可以用线性分类器解？**
  - 思路一
    - 对“类与类的非”进行线性分类
    - 只需要C－1个线性分类器就可以



5.1.2 解决方案

- **多类问题是否可以用线性分类器解？**
  - 思路二
    - “两两分类”进行线性分类
    - 需要C(C – 1)/2个线性分类器就可以

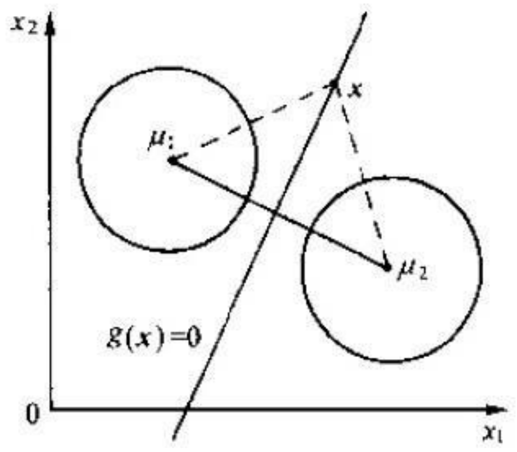


5.2 最小距离分类器

- 5.2.1 最小距离分类器原理
- 5.2.2 分段最小距离分类器
- 5.2.3 特点

5.2.1 最小距离分类器原理

- **回顾两类单峰线性分类器**
  - 垂直平分 / 最小距离分类器
  - 基于两类样本均值点作垂直平分线

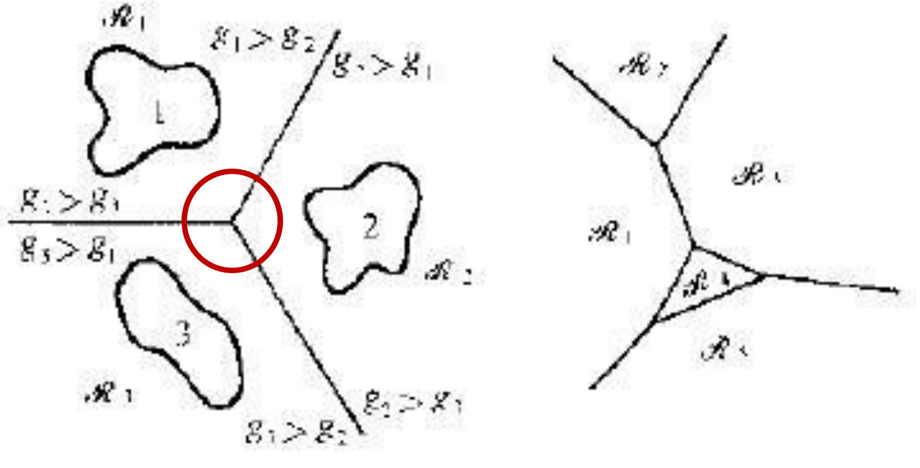


5.2.1 最小距离分类器原理

- **其最小距离形式**
  - 判别函数
    - $G_1(x) = d_1(x) = \|x - m_1\|$
    - $G_2(x) = d_2(x) = \|x - m_2\|$
  - 决策规则
    - 对于未知样本x，若$d_1(x) < d_2(x)$，则x决策为ω_1类
    - 若$d_1(x) > d_2(x)$，则x决策为ω_2类

5.2.1 最小距离分类器原理

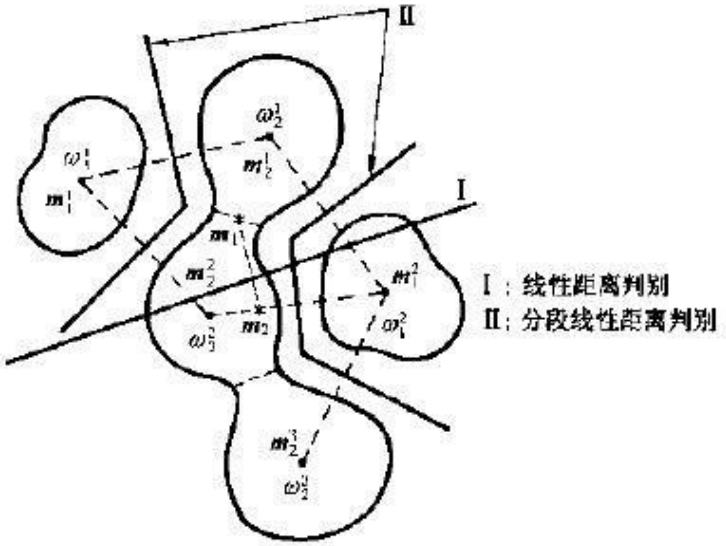
- **直接使用可以解决多类问题**
  - 解决C类单峰问题



(a) 三类　　(b) 五类

5.2.1 最小距离分类器原理

- **直接使用可以解决多类问题**
  - 解决两类多峰问题



5.2.2 分段最小距离分类器

- **问题**
  - 已知各类及其子类
  - 求分段最小距离分类器

5.2.2 分段最小距离分类器

- **分类器设计**
  - 先求各子类均值
    - m_{ij} （ω_i类的第j子类）
  - 定义各类判别函数
    - $G_i(x) = \min\limits_{j} \|x - m_{ij}\|$
  - 决策规则
    - 对于未知样本x，若$G_k(x) = \min\limits_{i} G_i(x)$，则x决策为$\omega_k$类

5.2.3 特点

- **分类器特点**
  - 解决两类多峰或多类问题的分段线性分类器
  - 可以解决几乎所有分类问题但要已知各类子类
  - 概念直观简单，未经优化
  - 分类器设计简单容易
  - （无重叠区或空白区）

5.3 分段线性分类器概述

- **5.3.1 问题与思路**
- **5.3.2 设计说明**

5.3.1 问题与思路

- **思路**
  - 参考分段最小距离分类器
    - 定义判别函数
    - 定义决策规则

5.3.1 问题与思路

- **针对不同已知条件**
  - 1、已知各类子类个数及子类分布区域
  - 2、已知各类子类个数（分布区域不知）
  - 3、一般情况（子类个数和分布区域均不知）

5.3.2 设计说明

- **两种判别函数的区别**
  - 小写g函数
    - 每段设计一个g函数，容易做
    - 多个分段，如何判断正负侧，需要特殊规则
  - 大写G函数
    - 每个子类设计一个G函数，需要知道类别分布区域
    - 直接计算Max或Min，判别规则简单

5.3.2 设计说明

- **设计关键**
  - 如何确定各类的子类个数
  - 如何确定子类的分布区域
  - 如何求解各子类的权向量和阈值权
  - 若采用小写g函数，决策规则如何

THE END !