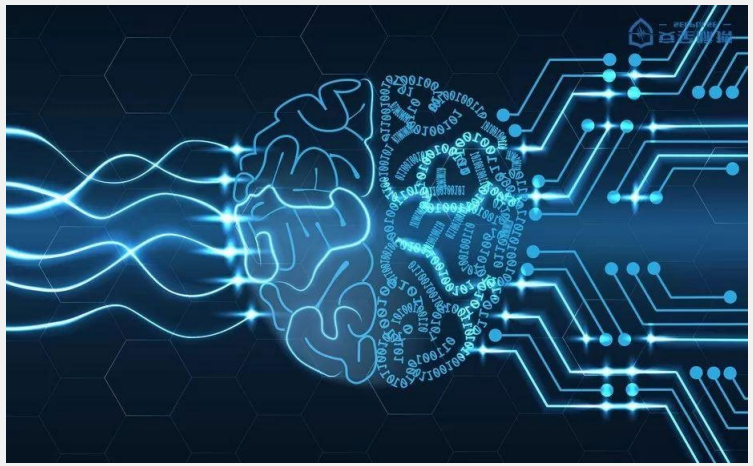


机器学习

Machine Learning

授课老师：谭毅华
电　　话：13886021197
办 公 室：科技楼1102
邮　　箱：*yhtan@hust.edu.cn*

第十一章、半监督学习

目录
CONTENTS

为人师表

1.未标记样本

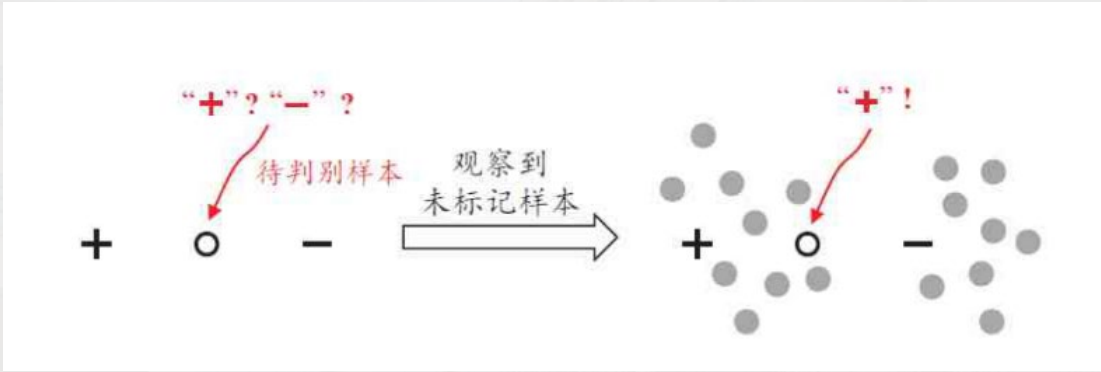
- 半监督学习问题的提出：在有标签样本较少时，如何利用无标签样本提升学习性能已成为机器学习及其应用中的重要研究问题。

$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l), \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_m\}$ **训练集含有标记样本D_l和未标记样本D_u**

有监督的学习：学习器通过对大量有标记的训练例进行学习，从而建立模型用于预测未见示例的标记(label)。很难获得大量的标记样本。

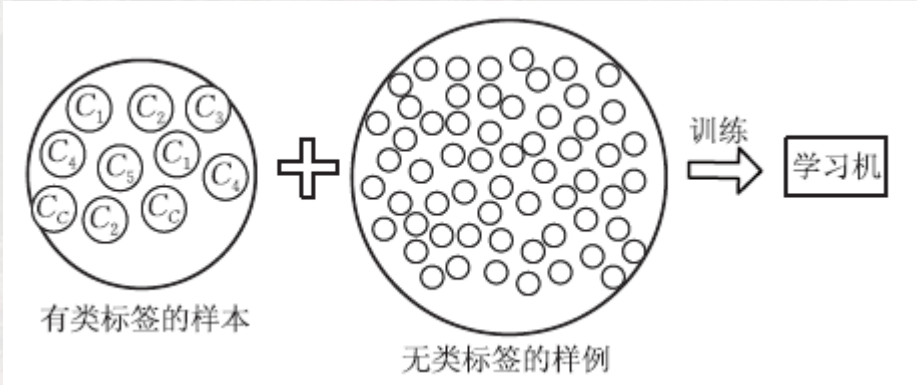
无监督的学习：无训练样本，仅根据测试样本的在特征空间分布情况来进行标记，准确性差。

半监督的学习：有少量训练样本，学习机以从训练样本获得的知识为基础，结合测试样本的分布情况逐步修正已有知识，并判断测试样本的类别。



未标记样本应用示例

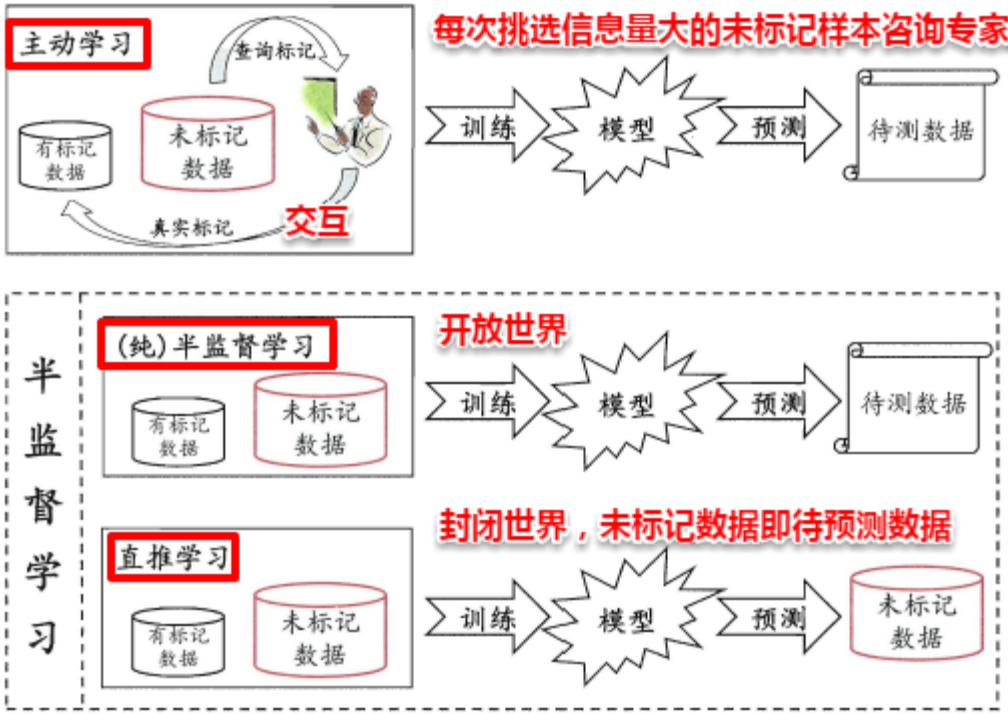
- 半监督学习特点：
  - 有标记数据少，无标记数据多
  - 利用数据分布与类别标记相联系的假设
  - 性能依赖于所用的半监督假设



未标记样本应用示例

1.未标记样本

- 主动学习需要与外界进行交互/查询/打标，其本质上仍然属于一种监督学习
- 半监督学习让学习过程不依赖外界的咨询交互,训练集同时包含有标记样本数据和未标记样本数据



- **数据分布信息与类别标记关系的假设：相似的样本有相似的输出**
  - **聚类假设：同一聚类中的样本点很可能具有同样的类别标记。关注样本空间的整体特征，探测样本分布稠密和稀疏的区域，从而更好地约束决策边界穿过稀疏区域**
  - **流形假设：处于一个很小的局部流形邻域内的示例具有相似的性质。利用大量的无标签样增加样本空间的密度，从而更准确地获取样本的局部近邻关系**
- **半监督学习分为两类**
  - **纯半监督学习：假定训练数据集中的未标记数据并非待预测数据。**
  - **直推式学习：学习过程中的未标记数据就是待预测数据**

2.生成式方法

- **假设**：所有样本数据（无论是否标记）服从一个潜在的分布。该分布由先验知识作出假设。以高斯混合模型为例，假定样本的概率由多个高斯分布组合形成，从而一个子高斯分布就代表一个类簇（类别）。数据样本的概率密度如下:

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

混合系数　均值向量　协方差矩阵

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

高斯分布子分量:

$$p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)}$$

2.生成式方法

分类：则可以根据有标签样本和无标签样本估计分布的参数，进而由后验概率判别类别：

$$f(\boldsymbol{x}) = \underset{j \in \mathcal{Y}}{\operatorname{argmax}} p(y = j|\boldsymbol{x})$$

$$= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y = j, \Theta = i|\boldsymbol{x})$$

$$= \underset{j \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{k} p(y = j|\Theta = i, \boldsymbol{x}) \cdot p(\Theta = i|\boldsymbol{x})$$

$p(y = j|\Theta = i)$

仅与样本x所属的高斯成分有关

其中属于第i个高斯成分的概率：

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

2.生成式方法

- **参数估计**：根据所有样本极大似然法估计高斯混合模型的参数 $D=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_l,y_l),\boldsymbol{x}_{l+1},\ldots,\boldsymbol{x}_m\}$

$$LL(D_l\cup D_u)=\sum_{(\boldsymbol{x}_j,y_j)\in D_l}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\,p(y_j\mid\Theta=i,\boldsymbol{x}_j)\right)+\sum_{\boldsymbol{x}_j\in D_u}\ln\left(\sum_{i=1}^{N}\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)$$

当且仅当i=j时取1

有类标样本只在特定类簇中出现

无类标样本可能在所有类簇中出现

- **迭代优化**：基于标记数据的有监督项和无标记数据的无监督项以EM算法求解
  - **期望步(E步）：未标记样本属于各高斯混合成分的概率**

$$\gamma_{ji}=\frac{\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N}\alpha_i\cdot p(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i}$$

  - **最大化期望(M步），更新模型参数，l_i表示第i类的有标记样本数目**

$$\boldsymbol{\mu}_i=\frac{1}{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\boldsymbol{x}_j+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j\right)$$

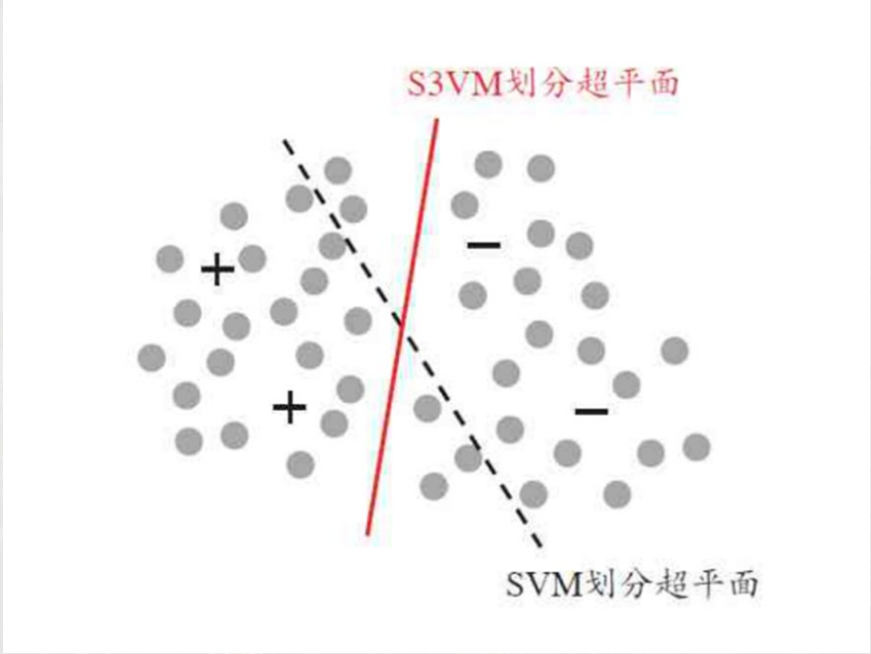
li指的是第i类有标记样本数目

$$\boldsymbol{\Sigma}_i=\frac{1}{\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\mathrm{T}}\right)$$

$$\alpha_i=\frac{1}{m}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i\right)$$

3.半监督SVM

- **考虑未标记样本**：目标是找到最大间隔划分超平面、且穿过数据低密度区域将标记样本分开



- **TSVM(Transductive SVM)**：尝试为未标记样本找到合适的标记指派，使得超平面划分后的间隔最大化

3. 半监督SVM

$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l), \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_m\}$ $\{D_l, D_u\}$

- TSVM目标：为D_u中的样本给出预测标记$\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \cdots, \hat{y}_m)$，$\hat{y}_i \in \{-1, +1\}$,满足：

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

松弛变量 hinge损失

$$\text{s.t.} \ \ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \ldots, l,$$

对应标记样本

$$\hat{y}_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i, \quad i = l+1, l+2, \ldots, m,$$

对应未标记样本

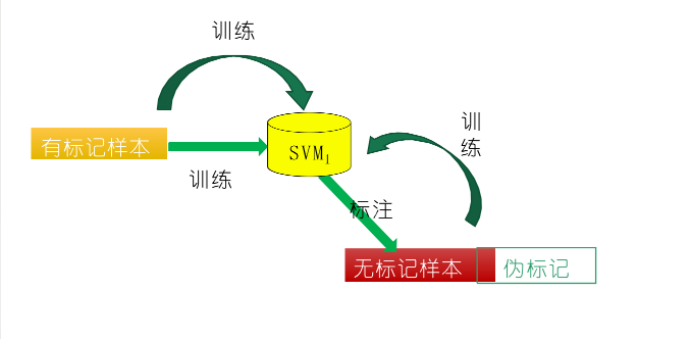
$$\xi_i \geqslant 0, \quad i = 1, 2, \ldots, m,$$

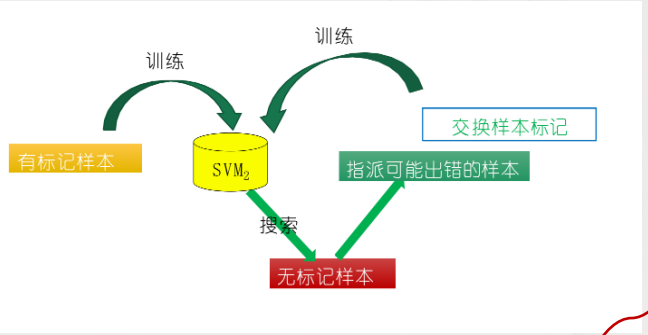
C_l与C_u用于平衡模型复杂度、有标记样本与未标记样本重要程度的折中参数

3. 半监督SVM

□ **TSVM求解**：迭代地进行局部搜索。分两个环节：分配伪标记并获得初始SVM、交换错分的未标记样本更新SVM

- **计算初始SVM**
  - **有标记样本训练SVM**
  - **获得未标记样本的伪标签后，再训练SVM**





- **交换标记重复训练SVM**
  - **挑选两个可能错分的异类未标记样本**
  - **交换其标签，重新训练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：

1: 用 D_l 训练一个 SVM_l; 初始SVM
2: 用 SVM_l 对 D_u 中样本进行预测，得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$; 伪标记
3: 初始化 $C_u \ll C_l$;
4: **while** $C_u < C_l$ **do**
5: 基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 求解式(13.9)，得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$;
6: **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**
7: $\hat{y}_i = -\hat{y}_i$;
8: $\hat{y}_j = -\hat{y}_j$;
9: 基于 $D_l, D_u, \hat{\boldsymbol{y}}, C_l, C_u$ 重新求解式(13.9)，得到 $(\boldsymbol{w}, b)$, $\boldsymbol{\xi}$
10: **end while**
11: $C_u = \min\{2C_u, C_l\}$ 逐渐增大Cu
12: **end while**

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

松弛变量越大表示离超平面越近，越容易分错

挑选准则

- **减轻类别不平衡的影响，基于伪标记而当作正反例的未标记样本数，将C_u拆为2项**

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

4.图半监督学习

- **二分类问题节点标签扩散**：将每个样本视为一个节点构建图，将已标记样本根据样本特征的相似性进行扩散，获得未标记样本的标记。

$$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l), \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_m\} \quad \{D_l, D_u\}$$

- **图G(V,E)，高斯函数定义邻接矩阵，表达样本间的相似度**

$$\mathbf{W}_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise}, \end{cases}$$

- **从图学得实值函数$f: V \to R$,分类规则**

$$y_i = sign\big(f(x_i)\big), y_i \in \{-1, +1\}$$

- **相似的样本应有相似的标记，定义关于f的能量函数，其中$\boldsymbol{f} = (\boldsymbol{f}_l^T, \boldsymbol{f}_u^T)$,分别为标记样本和未标记样本的预测结果，D为对角阵，其元素值为W每一行之和**

$$E(f) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\boldsymbol{W}_{ij}(f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j))^2$$

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

$$= \boldsymbol{f}^T(\boldsymbol{D} - \boldsymbol{W})\boldsymbol{f}$$

4.图半监督学习

- 若分块表示对角矩阵D和邻接矩阵W，则最小能量公式可重写为：

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

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

- 令最小能量公式取最小值，由$\frac{\partial E(f)}{\partial f_u}=0$得：

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

4.图半监督学习

$$f_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1}\mathbf{W}_{ul}f_l$$

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

$$\mathbf{P}_{uu} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{uu},\ \mathbf{P}_{ul} = \mathbf{D}_{uu}^{-1}\mathbf{W}_{ul}$$

简化

无标签样本计算标签公式：

$$y_i = sign\big(f(x_i)\big), y_i \in \{-1, +1\}$$

$$\begin{aligned} f_u &= (\boldsymbol{D}_{uu}(\boldsymbol{I} - \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu}))^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l \\ &= (\boldsymbol{I} - \boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{uu})^{-1}\boldsymbol{D}_{uu}^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l \\ &= (\boldsymbol{I} - \boldsymbol{P}_{uu})^{-1}\boldsymbol{P}_{ul}\boldsymbol{f}_l \end{aligned}$$

4.图半监督学习

- 对多分类问题$y_i \in |Y|$，构建图G=(V,E)方式与二类问题相同，W矩阵定义相同。但定义$(l+u)\times|Y|$非负标记矩阵$F=(F_1^T,F_2^T,\cdots,F_{l+u}^T)^T$,每一行对应示例$x_i$的标记向量，且分类规则为：

$$y_i = argmax_{1\le j\le |Y|}(F)_{ij}$$

- 对标记矩阵进行初始化

$$\mathbf{F}(0)=\mathbf{Y}_{ij}=\begin{cases}1, & \text{if } (1\le i\le l)\wedge(y_i=j)\ ;\\ 0, & \text{otherwise}\ ,\end{cases}$$

- 基于W构造一个标记 传播矩阵$\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$,其中$\boldsymbol{D}^{-1/2}=\boldsymbol{diag}(\frac{1}{\sqrt{d_1}},\frac{1}{\sqrt{d_2}},\ldots,\frac{1}{\sqrt{d_{l+u}}})$

- 可得迭代公式

$$\boldsymbol{F}(\boldsymbol{t}+\boldsymbol{1})=\boldsymbol{\alpha}\boldsymbol{S}\cdot\boldsymbol{F}(\boldsymbol{t})+(\boldsymbol{1}-\boldsymbol{\alpha})\boldsymbol{Y} \qquad (13.19)$$

4.图半监督学习

- 基于式（13.19）可迭代收敛到

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

- 上式对应正则化框架

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

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 σ;
折中参数 α.

过程：

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

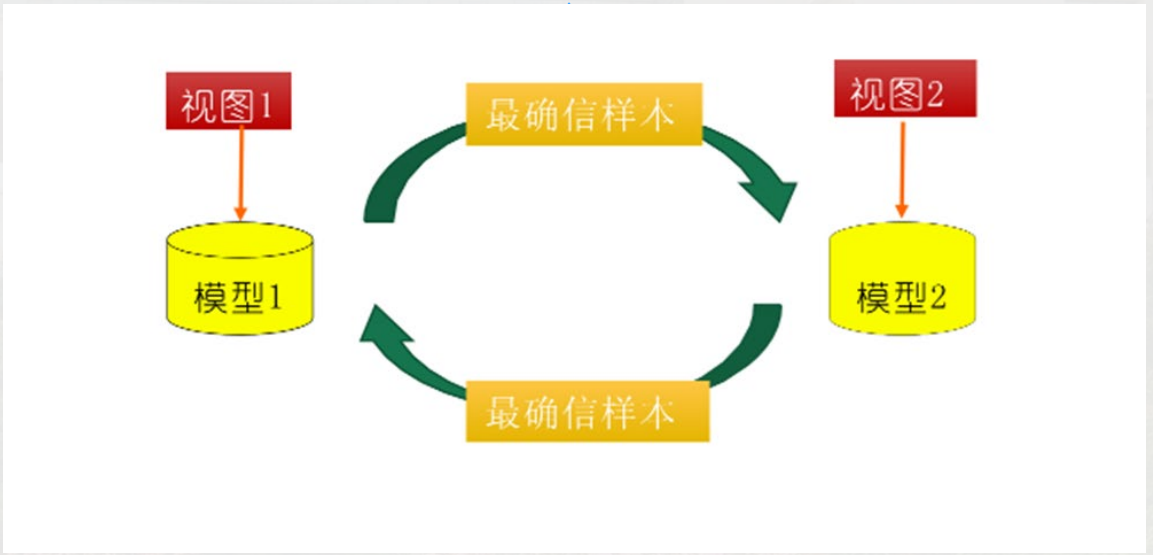
5.基于分歧的方法

- 多视图数据：一个数据对象有多个属性集，每个属性集构成了视图。
  - 样本（$\langle x^1, x^2 \rangle, y$），其中$x^i$为样本在视图i中的示例，y为标记
  - 例如电影中的声音和视频分别对应一个视图，类型则为“动作片”、“爱情片”等
- 多视图具有相容性，进而具有互补性
  - 不同视图输出空间是一致的，以电影为例，类别均应为{爱情片、动作片}
  - 故可利用多视图的互补性加强分类的准确性

5.基于分歧的方法

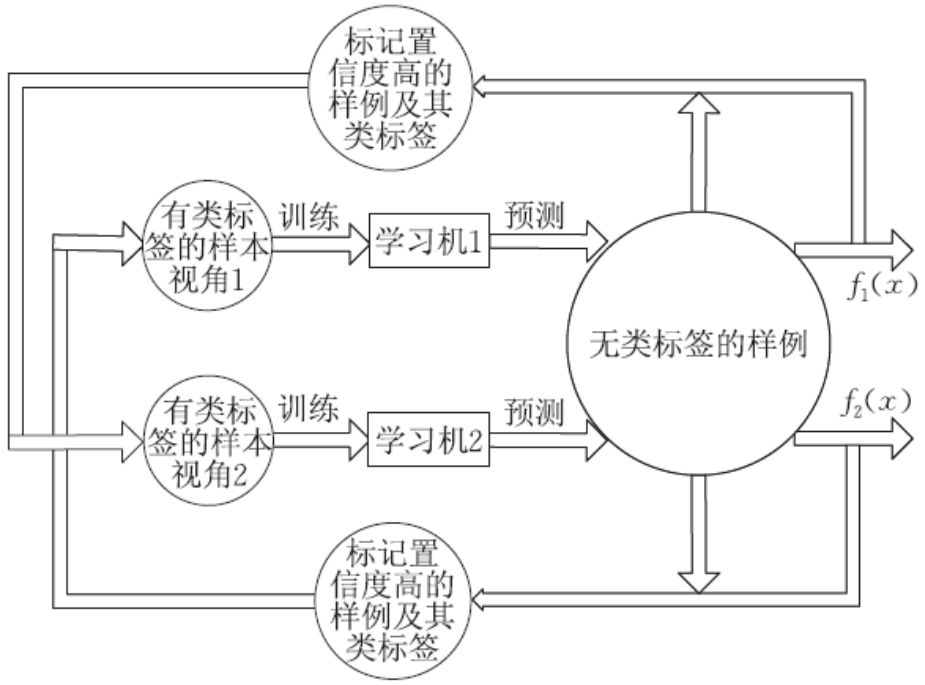
- 协同训练：基于两个充分且条件独立的视图，利用未标记数据相互促进。
  - 充分：每个视图均包含产生最优学习器的信息
  - 条件独立：给定类别标记条件下，两个视图独立

- 基于标记样本训练视图模型1
- 以模型1挑选出该视图最确信的未标记样本赋予其伪标记，并将以上伪标记样本作为新的标记样本加至视图模型2的训练集
- 对视图2模型进行训练，进而挑出该视图模型最确信的未标记样本赋予其伪标记，将其加至视图模型1的训练集
- 重复以上两步，直到两个分类器不变



5.基于分歧的方法

□ 协同训练算法



伪代码

输入: 有标记样本集 $D_l = \{(\langle \boldsymbol{x}_1^1, \boldsymbol{x}_1^2 \rangle, y_1), \ldots, (\langle \boldsymbol{x}_l^1, \boldsymbol{x}_l^2 \rangle, y_l)\}$;
未标记样本集 $D_u = \{\langle \boldsymbol{x}_{l+1}^1, \boldsymbol{x}_{l+1}^2 \rangle, \ldots, \langle \boldsymbol{x}_{l+u}^1, \boldsymbol{x}_{l+u}^2 \rangle\}$;
缓冲池大小 s;
每轮挑选的正例数 p;
每轮挑选的反例数 n;
基学习算法 $\mathfrak{L}$;
学习轮数 T.

设置缓冲池，减少了每轮计算置信度的次数

过程:

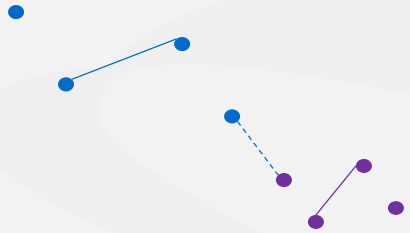
1: 从 D_u 中随机抽取 s 个样本构成缓冲池 D_s;
2: $D_u = D_u \setminus D_s$;
3: **for** $j = 1, 2$ **do**
4: $D_l^j = \{(\boldsymbol{x}_i^j, y_i) \mid (\langle \boldsymbol{x}_i^j, \boldsymbol{x}_i^{3-j} \rangle, y_i) \in D_l\}$; 各视图的有标记样本
5: **end for**
6: **for** $t = 1, 2, \ldots, T$ **do**
7: **for** $j = 1, 2$ **do**
8: $h_j \leftarrow \mathfrak{L}(D_l^j)$; 基于每个视图训练初始学习器
9: 考察 h_j 在 $D_s^j = \{\boldsymbol{x}_i^j \mid \langle \boldsymbol{x}_i^j, \boldsymbol{x}_i^{3-j} \rangle \in D_s\}$ 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 $D_n \subset D_s$;
10: 由 D_p^j 生成伪标记正例 $\tilde{D}_p^{3-j} = \{(\boldsymbol{x}_i^{3-j}, +1) \mid \boldsymbol{x}_i^j \in D_p^j\}$;
11: 由 D_n^j 生成伪标记反例 $\tilde{D}_n^{3-j} = \{(\boldsymbol{x}_i^{3-j}, -1) \mid \boldsymbol{x}_i^j \in D_n^j\}$;
12: $D_s = D_s \setminus (D_p \bigcup D_n)$; 两个学习器挑选的不会有重复
13: **end for**
14: **if** h_1, h_2 均未发生改变 **then**
15: **break**
16: **else**
17: **for** $j = 1, 2$ **do**
18: $D_l^j = D_l^j \bigcup \left(\tilde{D}_p^j \bigcup \tilde{D}_n^j\right)$; 加入打过伪标的未标记样本
19: **end for**
20: 从 D_u 中随机抽取 $2p + 2n$ 个样本加入 D_s 补充缓冲池
21: **end if**
22: **end for**

输出: 分类器 h_1, h_2 最终输出两个分类器做集成

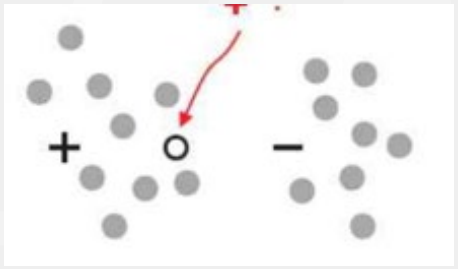
6.半监督聚类

口 拥有部分额外监督信息时：可利用监督信息改善聚类效果。

- 约束：必连（实线）与勿连（虚线）



- 少量有标签样本



6.半监督聚类

□ 约束K-均值算法

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.

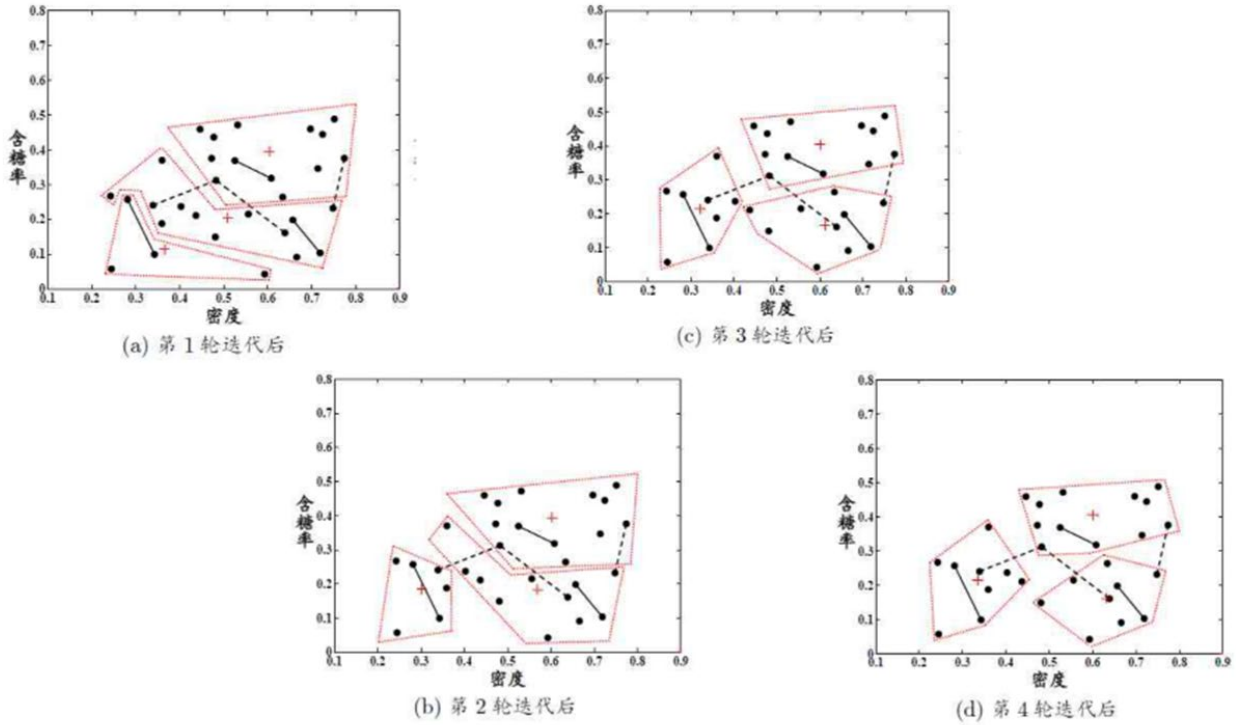
过程:

1: 从 D 中随机选取 k 个样本作为初始均值向量$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$;
2: **repeat**
3: $C_j = \varnothing$ $(1 \leqslant j \leqslant k)$;
4: **for** $i = 1, 2, \ldots, m$ **do**
5: 计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j$ $(1 \leqslant j \leqslant k)$ 的距离: $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2$;
6: $\mathcal{K} = \{1, 2, \ldots, k\}$;
7: is_merged=false;
8: **while** ¬ is_merged **do**
9: 基于 $\mathcal{K}$ 找出与样本 $\boldsymbol{x}_i$ 距离最近的簇: $r = \arg\min_{j \in \mathcal{K}} d_{ij}$;
10: 检测将 $\boldsymbol{x}_i$ 划入聚类簇 C_r 是否会违背 $\mathcal{M}$ 与 $\mathcal{C}$ 中的约束;
11: **if** ¬ is_voilated **then**
12: $C_r = C_r \bigcup \{\boldsymbol{x}_i\}$;
13: is_merged=true
14: **else**
15: $\mathcal{K} = \mathcal{K} \setminus \{r\}$; 若不满足则寻找距离次小的类簇
16: **if** $\mathcal{K} = \varnothing$ **then**
17: **break**并返回错误提示
18: **end if**
19: **end if**
20: **end while**
21: **end for**
22: **for** $j = 1, 2, \ldots, k$ **do**
23: $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$;
24: **end for**
25: **until** 均值向量均未更新

输出: 簇划分 $\{C_1, C_2, \ldots, C_k\}$

对样本进行划分时，需检测是否满足约束关系，其它步骤均相同

- 西瓜数据集4.0,施加若干必连和勿连约束，聚类迭代过程



(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

6.半监督聚类

口 少量有标记样本的聚类算法

输入：样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$；
少量有标记样本 $S=\bigcup_{j=1}^{k}S_j$；
聚类簇数 k.

过程：

```
for j = 1,2,...,k do
    μ_j = (1/|S_j|) Σ_{x∈S_j} x
end for
repeat
    C_j = ∅ (1 ⩽ j ⩽ k);
    for j = 1,2,...,k do
        for all x ∈ S_j do
            C_j = C_j ⋃ {x}
        end for
    end for
    for all x_i ∈ D \ S do
        计算样本 x_i 与各均值向量 μ_j (1 ⩽ j ⩽ k) 的距离: d_ij = ||x_i − μ_j||_2 ;
        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
```

输出：簇划分 $\{C_1,C_2,\ldots,C_k\}$

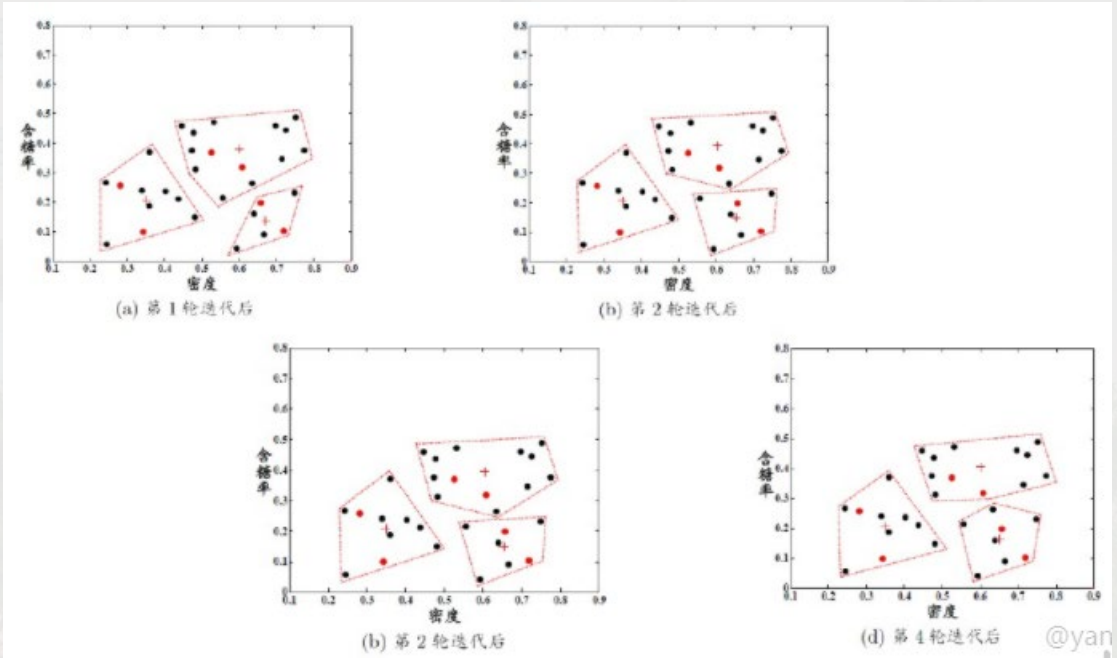
使用带标记样本各类别的均值向量作为初始类中心

带标记样本直接划入对应类簇

划分无标记样本

重新计算类中心

- 西瓜数据集4.0,假定部分样本类别已知



本章小结

- 建立无标记样本利用的概念
- 高斯混合模型实现赋予未标记样本伪标签
- TSVM法利用局部搜索异类错分未标记样本迭代求近似解
- 图半监督学习：二分类的扩散方法和多类情况的迭代传播算法
- 基于分歧的方法：多视图协同训练
- 半监督聚类：先验约束和部分标记样本