十三、半监督学习

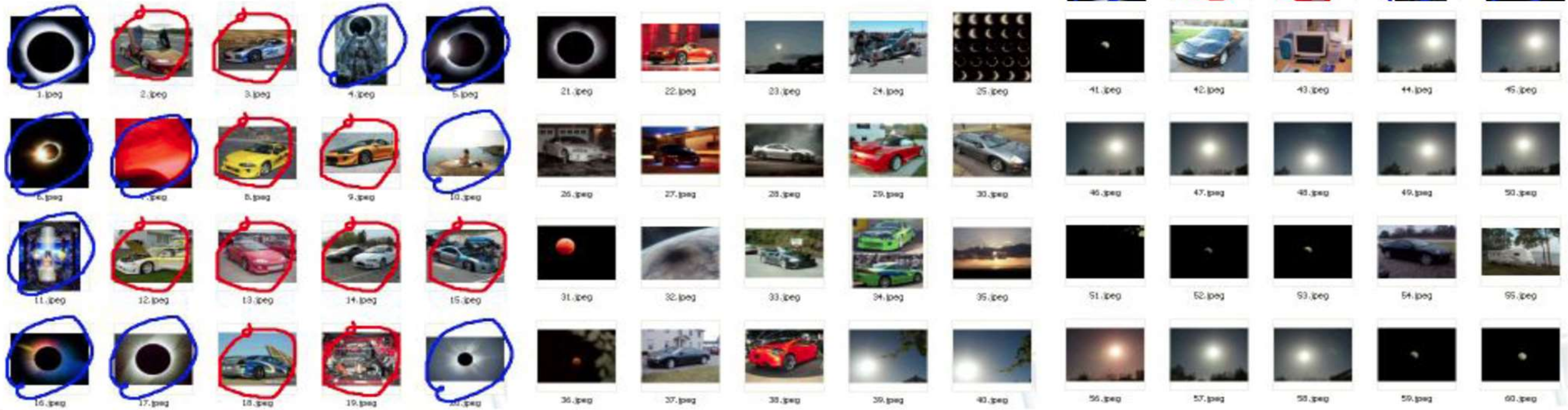
主讲教师：詹德川

什么是半监督学习？为什么要半监督学习？

- 因为在很多实际任务中，标签数据是有限的
- 人的观点：
  - 未标记样本容易取得
  - 标记样本很难取得

例如：

- Task: speech analysis
  - Switchboard dataset
  - telephone conversation transcription
  - 400 hours annotation time for each hour of speech
- Task: natural language parsing
  - Penn Chinese Treebank
  - 2 years for 4000 sentences
- Task: image categorization of “eclipse”



半监督学习之目的

- 同时利用标记样本和未标记样本，而不是只利用其一

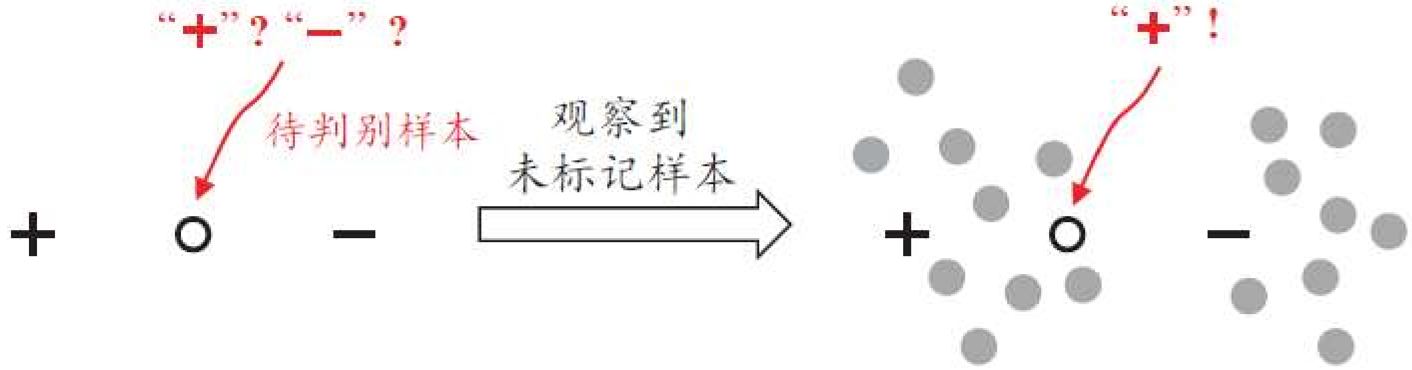


图 13.1　未标记样本效用的例示. 右边的灰色点表示未标记样本

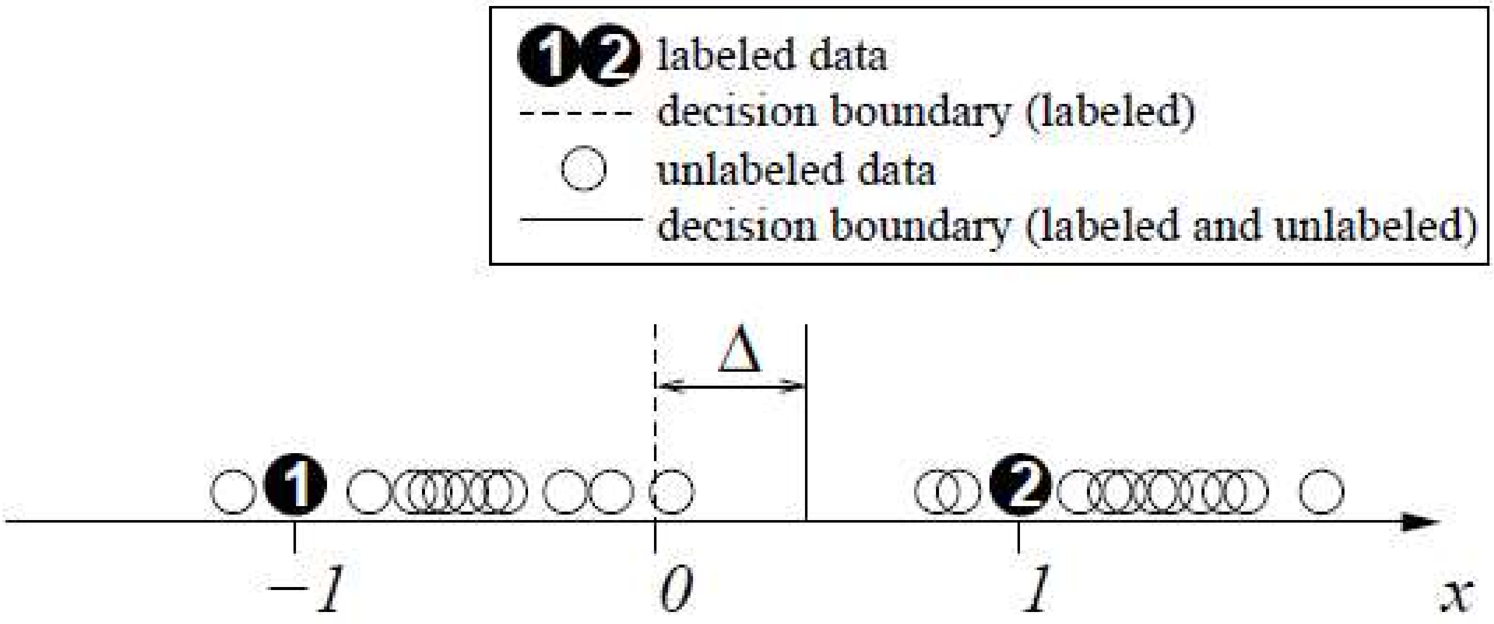
一种容易混淆的概念—直推学习

- 标记样本 $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- 未标记样本 $X_u = \{x_{l+1:n}\}$, 训练时可用
- 测试样本 $X_{test} = \{x_{n+1:}\}$, 只有测试时可以看到

- 半监督学习是一种归纳学习（inductive），可以对测试样本进行预测

- 直推学习是transductive学习，仅仅可以对未标记样本进行标记，模型不具备对测试样本进行泛化的能力！

半监督学习的一个简单的例子



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）：

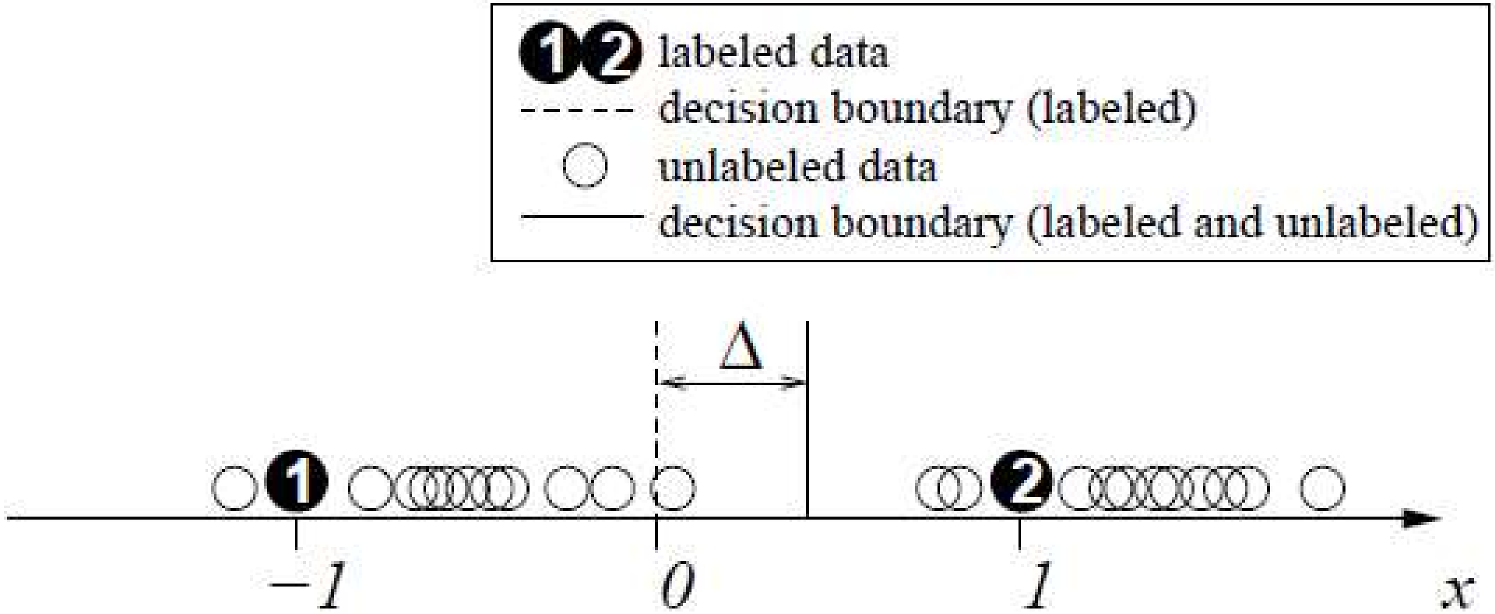
    假设数据分布在一个流形结构上，邻近的样本具有相似的输出值。

流形假设可看做聚类假设的推广

大纲

- 未标记样本
- 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



1 2 labeled data
decision boundary (labeled)
unlabeled data
decision boundary (labeled and unlabeled)
Δ
1
2
−1
0
1
x

生成式方法

- 假设样本由这个假设意味着混合成分 高斯混合模型生成，且每个类别对应一个高斯混合成分：

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

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

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

生成式方法

- 由最大化后验概率可知：

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

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

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

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

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

生成式方法

- 假设样本独立同分布，且由同一个高斯混合模型生成，则对数似然函数是：

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

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

生成式方法

- 高斯混合的参数估计可以采用EM算法求解，迭代更新式如下：

- E步：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。

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

生成式方法

- M步：基于 γ_{ji} 更新模型参数

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

$$\begin{aligned} \boldsymbol{\Sigma}_i = \frac{1}{\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \Big(& \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \\ & + \sum_{(\boldsymbol{x}_i, y_i) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \Big) \end{aligned}$$

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

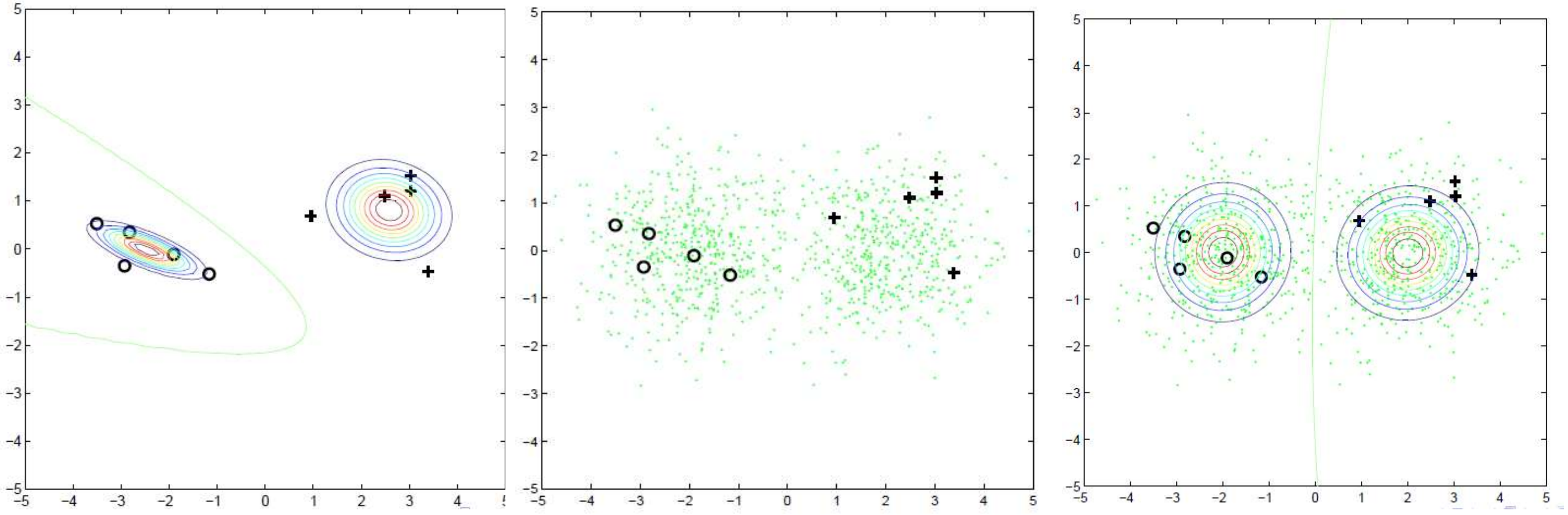
生成式方法

- 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。

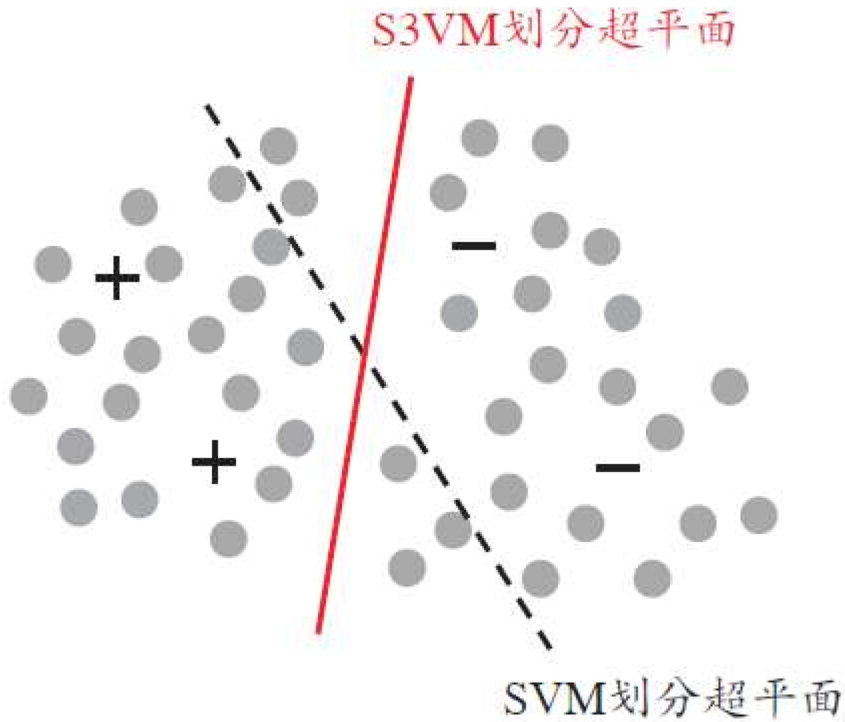
生成式方法：一个二维的例子



大纲

- 未标记样本
- 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

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)

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

半监督SVM

未标记样本的伪标记不准确

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 $\boxed{C_u \ll C_l;}$
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

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：

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

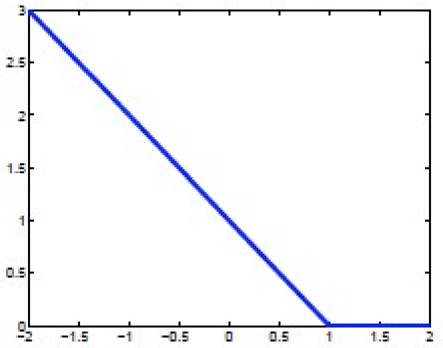
半监督SVM

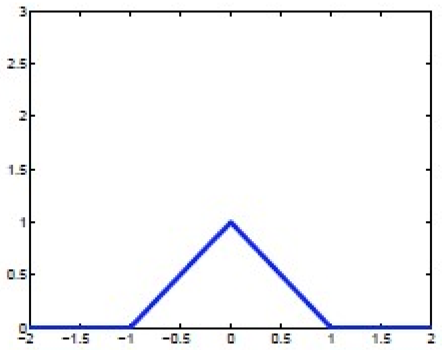
- 显然，搜寻标记指派可能出错的每一对未标记样本进行调整，仍是一个涉及巨大计算开销的大规模优化问题。

- 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.

为什么TSVM开销巨大？





大纲

- 未标记样本
- 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

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)。

- 我们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则尚未染色. 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。

图半监督学习

- 我们先基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$ ，其中结点集

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

- 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:

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

图半监督学习

- 假定从图 $G = (V, E)$ 将学得一个实值函数 $f : V \rightarrow \mathbb{R}$ 。
- 直观上讲相似的样本应具有相似的标记,即得到最优结果于是可定义关于*f*的 “能量函数” (energy function)[Zhu et al., 2003]:

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

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

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

图半监督学习

- 采用分块矩阵表示方式:

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

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

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

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

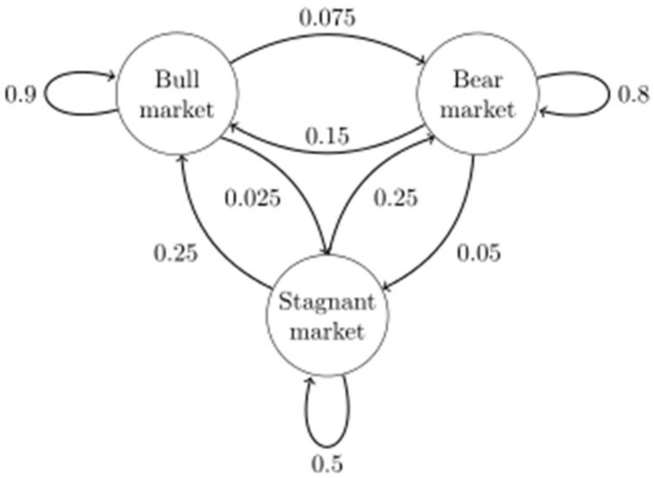
图半监督学习

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

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

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

图半监督学习与随机游走



Markov Chain

Transformation Prob. Matrix: T

$$\sum_q \mathbf{T}_{pq} = 1$$

Hitting Time:

$$HT_{p \to q} = \min \{t : v_0 = p, v_t = q\}$$

How about the expected hitting time?

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].
- 仍基于 $D_l \cup D_u$ 构建一个图 $G = (V, E)$ 其中结点集 $V = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_l, \boldsymbol{x}_{l+1}, ..., \boldsymbol{x}_{l+u}\}$
- 定义一个 $(l+u) \times |\mathcal{Y}|$ 的非负标记矩阵 $\mathbf{F} = (\mathbf{F}_1^\top, \ldots, \mathbf{F}_{|\mathcal{Y}|}^\top)^\top$， 其第i行元素

  $\boldsymbol{F}_i = (\boldsymbol{F}_{i1}, ..., \boldsymbol{F}_{i|\gamma|})$ 为示例 $\boldsymbol{x}_i$ 的标记向量, 相应的分类规则为:

$$y_i = \mathrm{argmax}_{1 \le j \le |\gamma|} \boldsymbol{F}_{ij}$$

- 将F初始化为:

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

图半监督学习

- 基于**W**构造一个标记传播矩阵 $\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$，其中 $\boldsymbol{D}^{-\frac{1}{2}} = \left(\frac{1}{\sqrt{d_1}}, \ldots, \frac{1}{\sqrt{d_{l+u}}}\right)$，于是有迭代计算式:

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

- 基于迭代至收敛可得:

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

图半监督学习

- 事实上，算法对应于正则化框架[Zhou et al., 2004]:

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

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

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 σ;
折中参数 α.

过程:

```
基于式(13.11)和参数 σ 得到 W;
基于 W 构造标记传播矩阵 S = D^{-1/2} W D^{-1/2};
根据式(13.18)初始化 F(0);
t = 0;
repeat
    F(t+1) = αSF(t) + (1-α)Y;
    t = t + 1
until 迭代收敛至 F*
for i = l+1, l+2, ..., l+u do
    y_i = arg max_{1≤j≤|𝒴|} (F*)_{ij}
end for
```

$$\mathbf{S} = \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}, \quad \mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t) + (1-\alpha)\mathbf{Y}, \quad y_i = \arg\max_{1\le j\le|\mathcal{Y}|}(\mathbf{F}^*)_{ij}$$

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

- 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．首先是在存储开销高。

- 另一方面，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。

大纲

- 未标记样本
- 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

基于分歧的方法

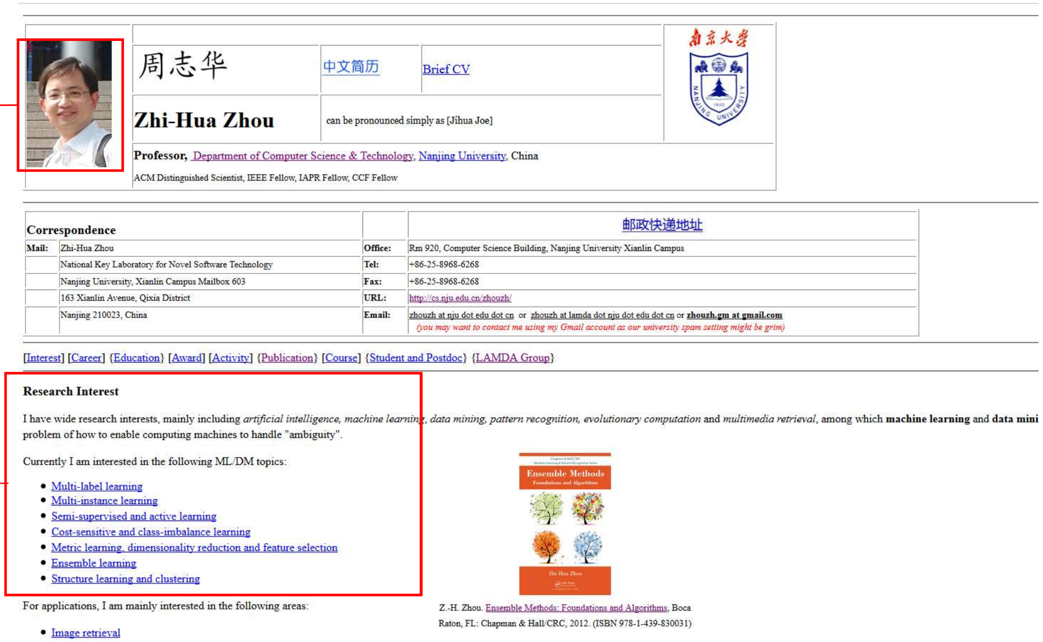
- 基于分歧的方法(disagreement-based methods)使用多学习器，而学disagreement亦称diversity 习器之间的 “分歧” (disagreement)对未标记数据的利用至关重要。

- 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

图片视图

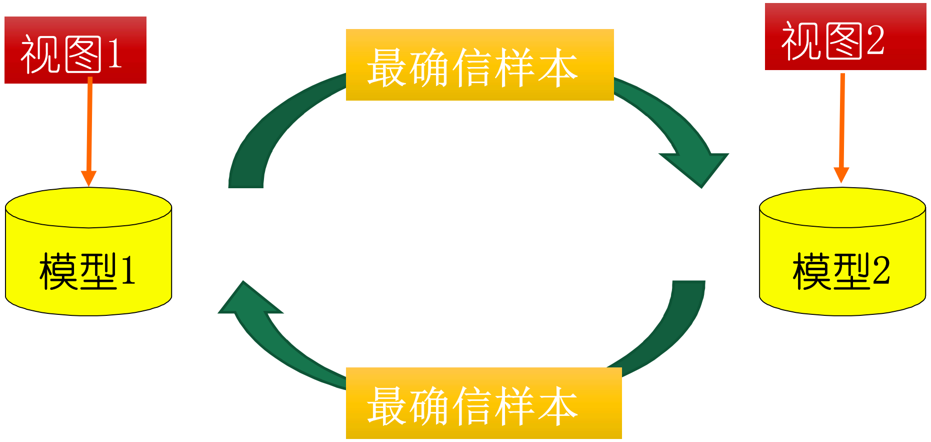
文字视图



网页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。

令 $p, n \ll s$.

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;
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;

在视图 j 上用有标记样本训练 h_j.

```
for j = 1,2 do
    h_j ← 𝔏(D_l^j);
    考察 h_j 在 D_s^j = {x_i^j | ⟨x_i^j, x_i^{3-j}⟩ ∈ 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 ⊂ D_s、n 个反例置信度最高的样本 D_n ⊂ D_s;
    由 D_p^j 生成伪标记正例 D̃_p^{3-j} = {(x_i^{3-j}, +1) | x_i^j ∈ D_p^j};
    由 D_n^j 生成伪标记反例 D̃_n^{3-j} = {(x_i^{3-j}, -1) | x_i^j ∈ D_n^j};
    D_s = D_s \ (D_p ⋃ 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 ⋃ (D̃_p^j ⋃ D̃_n^j);
    end for
  end if
end for
```

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].

- 不过，视图的条件独立性在现实任务中通常很难满足,不会是条件独立的因此性能提升幅度不会那么大，但研究表明，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].

大纲

- 未标记样本
- 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

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 $D=\{x_1,x_2,\ldots,x_m\}$;
必连约束集合 $\mathcal{M}$;
勿连约束集合 $\mathcal{C}$;
聚类簇数 k.

while $\neg$ is_merged do
基于 $\mathcal{K}$ 找出与样本 x_i 距离最近的簇：$r=\arg\min_{j\in\mathcal{K}} d_{ij}$;
检测将 x_i 划入聚类簇 C_r 是否会违背 $\mathcal{M}$ 与 $\mathcal{C}$ 中的约束;
if $\neg$ is_voilated then
$C_r=C_r\bigcup\{x_i\}$;
is_merged=true
else
$\mathcal{K}=\mathcal{K}\setminus\{r\}$;
if $\mathcal{K}=\varnothing$ then
break并返回错误提示
end if
end if
end while

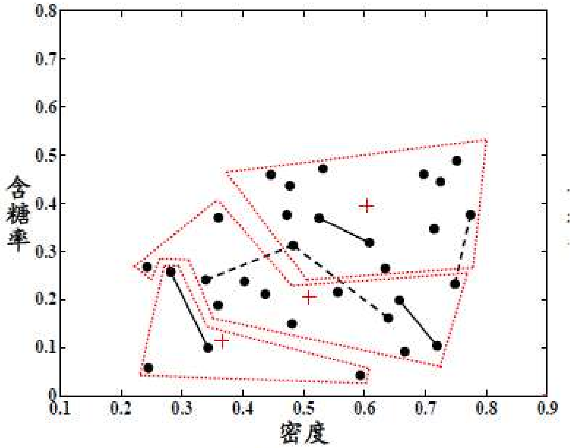
不冲突，选择最近的簇

冲突，尝试次近的簇

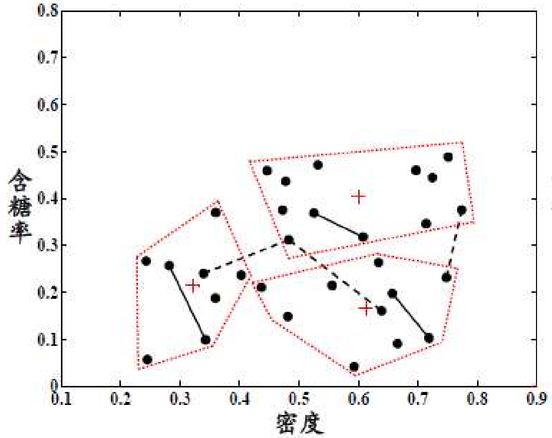
end for
until 均值向量均未更新
输出：簇划分 $\{C_1,C_2,\ldots,C_k\}$

图 13.7 约束 k 均值算法

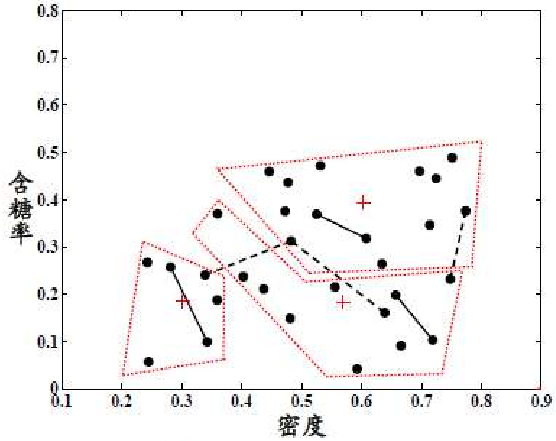
半监督聚类



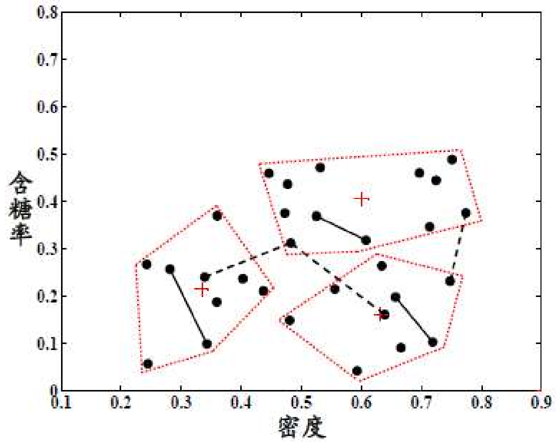
(a) 第 1 轮迭代后



(c) 第 3 轮迭代后



(b) 第 2 轮迭代后



(d) 第 4 轮迭代后

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。

- 这样的监督信息利用起来很容易: 直接将它们作为“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 $j = 1, 2, \ldots, k$ do
	$\quad \boldsymbol{\mu}_j = \frac{1}{\lvert S_j \rvert} \sum_{\boldsymbol{x} \in S_j} \boldsymbol{x}$
	end for

用有标记样本初始化 k 个簇.	for $j = 1, 2, \ldots, k$ do $\quad$ for all $\boldsymbol{x} \in S_j$ do $\quad\quad C_j = C_j \bigcup \{\boldsymbol{x}\}$ $\quad$ end for

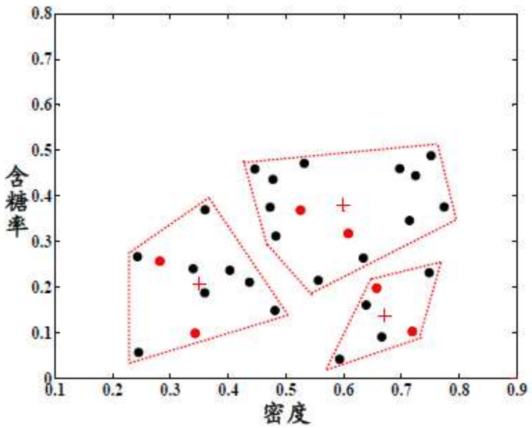
```
        找出与样本 x_i 距离最近的簇: r = arg min_{j∈{1,2,...,k}} d_ij ;
        将样本 x_i 划入相应的簇: C_r = C_r ⋃ {x_i}
    end for
    for j = 1, 2, ..., k do
        μ_j = 1/|C_j| Σ_{x∈C_j} x;
    end for
until 均值向量均未更新
```

更新均值向量.

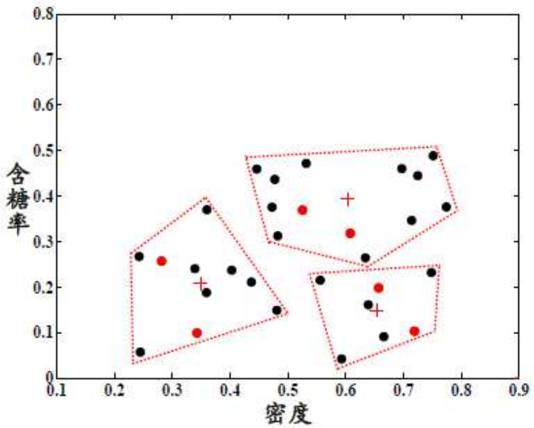
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.9 约束种子 k 均值算法

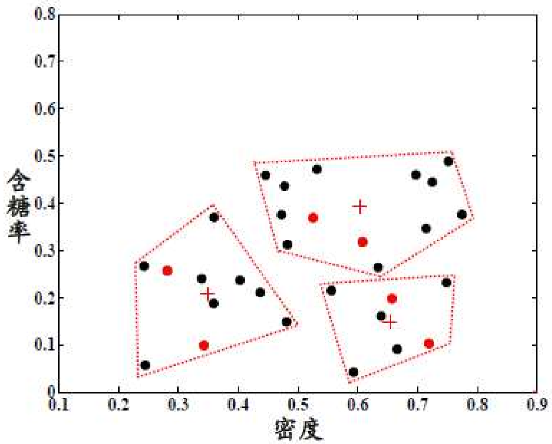
半监督聚类



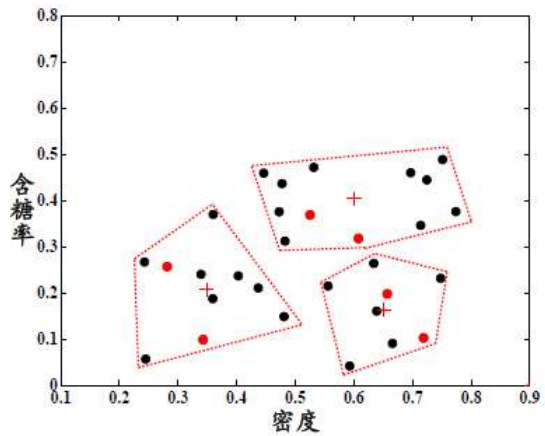
(a) 第 1 轮迭代后



(b) 第 2 轮迭代后



(b) 第 2 轮迭代后



(d) 第 4 轮迭代后

半监督学习的衍生

- Positive Unlabeled Learning (PU Learning)
- Weak Label Learning
- Multi-Instance Learning

多示例学习谈开去

（a） 传统监督学习（单示例、单标记）

（b） 多示例学习（多示例、单标记）

（c） 多标记学习（单示例、多标记）

（d） 多示例多标记学习

前往……