# 第十三章：半监督学习

主讲：连德富 特任教授 | 博士生导师

邮箱：liandefu@ustc.edu.cn

手机：13739227137

主页：http://staff.ustc.edu.cn/~liandefu

# 背景（半监督学习）



品瓜师


# 背景（半监督学习）



训练数据中的未标记样本并非待预测数据


假定学习过程中所考虑的未标记样本恰是待预测数据

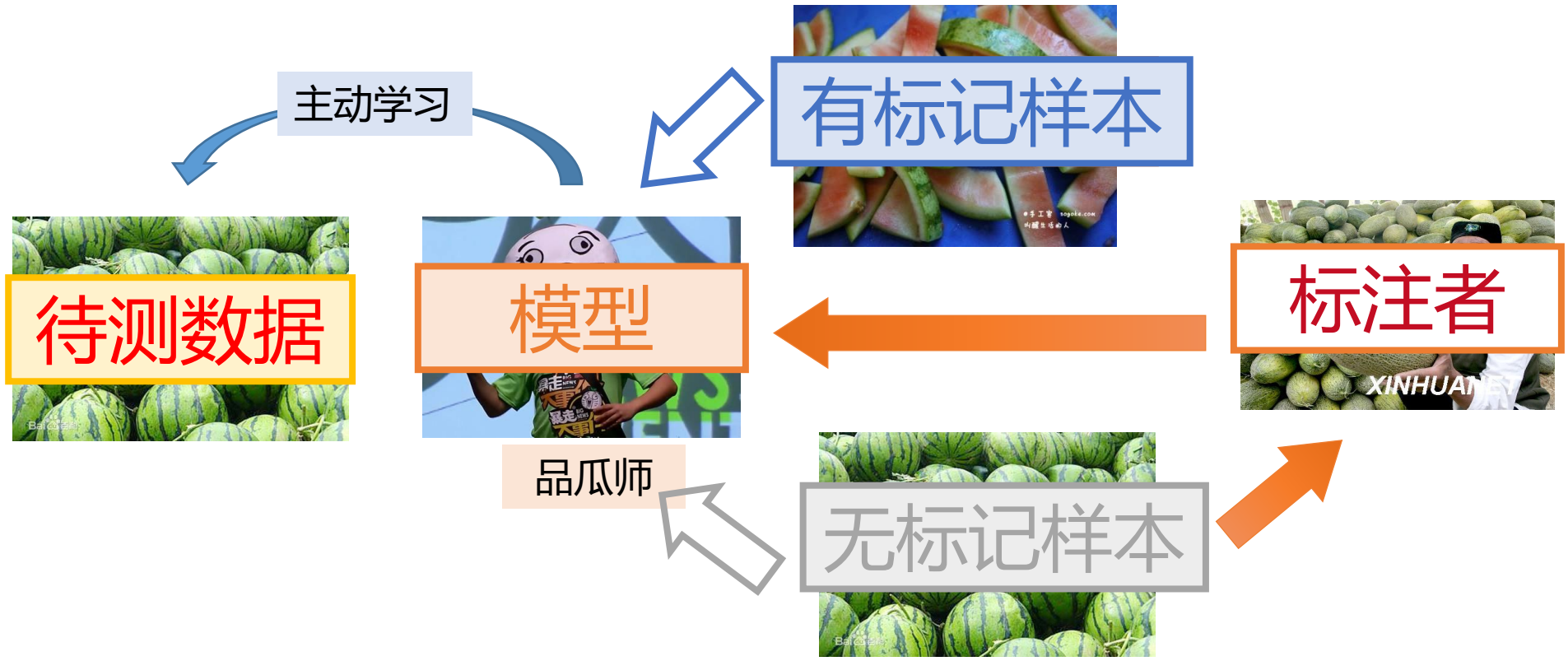
# 背景（主动学习）



品瓜师

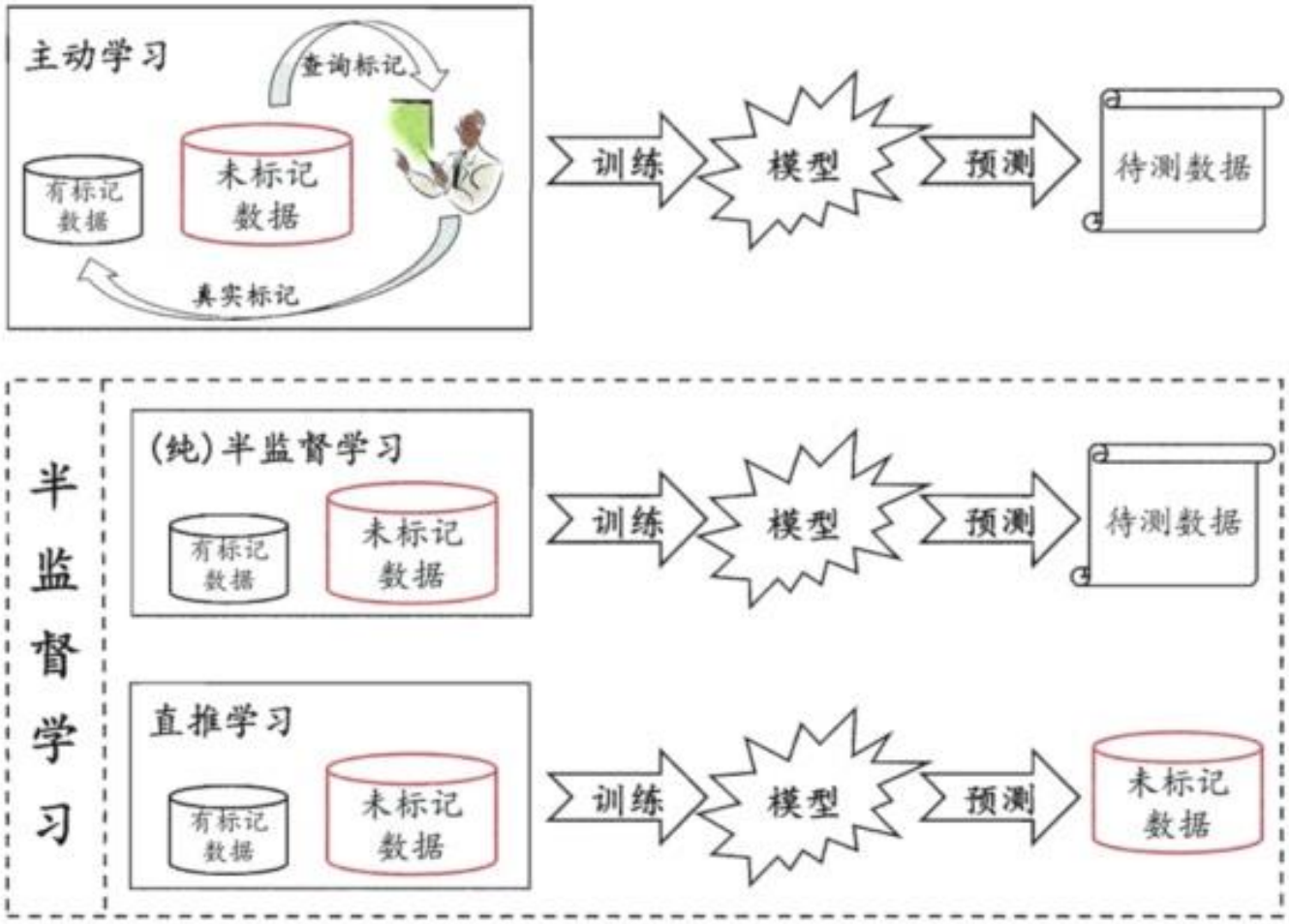

# 背景（主动学习）





# 背景（半监督学习）




主动学习
查询标记
有标记数据
未标记数据
真实标记
训练
模型
预测
待测数据
半监督学习
(纯)半监督学习
有标记数据
未标记数据
训练
模型
预测
待测数据
直推学习
有标记数据
未标记数据
训练
模型
预测
未标记数据

# 未标记样本的效用



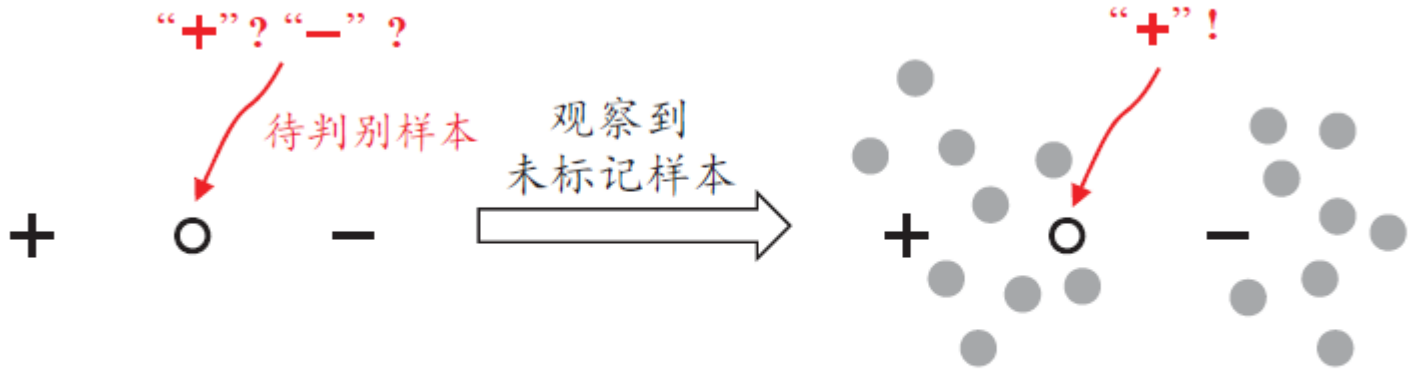



图 13.1　未标记样本效用的例示. 右边的灰色点表示未标记样本

# 未标记样本的假设



- 要利用未标记样本，必然要做一些将未标记样本所揭示的数据分布信息与类别标记相联系的假设，其中有两种常见的假设。

- 聚类假设（clustering assumption）
  假设数据存在簇结构，同一簇的样本属于同一类别

- 流形假设（manifold assumption）
  假设数据分布在一个流形结构上，邻近的样本具有相似的输出值

流形假设可看做聚类假设的推广

# 半监督算法



- 生成式方法
- 半监督SVM
- 图半监督学习
- 基于分歧的方法
- 半监督聚类

# 生成式方法



- 生成式方法是直接基于生成式模型的方法。
- 假设所有数据都是由同一个潜在模型生成的
- 假设样本由混合高斯混合模型生成, 且每个类别对应一个高斯混合成分：

$$p(\boldsymbol{x}) = \sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

其中 $\alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1$

$$p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right)$$

# 生成式方法



- 由最大化后验概率可知：

$$f(\boldsymbol{x}) = \arg\max_{j \in \mathcal{Y}} P(y = j|\boldsymbol{x})$$

$$= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{k} P(y = j, \mathrm{z} = i|\boldsymbol{x})$$

$$= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{k} P(y = j|\mathrm{z} = i, \boldsymbol{x}) P(\mathrm{z} = i|\boldsymbol{x})$$

$$P(y = j|\mathrm{z} = i) = \begin{cases} 1, \text{if } i = j \\ 0, \text{otherise} \end{cases}$$

不涉及标记

$$P(\mathrm{z} = i|\boldsymbol{x}) = \frac{\alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}$$

# 生成式方法

$$D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_l, y_l)\}$$

$$D_u = \{\boldsymbol{x}_{1+l}, \boldsymbol{x}_{2+l}, \boldsymbol{x}_{u+l}\}$$



- 假设样本独立同分布，且由同一个高斯混合模型生成，对数似然函数：

$$\ln P(D_l \cup D_u) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln P(y_j, z_j, \boldsymbol{x}_j) + \sum_{\boldsymbol{x}_j \in D_u} \ln P(z_j, \boldsymbol{x}_j)$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln P(z_j, \boldsymbol{x}_j) P(y_j | z_j, \boldsymbol{x}_j) + \sum_{\boldsymbol{x}_j \in D_u} \ln P(z_j, \boldsymbol{x}_j)$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \sum_{i=1}^{K} P(z_j = i) P(\boldsymbol{x}_j | z_j = i) P(y_j | z_j = i, \boldsymbol{x}_j) + \sum_{\boldsymbol{x}_j \in D_u} \ln P(z_j, \boldsymbol{x}_j)$$

$$= \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left( \sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(y_j | z_j = i) \right)$$

$$+ \sum_{\boldsymbol{x}_j \in D_u} \ln \left( \sum_{i=1}^{k} \alpha_i p(\boldsymbol{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)$$

# 生成式方法



E 基于$\Theta^t$推断隐变量$z$的分布$p(z \mid x, \Theta^t)$，并计算对数似然$LL(\Theta|x,z)$关于$z$的期望;

$$\mathcal{L}(p(z|x,\Theta),\Theta) = \mathbb{E}_{z\sim p(z\mid x,\Theta^t)} LL(\Theta|x,z) \quad = Q(\Theta|\Theta^t)$$

根据当前模型参数计算未标记样本属于各高斯混合成分的概率

$$r_{ji} = \frac{\alpha_i p(x_j|\mu_i,\Sigma_i)}{\sum_{i=1}^{k}\alpha_i p(x_j|\mu_i,\Sigma_i)}$$

M 寻找参数最大化期望似然;

$$\Theta^{t+1} = \arg\max_{\Theta} Q(\Theta|\Theta^t)$$

$$\mu_i = \frac{1}{\sum_{x_j\in D_u} r_{ji} + l_i}\left(\sum_{x_j\in D_u} r_{ji}x_j + \sum_{(x_j,y_j)\in D_l \wedge y_j = i} x_j\right) \qquad \alpha_i = \frac{1}{m}\sum_{x_j\in D_u} r_{ji} + l_i$$

$$\Sigma_i = \frac{1}{\sum_{x_j\in D_u} r_{ji} + l_i}\left(\sum_{x_j\in D_u} r_{ji}(x_j-\mu_i)(x_j-\mu_i)^{\mathsf{T}} + \sum_{(x_j,y_j)\in D_l \wedge y_j = i} (x_j-\mu_i)(x_j-\mu_i)^{\mathsf{T}}\right)$$

# 生成式方法



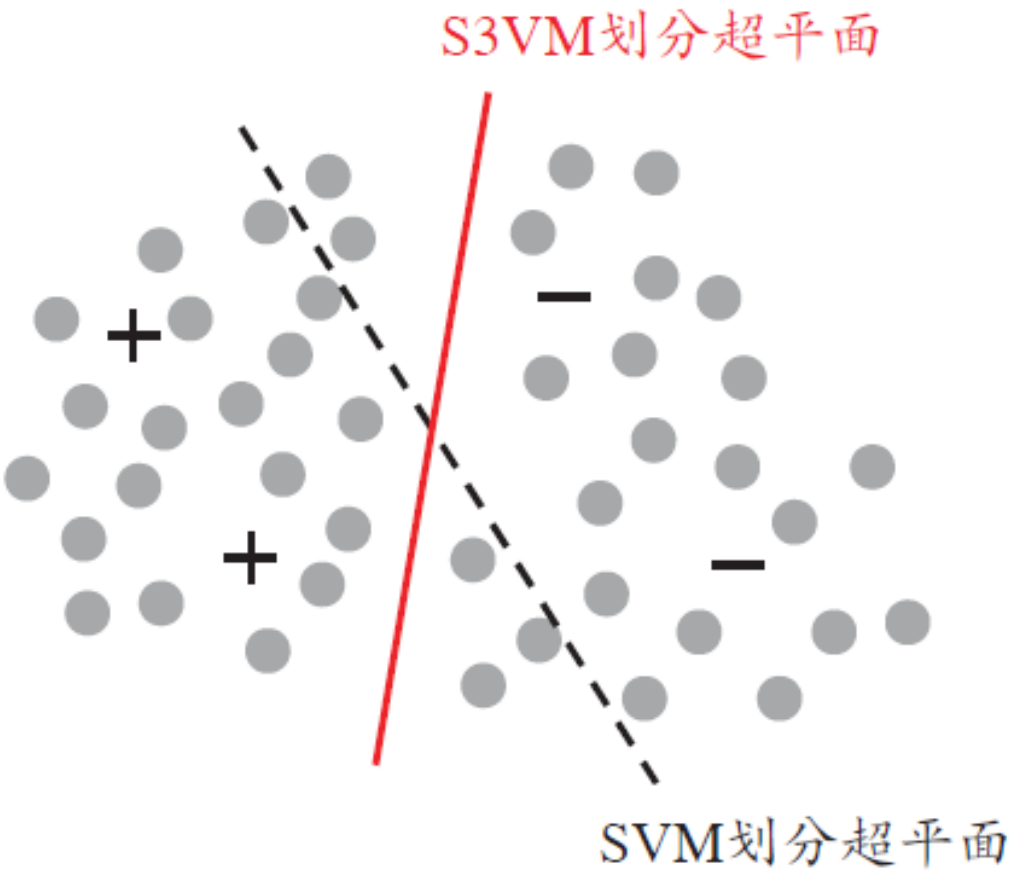
- 将上述过程中的高斯混合模型换成混合专家模型，朴素贝叶斯模型等即可推导出其他的生成式半监督学习算法。

- 此类方法简单、易于实现, 在有标记数据极少的情形下往往比其他方法性能更好。

- 然而, 此类方法有一个关键: 模型假设必须准确, 即假设的生成式模型必须与真实数据分布吻合; 否则利用未标记数据反而会显著降低泛化性能。

# 半监督SVM





# 半监督SVM



- 半监督支持向量机中最著名的是TSVM(Transductive Support Vector Machine)

$$\min_{\boldsymbol{w},b,\hat{\boldsymbol{y}},\boldsymbol{\xi}} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l,$$

$$\hat{y}_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = l+1, \ldots, m,$$

$$\xi_i \geq 0, \quad i = 1, \ldots, m,$$

试图考虑对未标记样本进行各种可能的标记指派

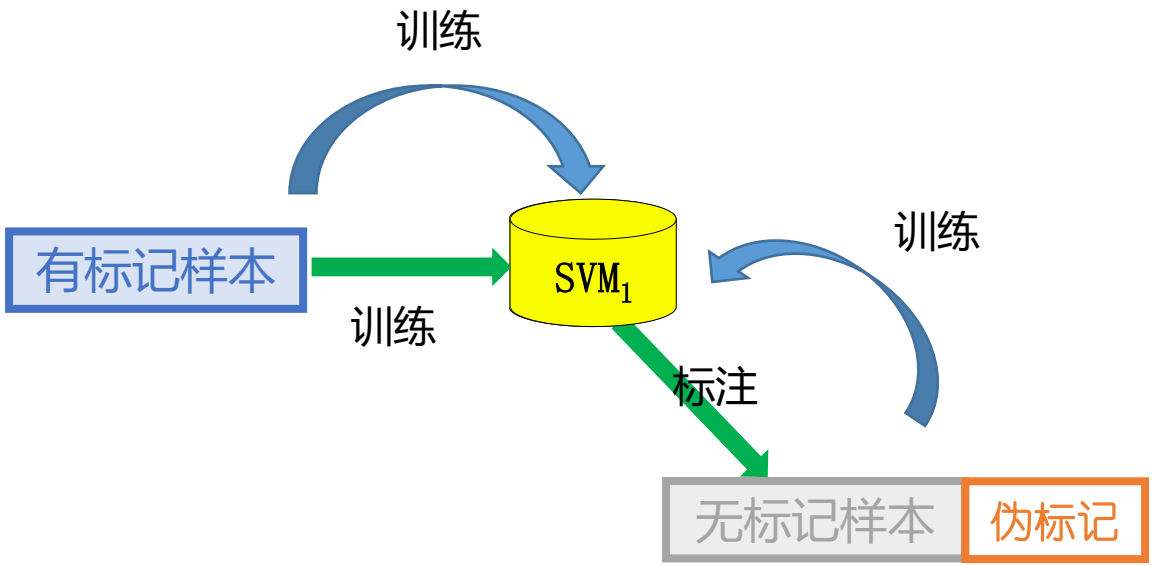
即尝试将每个未标记样本分别作为正例或者负例，然后在所有这些结果中，寻求一个在所有样本上间隔最大化的划分超平面

# 半监督SVM



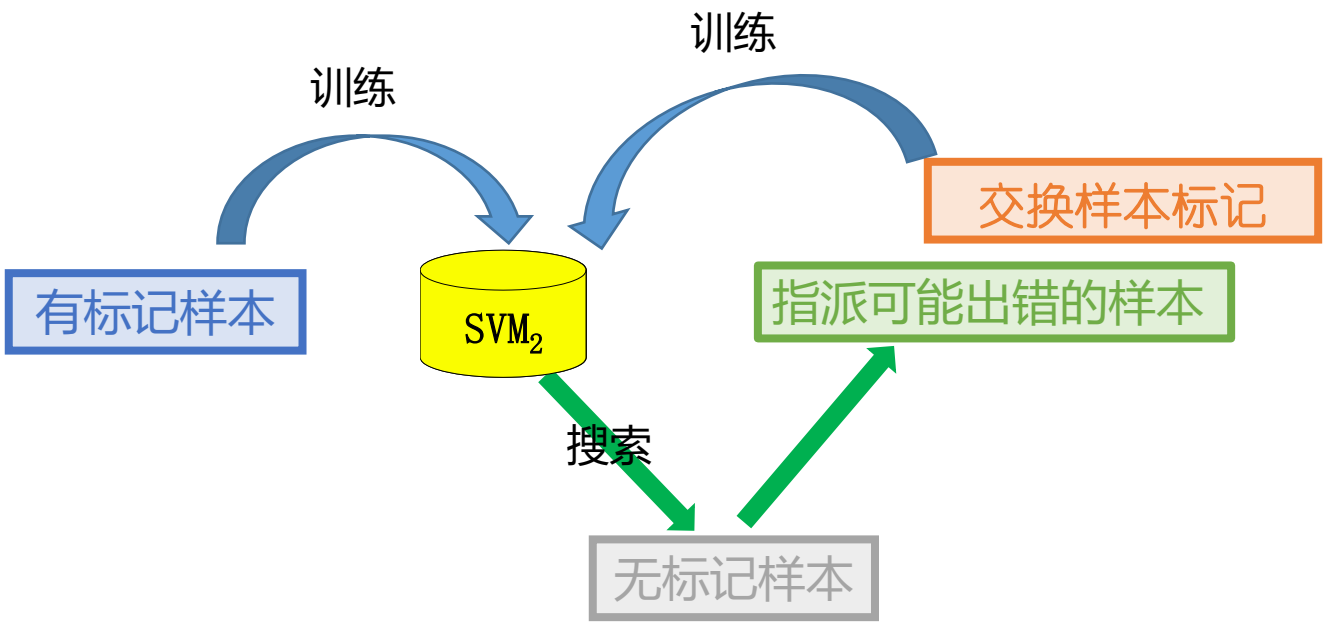
- 尝试未标记样本的各种标记指派是一个穷举过程，仅当为标记样本很少是才有可能直接求解

- TSVM采用局部搜索来迭代地寻找近似解



# 半监督SVM





# 半监督SVM



**输入:** 有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$;
未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
折中参数 $C_l$, $C_u$.

**过程:**

用 $D_l$ 训练一个 $\text{SVM}_l$;
用 $\text{SVM}_l$ 对 $D_u$ 中样本进行预测, 得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$;
初始化 $C_u \ll C_l$;
**while** $C_u < C_l$ **do**
  基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$;
  **while** $\exists\{i, j \mid (\hat{y}_i \hat{y}_j < 0) \wedge (\xi_i > 0) \wedge (\xi_j > 0) \wedge (\xi_i + \xi_j > 2)\}$ **do**
    $\hat{y}_i = -\hat{y}_i$;
    $\hat{y}_j = -\hat{y}_j$;
    基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 重新求解式(13.9), 得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$
  **end while**
  $C_u = \min\{2C_u, C_l\}$
**end while**

**输出:** 未标记样本的预测结果: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

图 13.4 TSVM 算法

未标记样本的伪标记不准确

可使得每轮迭代后目标函数值下降

# 半监督SVM



- 未标记样本进行标记指派及调整的过程中, 有可能出现类别不平衡问题,即某类的样本远多于另一类。

- 为了减轻类别不平衡性所造成的不利影响, 可对算法稍加改进:
  将优化目标中的$C_u$项拆分为$C_u^+$与$C_u^-$两项,并在初始化时令:

$$C_u^+ = \frac{u_-}{u_+} C_u^-$$

$u_+$和 $u_-$为基于伪标记而当做正、反例使用的未标记样本

# 半监督SVM



- 显然, 搜寻标记指派可能出错的每一对未标记样本进行调整, 仍是一个涉及巨大计算开销的大规模优化问题。

- 因此, 半监督SVM研究的一个重点是如何设计出高效的优化求解策略

- 例如基于图核(graph kernel)函数梯度下降的Laplacian SVM [Chapelle and Zien, 2005]、基于标记均值估计的meanS3VM[Li et al., 2009]等.

# 图半监督学习



- 给定一个数据集, 我们可将其映射为一个图, 数据集中每个样本对应于图中一个结点, 若两个样本之间的相似度很高(或相关性很强), 则对应的结点之间存在一条边, 边的“强度”(strength)正比于样本之间的相似度(或相关性)

- 将有标记样本所对应的结点想象为染过色, 而未标记样本所对应的结点则尚未染色。半监督学习就对应于“颜色”在图上扩散或传播的过程

- 由于一个图对应了一个矩阵, 这就使得我们能基于矩阵运算来进行半监督学习算法的推导与分析

# 图半监督学习



- 先基于$D_l \cup D_u$构建图$G = (V, E)$，其中节点集

$$V = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+u}\}$$

- 边集$E$可表示为一个相似度矩阵，常基于高斯函数定义为

$$w_{ij} = \begin{cases} \exp\dfrac{-\|x_i - x_j\|^2}{2\sigma^2}, \text{if } i \neq j \\ 0, \text{ otherwise} \end{cases}$$

# 图半监督学习



- 假定从图$G=(V,E)$将学得一个实值函数$f: V \to \mathbb{R}$

- 直观上讲相似的样本应具有相似的标记,于是可定义关于$f$的“能量函数”(energy function)[Zhu et al., 2003]:

$$E(f)=\frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m}w_{ij}\left(f(\boldsymbol{x}_i)-f\left(\boldsymbol{x}_j\right)\right)^2$$

$$=\frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m}w_{ij}\left(f(\boldsymbol{x}_i)^2-2f(\boldsymbol{x}_i)f\left(\boldsymbol{x}_j\right)+f\left(\boldsymbol{x}_j\right)^2\right)$$

$$d_i=\sum_{j=1}^{m}w_{ij} \qquad =\sum_{i=1}^{m}d_i f(\boldsymbol{x}_i)^2-\sum_{i=1}^{m}\sum_{i=1}^{m}w_{ij}f(\boldsymbol{x}_i)f\left(\boldsymbol{x}_j\right)$$

$$\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\ldots,d_m) \qquad =\boldsymbol{f}^{\top}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}$$

$$\boldsymbol{f}=\left(f(\boldsymbol{x}_1),\ldots,f(\boldsymbol{x}_l),f(\boldsymbol{x}_{l+1}),\ldots,f(\boldsymbol{x}_{l+u})\right)$$

# 监督学习



- 能量函数$E(f) = \boldsymbol{f}^{\top}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{f}$
- 能量函数最小化时即得到最优结果
- 具有最小能量的函数f
  - 在有标记样本上满足$f(\boldsymbol{x}_i) = y_i \ (i = 1,2, \dots, l)$
  - 在未标记样本上满足 $\boldsymbol{\Delta f} = 0$

$\boldsymbol{\Delta} = \boldsymbol{D} - \boldsymbol{W}$ 称为拉普拉斯矩阵

# 图半监督学习



- 采用分块矩阵表示方式:

$$E(f) = \boldsymbol{f}^{\top}(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{f} = (\boldsymbol{f}_l^{\top}\ \boldsymbol{f}_u^{\top})\left(\begin{bmatrix}\boldsymbol{D}_{ll} & 0\\ 0 & \boldsymbol{D}_{uu}\end{bmatrix} - \begin{bmatrix}\boldsymbol{W}_{ll} & \boldsymbol{W}_{lu}\\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu}\end{bmatrix}\right)\begin{pmatrix}\boldsymbol{f}_l\\ \boldsymbol{f}_u\end{pmatrix}$$

$$= \boldsymbol{f}_l^{\top}(\boldsymbol{D}_{ll} - \boldsymbol{W}_{ll})\boldsymbol{f}_l - 2\boldsymbol{f}_u^{\top}\boldsymbol{W}_{ul}\boldsymbol{f}_l + \boldsymbol{f}_u^{\top}(\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu})\boldsymbol{f}_u$$

- 令$\frac{\partial E(f)}{\boldsymbol{f}_u} = 0$可得

$$\boldsymbol{f}_u = (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu})^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

# 图半监督学习



- 若对$\boldsymbol{W}$矩阵归一化 $\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{W}$

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{D}_{ll}^{-1} & 0 \\ 0 & \boldsymbol{D}_{uu}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{ll} & \boldsymbol{W}_{lu} \\ \boldsymbol{W}_{ul} & \boldsymbol{W}_{uu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_{ll}^{-1}\boldsymbol{W}_{ll} & \boldsymbol{D}_{ll}^{-1}\boldsymbol{W}_{lu} \\ \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul} & \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu} \end{bmatrix}$$

令$\frac{\partial E(f)}{\boldsymbol{f}_u} = 0$可得

$$\boldsymbol{f}_u = (\boldsymbol{D}_{uu} - \boldsymbol{W}_{uu})^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

$$= (\boldsymbol{I} - \boldsymbol{D}_{uu}\boldsymbol{W}_{uu})^{-1}\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

$$= (\boldsymbol{I} - \boldsymbol{P}_{uu})^{-1}\boldsymbol{P}_{ul}\boldsymbol{f}_l$$

# 图半监督学习



- 上面描述的是一个针对二分类问题的“单步式”标记传播(label propagation)方法, 下面我们来看一个适用于多分类问题的“迭代式”标记传播方法[Zhou et al., 2004].

- 先基于$D_l \cup D_u$构建图$G = (V, E)$，其中节点集 $V = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+u}\}$

- 定义一个$(l+u) \times |\mathcal{Y}|$的非负标记矩阵 $\boldsymbol{F} = (\boldsymbol{F}_1^\top, \ldots, \boldsymbol{F}_m^\top)^\top$，其第$i$行$\boldsymbol{F}_i = (F_{i1}, \ldots, F_{i|\mathcal{Y}|})$为示例$\boldsymbol{x}_i$的标记向量, 相应的分类规则为

$$y_i = \underset{1 \le j \le |\mathcal{Y}|}{\operatorname{argmax}} F_{ij}$$

- 将$\boldsymbol{F}$ 初始化为

$$\boldsymbol{F}(0) = \boldsymbol{Y} \qquad Y_{ij} = \begin{cases} 1, \text{if } (1 \le i \le l) \wedge (y_i = j) \\ 0, \text{otherwise} \end{cases}$$

$\boldsymbol{Y}$的前$l$行 为$l$个有标记样本的标记向量

# 图半监督学习



- 基于$\boldsymbol{W}$构造一个标记传播矩阵$\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$

  其中$\boldsymbol{D}^{-\frac{1}{2}} = \mathrm{diag}\left(\frac{1}{\sqrt{d_1}}, \ldots, \frac{1}{\sqrt{d_m}}\right)$

- 有迭代计算式:

$$\boldsymbol{F}(t+1) = \alpha\, \boldsymbol{S}\, \boldsymbol{F}(t) + (1-\alpha)\, \boldsymbol{Y}$$

- 基于迭代至收敛可得:

$$\boldsymbol{F}^* = \lim_{t\to\infty} \boldsymbol{F}(t) = (1-\alpha)(\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\boldsymbol{Y}$$

# 图半监督学习



**输入:** 有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$;
未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$;
构图参数 $\sigma$;
折中参数 $\alpha$.

**过程:**

1: 基于式(13.11)和参数 $\sigma$ 得到 $\mathbf{W}$;
2: 基于 $\mathbf{W}$ 构造标记传播矩阵 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$;
3: 根据式(13.18)初始化 $\mathbf{F}(0)$;
4: $t = 0$;
5: **repeat**
6: $\quad \mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t) + (1-\alpha)\mathbf{Y}$;
7: $\quad t = t + 1$
8: **until** 迭代收敛至 $\mathbf{F}^*$
9: **for** $i = l+1, l+2, \ldots, l+u$ **do**
10: $\quad y_i = \arg\max_{1 \le j \le |\mathcal{Y}|} (\mathbf{F}^*)_{ij}$
11: **end for**

**输出:** 未标记样本的预测结果: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$

图 13.5 迭代式标记传播算法

# 图半监督学习



- 事实上, 算法对应于正则化框架[Zhou et al., 2004]:

$$\min_{\boldsymbol{F}} \frac{1}{2}\left(\sum_{i,j=1}^{m} w_{ij}\left\|\frac{1}{\sqrt{d_i}}\boldsymbol{F}_i-\frac{1}{\sqrt{d_j}}\boldsymbol{F}_j\right\|^2\right)+\mu\sum_{i=1}^{l}\|\boldsymbol{F}_i-\boldsymbol{Y}_i\|^2$$

$$\Rightarrow \quad \min_{\boldsymbol{F}} \operatorname{tr}(\boldsymbol{F}^{\top}(\boldsymbol{I}-\boldsymbol{S})\boldsymbol{F})+\mu\|\boldsymbol{F}-\boldsymbol{Y}\|_F^2$$

- 当$\mu=\frac{1-\alpha}{\alpha}$时，最优解恰为迭代算法的收敛解 $\boldsymbol{F}^*$

$$\frac{1}{2}\left(\sum_{i,j=1}^{m}\frac{w_{ij}}{d_i}\|\boldsymbol{F}_i\|^2+\frac{w_{ij}}{d_j}\|\boldsymbol{F}_j\|^2-\frac{2w_{ij}}{\sqrt{d_i d_j}}\boldsymbol{F}_i^{\top}\boldsymbol{F}_j\right)$$

$$=\sum_{i}^{m}\|\boldsymbol{F}_i\|^2-\sum_{i,j=1}^{m}\frac{w_{ij}}{\sqrt{d_i d_j}}\boldsymbol{F}_i^{\top}\boldsymbol{F}_j \quad =\|\boldsymbol{F}\|_F^2-\operatorname{tr}(\boldsymbol{F}^{\top}\boldsymbol{S}\boldsymbol{F})$$

# 图半监督学习



- 图半监督学习方法在概念上相当清晰，且易于通过对所涉矩阵运算的分析来探索算法性质。

- 但此类算法的缺陷也相当明显

  - 首先，存储开销高

  - 其次，由于构图过程仅能考虑训练样本集，难以判知新样本在图中的位置，因此，在接收到新样本时，或是将其加入原数据集对图进行重构并重新进行标记传播，或是需引入额外的预测机制

# 基于分歧的方法



- 生成式、半监督SVM、图半监督学习等基于单学习器
- 基于分歧的方法(disagreement-based methods)使用多学习器，disagreement亦称diversity
- 学习器之间的“分歧”(disagreement)对未标记数据的利用至关重要
- 协同训练(co-training)[Blum and Mitchell, 1998]是基于分歧的方法的重要代表, 它最初是针对“多视图”(multi-view)数据设计的, 因此也被看作“多视图学习”(multi-view learning)的代表.

# 基于分歧的方法



文字视图

从目前的情况来看，1月6日华盛顿特区的大游行，充其量只能为特朗普积攒人气，对于改变选举结果没有丝毫意义。

**作者 | 南风窗常务副主编 谢奕秋**

图片视图

1月6日，美国首都所在地华盛顿特区，会发生大事。

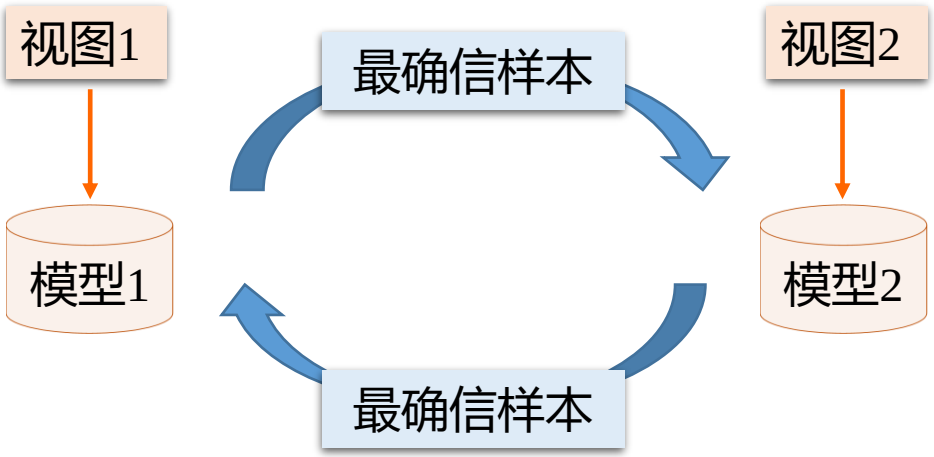
去年12月28日一大早，特朗普发了一条推特，内容是这样的：“1月6日华盛顿特区见，不见不散。且待后续！”12月31日，他提前结束在佛罗里达的度假，回到白宫。

新闻分类任务中的双视图

# 基于分歧的方法



- 协同训练正是很好地利用了多视图的“相容互补性”。 假设数据拥有两个“充分”(sufficient)且“条件独立”视图。



# 基于分歧的方法



$x_i$ 的上标仅用于指代两个视图，不表示序关系，即 $\langle x_i^1, x_i^2\rangle$ 与 $\langle x_i^2, x_i^1\rangle$ 表示的是同一个样本.

**输入:** 有标记样本集 $D_l = \{(\langle x_1^1, x_1^2\rangle, y_1), \ldots, (\langle x_l^1, x_l^2\rangle, y_l)\}$;
未标记样本集 $D_u = \{\langle x_{l+1}^1, x_{l+1}^2\rangle, \ldots, \langle x_{l+u}^1, x_{l+u}^2\rangle\}$;
缓冲池大小 $s$;
每轮挑选的正例数 $p$;

在每个试图上基于有标记样本分别训练出一个分类器，然后让每个分类器分别去挑选自己最有把握的未标记样本赋予伪标记

在视图 $j$ 上用有标记样本训练 $h_j$.

**for** $j = 1, 2$ **do**
  $h_j \leftarrow \mathfrak{L}(D_l^j)$;
  考察 $h_j$ 在 $D_s^j = \{x_i^j \mid \langle x_i^j, x_i^{3-j}\rangle \in D_s\}$ 上的分类置信度，挑选 $p$ 个正例置信度最高的样本 $D_p \subset D_s$、$n$ 个反例置信度最高的样本 $D_n \subset D_s$;
  由 $D_p^j$ 生成伪标记正例 $\tilde{D}_p^{3-j} = \{(x_i^{3-j}, +1) \mid x_i^j \in D_p^j\}$;
  由 $D_n^j$ 生成伪标记反例 $\tilde{D}_n^{3-j} = \{(x_i^{3-j}, -1) \mid x_i^j \in D_n^j\}$;
  $D_s = D_s \setminus (D_p \bigcup D_n)$;
**end for**

**if** $h_1, h_2$ 均未发生改变 **then**

扩充有标记数据集.

  **for** $j = 1, 2$ **do**
    $D_l^j = D_l^j \bigcup \left(\tilde{D}_p^j \bigcup \tilde{D}_n^j\right)$;
  **end for**

将伪标记样本提供给另外一个分类器作为新增的有标记样本用于训练更新

**输出:** 分类器 $h_1, h_2$

图 13.6 协同训练算法

# 基于分歧的方法



- 协同训练过程虽简单, 但令人惊讶的是, 理论证明显示出, 若两个视图充分且条件独立, 则可利用未标记样本通过协同训练将弱分类器的泛化性能提升到任意高[Blum and Mitchell, 1998]

- 视图的条件独立性在现实任务中通常很难满足,不会是条件独立的，性能提升幅度不会那么大

- 研究表明, 即使在更弱的条件下,协同训练仍可有效地提升弱分类器的性能[周志华, 2013]

# 基于分歧的方法



- 协同训练算法本身是为多视图数据而设计的, 但此后出现了一些能在单视图数据上使用的变体算法

- 它们或是使用不同的学习算法[Goldman and Zhou,2000]、或使用不同的数据采样[Zhou and Li, 2005b]、甚至使用不同的参数设置[Zhou and Li, 2005a]来产生不同的学习器, 也能有效地利用未标记数据来提升性能

- 后续理论研究发现, 此类算法事实上无需数据拥有多视图, 仅需弱学习器之间具有显著的分歧(或差异), 即可通过相互提供伪标记样本的方式来提高泛化性能[周志华, 2013]

# 基于分歧的方法



- 基于分歧的方法只需采用合适的基学习器, 就较少受到模型假设、损失函数非凸性和数据规模问题的影响, 学习方法简单有效、理论基础相对坚实、适用范围较为广泛。

- 为了使用此类方法, 需能生成具有显著分歧、性能尚可的多个学习器, 但当有标记样本很少、尤其是数据不具有多视图时, 要做到这一点并不容易。

# 半监督聚类



- 聚类是一种典型的无监督学习任务
- 在现实聚类任务中我们往往能获得一些额外的监督信息, 于是可通过“半监督聚类”(semi-supervised clustering)来利用监督信息以获得更好的聚类效果
- 聚类任务中获得的监督信息大致有两种类型:
  - 第一种类型是“必连”(must-link)与“勿连”(cannot-link)约束, 前者是指样本必属于同一个簇, 后者则是指样本必不属于同一个簇;
  - 第二种类型的监督信息则是少量的有标记样本

# 半监督聚类



- 约束*k*均值(Constrained *k*-means)算法[Wagstaff et al., 2001]是利用第一类监督信息的代表。

- 该算法是*k*均值算法的扩展,它在聚类过程中要确保“必连”关系集合与“勿连”关系集合中的约束得以满足, 否则将返回错误提示。

# 半监督聚类



输入：样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$；
必连约束集合 $\mathcal{M}$；
勿连约束集合 $\mathcal{C}$；
聚类簇数 $k$.

```
while ¬ is_merged do
  基于 K 找出与样本 x_i 距离最近的簇: r = arg min_{j∈K} d_ij ;
  检测将 x_i 划入聚类簇 C_r 是否会违背 M 与 C 中的约束;
  if ¬ is_voilated then
    C_r = C_r ⋃ {x_i};
    is_merged=true
  else
    K = K \ {r};
    if K = ∅ then
      break并返回错误提示
    end if
  end if
end while
```

end for
until 均值向量均未更新
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

不冲突，选择最近的簇

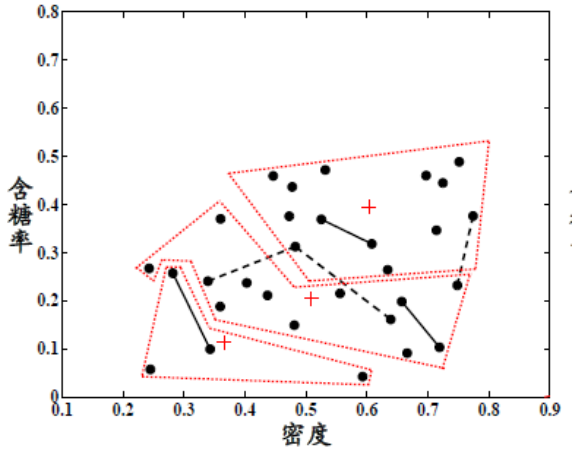
冲突，尝试次近的簇

找不到满足条件的簇，报错

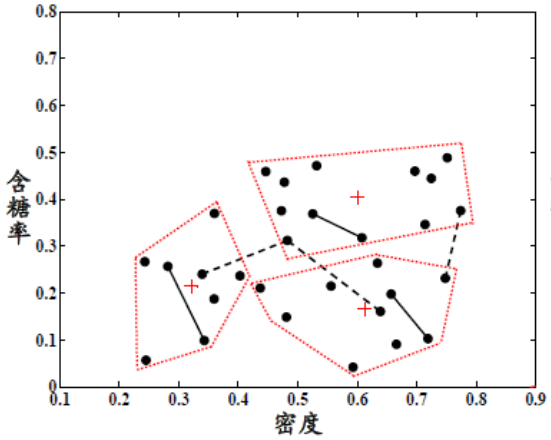
图 13.7 约束 $k$ 均值算法

# 半监督聚类

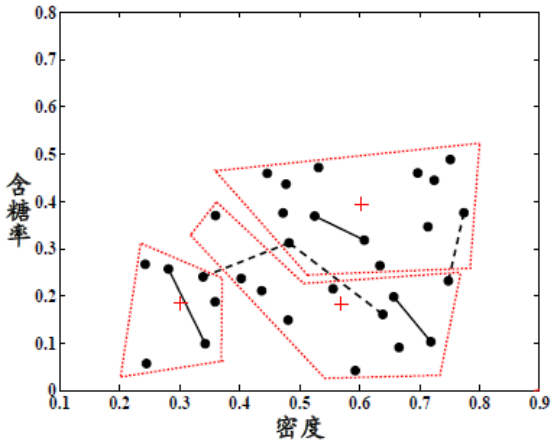




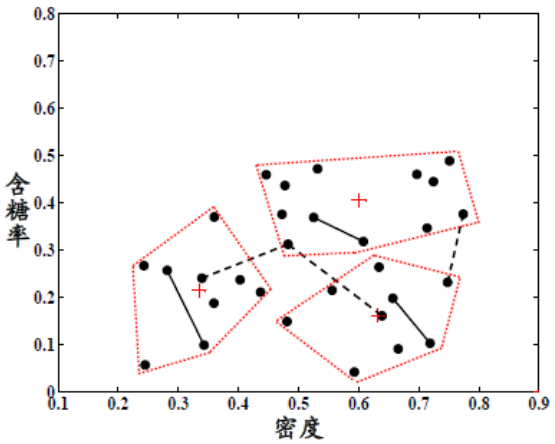

(a) 第 1 轮迭代后




(c) 第 3 轮迭代后




(b) 第 2 轮迭代后




(d) 第 4 轮迭代后

在西瓜数据集4.0上运行结果

# 半监督聚类



- 第二种监督信息是少量有标记样本。即假设少量有标记样本属于k个聚类簇。

- 将它们作为“种子”, 用它们初始化*k*均值算法的*k*个聚类中心, 并且在聚类簇迭代更新过程中不改变种子样本的簇隶属关系

- 这样就得到了约束种子*k*均值(Constrained Seed *k*-means)算法[Basu et al., 2002]。

# 半监督聚类



输入：样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$；

用有标记样本初始化簇中心

```
for j = 1, 2, ..., k do
    μ_j = (1/|S_j|) Σ_{x∈S_j} x
end for
```

用有标记样本初始化k个簇

```
    for j = 1, 2, ..., k do
        for all x ∈ S_j do
            C_j = C_j ⋃ {x}
        end for
```

```
        找出与样本 x_i 距离最近的簇：r = arg min_{j∈{1,2,...,k}} d_ij ；
        将样本 x_i 划入相应的簇：C_r = C_r ⋃ {x_i}
    end for
    for j = 1, 2, ..., k do
        μ_j = (1/|C_j|) Σ_{x∈C_j} x;
    end for
until 均值向量均未更新
```

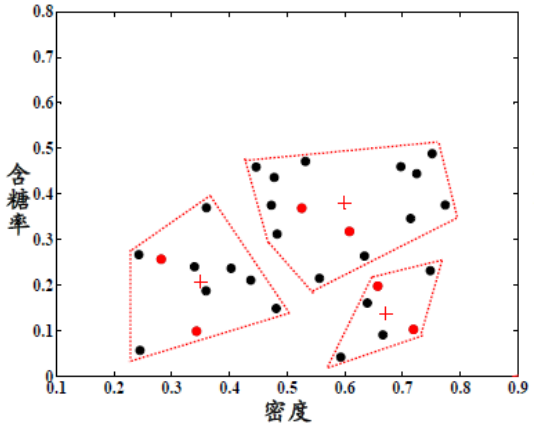
更新均值向量.

输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

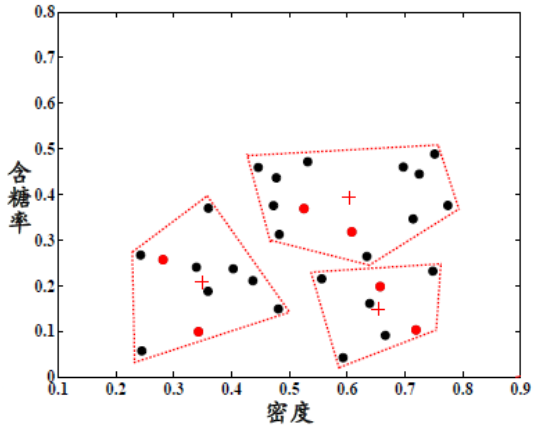
图 13.9　约束种子 $k$ 均值算法

# 半监督聚类

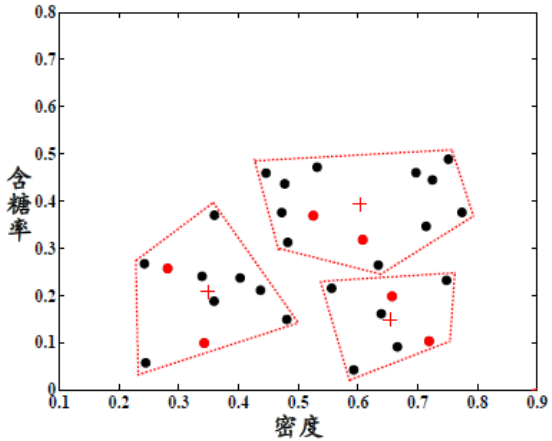




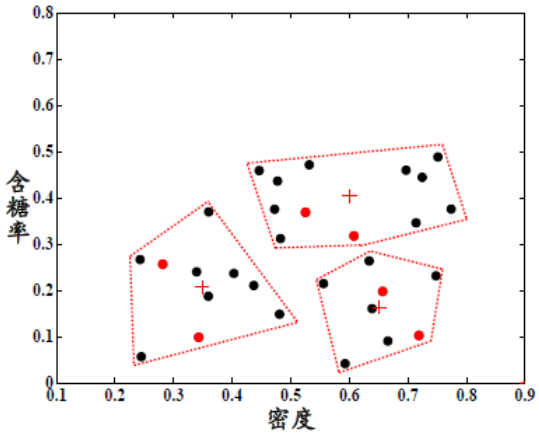

(a) 第 1 轮迭代后




(b) 第 2 轮迭代后




(b) 第 2 轮迭代后




(d) 第 4 轮迭代后

在西瓜数据集4.0上运行结果

# 课堂测验



- PPT 第31页