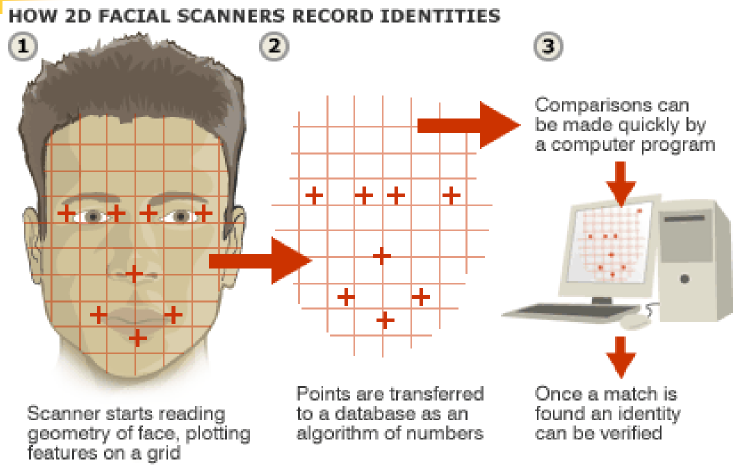
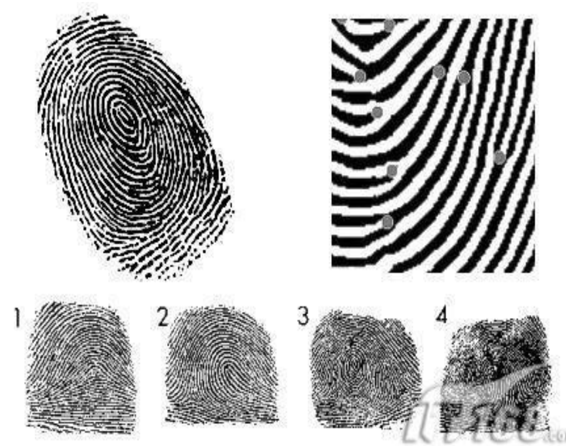
第8章 文本分类与倾向性分析（1/2）

北京市海淀区中关村东路95号
邮编：100190

电话： +86-10-8254 4588
邮件：jjzhang@nlpr.ia.ac.cn

真实生活中的模式识别问题

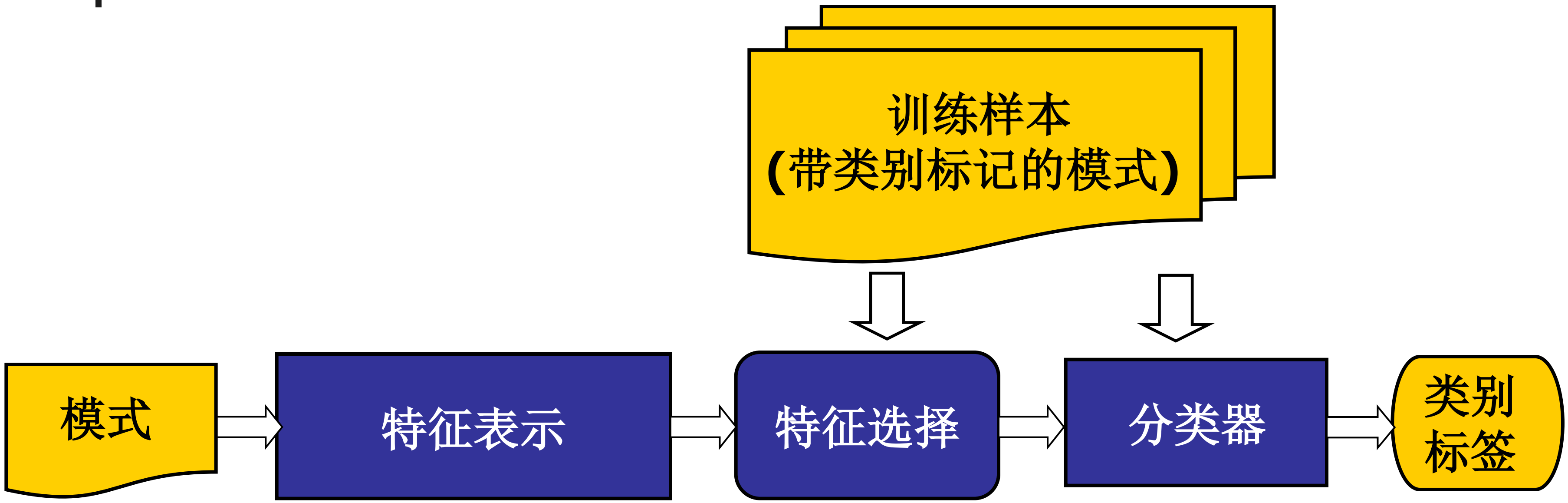





Sent Mail
Spam (372)
Trash

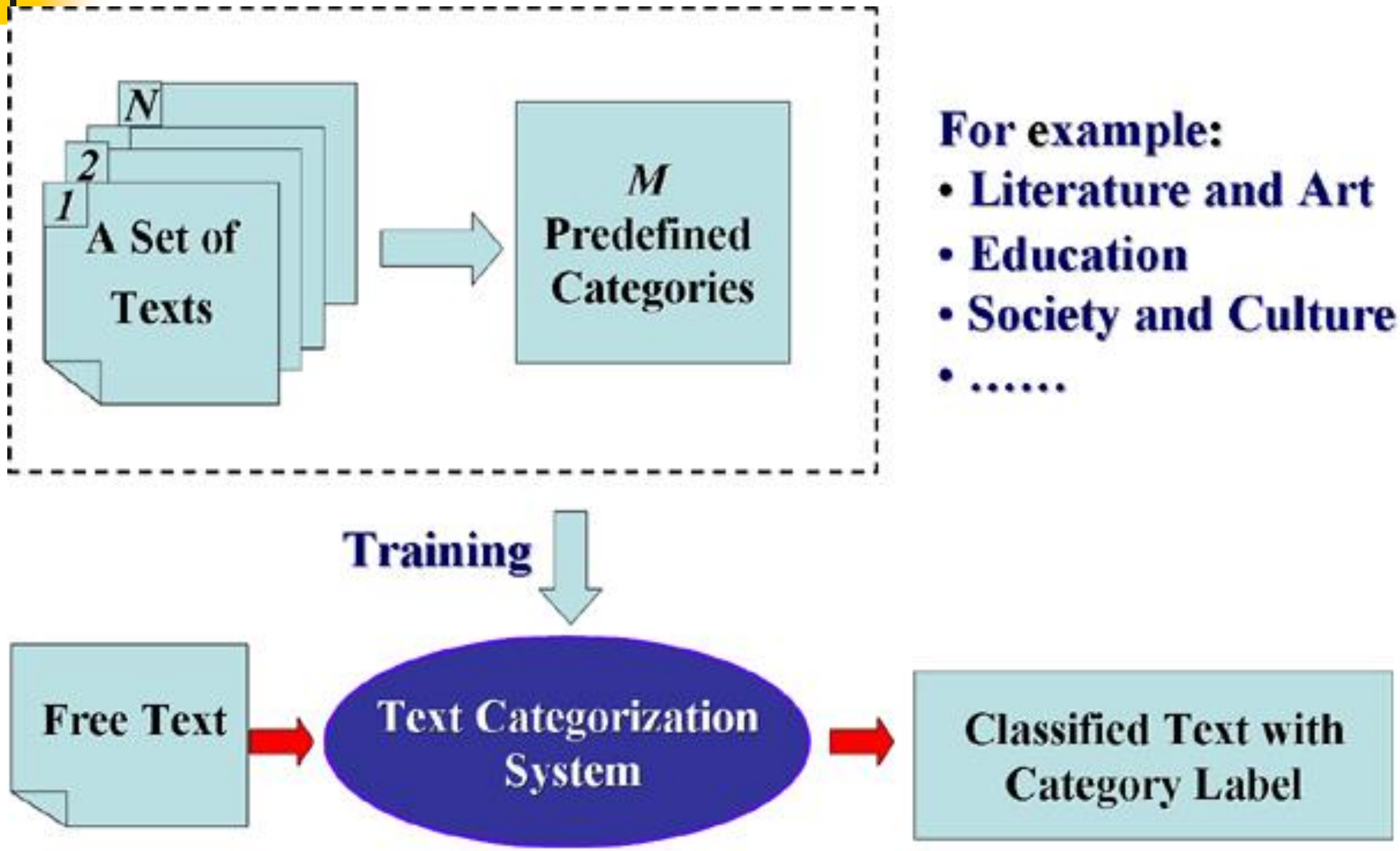

●●○○○ 中国移动 4G
22:51
29%
科技 体育 财经 军事 文化 旅游 ＋
俄T-90坦克在叙利亚被导弹打爆 乘员逃生
图片 手机和讯网 276评论
许世友因轻敌在越南遭受重创，战后他发誓不再进北京
百代旅行家 341评论 20分钟前
中国东风-21D导弹到底有多厉害？外媒一张图把国人惊呆了
热 迷彩先生 84评论 30分钟前
为何打仗需要它：解放军开始佩戴新型身份识别牌全面与美军接轨
多图
战略吐槽秀 3100评论 40分钟前
中国为何突然曝光绝密战机？俄罗斯空军表示望尘莫及
首页
视频
话题
我的

模式识别系统的基本框架



$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{x} \qquad \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_m \end{bmatrix} = \hat{\mathbf{x}} \qquad y$$

文本分类系统的基本框架



主要内容

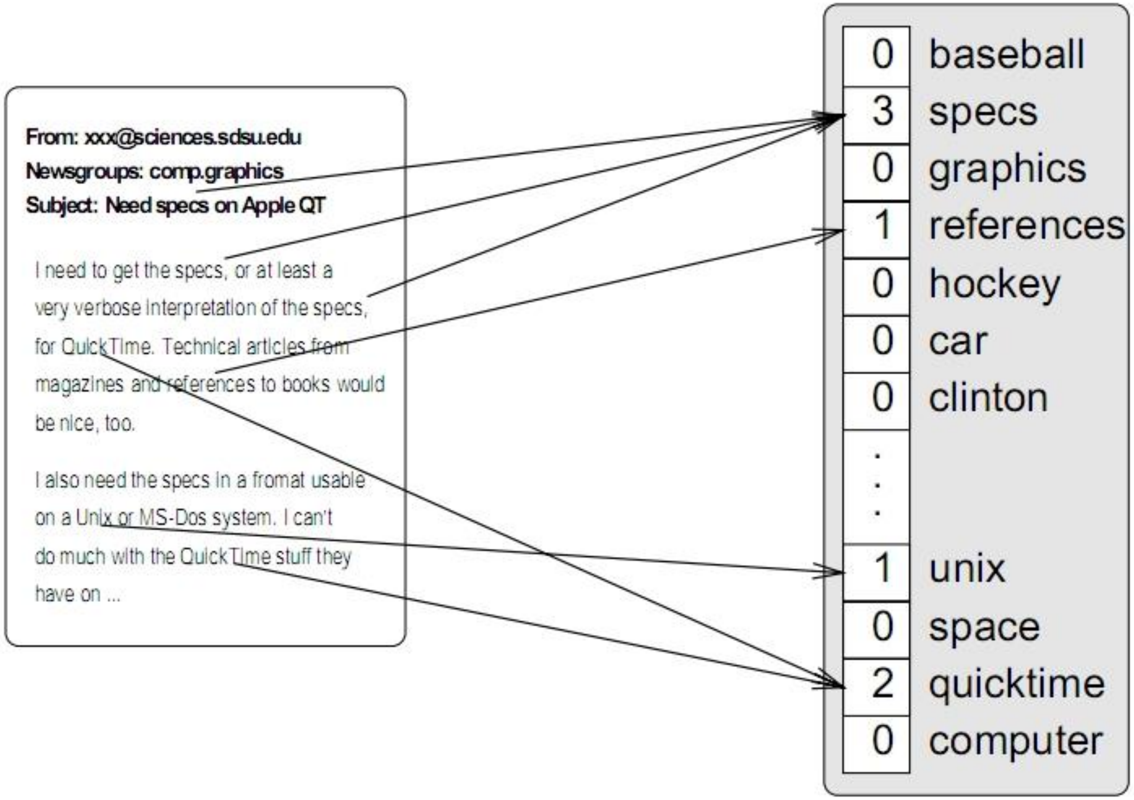
- 文本分类
  - 文本表示
  - 特征选择
  - 分类算法

文本表示-离散表示

- 向量空间模型（Vector Space Model, VSM）
  - 也称为词袋模型（Bag-of-Words Model, BOW）

文本表示-离散表示

- 向量空间模型（Vector Space Model，VSM）
  - 也称为词袋模型（Bag-of-Words Model，BOW）



词的权重

- 词频（Term Frequency，TF）

$$\omega_{ki} = tf_{ki}$$

- 布尔变量（是否出现）

$$\omega_{ki} = \begin{cases} 1, & if\ \ t_i\ exists\ in\ \mathbf{d}_k \\ 0, & otherwise \end{cases}$$

- 逆文档频率（Inverse Document Frequency, IDF）

$$\omega_i = \log \frac{N}{df_i}$$

- TF-IDF

$$\omega_{ki} = tf_{ki} \cdot \log \frac{N}{df_i}$$

一个文本表示的例子

- **训练数据（带类别标签的文档）**

教育	体育
北京 理工 大学 计算机 专业 创建 于 1958 年 是 中国 最早 设立 计算 机 专业 的 高校 之一	北京 理工 大学 体育馆 是 2008 年 中国 北京 奥 林匹克 运动会 的 排球 预赛 场地
北京 理工 大学 学子 在 第四 届 中国 计算机 博 弈 锦标赛 中 夺冠	第五 届 东亚 运动会 中 国 军团 奖牌 总数 创 新 高 男女 排球 双双 夺冠

一个文本表示的例子

- **词袋表示（含40个词，即词表大小为40）**

1958 2008 奥林匹克 北京 博弈 场地 创 创建 大学 的 第四 第五 东亚 夺冠 高校 计算机 奖牌 届 锦标赛 军团 理工 男女 年 排球 设立 是 双双 体育馆 新高 学子 于 预赛 运动会 在 之一 中 中国 专业 总数 最早

主要内容

- 文本分类
  - 文本表示
  - 特征选择
  - 分类算法

特征选择（特征过滤）

- 文本分类
  - 文本表示
  - 特征选择
    - 文档频率（Document Frequency，DF）
    - 互信息（Mutual Information，MI）
    - 信息增益（Information Gain，IG）
    - Chi-Square统计（Chi-Square Statistics，CHI）
  - 分类器设计

相关概率估计

- A 关于特征 t_i 与类别 c_j的统计表

特征 \ 类别	c_j	$\overline{c}_j$
t_i	A_{ij}	B_{ij}
$\overline{t}_i$	C_{ij}	D_{ij}

$$P(c_j) \approx (A_{ij} + C_{ij}) / N_{all}$$

$$P(t_i) \approx (A_{ij} + B_{ij}) / N_{all}$$

$$P(\overline{t}_i) \approx (C_{ij} + D_{ij}) / N_{all}$$

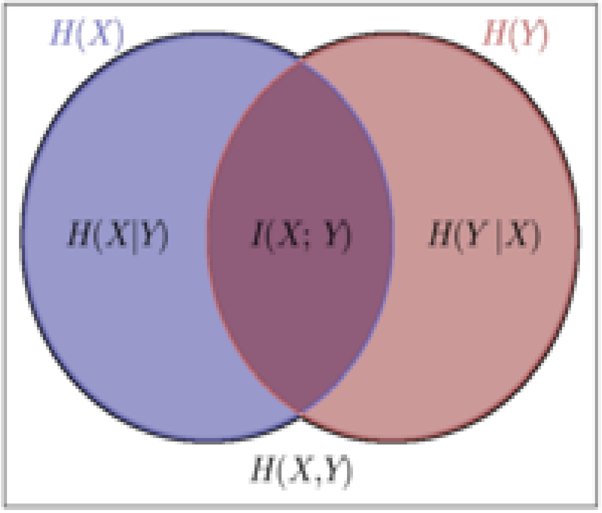
$$P(c_j \mid t_i) \approx \frac{A_{ij} + 1}{A_{ij} + B_{ij} + C}$$

$$P(c_j \mid \overline{t}_i) \approx \frac{C_{ij} + 1}{C_{ij} + D_{ij} + C}$$

相关信息论概念

- **熵（Entropy）**

$$H(X) = -\sum_{x} p(x)\log p(x)$$



- **联合熵（Joint Entropy）**

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y)$$

- **条件熵（Conditional Entropy）**

$$H(Y|X) = \sum_{x} p(x)H(Y|X=x) = -\sum_{x}\sum_{y} p(x,y)\log p(y|x)$$

$$H(Y|X) = H(X,Y) - H(X)$$

特征选择–互信息

- 互信息（Mutual Information，MI）

互信息是关于两个随机变量互相依赖程度的一种度量

$$I(X,Y)=H(X)-H(X\mid Y)=\sum_{y}\sum_{x}P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

$$MI(t_i,c_j)=\log\frac{P(t_i,c_j)}{P(t_i)P(c_j)}\approx\log\frac{A_{ij}N_{all}}{(A_{ij}+C_{ij})(A_{ij}+B_{ij})}$$

$$MI_{avg}(t_i)=\sum_{j=1}^{C}P(c_j)MI(t_i,c_j)$$

特征选择–信息增益

- **信息增益 (IG)**

$$
\begin{aligned}
&IG(t_i)\\
&=\{-\sum_{j=1}^{C} P(c_j)\log P(c_j)\}\\
&\quad+\{P(t_i)[\sum_{i=1}^{C} P(c_j \mid t_i)\log P(c_j \mid t_i)]\\
&\quad+P(\overline{t_i})[\sum_{j=1}^{C} P(c_j \mid \overline{t_i})\log P(c_j \mid \overline{t_i})]\}
\end{aligned}
$$

IG衡量特征能够为分类系统带来多少信息

- **信息增益与互信息的关系**

$$
IG(t_i)=\sum_{j=1}^{C} P(t_i,c_j)MI(t_i,c_j)+\sum_{j=1}^{C} P(\overline{t_i},c_j)MI(\overline{t_i},c_j)
$$

“计算机”的信息增益

特征 \ 类别	教育	体育
计算机	2	0
$\overline{\text{计算机}}$	0	2

$P(\text{计算机})=1/2 \quad P(\overline{\text{计算机}})=1/2$

$P(\text{教育}\mid\text{计算机})=(2+1)/(2+2)=3/4$

$P(\text{体育}\mid\text{计算机})=1/(2+2)=1/4$

$P(\text{教育}\mid\overline{\text{计算机}})=1/(2+2)=1/4$

$P(\text{体育}\mid\overline{\text{计算机}})=(2+1)/(2+2)=3/4$

$$\begin{aligned}\text{IG}(\text{算机}) = & -0.5\log 0.5-0.5\log 0.5 \\ & +0.5(0.75\log 0.75+0.25\log 0.25) \\ & +0.5(0.75\log 0.75+0.25\log 0.25) \\ = & -\log 0.5+0.75\log 0.75+0.25\log 0.25=0.1308\end{aligned}$$

"北京"的信息增益

class / feature	教育	体育
北京	*2*	*1*
$\overline{\text{北京}}$	*0*	*1*

$$P(\text{北京}) = (1+2)/4 = 3/4$$

$$P(\overline{\text{北京}}) = 1/4$$

$$P(\text{教育} \mid \text{北京}) = (2+1)/(3+2) = 3/5$$

$$P(\text{体育} \mid \text{北京}) = (1+1)/(3+2) = 1/5$$

$$P(\text{教育} \mid \overline{\text{北京}}) = 1/(1+2) = 1/3$$

$$P(\text{体育} \mid \overline{\text{北京}}) = (1+1)/(1+2) = 2/3$$

$$\begin{aligned} & IG(\text{北京}) \\ & = -0.5\log 0.5 - 0.5\log 0.5 \\ & \quad +0.75(0.6\log 0.6 + 0.4\log 0.4) \\ & \quad +0.25(0.667\log 0.667 + 0.333\log 0.333) \\ & = 0.0293 \end{aligned}$$

信息增益的例子

根据信息增益的特征排序

Features	IG
计算机 排球 运动会	0.1308
1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 高校 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 大学 理工	0.0293
的 夺冠 届 年 是 中国	0.0000

信息增益的例子

- 选择的特征

计算机 排球 运动会 高校 大学 1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 理工

- 精简后的训练数据

教育	体育
大学 计算机 计算机 高校	大学 运动会 排球
大学 计算机	运动会 排球

主要内容

- 文本分类
  - 文本表示
  - 特征选择
  - 分类算法
    - 朴素贝叶斯（Naïve Bayes）
    - 线性判别函数（Linear Discriminate Function）

分类算法

- 监督学习
  - 生成式模型
    - 朴素贝叶斯（Naïve Bayes）
  - 判别式模型
    - 线性判别函数（Linear Discriminate Function）
    - 支持向量机（Support Vector Machine）
    - 最大熵模型（Maximum Entropy）
- 无监督、半监督学习

分类算法相关概念

- 模型表示
  - 用参数进行建模（构建目标函数）
- 学习算法
  - 最大似然、最大后验（生成式模型）
  - 梯度下降、牛顿法（判别式模型）
- 推断
  - 决策/预测规则

监督学习过程

- 我们有什么？
  - 训练数据
- 我们的任务是什么？
  - 利用**参数**构建**模型（目标函数）**
  - **参数**需要**估计**
- 如何进行参数估计？
  - 根据某个**准则**从训练数据中学习
  - 学习在训练数据上**准则**最优的**参数**

$$y = f(x;\theta)$$

$$\uparrow$$

$$\theta?$$

$$\uparrow$$

$$\theta := \theta + \nabla f$$

贝叶斯决策理论

- 贝叶斯理论

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

- 贝叶斯决策理论

$$P(c_j|\mathbf{x})=\frac{P(c_j,\mathbf{x})}{P(\mathbf{x})}=\frac{P(c_j)P(\mathbf{x}|c_j)}{P(\mathbf{x})}$$

$$c^*=\arg\max_{j=1,\ldots,C}P(c_j|\mathbf{x})=\arg\max_{j=1,\ldots,C}P(c_j)P(\mathbf{x}|c_j)$$

贝叶斯模型是理论上最优的分类器!

朴素贝叶斯分类器

- 学习难点

$$P(\mathbf{x} \mid c_j) = ???$$

- 朴素贝叶斯假设

$$P(\boldsymbol{X}|c_j) \approx P([w_1, \cdots, w_n]|c_j) \approx \prod_{k=1}^{N} P(w_k|c_j) = \prod_{i=1}^{M} P(w_i|c_j)^{N(w_i)}$$

- 朴素贝叶斯决策规则 (模型)

$$P(c_j \mid \mathbf{x}) = \frac{P(\mathbf{x}, c_j)}{P(\mathbf{x})} \propto P(\mathbf{x}, c_j) = P(c_j) \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

$$c^* = \arg\max_{j=1,\ldots,C} P(c_j) \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

为什么是生成式!

NB模型中的参数估计

- 最大似然估计

$$P(c_j) \approx \frac{1+N(c_j)}{C+N_{all}} \qquad P(w_i \mid c_j) \approx \frac{1+N(w_i,c_j)}{M+\sum_{i'=1}^{M} N(w_{i'},c_j)}$$

- NB模型一个例子

$P(c_j)$	P(教育)=0.5	P(体育)=0.5
$P(w_i\|c_j)$	P(计算机|教育)=0.3	P(计算机|体育)=0.1
	P(排球|教育)=0.1	P(排球|体育)=0.3
	P(运动会|教育)=0.1	P(运动会|体育)=0.3
	P(高校|教育)=0.2	P(高校|体育)=0.1
	P(大学|教育)=0.3	P(大学|体育)=0.2

NB决策的例子

- "北京 理工 大学 是 理工 为主 工理文 协调 发展 的 全国 重点 高校"

Feature Set = [计算机， 排球， 运动会， 高校， 大学]

$$\mathbf{x} = [0,0,0,1,1]^T$$

$$P(\text{教育})P(\mathbf{x} \mid \text{教育}) = 0.5 \times 0.3 \times 0.2 = 0.03$$

$$P(\text{体育})P(\mathbf{x} \mid \text{体育}) = 0.5 \times 0.1 \times 0.2 = 0.01$$

$$P(\text{教育} \mid \mathbf{x}) = \frac{0.03}{0.03 + 0.01} = 0.75$$

$$P(\text{体育} \mid \mathbf{x}) = 0.25$$

NB决策的例子

“复旦 大学 排球 队 获得 本届 大学生 运动会 排球 比赛 冠军”

Feature Set = [计算机， 排球， 运动会， 高校， 大学]

$$\mathbf{x} = [0,1,1,0,1]^T$$

$$P(教育)P(\mathbf{x} \mid 教育) = 0.5 \times 0.1 \times 0.1 \times 0.3 = 0.0015$$

$$P(体育)P(\mathbf{x} \mid 体育) = 0.5 \times 0.3 \times 0.3 \times 0.2 = 0.0090$$

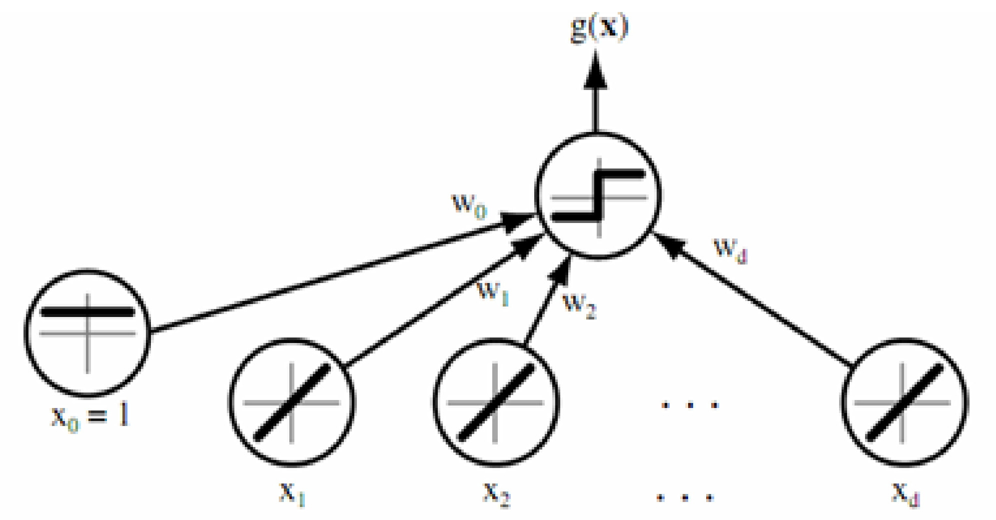
$$P(教育 \mid \mathbf{x}) = \frac{0.0015}{0.0015 + 0.0090} = 0.1429$$

$$P(体育 \mid \mathbf{x}) = 0.8571$$

线性判别函数

- 模型表示

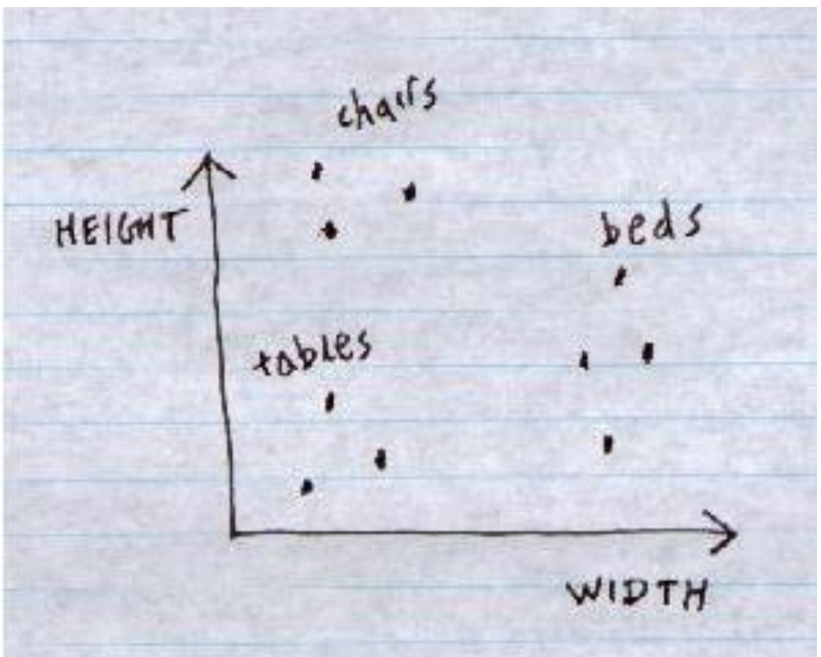
$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = \sum_{l=1}^{M} w_l x_l + w_0$$

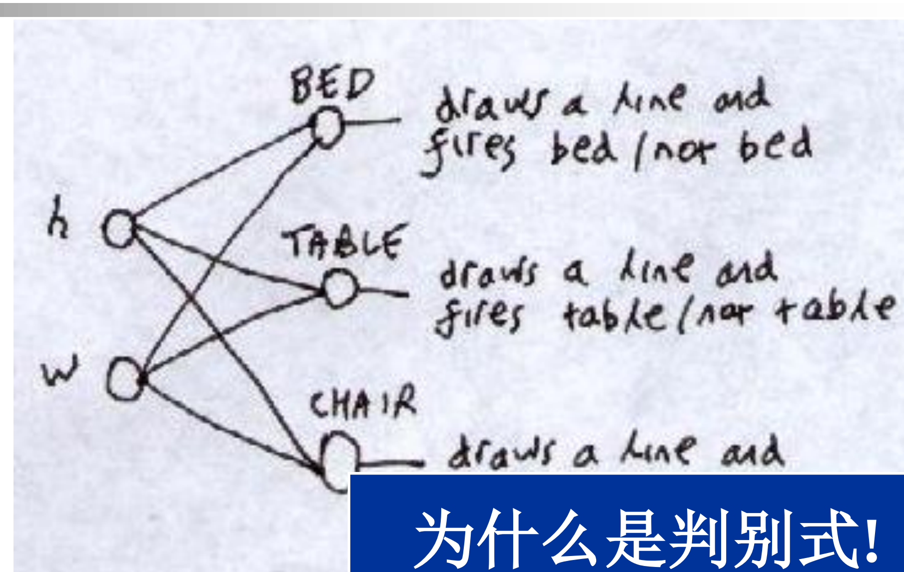


线性判别函数对应一个线性决策面

线性判别函数的一个例子

特征: ***height, width***
类别: ***bed, table, chair***





为什么是判别式!

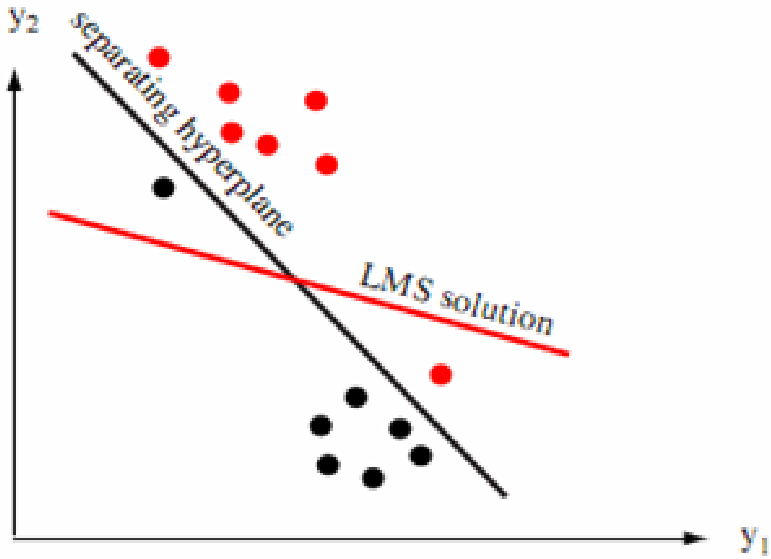
判别函数:

$$g_j(\mathbf{x}) = p(c_j \mid \mathbf{x})$$

$$= \sum_{l=1}^{2} w_{jl} x_l + w_{j0}$$

$$= \sum_{l=0}^{2} w_{jl} x_l, \ j = 0, 1, 2$$

线性判别函数的学习准则



- 感知器准则
- 最小均方差 (LMS)
- 交叉熵 (CE)
- 最小分类错误率 (MCE)
- …

哪个分类面更优?

选择哪个学习准则?

参数优化方法

- **模型** $g_j(\mathbf{x}) = \sum_{l=0}^{M} w_{jl} x_l$
- **准则**

$$J_{lms} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left[I(y_i = j) - f_j(\mathbf{x}_i)\right]^2,$$

$$\text{where } f_j(\mathbf{x}) = \frac{1}{1+\exp\{-(\sum_{l=0}^{M} w_{jl}\mathbf{x}_l)\}}$$

- **（随机）梯度下降**

$$w_{mn}(k+1) = w_{mn}(k) - \eta(k)\frac{\partial J}{\partial w_{mn}}$$

线性支持向量机

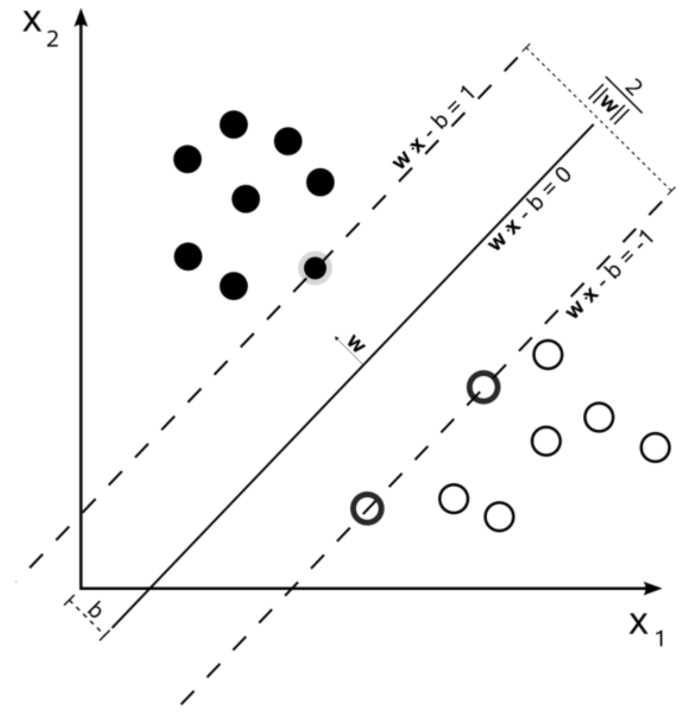
- **判别函数**

$$y = \mathbf{w}^T \mathbf{x} + b$$

- **最大间隔准则**

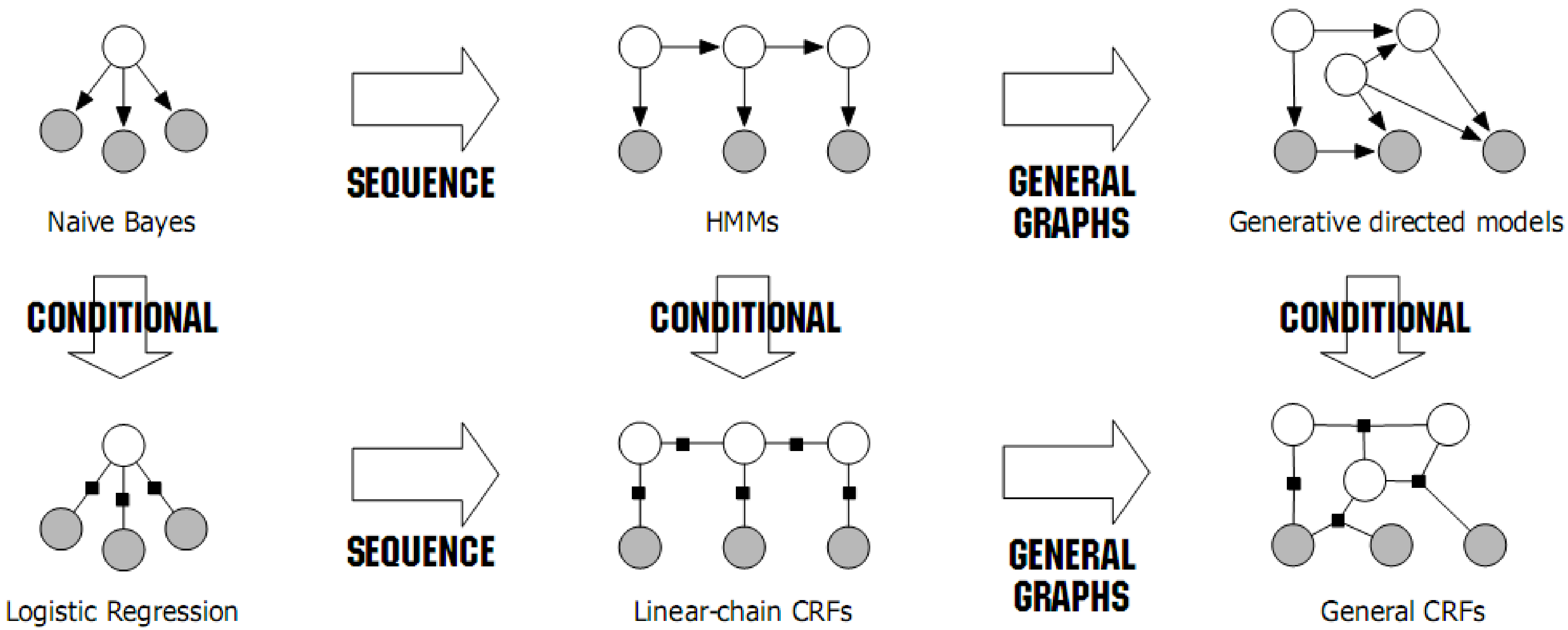
$$\min \frac{1}{2} \| \mathbf{w} \|^2$$

$$s.t.\ \ y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1,\ i = 1, \dots, n$$



线性支持向量机可视为一种线性判别函数!

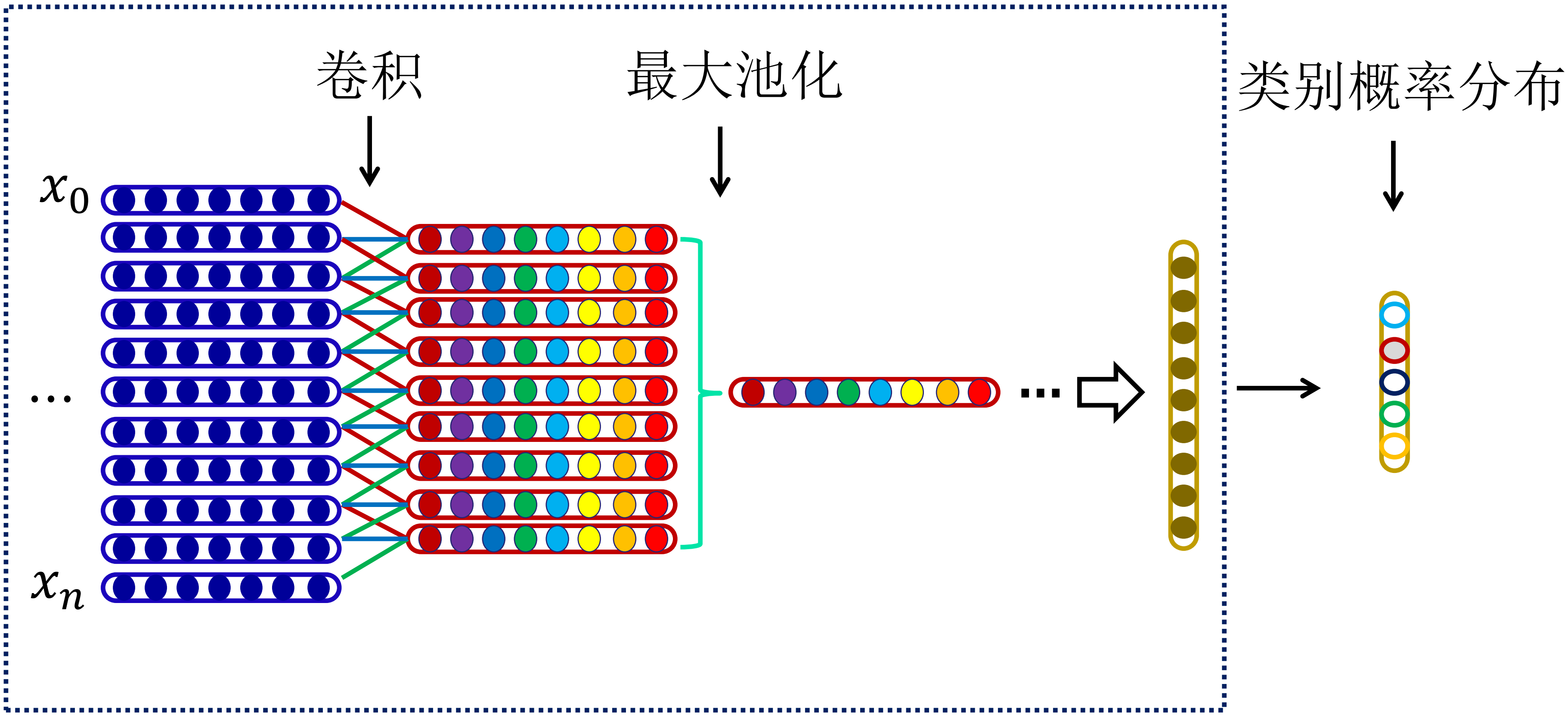
机器学习算法



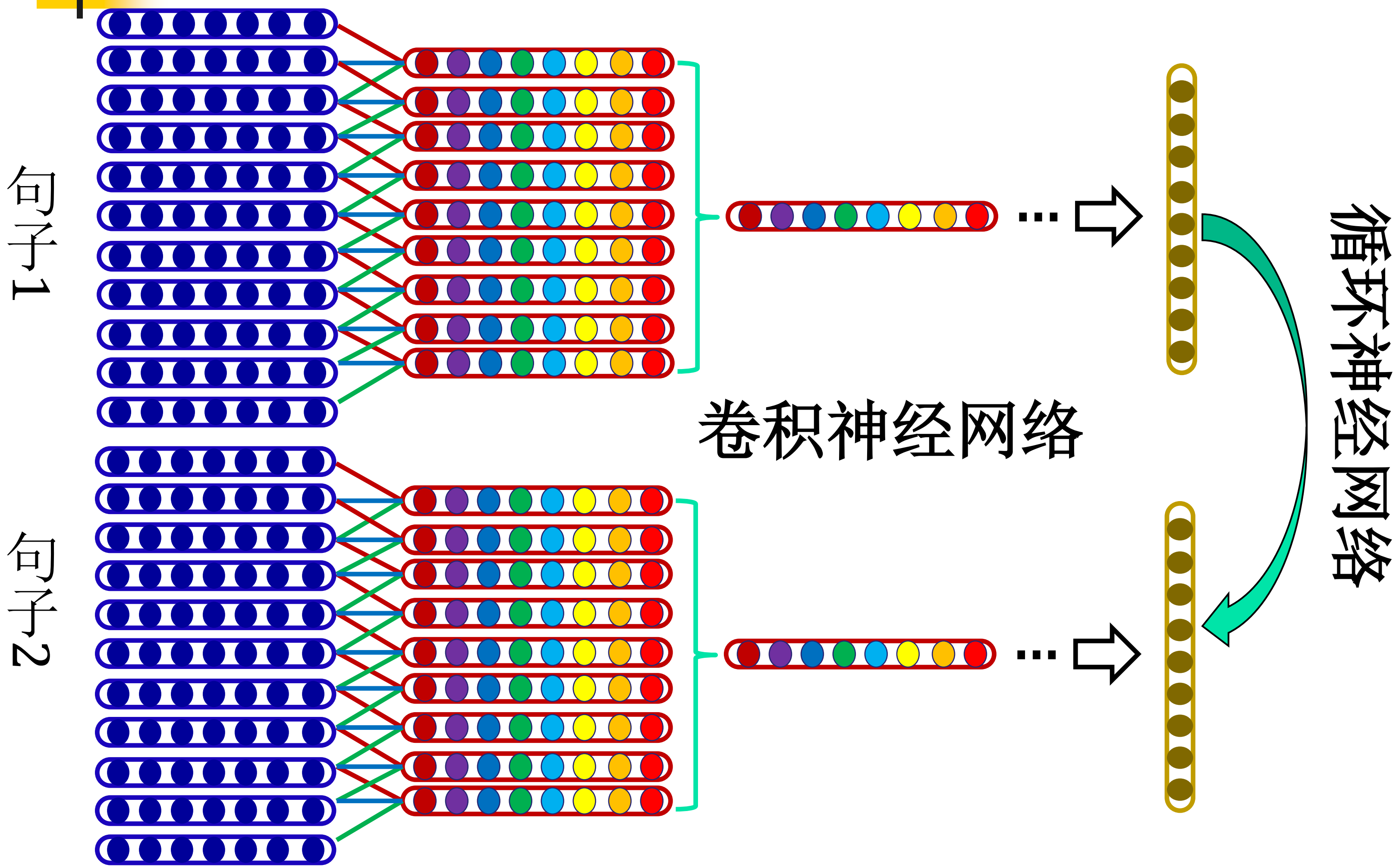
学习策略：点 → 序列 → 图

文本表示-分布式表示

- 分布式表示（Distributed Representation）



文本表示-分布式表示



Thanks

谢谢!

特征选择-CHI

- **Chi-Square 统计量 (CHI)**

CHI统计量用于检验两个事件之间的独立性，CHI度量了期望计数E和观察计数N之间相互之间关系。

$$\chi^2(t,c)=\sum_{It\in\{0,1\}}\sum_{Ic\in\{0,1\}}\frac{(N_{It,Ic}-E_{It,Ic})^2}{E_{It,Ic}}$$

$$\chi^2(t_i,c_j)=\frac{N_{all}\cdot(A_{ij}D_{ij}-C_{ij}B_{ij})^2}{(A_{ij}+C_{ij})\cdot(B_{ij}+D_{ij})\cdot(A_{ij}+B_{ij})\cdot(C_{ij}+D_{ij})}$$

$$CHI_{avg}(t_i)=\sum_{j=1}^{C}P(c_j)\chi^2(t_i,c_j)$$