机器学习与数据挖掘

Machine Learning & Data Mining

权小军 教授

中山大学数据科学与计算机学院

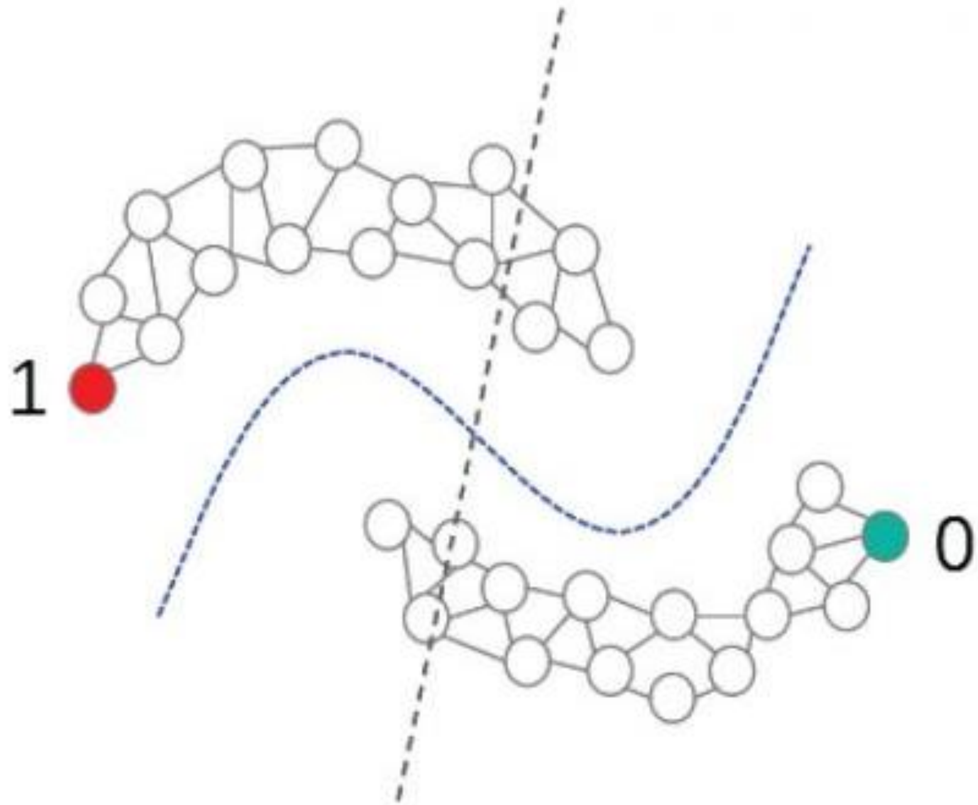
quanxj3@mail.sysu.edu.cn

Preface

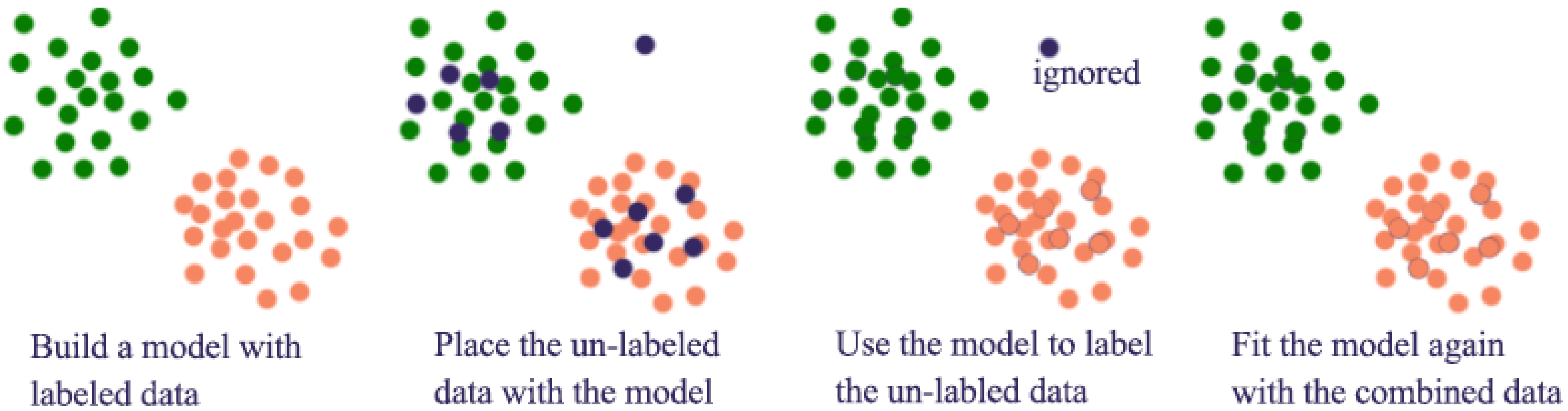
Preface

Labeled data is expensive!

前言



前言



Active Learning

前言

- Other solutions
  - Argumentation
  - Zero-shot Learning
  - One-shot Learning
  - Few-shot Learning
  - Semi-supervised Learning

前言

- Other solutions
  - Argumentation
  - Zero-shot Learning
  - One-shot Learning
  - Few-shot Learning
  - Semi-supervised Learning

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

Lecture 17
Semi-supervised Learning

大纲

- 生成式方法
- 半监督SVM
- 图半监督学习
- 半监督聚类

生成式方法

- 假设样本由高斯混合模型生成, 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:

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

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

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

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

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

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

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

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

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

生成式方法

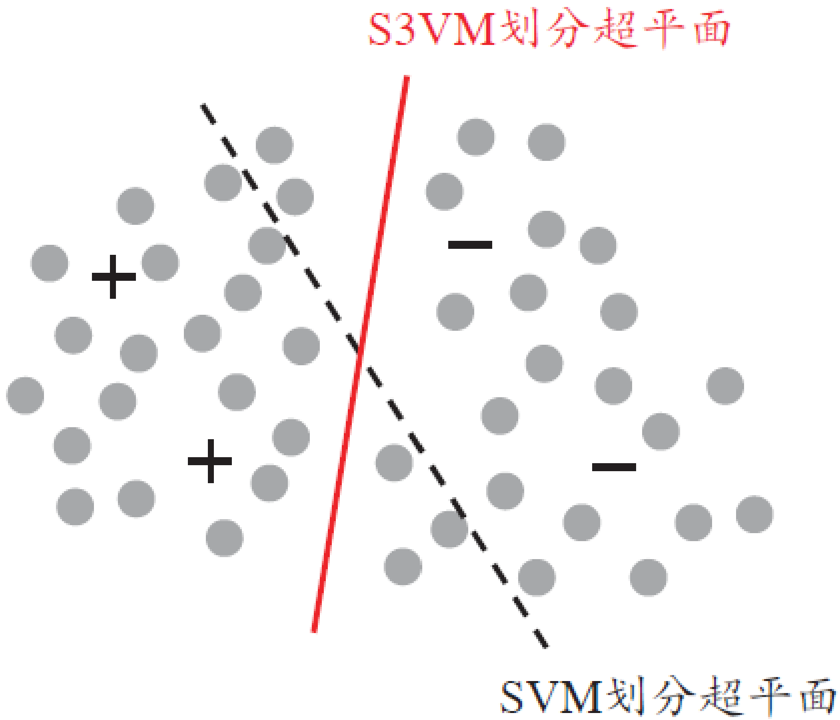
- 此类方法简单、易于实现，在有标记数据极少的情形下往往比其他方法性能更好。
- 然而，此类方法有一个关键：模型假设必须准确，即假设的生成式模型必须与真实数据分布吻合；否则利用未标记数据反而会显著降低泛化性能。

大纲

- 生成式方法
- 半监督SVM
- 图半监督学习
- 半监督聚类

半监督SVM

- 半监督支持向量机（Semi-Supervised Support Vector Machine）



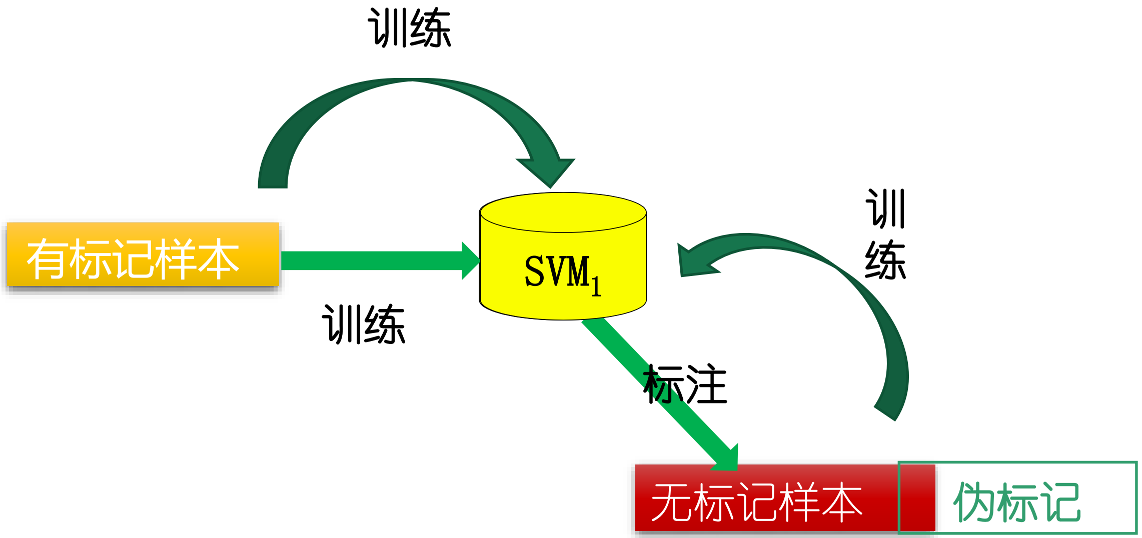
半监督SVM

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

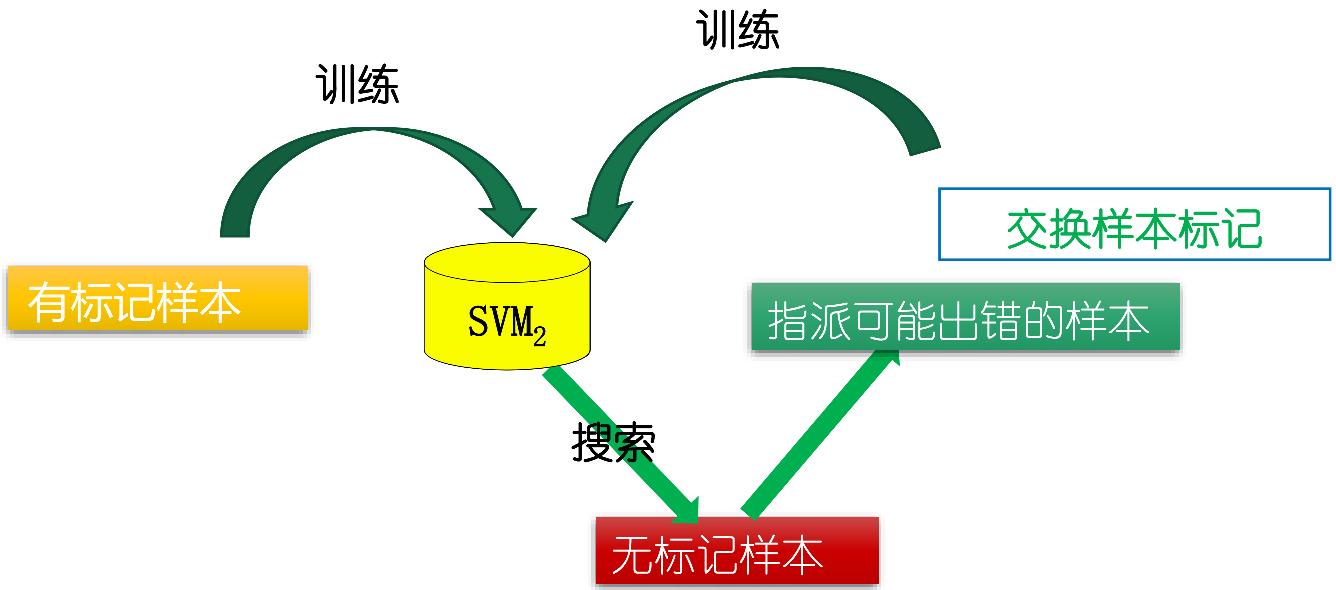
$$
\begin{aligned}
\min_{\boldsymbol{w},b,\hat{\boldsymbol{y}},\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C_l \sum_{i=1}^{l} \xi_i + C_u \sum_{i=l+1}^{m} \xi_i \\
\text{s.t.} \quad & y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l, \\
& \hat{y}_i(\boldsymbol{w}^\top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

半监督SVM

- 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 SVM_l;
用 SVM_l 对 D_u 中样本进行预测, 得到 ŷ = (ŷ_{l+1}, ŷ_{l+2}, ..., ŷ_{l+u});
初始化 C_u ≪ C_l;
while C_u < C_l do
    基于 D_l, D_u, ŷ, C_l, C_u 求解式(13.9), 得到 (w, b), ξ;
    while ∃{i, j | (ŷ_i ŷ_j < 0) ∧ (ξ_i > 0) ∧ (ξ_j > 0) ∧ (ξ_i + ξ_j > 2)} do
        ŷ_i = −ŷ_i;
        ŷ_j = −ŷ_j;
        基于 D_l, D_u, ŷ, C_l, C_u 重新求解式(13.9), 得到 (w, b), ξ
    end while
    C_u = min{2C_u, C_l}
end while
```

未标记样本的伪标记不准确

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：

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

大纲

- 生成式方法
- 半监督SVM
- 图半监督学习
- 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

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

- 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:

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

图半监督学习

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

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

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

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

大纲

- 生成式方法
- 半监督SVM
- 图半监督学习
- 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 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
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$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$;
repeat
  $C_j = \varnothing \ (1 \leq j \leq k)$; 初始化 k 个空簇.
  for $i = 1, 2, \ldots, m$ **do**
    计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j \ (1 \leq j \leq k)$ 的距离: $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{\mu}_j\|_2$;
    $\mathcal{K} = \{1, 2, \ldots, k\}$;
    `is_merged=false`;
    while $\neg$ `is_merged` **do**
      基于 $\mathcal{K}$ 找出与样本 $\boldsymbol{x}_i$ 距离最近的簇: $r = \arg\min_{j \in \mathcal{K}} d_{ij}$;
      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;
      if $\neg$ `is_voilated` **then**
        $C_r = C_r \bigcup \{\boldsymbol{x}_i\}$;
        `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
  end for
  for $j = 1, 2, \ldots, k$ **do**
    $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$; 更新均值向量.
  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

输入：样本集 $D = \{x_1, x_2, \ldots, x_m\}$;
　　　必连约束集合 $\mathcal{M}$;

```
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
```

不冲突，选择最近的簇

冲突，尝试次近的簇

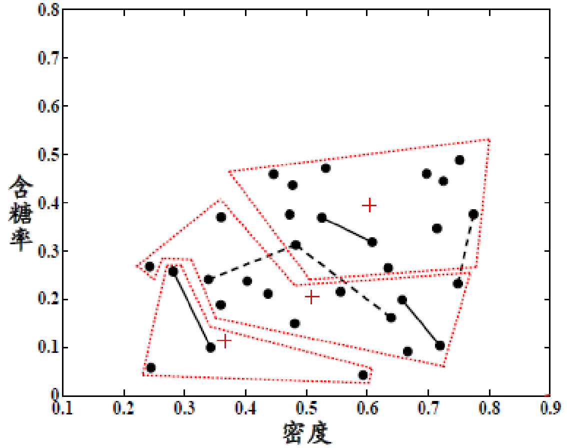
更新均值向量.

```
    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
```

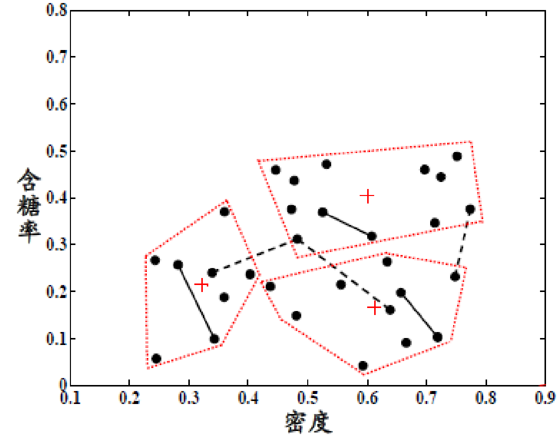
输出：簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.7　约束 k 均值算法

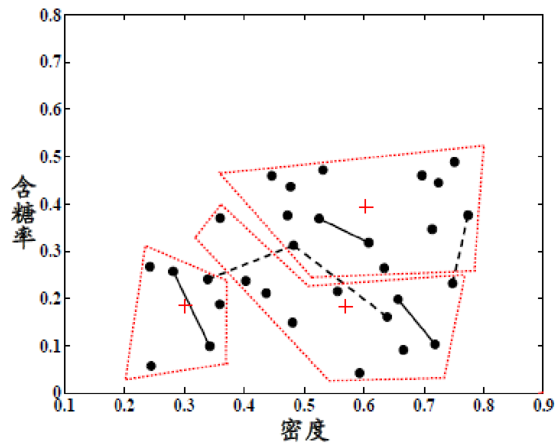
半监督聚类



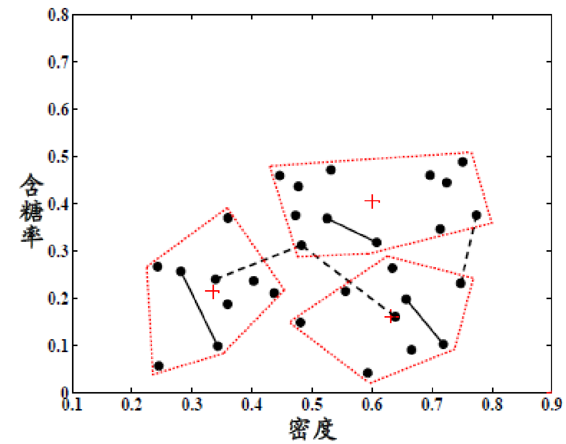
(a) 第 1 轮迭代后



(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

$S \subset D$, $|S| \ll |D|$.

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：

用有标记样本初始化簇中心.

for $j = 1, 2, \ldots, k$ **do**
  $\boldsymbol{\mu}_j = \frac{1}{|S_j|} \sum_{\boldsymbol{x} \in S_j} \boldsymbol{x}$
end for
repeat
  $C_j = \varnothing \ (1 \leq j \leq k)$;

用有标记样本初始化 k 个簇.

  for $j = 1, 2, \ldots, k$ **do**
    for all $\boldsymbol{x} \in S_j$ **do**
      $C_j = C_j \bigcup \{\boldsymbol{x}\}$
    end for
  end for
  for all $\boldsymbol{x}_i \in D \setminus S$ **do**
    计算样本 $\boldsymbol{x}_i$ 与各均值向量 $\boldsymbol{\mu}_j \ (1 \leq j \leq k)$ 的距离：$d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||_2$；
    找出与样本 $\boldsymbol{x}_i$ 距离最近的簇：$r = \arg\min_{j \in \{1,2,\ldots,k\}} d_{ij}$；
    将样本 $\boldsymbol{x}_i$ 划入相应的簇：$C_r = C_r \bigcup \{\boldsymbol{x}_i\}$
  end for

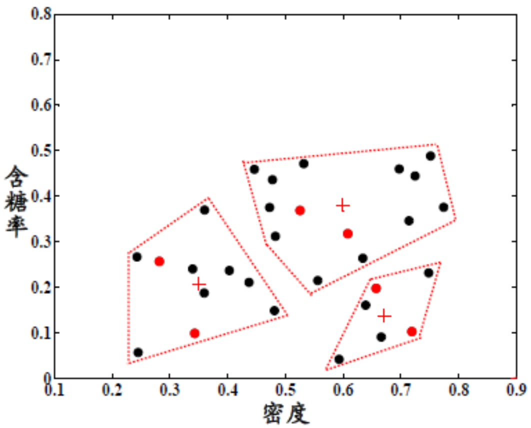
更新均值向量.

  for $j = 1, 2, \ldots, k$ **do**
    $\boldsymbol{\mu}_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$;
  end for
until 均值向量均未更新

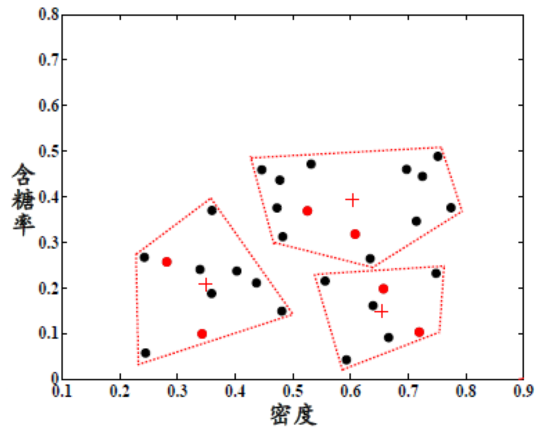
输出： 簇划分 $\{C_1, C_2, \ldots, C_k\}$

图 13.9 约束种子 k 均值算法

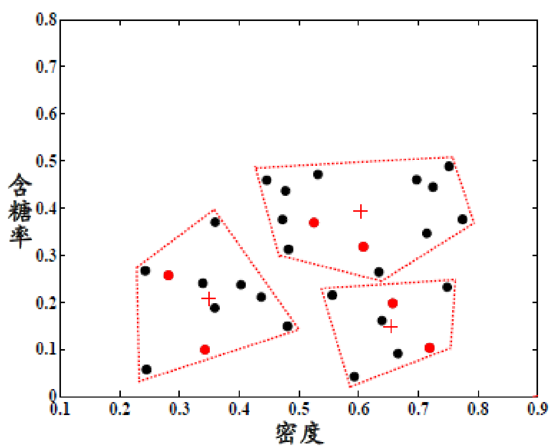
半监督聚类



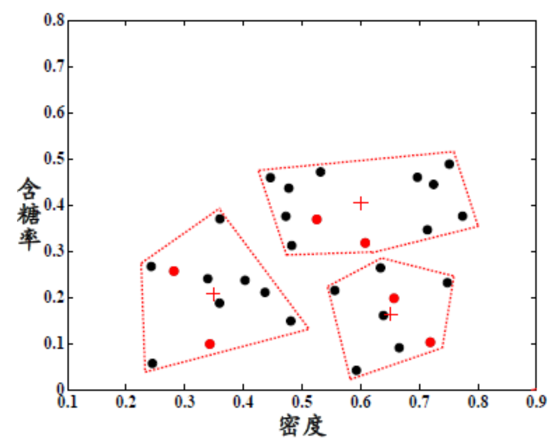
(a) 第 1 轮迭代后



(b) 第 2 轮迭代后



(b) 第 2 轮迭代后



(d) 第 4 轮迭代后

阅读材料

- 半监督学习的研究一般认为始于[Shahshahani and Landgrebe, 1994], 该领域在上世纪末、本世纪初蓬勃发展. 国际机器学习大会(ICML) 从2008年开始评选“十年最佳论文”, 在短短6年中, 半监督学习四大范型(paradim)中基于分歧的方法、半监督SVM、图半监督学习的代表性工作先后于2008年[Blumand Mitchell, 1998]、2009年[Joachims, 1999]、2013年[Zhu et al., 2003]获奖.

阅读材料

- 半监督学习在利用未标记样本后并非必然提升泛化性能，在有些情形下甚至会导致性能下降. 对生成式方法，其成因被认为是模型假设不准确 [Cozmanand Cohen, 2002]，因此需依赖充分可靠的领域知识来设计模型. 对半监督SVM, 其成因被认为是训练数据中存在多个“低密度划分”, 而学习算法有可能做出不利的选择; S4VM [Li and Zhou, 2015] 通过优化最坏情形性能来综“安全”指利用未标记数 合利用多个低密度划分，提升了此类技术的安全性.据之后, 确保泛化性能至少不差于仅利用有标记数据更一般的“安全”(safe)半监督学习仍是一个未决问题.

In-class exercise

问题：分析生成式方法、半监督SVM、图半监督学习、半监督聚类等半监督方法的相同和不同之处。

今天中午12点前将答案发送到邮箱: sysucsers@163.com

邮件标题和文件名称：学号-姓名-Lecture 17(semi)，例如: 01111-张三- Lecture 17(semi).

Thank you!

权小军 中山大学数据科学与计算机学院