统计机器学习
Statistical Machine Learning

魏莱

上海海事大学信息工程学院

2019 年 4 月 4 日

第十二章：半监督学习

1. 无标记样本

在现实世界中，对样本进行标记是一件耗时耗力的工作，而无标记的数据往往容易得到。例如，标记一个网页属于新闻亦或文学的需要人们去阅读才能标记，但互联网上有大量无标记的网页很容易获取。

假设训练数据集 $\mathbf{D}_l = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)\}$，这里 $\forall \mathbf{x}_i$ 都有其相应的类别 y_i，称为**有标记（labeled）样本**。而一部分数据 $\mathbf{D}_u = \{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$ 没有标签，称为**无标记（unlabeled）样本**。通常情况下 $u \gg l$。

1. 无标记样本

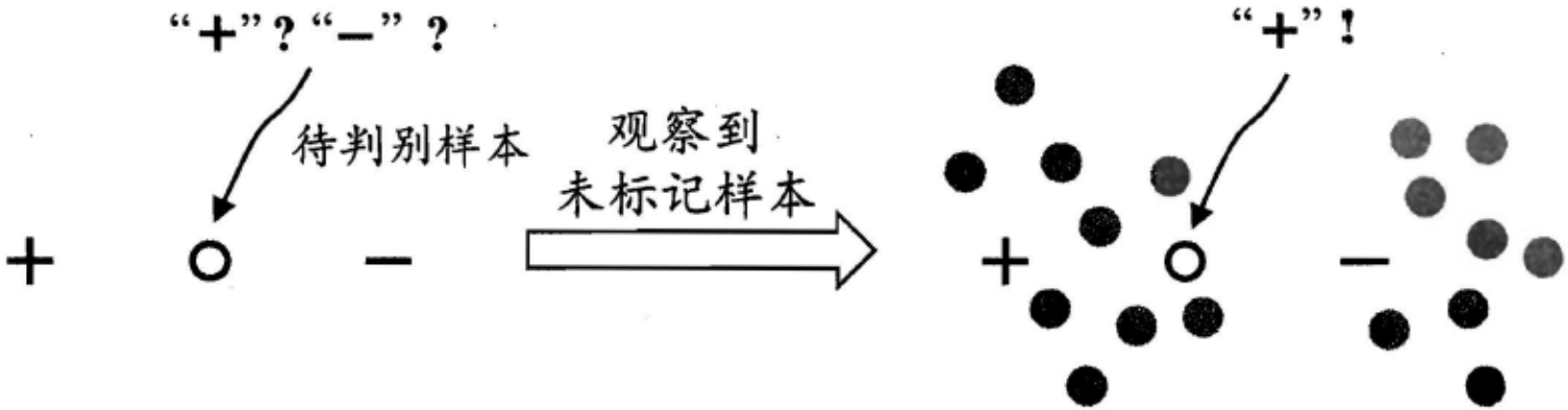
若直接使用传统监督学习技术，则仅有 $\mathbf{D}_l$ 能用于构建模型，$\mathbf{D}_u$ 所包含的信息被浪费了. 另一方面，若 $\mathbf{D}_l$ 较小，则由于训练样本不足，学得模型的泛化能力往往不佳。那么，能否在构建模型的过程中将 $\mathbf{D}_u$ 利用起来呢？

1. 无标记样本

利用 $\mathbf{D}_u$ 的一个简单方法是：通过 $\mathbf{D}_l$ 训练一个分类器，然后将 $\mathbf{D}_u$ 中的样本进行分类，选择一个分类最可靠的样本加入 $\mathbf{D}_l$，然后重新训练分类器，再对 $\mathbf{D}_u$ 中样本分类，再选择一个分类最可靠的样本加入 $\mathbf{D}_l$，$\cdots$，通过不断扩大 $\mathbf{D}_l$，减小 $\mathbf{D}_u$ 来最终完成分类器学习。这种方法称为**主动学习 (active learning)**。

1. 无标记样本

主动学习引入了额外的专家知识，通过与外界的交互来将部分未标记样本转变为有标记样本。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来样而来，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。



1. 无标记样本

让学习器不依赖外界交互、自动地利用未标记样本来提升学习性能，就是**半监督学习 (semi-supervised learning)**。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。最常见的假设包括两个，一是**聚类假设 (cluster assumption)**，即假设数据存在簇结构，同一个簇的样本属于同一个类别。另一是**流形假设 (manifold assumption)**，即假设数据分布在一个流形结构上，邻近的样本拥有相似的输出。

2. 生成式方法

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。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。

2. 生成式方法

给定样本 $\mathbf{x}$，其真实类别标记为 $y \in \{1, 2, \cdots, C\}$。假设样本由高斯混合模型生成，且每一类别对应一个高斯混合成分。即数据样本的概率密度可以写成：

$$p(\mathbf{x}) = \sum_{i=1}^{C} w_i p(\mathbf{x}|\mu_i, \Sigma_i), \tag{1}$$

其中，混合系数 $w_i \geq 0, \sum_{i=1}^{C} w_i = 1$，$p(\mathbf{x}|\mu_i, \Sigma_i)$ 是样本 $\mathbf{x}$ 属于第 i 个高斯混合成分的概率，μ_i, Σ_i 为该高斯混合成分的参数。

令 $f(\mathbf{x})$ 表示模型 f 对 $\mathbf{x}$ 的预测标记，则最大化后验概率可以表示为：

$$\begin{aligned} f(\mathbf{x}) &= \arg\max p(y=j|\mathbf{x}) \\ &= \arg\max \textstyle\sum_{i=1}^{C} p(y=j,\Theta=i|\mathbf{x}) \\ &= \arg\max \textstyle\sum_{i=1}^{C} p(y=j|\Theta=i,\mathbf{x})p(\Theta=i|\mathbf{x}) \end{aligned} \tag{2}$$

其中 $\Theta \in \{1,2,\cdots,C\}$, $p(\Theta=i|\mathbf{x}) = \frac{w_i p(\mathbf{x}|\mu_i,\Sigma_i)}{\sum_{i=1}^{C} w_i p(\mathbf{x}|\mu_i,\Sigma_i)}$ 表示样本 $\mathbf{x}$ 由第 i 个高斯混合成分生成的后验概率，$p(y=j|\Theta=i,\mathbf{x})$ 为 $\mathbf{x}$ 由第 i 个高斯混成成分生成且其类别为 j 的概率。

可以发现式（2）中，$p(y=j|\Theta=i,\mathbf{x})$ 需要知道样本标记，而 $p(\Theta_i|\mathbf{x})$ 不涉及样本标记，因此可以利用有标记和未标记数据。通过引入大量的未标记数据，对这一项的估计可望由于数据量的增长而更为准确。于是式（2）整体的估计可能会更准确。

对标记样本集 $\mathbf{D}_l$ 和未标记样本集 $\mathbf{D}_u$ 中样本独立同分布，且由同一高斯混合模型生成，用极大似然估计高斯混合模型的参数 $\{(w_i, \mu_i, \Sigma_i)|1 \leq i \leq C\}$，则 $\mathbf{D}_l \bigcup \mathbf{D}_u$ 的对数似然为

$$\begin{aligned} LL(\mathbf{D}_l \bigcup \mathbf{D}_u) &= \sum_{\mathbf{x}_j \in \mathbf{D}_l} \ln \left(\sum_{i=1}^{C} w_i p(\mathbf{x}|\mu_i, \Sigma_i) p(y_j|\Theta = i, \mathbf{x}_j) \right) \\ &+ \sum_{\mathbf{x}_j \in \mathbf{D}_u} \ln \left(\sum_{i=1}^{C} w_i p(\mathbf{x}|\mu_i, \Sigma_i) \right) \end{aligned} \tag{3}$$

上式由两项组成：基于有标记数据 $\mathbf{D}_l$ 的有监督项和基于未标记数据 $\mathbf{D}_u$ 的无监督项。

按照第八章聚类内容，上式可以通过 EM 算法来求解 w_j, μ_j, Σ_j。

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

$$\gamma_{ji} = \frac{w_i p(\mathbf{x}|\mu_i, \Sigma_i)}{\sum_{i=1}^{C} w_i p(\mathbf{x}|\mu_i, \Sigma_i)} \tag{4}$$

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

$$\mu_i = \frac{1}{\sum_{\mathbf{x}_j \in \mathbf{D}_u} \gamma_{ji} + l_i} \Big(\sum_{\mathbf{x}_j \in \mathbf{D}_u} \gamma_{ji} \mathbf{x}_j + \sum_{\mathbf{x}_j \in \mathbf{D}_l \bigwedge y_j = i} \mathbf{x}_j \Big) \tag{5}$$

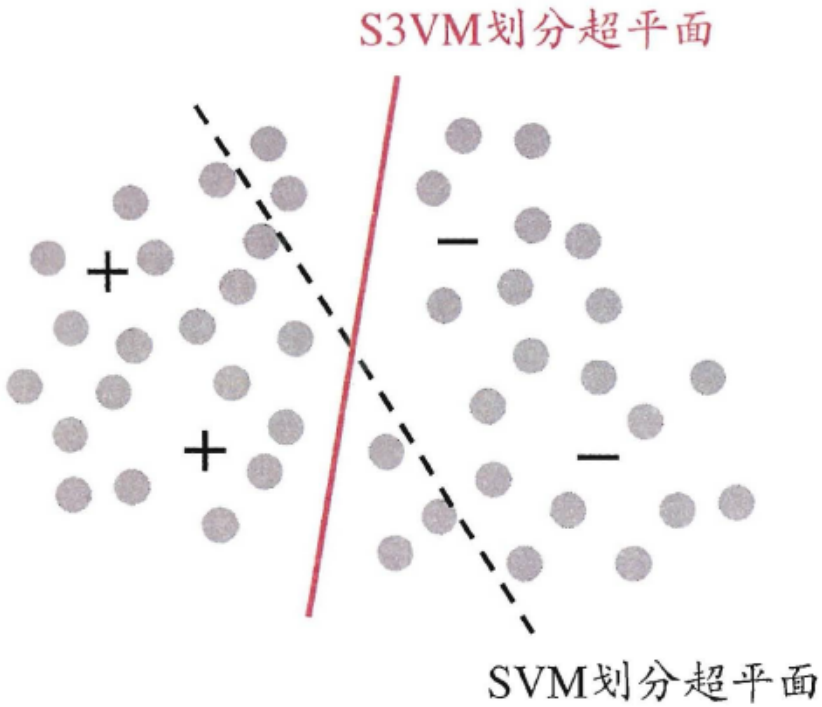
$$\begin{aligned} \Sigma_i \quad &= \frac{1}{\sum_{\mathbf{x}_j \in \mathbf{D}_u} \gamma_{ji} + l_i} \Big(\sum_{\mathbf{x}_j \in \mathbf{D}_u} \gamma_{ij} (\mathbf{x}_i - \mu_i)(\mathbf{x}_i - \mu_i)^t \\ &+ \sum_{\mathbf{x}_j \in \mathbf{D}_l \bigwedge y_j = i} (\mathbf{x}_i - \mu_i)(\mathbf{x}_i - \mu_i)^t \Big) \end{aligned} \tag{6}$$

$$w_i = \frac{1}{n}\Big(\sum_{\mathbf{x}_j \in \mathbf{D}_u} \gamma_{ji} + l_i \Big) \tag{7}$$

以上过程不断迭代直至收敛，即时获得模型参数。

3. 半监督 SVM

半监督支持向量机 (Semi-Supervised Support Vector Machine，简称 S3VM)是支持向量机在半监督学习上的推广。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，而在考虑未标记样本后，S3VM 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，如下图所示。



3. 半监督 SVM

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

形式化的说，TSVM 的学习目标是为 $\mathbf{D}_u$ 中的样本给出预测标记 $\mathbf{y} = (y_{l+1}, \cdots, y_{l+u}), y_i \in \{-1, +1\}$，使得：

$$\begin{aligned} \min_{\mathbf{w},\mathbf{y},\varepsilon,b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda_l \sum_{i=1}^{l} \varepsilon_i + \lambda_u \sum_{i=l+1}^{l+u} \varepsilon_i \\ s.t. \quad & y_i(\mathbf{w^t x_i} + \mathbf{b}) \geq 1 - \varepsilon_i, i = 1, 2, \cdots, l \\ & y_i(\mathbf{w^t x_i} + \mathbf{b}) \geq 1 - \varepsilon_i, i = l+1, l+2, \cdots, l+u \\ & \varepsilon_i \geq 0, i = 1, 2, \cdots, l+u \end{aligned} \tag{8}$$

其中 $(\mathbf{w}, b)$ 确定一个划分超平面；ε_i 为松弛变量，λ_l, λ_u 为两个参数。

3. 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，仅当未标记样本很少时才有可能直接求解。在一般情形下必须考虑更高效的优化策略。TSVM 采用局部搜索来迭代地寻找式 (8) 的近似解。即，

1. 利用带标记数据训练标准 SVM 模型；
2. 利用学习得到的 SVM 对未标记样本分类，得到“伪标记”；
3. 利用所有数据重新训练 SVM，令 $\lambda_l > \lambda_u$;
4. 交换伪标记为异类的两个可能错分的样本标记重新训练 SVM，逐步增大 λ_u，循环此步骤，直到 $\lambda_l = \lambda_u$。

上述算法中一个关键问题是：如何寻找伪标记为异类的两个可能错分的样本。实际上，若两个样本 $\mathbf{x}_i, \mathbf{x}_j$ 标记为异类，如果有 $\varepsilon_i + \varepsilon_j > 2$，则他们很可能被错分。因此可以按照这个策略选择一对错分样本。

4. 图半监督学习

谱图学习（spectral learning）是希望将一个数据集映射成一个谱图（spectral），通过谱图上的一些性质来发现数据集的某些结构。具体说来，构建谱图就是讲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)。

可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尚未染色。于是，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。

基于 $\mathbf{D}_l \bigcup \mathbf{D}_u$ 构建一个图 $\mathbf{G} = (\mathbf{V}, \mathbf{E})$，其中结点集 $\mathbf{V} = \{\mathbf{x}_1, \cdots, \mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}\}$，边集合 $\mathbf{E}$ 可以表示成一个**亲和度矩阵（affinity matrix）**，常常基于高斯函数定义，即：

$$[\mathbf{W}]_{ij} = \begin{cases} \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\delta^2}\right), & \text{if } i \neq j; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

其中 $\delta > 0$ 是带宽（bandwidth）参数。

假定可以从图 $\mathbf{G}=(\mathbf{V},\mathbf{E})$ 可以学到一个实值函数 $f:\mathbf{V}\to\mathbb{R}$，其对应分类规则为 $y_i=sign(f(\mathbf{x}_i)),y_i\in\{-1,+1\}$。可以想象，相似的样本应该具有相似的标记，于是可以定义能量函数：

$$\begin{aligned}E(f) &= \frac{1}{2}\sum_{i=1}^{l+u}\sum_{j=1}^{l+u}[\mathbf{W}]_{ij}(f(\mathbf{x}_i-f(\mathbf{x}_j)))^2\\ &= \frac{1}{2}\Big(\sum_{i=1}^{l+u}f^2(\mathbf{x}_i)+\sum_{j=1}^{l+u}d_jf^2(\mathbf{x}_j)\\ &-2\sum_{i=1}^{l+u}\sum_{j=1}^{l+u}[\mathbf{W}]_{ij}f(\mathbf{x}_i)f(\mathbf{x}_j)\Big)\\ &= \mathbf{f}^t(\mathbf{D}-\mathbf{W})\mathbf{f}\end{aligned}\tag{10}$$

其中 $\mathbf{f}=(\mathbf{f}_l^t,\mathbf{f}_u^t)^t$，$\mathbf{f}_l=(f(\mathbf{x}_1),\cdots,f(\mathbf{x}_l))$，$\mathbf{f}_u=(f(\mathbf{x}_{l+1}),\cdots,f(\mathbf{x}_{l+u}))$，$\mathbf{D}$ 为一个对角阵，满足 $[\mathbf{D}]_{ii}=\sum_{i=1}^{l+u}[\mathbf{W}]_{ij}$。

具有最小能量的函数 f 在有标记样本上满足 $f(\mathbf{x}_i) = y_i$，在未标记的样本上满足 $\Delta\mathbf{f} = \mathbf{0}$，其中 $\Delta = \mathbf{D} - \mathbf{W}$，称为**拉普拉斯矩阵（Laplacian matrix）**。如果采用分块矩阵表示 $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{uu} \end{bmatrix}$，则式（10）可以改写为：

$$E(f) = \mathbf{f}_l^t(\mathbf{D}_{ll} - \mathbf{W}_{ll})\mathbf{f}_l - 2\mathbf{f}_u^t\mathbf{W}_{ul}\mathbf{f}_l + \mathbf{f}_u^t(\mathbf{D}_{uu} - \mathbf{W}_{uu})\mathbf{f}_u \tag{11}$$

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

$$\mathbf{f}_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1}\mathbf{W}_{ul}\mathbf{f}_l \tag{12}$$

于是，将 $\mathbf{D}_l$ 上的标记信息作为 $\mathbf{f}_l$ 代入上式，即可利用求得的 $\mathbf{f}_u$ 对未标记样本进行预测。

5. 半监督聚类

聚类是一种典型的无监督学习任务，然而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。

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。算法流程：

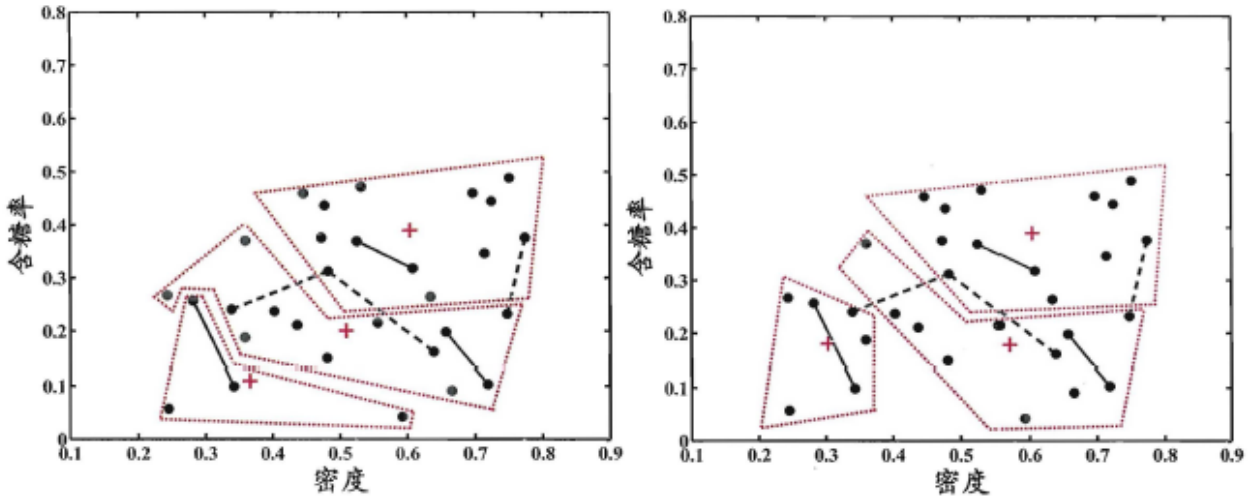
STEP 1. 随机选择 k 个样本作为聚类中心；

STEP 2. 计算剩余样本分别距离 k 个中心的距离，对每个样本按照离 k 个中心的距离进行聚类；

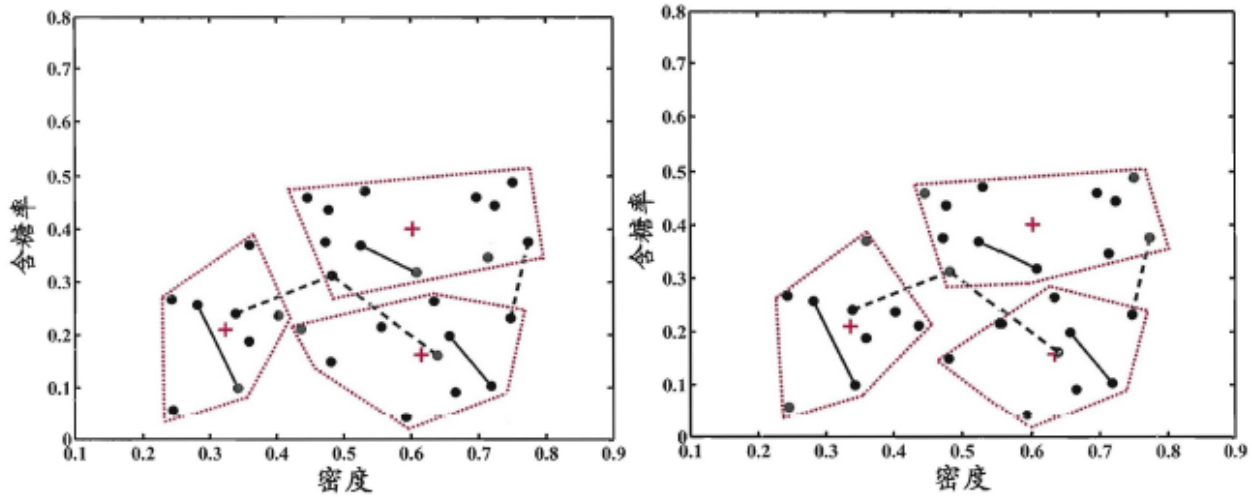
STEP 3. 检查是否违反约束，若违反，则选择下一个中心，若所有簇都不满足约束，则报错；

STEP 4. 重新计算 k 个中心，跳转 STEP 2。

5. 半监督聚类



(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

5. 半监督聚类

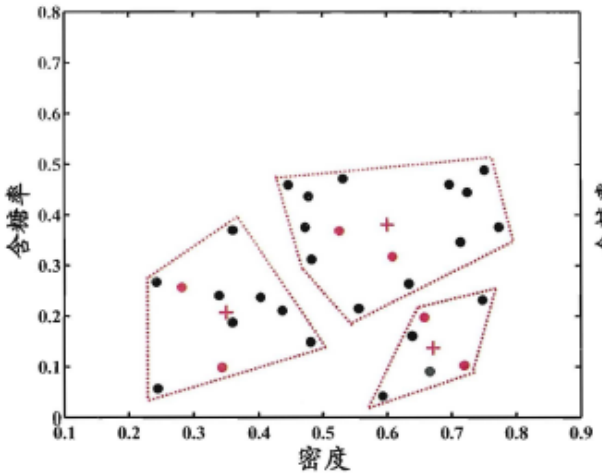
约束种子 k 均值 (Constrained Seed k-means) 算法是利用第二类监督信息的代表。算法流程：

STEP 1. 用有标记样本初始化 k 个中心

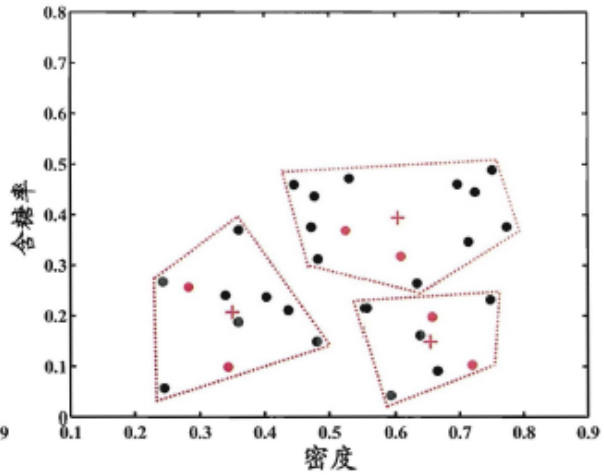
STEP 2. 计算无标记样本分别距离 k 个中心的距离，对每个样本按照离 k 个中心的距离进行聚类；

STEP 3. 利用所有样本重新计算 k 个中心，跳转 STEP 2。

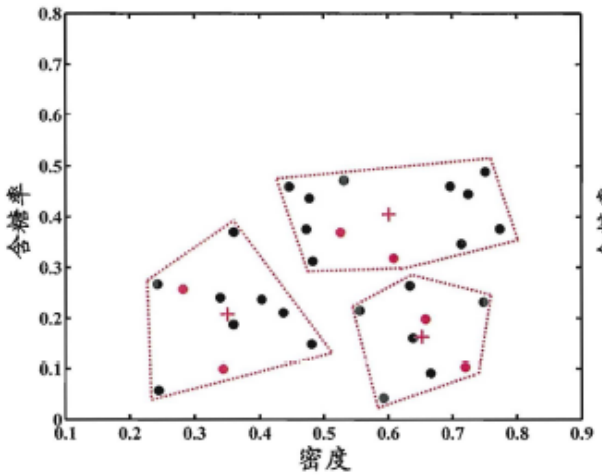
5. 半监督聚类



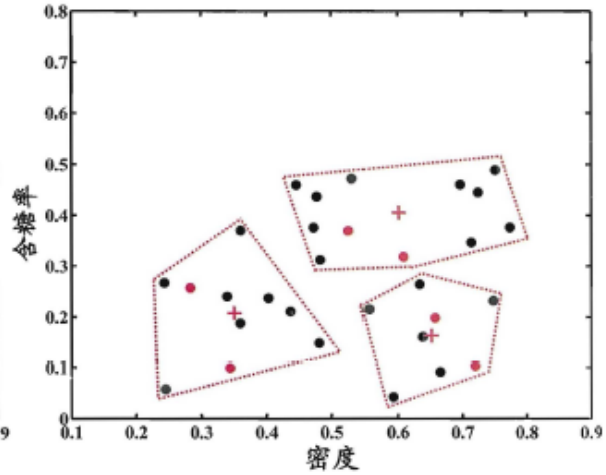
(a) 第 1 轮迭代后



(b) 第 2 轮迭代后



(c) 第 3 轮迭代后



(d) 第 4 轮迭代后

Thanks !