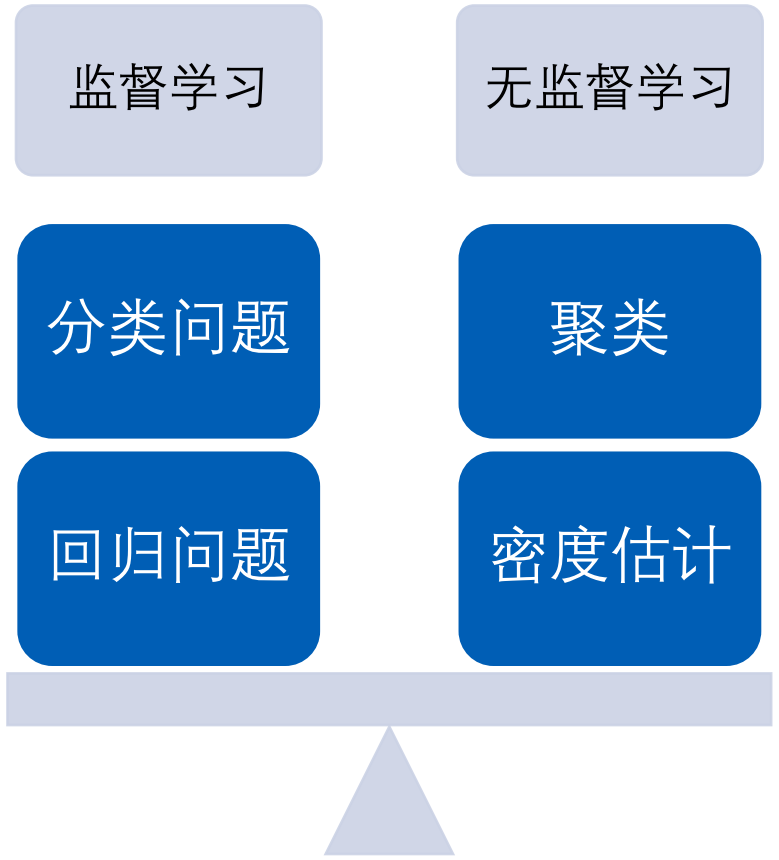
半监督学习

从“西瓜书”开始

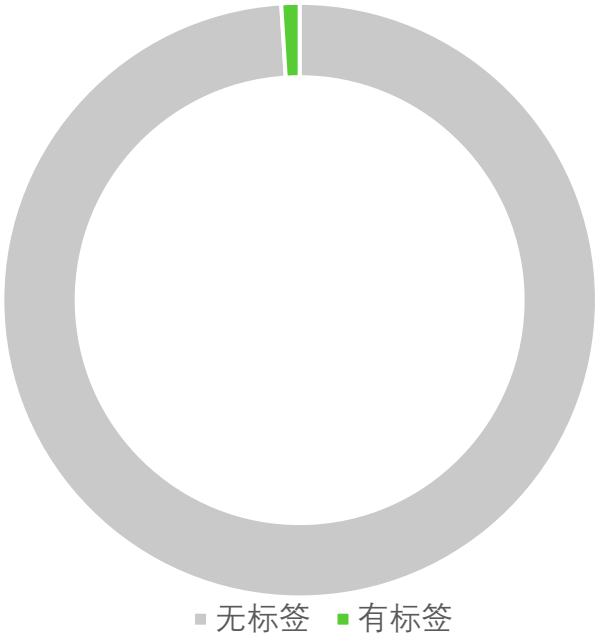
13.1 未标记样本
13.2 生成式方法
13.3 半监督SVM
13.4 图半监督学习
13.5 基于分歧的方法
13.6 半监督聚类

潘贤润
2025/02/23

13.1 未标记样本



半监督学习（semi-supervised learning），即训练集同时包含有标记样本数据和未标记样本数据。



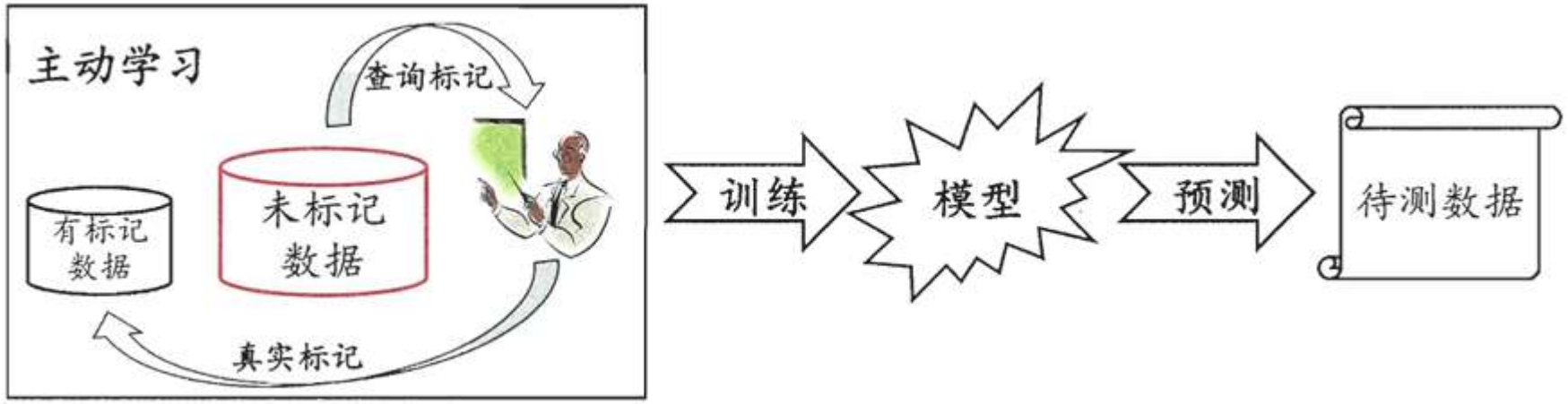
如何同时应用小部分的有标签数据和数量上较多的无标签数据

13.1 未标记样本

- 通过人力或者其他方法来赋予这些数据标签 ➡ 巨大的人力耗费

- **主动学习（active learning）：**

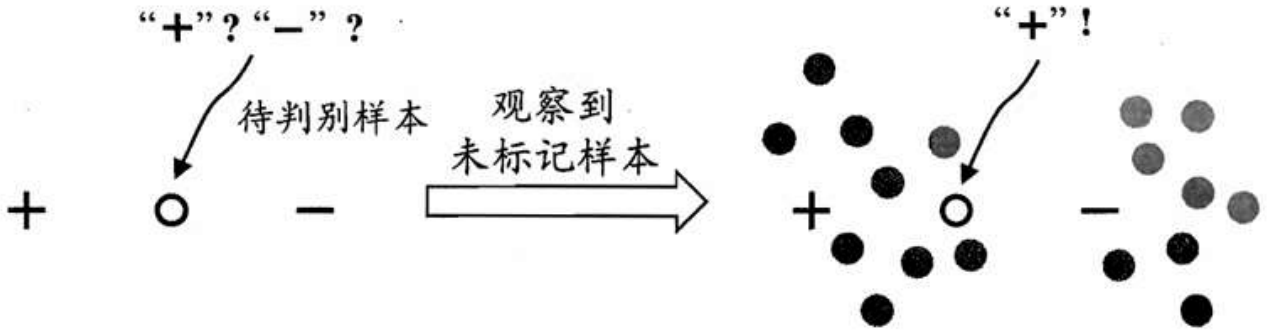
核心思想是从无标签数据中挑出那些对改善模型性能帮助大的数据来打上标签，这样使用尽量少的“查询（query）”来获得尽量好的性能。



本质上仍然属于一种监督学习

13.1 未标记样本

事实上，未标记样本虽未直接包含标记信息，但若它们与有标记样本是从同样的数据源独立同分布采样而来，则它们所包含的关于数据分布的信息对建立模型将大有裨益. 图 13.1 给出了一个直观的例示. 若仅基于图中的一个正例和一个反例，则由于待判别样本恰位于两者正中间，大体上只能随机猜测；若能观察到图中的未标记样本，则将很有把握地判别为正例.



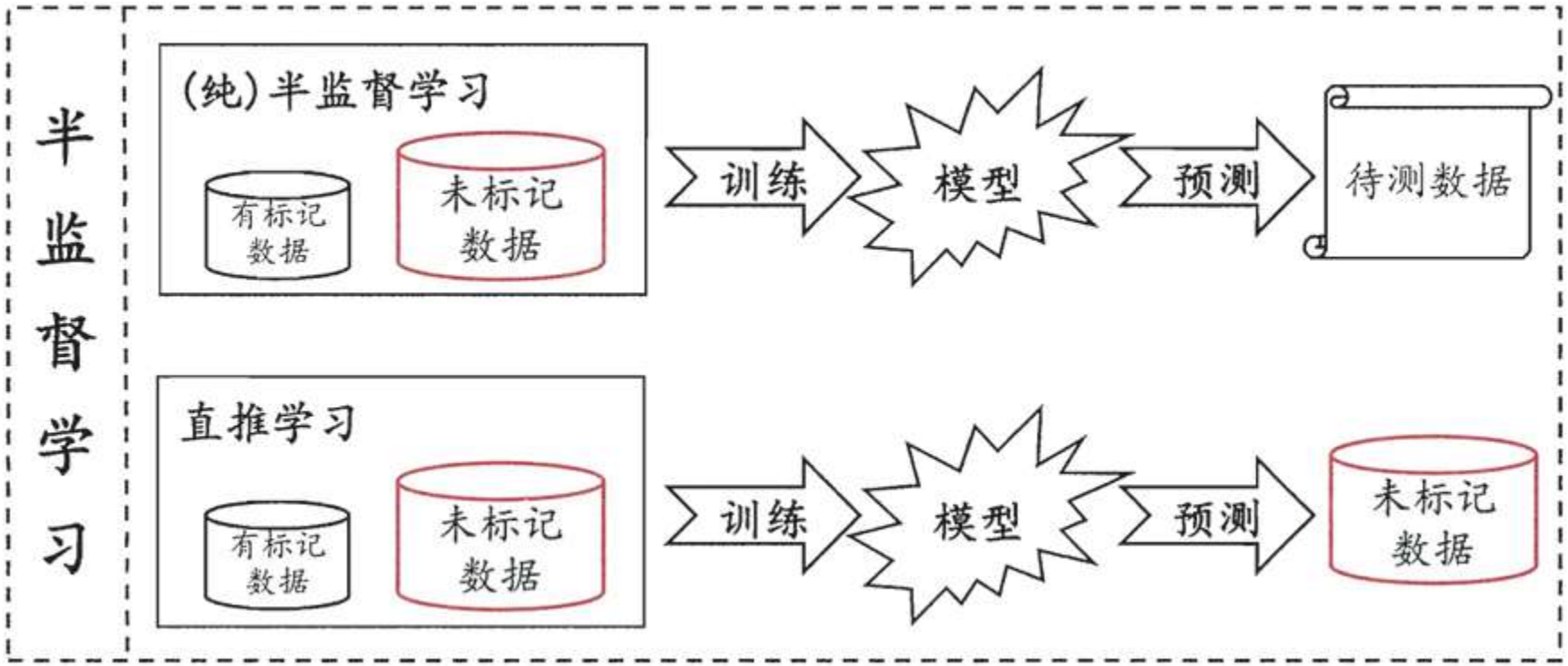
要利用这些未标记样本，需要做一些将未标记样本所揭示的数据分布信息与类别标记相联系的假设，首先我们默认下面讨论都有一个前提，即未标记样本和标记样本是（近似）**独立同分布**的：

- 聚类假设(cluster assumption): 假设数据存在簇结构，同一个簇的样本属于同一个类别。
- 流形假设(manifold assumption): 假设数据分布在一个流形机构上，邻近的样本有相似的输出值。

13.1 未标记样本

半监督学习可以进一步分为两类:

- 纯(pure)半监督学习: 假定训练数据中的未标记样本并非待预测的数据。
- 直推学习(transductive learning): 假定学习过程中所考虑的未标记样本恰好是待预测数据。



13.2 生成式方法

基于概率模型，通过EM算法优化联合似然

生成式方法(generative methods)是直接基于生成式模型的方法. 此类方法假设所有数据(无论是否有标记)都是由同一个潜在的模型“生成”的. 这个假设使得我们能通过潜在模型的参数将未标记数据与学习目标联系起来, 而未标记数据的标记则可看作模型的缺失参数, 通常可基于EM 算法进行极大似然估计求解. 此类方法的区别主要在于生成式模型的假设, 不同的模型假设将产生不同的方法.

EM算法（Expectation-Maximization Algorithm，期望最大化算法）：

通过交替执行**E步**和**M步**，逐步提高模型的对数似然值，最终收敛到一个局部最优解。

13.2 生成式方法

基于概率模型，通过EM算法优化联合似然

模型假设可以为高斯混合模型、混合专家模型［Miller and Uyar, 1997］、朴素贝叶斯模型［Nigam et al., 2000］等，不同模型可推导出不同的生成式半监督学习方法。

- 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。
- 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倒会降低泛化性能 [Cozman and Cohen, 2002]。

高斯混合模型生成方式：

假设样本由高斯混合模型生成，即模型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。

$$p(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \ , \tag{13.1}$$

其中，混合系数 $\alpha_i \geqslant 0$，$\sum_{i=1}^{N} \alpha_i = 1$；$p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ 是样本 $\boldsymbol{x}$ 属于第 i 个高斯混合成分的概率；$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Sigma}_i$ 为该高斯混合成分的参数.

13.2 生成式方法

基于概率模型，通过EM算法优化联合似然

令 $f(\boldsymbol{x}) \in \mathcal{Y}$ 表示模型 f 对 $\boldsymbol{x}$ 的预测标记, $\Theta \in \{1, 2, \ldots, N\}$ 表示样本 $\boldsymbol{x}$ 隶属的高斯混合成分. 由最大化后验概率可知

$$
\begin{aligned}
f(\boldsymbol{x}) &= \arg\max_{j \in \mathcal{Y}} p(y = j \mid \boldsymbol{x}) \\
&= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{N} p(y = j, \Theta = i \mid \boldsymbol{x}) \\
&= \arg\max_{j \in \mathcal{Y}} \sum_{i=1}^{N} p(y = j \mid \Theta = i, \boldsymbol{x}) \cdot p(\Theta = i \mid \boldsymbol{x}) \,, \qquad (13.2)
\end{aligned}
$$

$$
\begin{aligned}
p(y = j, \Theta = i|\boldsymbol{x}) &= \frac{p(y = j, \Theta = i, \boldsymbol{x})}{p(\boldsymbol{x})} \\
&= \frac{p(y = j, \Theta = i, \boldsymbol{x})}{p(\Theta = i, \boldsymbol{x})} \cdot \frac{p(\Theta = i, \boldsymbol{x})}{p(\boldsymbol{x})} \\
&= p(y = j|\Theta = i, \boldsymbol{x}) \cdot p(\Theta = i|\boldsymbol{x})
\end{aligned}
$$

其中

$$
p(y = j \mid \Theta = i, \boldsymbol{x}) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}
$$

$$
p(\Theta = i \mid \boldsymbol{x}) = \frac{\alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \qquad (13.3)
$$

$$
p(\Theta = i|x) = \frac{p(\Theta = i, x)}{p(x)}
$$

13.2 生成式方法

基于概率模型，通过EM算法优化联合似然

给定有标记样本集 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$ 和未标记样本集 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$, $l \ll u$, $l + u = m$. 假设所有样本独立同分布, 且都是由同一个高斯混合模型生成的. 用极大似然法来估计高斯混合模型的参数 $\{(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \mid 1 \leqslant i \leqslant N\}$, $D_l \cup D_u$ 的对数似然是

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

南瓜书P165
9.4.3 高斯混合聚类

- E 步: 根据当前模型参数计算未标记样本 $\boldsymbol{x}_j$ 属于各高斯混合成分的概率

$$\gamma_{ji} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} ; \tag{13.5}$$

- M 步: 基于 γ_{ji} 更新模型参数, 其中 l_i 表示第 i 类的有标记样本数目

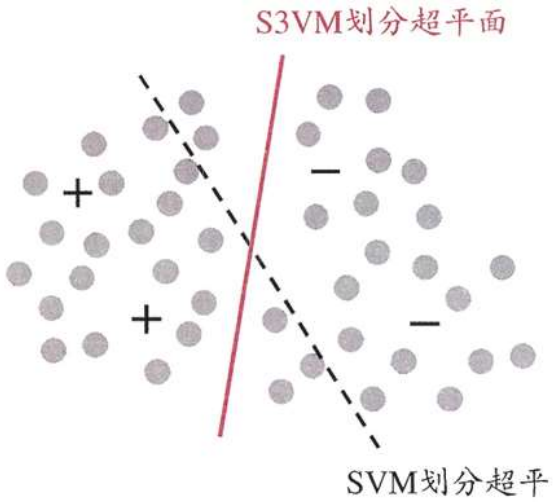
$$\boldsymbol{\mu}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j \right) , \tag{13.6}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}} + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^{\mathrm{T}} \right) , \tag{13.7}$$

$$\alpha_i = \frac{1}{m} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) . \tag{13.8}$$

13.3 半监督SVM

基于低密度分离假设，通过混合整数规划优化决策边界



- 在不考虑未标记样本时，支持向量机试图找到最大间隔划分超平面
- 在考虑未标记样本后，S3VM （Semi-Supervised Support Vector Machine，半监督支持向量机）试图找到能将两类有标记样本分开，且穿过数据低密度区域的划分超平面

的学习方法. TSVM 试图考虑对未标记样本进行各种可能的标记指派(label assignment), 即尝试将每个未标记样本分别作为正例或反例, 然后在所有这些结果中, 寻求一个在所有样本(包括有标记样本和进行了标记指派的未标记样本)上间隔最大化的划分超平面. 一旦划分超平面得以确定, 未标记样本的最终标记指派就是其预测结果.

13.3 半监督SVM

基于低密度分离假设，通过混合整数规划优化决策边界

形式化地说, 给定 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$ 和 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$, 其中 $y_i \in \{-1, +1\}$, $l \ll u$, $l + u = m$. TSVM 的学习目标是为 D_u 中的样本给出预测标记 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$, $\hat{y}_i \in \{-1, +1\}$, 使得

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \hat{\boldsymbol{y}}, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i \qquad (13.9) \\
\text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i, \quad i = 1, 2, \ldots, l, \\
& \hat{y}_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1 - \xi_i, \quad i = l+1, l+2, \ldots, m, \\
& \xi_i \geqslant 0, \quad i = 1, 2, \ldots, m,
\end{aligned}
$$

其中, $(\boldsymbol{w}, b)$ 确定了一个划分超平面; $\boldsymbol{\xi}$ 为松弛向量, ξ_i $(i = 1, 2, \ldots, l)$ 对应于有标记样本, ξ_i $(i = l+1, l+2, \ldots, m)$ 对应于未标记样本; C_l 与 C_u 是由用户指定的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.

标注数据: Dl
未标注数据: Du

超平面方程: $\boldsymbol{w}^{\top}\boldsymbol{x} + b = 0$

13.3 半监督SVM

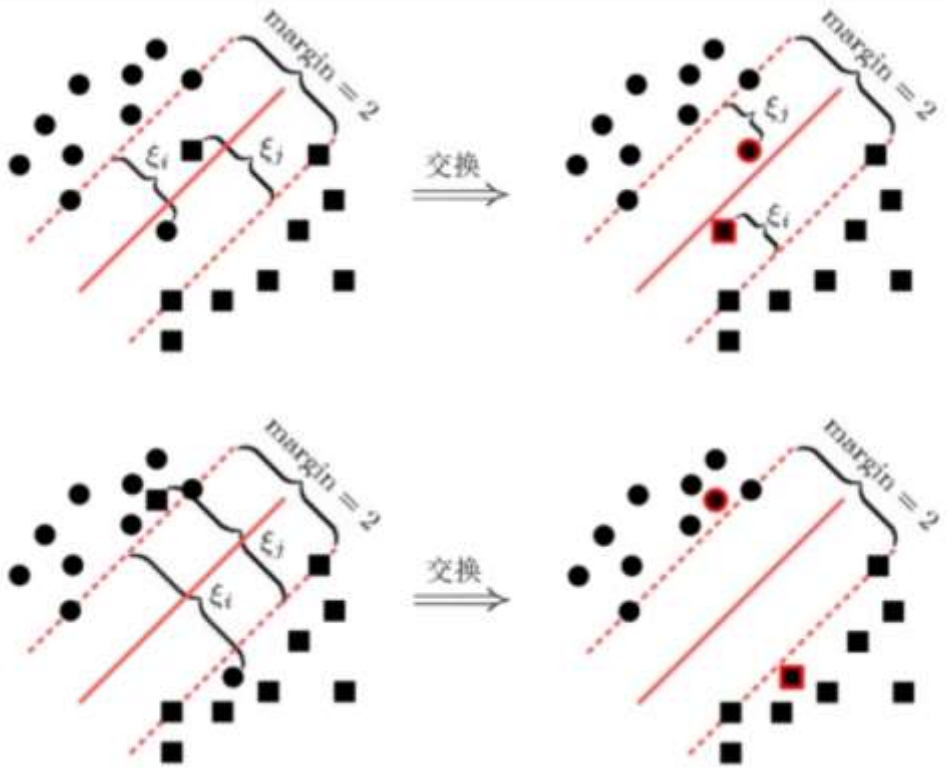
基于低密度分离假设，通过混合整数规划优化决策边界

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 SVM_l;
用 SVM_l 对 D_u 中样本进行预测, 得到 $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$;
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: $\hat{\boldsymbol{y}} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \ldots, \hat{y}_{l+u})$



13.3 图半监督学习

基于图结构，通过正则化目标函数实现标签传播

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，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。

给定 $D_l = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_l, y_l)\}$ 和 $D_u = \{\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}\}$, $l \ll u$, $l+u=m$. 我们先基于 $D_l \cup D_u$ 构建一个图 $G=(V,E)$, 其中结点集 $V = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_{l+u}\}$, 边集 E 可表示为一个亲和矩阵 (affinity matrix), 常基于高斯函数定义为

用于表示数据点之间相似性或关联性的矩阵，值越大表示相似性越高

$$(\mathbf{W})_{ij} = \begin{cases} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{2\sigma^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise}, \end{cases} \tag{13.11}$$

其中 $i, j \in \{1, 2, \ldots, m\}$, $\sigma > 0$ 是用户指定的高斯函数带宽参数.

13.3 图半监督学习

基于图结构，通过正则化目标函数实现标签传播

假定从图 $G=(V,E)$ 将学得一个实值函数 $f: V \to \mathbb{R}$, 其对应的分类规则为: $y_i = \mathrm{sign}(f(\boldsymbol{x}_i))$, $y_i \in \{-1,+1\}$. 直观上看, 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]:

$$
\begin{aligned}
E(f) &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right)^2 \\
&= \frac{1}{2}\left(\sum_{i=1}^{m} d_i f^2(\boldsymbol{x}_i) + \sum_{j=1}^{m} d_j f^2(\boldsymbol{x}_j) - 2\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j)\right) \\
&= \sum_{i=1}^{m} d_i f^2(\boldsymbol{x}_i) - \sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j) \\
&= \boldsymbol{f}^{\mathrm{T}}(\mathbf{D}-\mathbf{W})\boldsymbol{f}\ ,
\end{aligned}
$$

南瓜书P171

即式 (13.12) 的第 3 行, 其中第一项 $\sum_{i=1}^{m} d_i f^2(\boldsymbol{x}_i)$ 可以写为如下矩阵形式:

$$= \boldsymbol{f}^{\mathrm{T}}\mathbf{D}\boldsymbol{f}$$

第二项 $\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j)$ 也可以写为如下矩阵形式:

$$
\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij} f(\boldsymbol{x}_i) f(\boldsymbol{x}_j) \\
&= \begin{bmatrix} f(\boldsymbol{x}_1) & f(\boldsymbol{x}_2) & \cdots & f(\boldsymbol{x}_m) \end{bmatrix}
\begin{bmatrix} (\mathbf{W})_{11} & (\mathbf{W})_{12} & \cdots & (\mathbf{W})_{1m} \\ (\mathbf{W})_{21} & (\mathbf{W})_{22} & \cdots & (\mathbf{W})_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{W})_{m1} & (\mathbf{W})_{m2} & \cdots & (\mathbf{W})_{mm} \end{bmatrix}
\begin{bmatrix} f(\boldsymbol{x}_1) \\ f(\boldsymbol{x}_2) \\ \vdots \\ f(\boldsymbol{x}_m) \end{bmatrix} \\
&= \boldsymbol{f}^{\mathrm{T}}\mathbf{W}\boldsymbol{f}
\end{aligned}
$$

所以 $E(f) = \boldsymbol{f}^{\mathrm{T}}\mathbf{D}\boldsymbol{f} - \boldsymbol{f}^{\mathrm{T}}\mathbf{W}\boldsymbol{f} = \boldsymbol{f}^{\mathrm{T}}(\mathbf{D}-\mathbf{W})\boldsymbol{f}$, 即式 (13.12)。

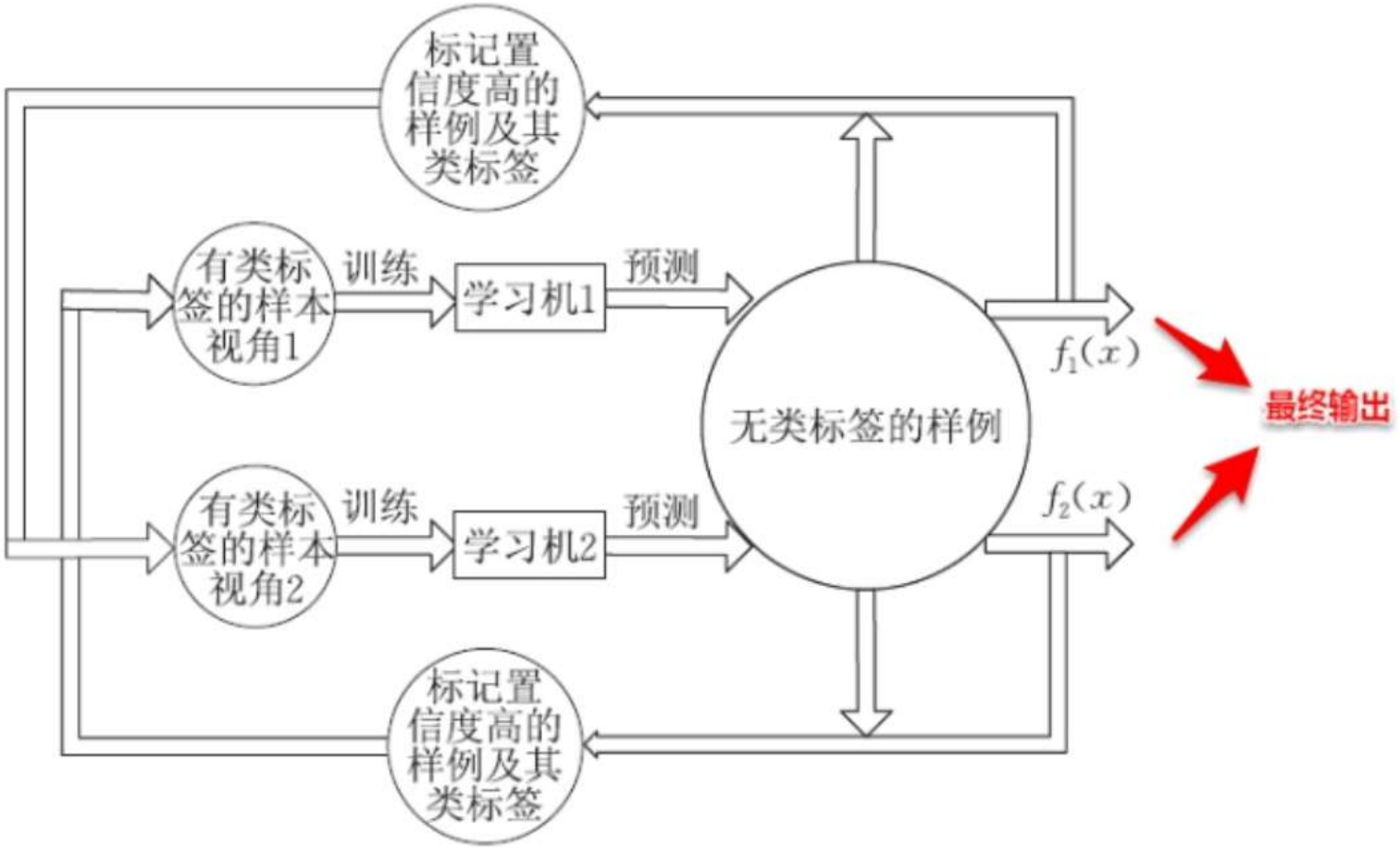
维度	生成式方法	半监督SVM（S3VM）	图半监督学习
核心思想	假设数据由概率模型生成，通过建模分布参数联合优化有标签和无标签数据。	基于低密度分离假设，调整决策边界使其位于无标签数据密度低的区域。	基于流形或平滑假设，利用图结构传播标签，使相似节点标签一致。
技术手段	概率模型（如高斯混合模型+EM算法）估计数据分布，利用贝叶斯定理分类。	扩展传统SVM，优化目标包含有标签数据的分类误差和无标签数据的间隔最大化。	构建图（节点为数据，边为相似度），通过标签传播或正则化约束（如拉普拉斯正则化）实现平滑预测。
关键假设	数据符合预设的生成模型（如高斯分布），模型结构正确性直接影响性能。	类别边界位于低密度区域，无标签数据能帮助确定边界位置。	数据在流形上分布，邻近节点标签相似。
计算复杂度	依赖EM等迭代优化，复杂度中等，但对模型假设敏感。	组合优化问题（需枚举无标签数据可能的标签），计算复杂度高，常用启发式近似。	构建图的存储和计算复杂度高（尤其大规模数据），但优化过程通常线性或凸。
适用场景	数据生成机制明确（如文本、图像生成）。	类别间存在明显低密度区域（如二分类边界清晰）。	数据具有局部结构或流形特性（如社交网络、生物信息数据）。

13.4 基于分歧的方法

多个学习器之间的分歧

与前边介绍的只基于单学习器的方法不同，基于分歧的方法（disagreement-based methods）通过多个学习器之间的分歧（disagreement）/多样性（diversity）来利用未标记样本数据。

协同训练就是其中的一种经典方法，它最初是针对于多视图（multi-view）数据而设计的



（1）在每个视图上基于有标记样本分别训练出一个分类器

（2）让每个分类器分别挑选自己“最有把握”的未标记样本赋予伪标记

（3）将伪标记样本提供给另一个分类器作为新增的有标记样本用于训练更新

（4）重复（1）、（2）直到两个学习器不再变化或达到预设的迭代次数时停止

多个学习器之间的分歧

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.

过程:

从 D_u 中随机抽取 s 个样本构成缓冲池 D_s;
$D_u = D_u \setminus D_s$;
for $j = 1, 2$ **do**
  $D_l^j = \{(\boldsymbol{x}_i^j, y_i) \mid (\langle \boldsymbol{x}_i^j, \boldsymbol{x}_i^{3-j} \rangle, y_i) \in D_l\}$;
end for
for $t = 1, 2, \ldots, T$ **do**
  for $j = 1, 2$ **do**
    $h_j \leftarrow \mathfrak{L}(D_l^j)$;
    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 $D_n \subset D_s$;
    由 D_p^j 生成伪标记正例 $\tilde{D}_p^{3-j} = \{(\boldsymbol{x}_i^{3-j}, +1) \mid \boldsymbol{x}_i^j \in D_p^j\}$;
    由 D_n^j 生成伪标记反例 $\tilde{D}_n^{3-j} = \{(\boldsymbol{x}_i^{3-j}, -1) \mid \boldsymbol{x}_i^j \in D_n^j\}$;
    $D_s = D_s \setminus (D_p \bigcup D_n)$;
  end for
  if h_1, h_2 均未发生改变 **then**
    break
  else
    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
    从 D_u 中随机抽取 $2p + 2n$ 个样本加入 D_s
  end if
end for

输出: 分类器 h_1, h_2

图 13.6 协同训练算法

缓冲池大小:每次从无标记数据中随机抽取的样本数量

13.5 半监督聚类

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.

13.5 半监督聚类

约束k均值（Constrained k-means）算法是利用第一类监督信息的代表，其在 k 均值算法的基础上考虑了“必连”和“勿连”约束。

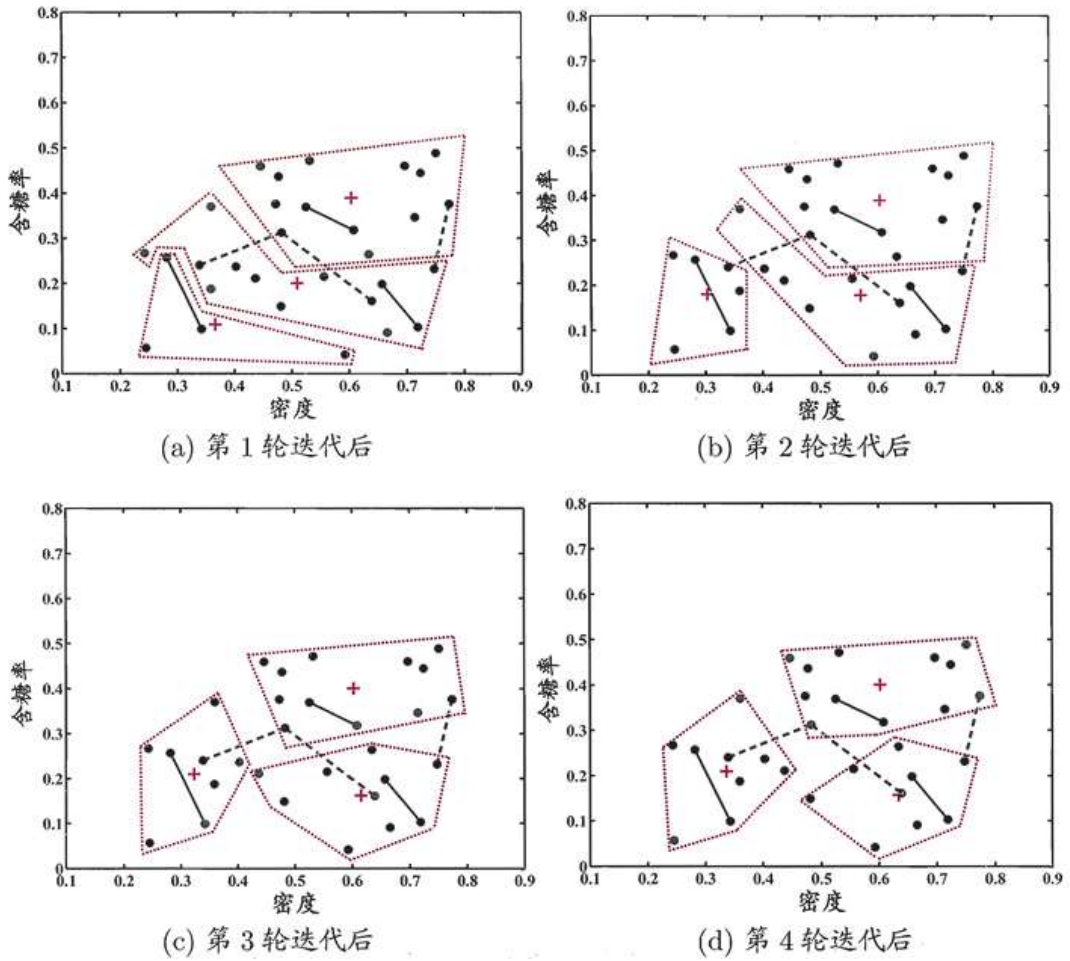
输入: 样本集 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$;
必连约束集合 $\mathcal{M}$;
勿连约束集合 $\mathcal{C}$;
聚类簇数 k.

过程:

```
从 D 中随机选取 k 个样本作为初始均值向量{μ_1, μ_2, ..., μ_k};
repeat
    C_j = ∅ (1 ⩽ j ⩽ k);
    for i = 1, 2, ..., m 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;
        K = {1, 2, ..., k};
        is_merged=false;
        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
    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 $\{C_1,C_2,\ldots,C_k\}$

图 13.7 约束 k 均值算法



(a) 第 1 轮迭代后　(b) 第 2 轮迭代后

(c) 第 3 轮迭代后　(d) 第 4 轮迭代后

13.5 半监督聚类

约束种子k均值（Constrained Seed k-means）算法是使用少量标记数据来辅助聚类计算。

与k均值算法一样，不同点在于利用有标记样本进行类中心的指定，同时在对样本进行划分时，不需要改变有标记样本的簇隶属关系，直接将其划分到对应类簇即可。

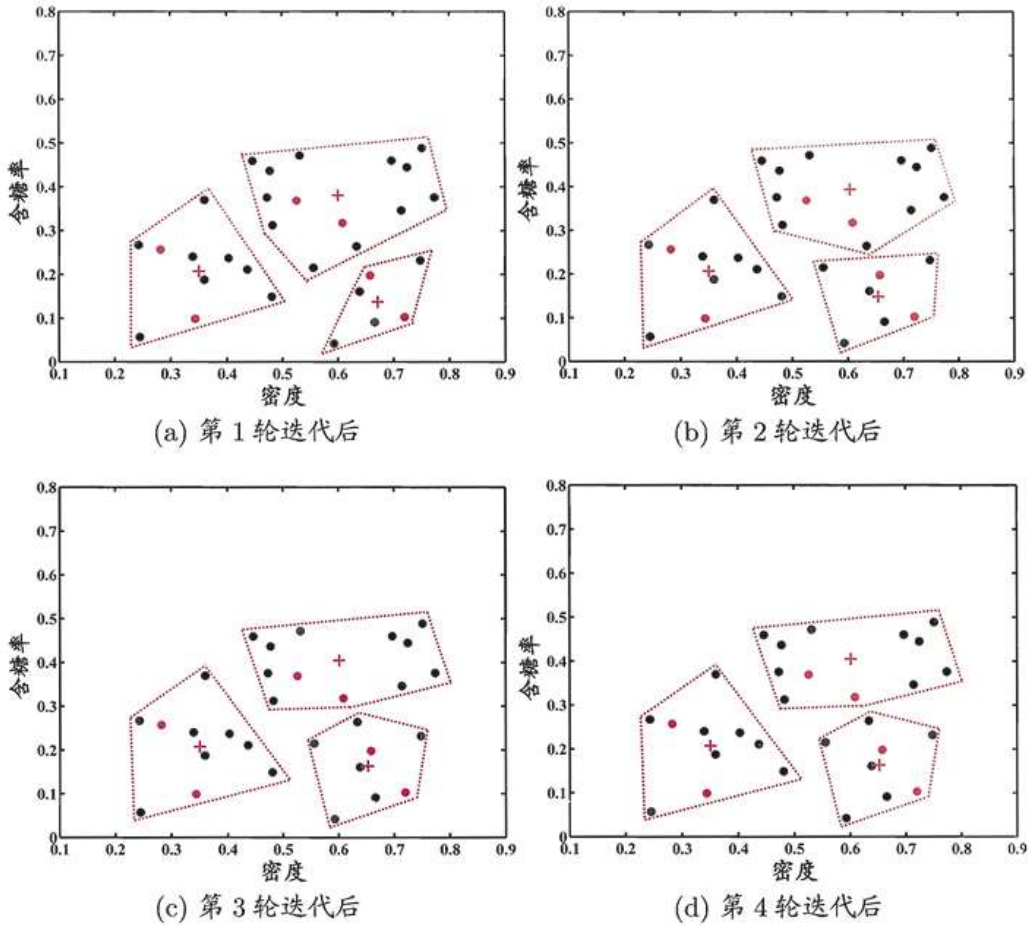
输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
少量有标记样本 $S = \bigcup_{j=1}^{k} S_j$;
聚类簇数 k.

过程:

1: **for** $j = 1, 2, \ldots, k$ **do**
2: $\quad \boldsymbol{\mu}_j = \frac{1}{|S_j|} \sum_{\boldsymbol{x} \in S_j} \boldsymbol{x}$
3: **end for**
4: **repeat**
5: $\quad C_j = \varnothing \ (1 \leqslant j \leqslant k)$;
6: $\quad$ **for** $j = 1, 2, \ldots, k$ **do**
7: $\quad\quad$ **for all** $\boldsymbol{x} \in S_j$ **do**
8: $\quad\quad\quad C_j = C_j \bigcup \{\boldsymbol{x}\}$
9: $\quad\quad$ **end for**
10: $\quad$ **end for**
11: $\quad$ **for all** $\boldsymbol{x}_i \in D \setminus S$ **do**
12: $\quad\quad$ 计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j \ (1 \leqslant j \leqslant k)$ 的距离: $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2$;
13: $\quad\quad$ 找出与样本 $\boldsymbol{x}_i$ 距离最近的簇: $r = \arg\min_{j \in \{1,2,\ldots,k\}} d_{ij}$;
14: $\quad\quad$ 将样本 $\boldsymbol{x}_i$ 划入相应的簇: $C_r = C_r \bigcup \{\boldsymbol{x}_i\}$
15: $\quad$ **end for**
16: $\quad$ **for** $j = 1, 2, \ldots, k$ **do**
17: $\quad\quad \boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$;
18: $\quad$ **end for**
19: **until** 均值向量均未更新

输出: 簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.9 约束种子 k 均值算法



(a) 第 1 轮迭代后　(b) 第 2 轮迭代后

(c) 第 3 轮迭代后　(d) 第 4 轮迭代后

感谢倾听

See you next.