

第十五章：半监督学习

Chapter 15: Semi-Supervised Learning

张永飞

2025年6月13日

内容提要

- 传统学习面临的问题
- 半监督学习的提出
- 两个基本假设
- 典型半监督学习算法
 - 自学习方法
 - 半监督支持向量机
 - 半监督聚类
- 存在问题及发展趋势

机器学习的主要研究问题

学习方式

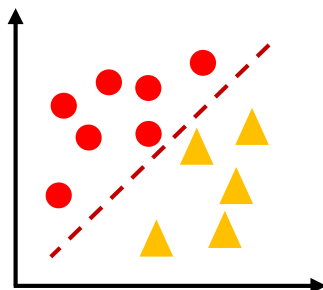
监督学习

无监督学习

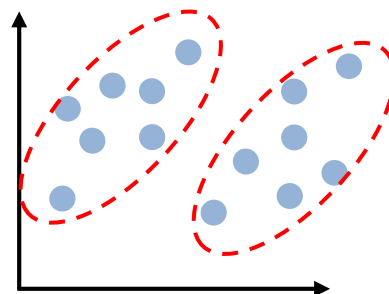
输出变量的类型

离散

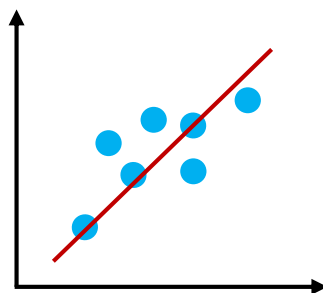
分类



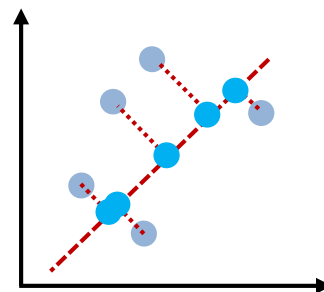
聚类



回归



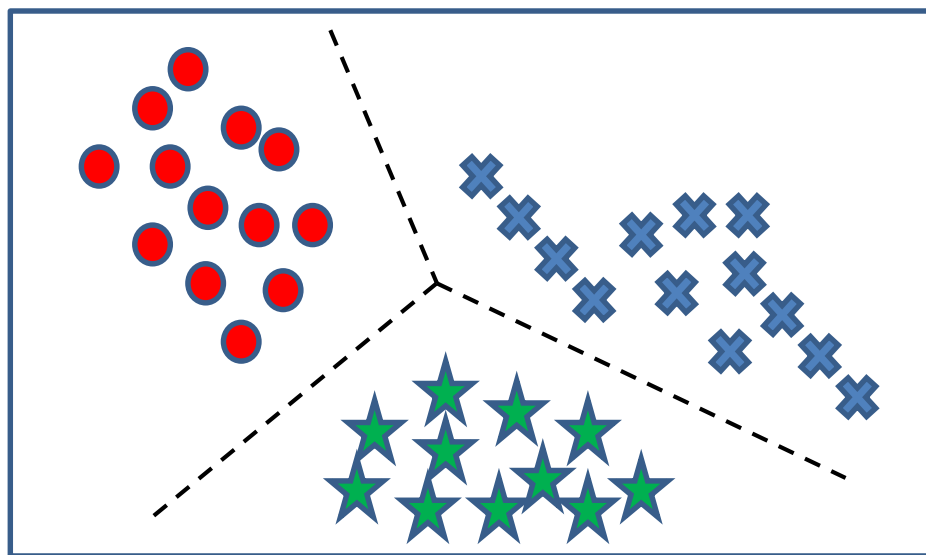
降维



连续

传统学习面临的问题

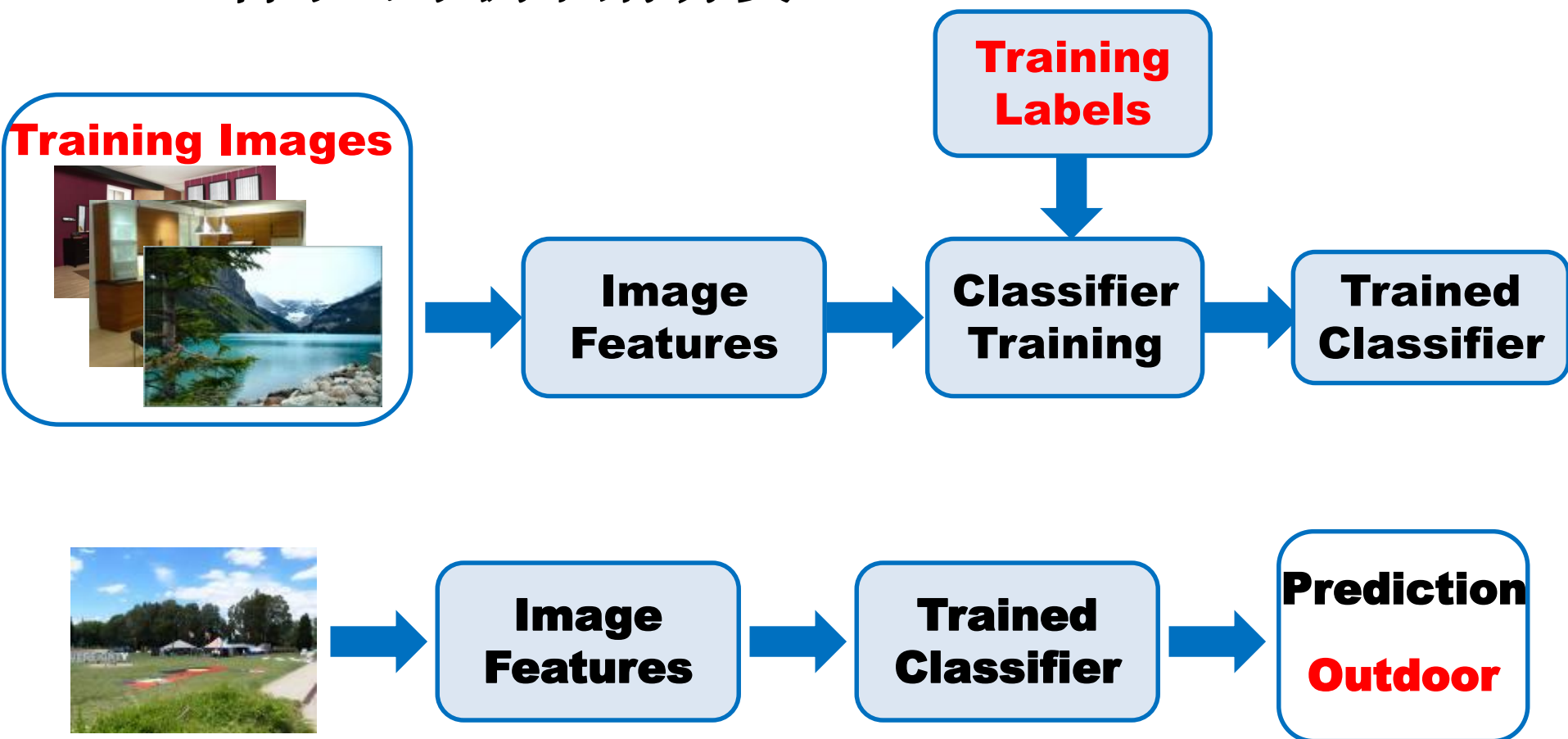
- 监督学习：利用已标记（监督信息）样本进行学习生成学习器，再对无标记样本进行测试



- 典型算法：分类和回归

传统学习面临的问题

- 监督学习示例-图像分类



传统学习面临的问题

● 监督学习示例-房价估计

训练样本

Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)
x_1	x_2	x_3	x_4	y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...

} N

测试样本

Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)
1500	3	2	30	?

传统学习面临的问题

- 监督学习-存在问题

- 需要大量已标记的训练样本（大模型、见多识广）

- 大规模人工标注耗时、难度大、代价高（隐私、安全）



ImageNet: 167个国家的
48,940标注员，两年时间



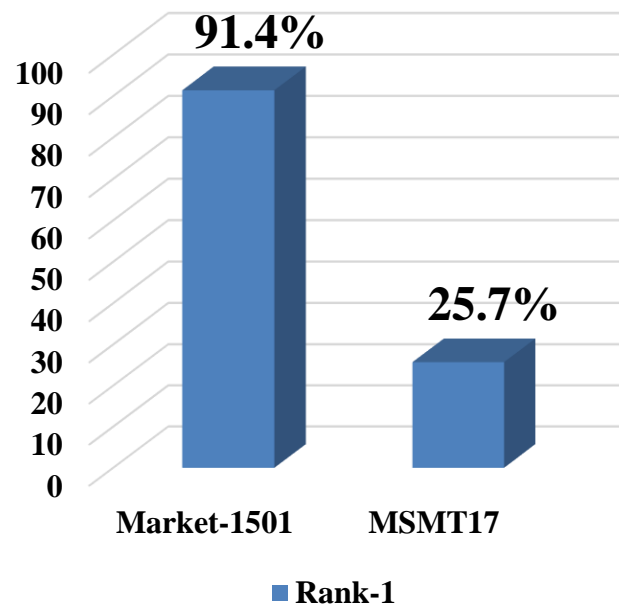
人工标注耗时、难度大

传统学习面临的问题

- 监督学习-存在问题

- 需要大量已标记的训练样本

- 标注数据通用性较弱



相同任务（行人再识别）
跨域性能

传统学习面临的问题


● 监督学习-存在问题

● 需要大量已标记的训练样本：非友好性

训练样本

测试样本

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜?
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	是
4	青绿	稍蜷	清脆	清晰	稍凹	软粘	是
5	乌黑	稍蜷	清脆	清晰	稍凹	软粘	是
10	青绿	硬挺	清脆	清晰	平坦	软粘	否
14	浅白	硬挺	清脆	清晰	凹陷	硬滑	否
15	乌黑	硬挺	清脆	清晰	稍凹	软粘	否
16	浅白	蜷缩	沉闷	模糊	平坦	硬滑	否
17	青绿	蜷缩	沉闷	模糊	稍凹	硬滑	否



编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜?
4	青绿	蜷缩	清脆	清晰	凹陷	硬滑	?
5	浅白	蜷缩	清脆	清晰	凹陷	硬滑	?
11	浅白	硬挺	清脆	清晰	平坦	硬滑	?

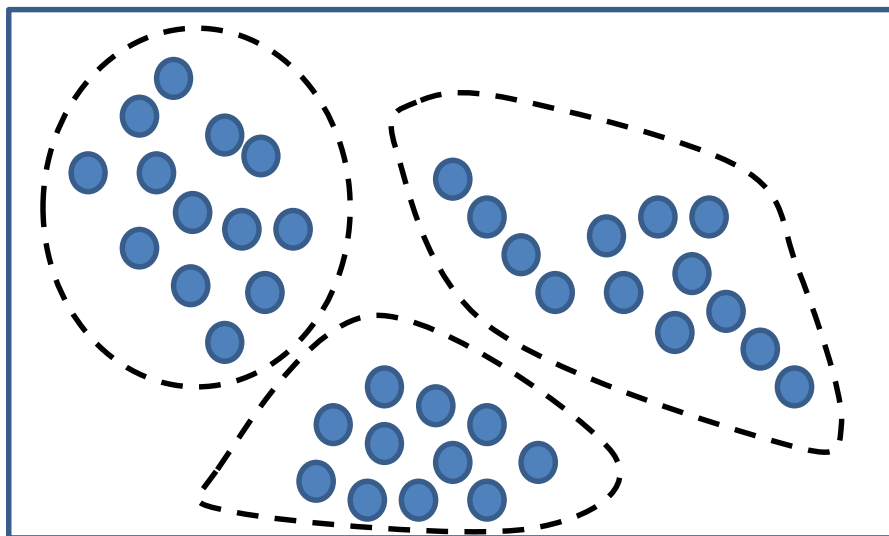
9

传统学习面临的问题

- 监督学习-存在问题
 - 需要大量已标记的训练样本!
 - 大规模的标注数据代价昂贵：常常需要人工标注或非友好性（西瓜）、耗时、代价高
 - 领域/应用：标注数据一般都是针对某个特定领域或特定应用，换个领域或应用就不适用了
 - 未标注数据很容易获得，但未利用

传统学习面临的问题

- 无监督学习：无已标记（监督信息）样本，基于未标记样本间的相似度，对样本进行类别归纳



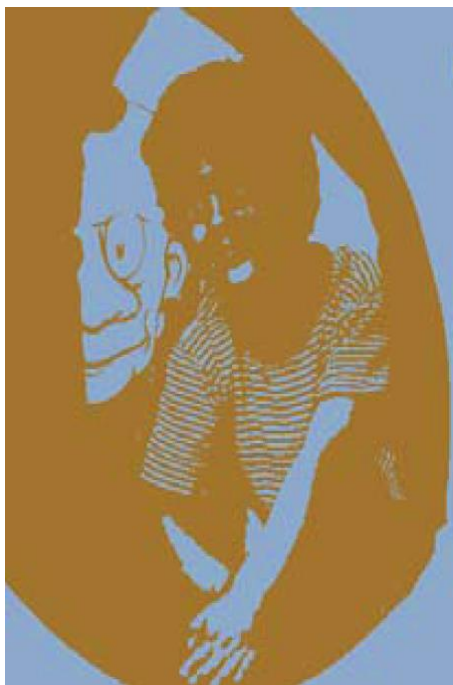
- 典型算法：聚类和降维

传统学习面临的问题

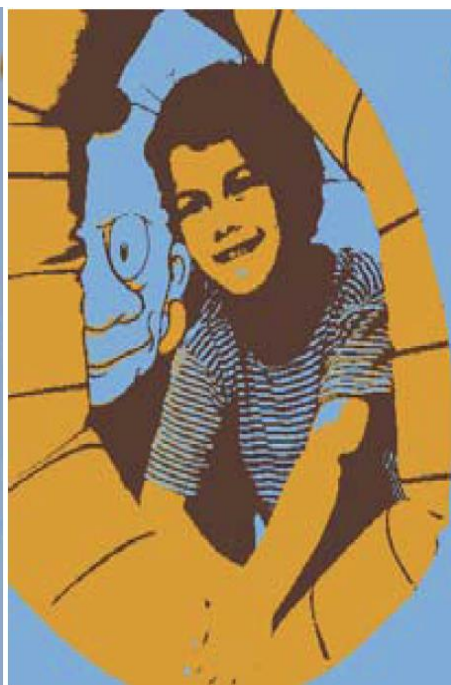
- 无监督学习-图像分割



原图



K=2



K=3



K=10

传统学习面临的问题

- 无监督学习-特征降维

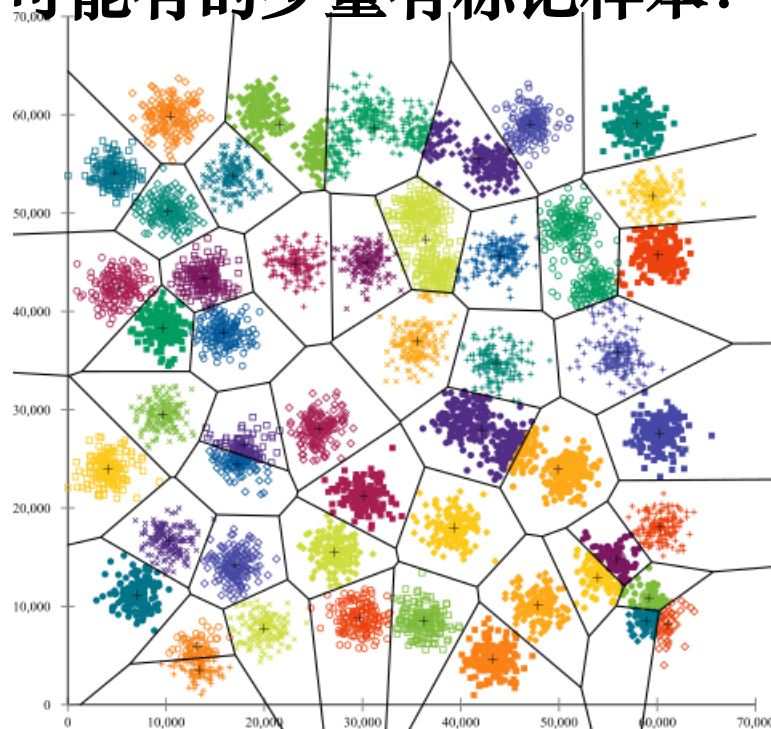


传统学习面临的问题

- 无监督学习-问题

- 完全基于无标记样本学习，精度难以保证

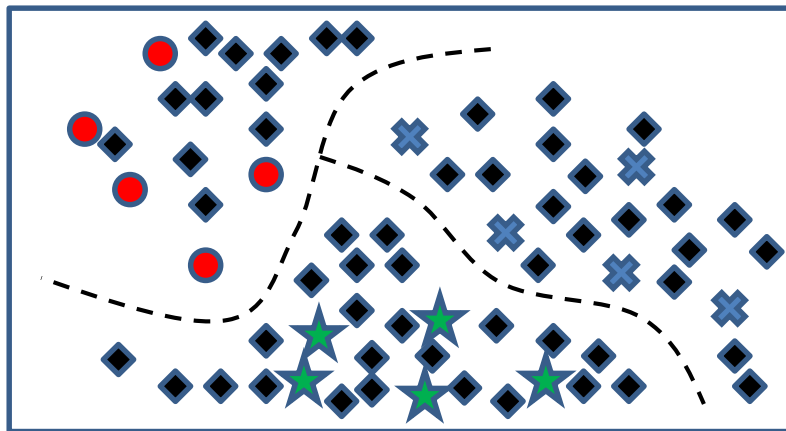
- 未利用有可能有的少量有标记样本！



传统学习面临的问题

- 实际情况

- 只有少量有标记样本，但只使用少量有标记样本，训练出来的学习系统往往难具有很好的强泛化能力
- 往往有大量未标记样本，仅使用少量“昂贵的”标记样本而不利用“廉价的”未标记样本对数据资源是一种浪费

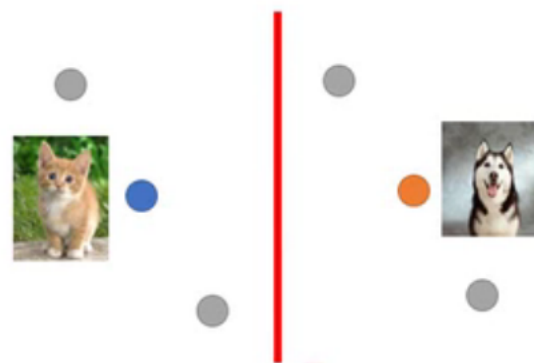


- 如何有效同时利用两种样本进行学习？

未标记样本的利用

- 人的认知可以认为是半监督的

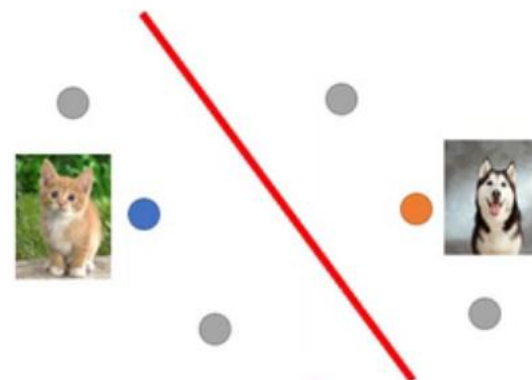
Labelled
data



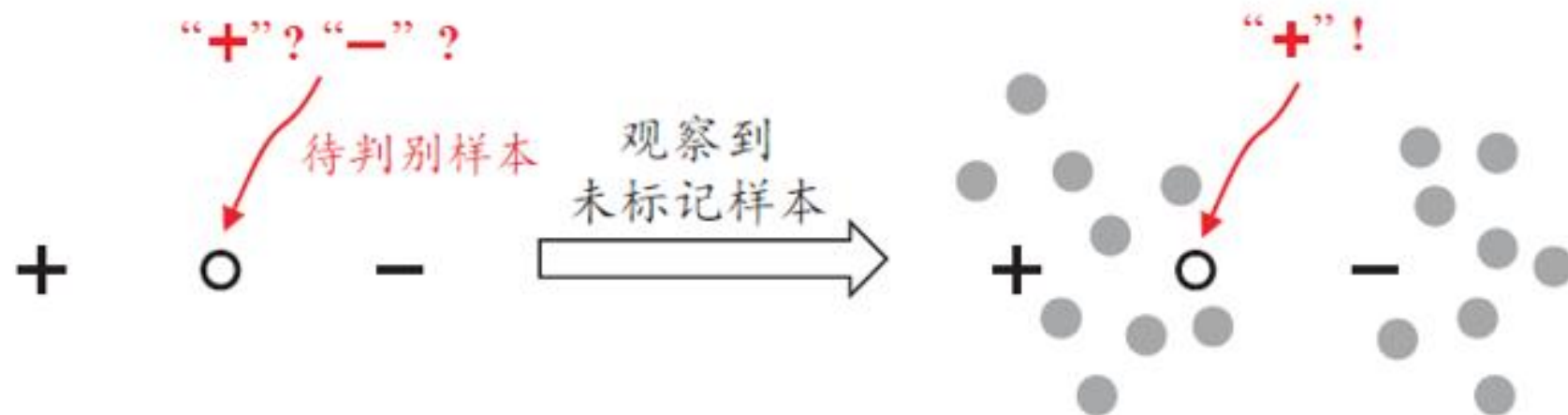
Unlabeled
data



(Image of cats and dogs without labeling)



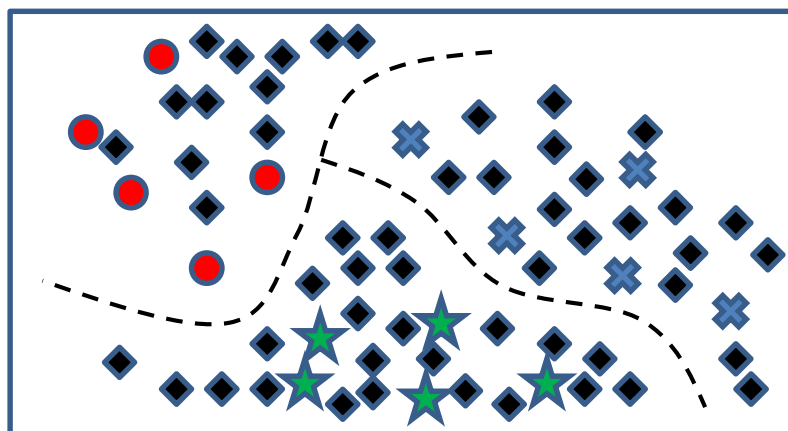
未标记样本的利用



半监督学习

- 如何有效同时利用两种样本进行学习？

当训练样本的部分信息缺失，特别是样本数据的类别标记缺失的情况下，如何获得具有良好**泛化能力**的学习器，即**利用大量的未标记样本辅助标记样本建立一个更好的学习器**



半监督学习

- 形式化描述

- 给定一个来自某未知分布的样本集 $S = D_l \cup D_u$

D_l 是已标记样本集

$$D_l = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_l, y_l)\}$$

D_u 是未标记样本集

$$D_u = \{\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \dots, \mathbf{x}_{l+u}\}$$

其中, x 为 d 维向量, $y_i \in Y$ 为 D_l 中样本 \mathbf{x}_i 的标记

$|D_l|$ 和 $|D_u|$ 分别为 D_l 和 D_u 的大小, 即所包含的样本数

半监督学习

- 形式化描述

- 半监督学习就是在样本集 $S = D_l \cup D_u$ 上寻找最优学习器，
即函数 $f: X \rightarrow Y$ ，可以准确地对样本 x 预测其标记 y

这个函数可能是参数的(如最大似然法)

也可能是非参数的(如最近邻法、神经网络法、SVM等)

也可能是非数值的(如决策树分类等)

如何综合利用所有样本(特别是未标记样本)建模，是半监督学习需要解决的关键问题

两个基本假设

- 要利用未标记样本，就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
- **基本假设：相似的样本具有相似的输出**
- 两种常见的假设
 - **聚类假设** (Clustering Assumption)
 - **流形假设** (Manifold Assumption)
 - 流形：局部具有欧几里得空间性质的空间

聚类假设

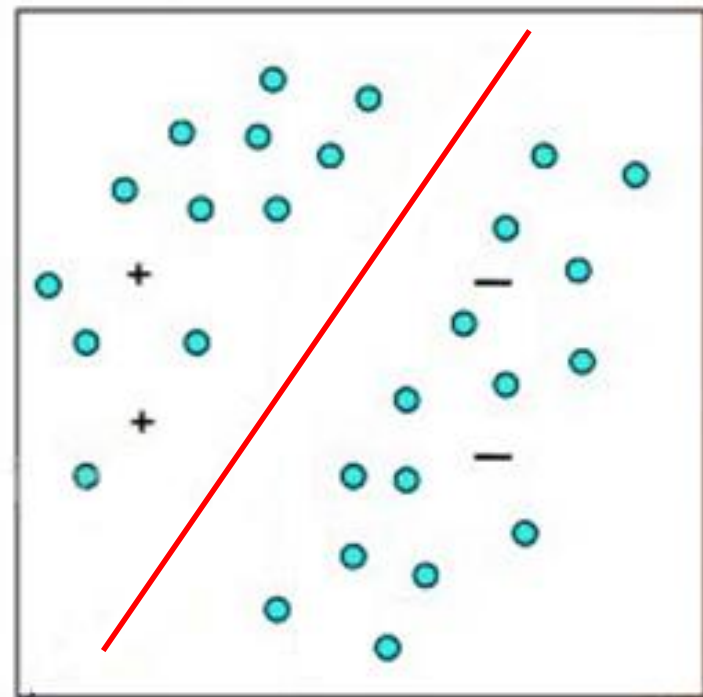
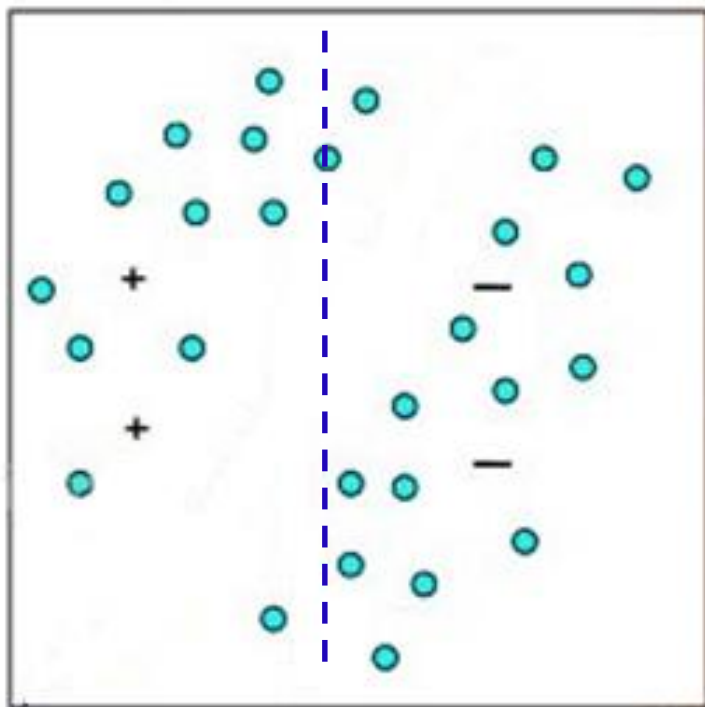
● 聚类假设（全局）

“数据存在簇结构，同一簇的样本属于同一类别”

- 处在相同聚类中的样本有较大可能拥有相同标记
- 决策边界应尽量通过数据较稀疏的地方，避免把稠密聚类中的数据分到决策边界两侧
- 大量未标记样本的作用就是帮助探明样本空间中数据分布的稠密和稀疏区域，指导对决策边界进行调整

聚类假设

- 聚类假设



两个基本假设

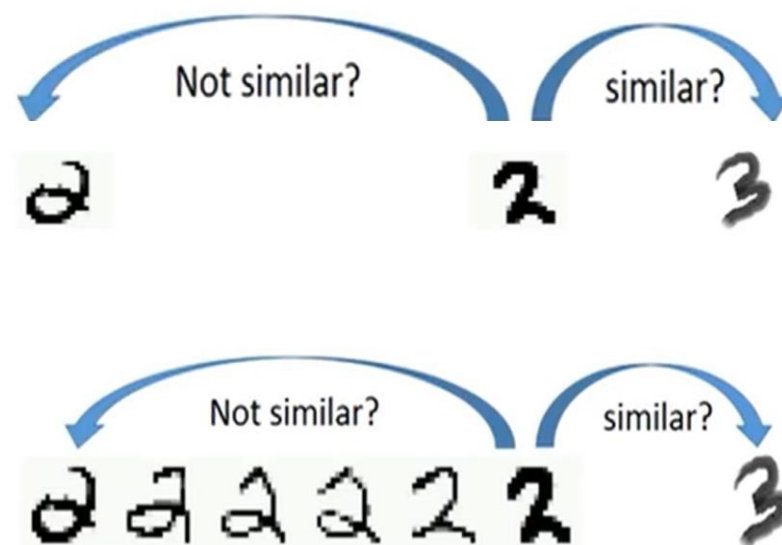
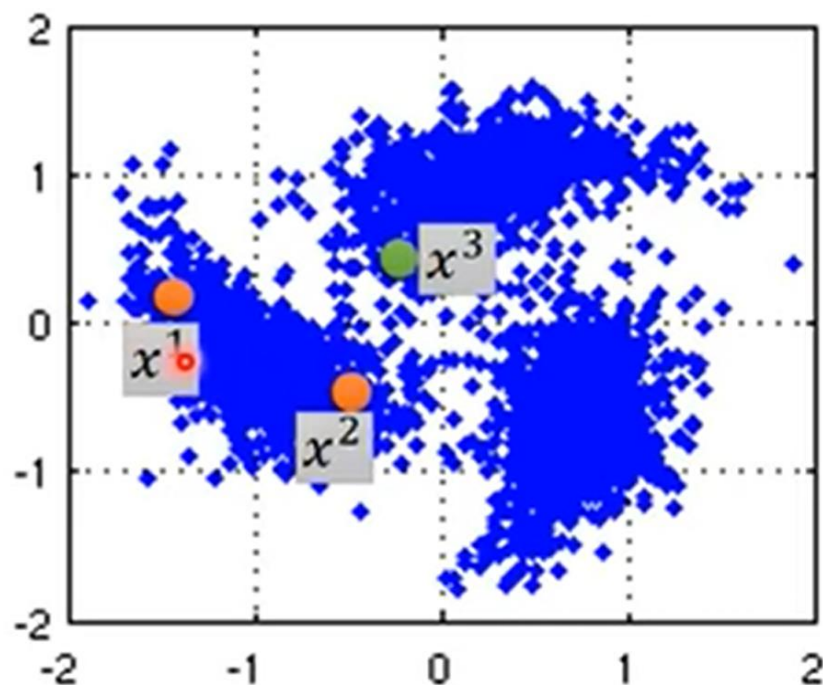
● 流形假设（局部）

“数据分布在一个流形结构上，邻近样本有相似输出值”

- 处于很小局部区域内的样本有相似性质，因此标记也相似
- (和聚类假设着眼整体特性不同)主要考虑模型的局部特性，反映决策函数的局部平滑性
- 大量未标记示例的作用就是让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性，使得决策函数能够更好地进行数据拟合
- 可看做是聚类假设的推广，对输出值没有限制（可以是相近的连续值），应用更广泛

两个基本假设

- 流形假设



半监督学习算法分类

- 直推学习(Transductive Learning)

- “封闭世界” 假设

课后习题

- 仅试图对学习过程中观察到的未标记样本进行预测

- 纯半监督学习(或归纳学习, Inductive Learning)

- “开放世界” 假设

- 学得模型能适用于训练过程中未观察到的未标记数据进行预测

闭卷考试

半监督学习算法分类

- 半监督分类

- 自学习方法 (Self-Training Methods)
- 直推式支持向量机 (Transductive SVM)
- 生成式方法 (Generative Methods)
- 图半监督学习 (Graph-based semi-supervised learning)
- 基于分歧的学习 (Disagreement-based Method)

- 半监督聚类

- 约束K均值 (Constrained K-means) 算法
- 约束种子K均值 (Constrained Seed K-means)

自学习方法 (Self-Training Methods)

●核心思想

分类器递归拟合时，每次递归仅将满足设定置信度阈值，即置信度高的样本纳入已标记样本集中，参与递归拟合

●算法流程

- Step1:用已标记的样本来训练得到一个初始分类器
- Step2:用初始分类器对未标记样本进行分类，将标记置信度高的未标记样本进行标记
- Step3:对所有样本进行重新训练，直到将所有未标记样本都标记为止

自学习方法 (Self-Training Methods)

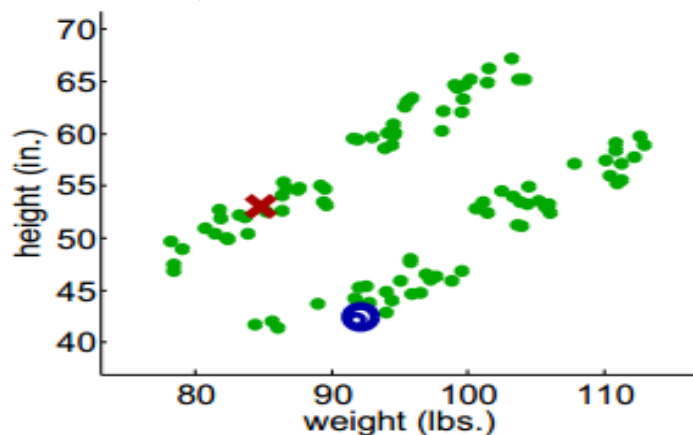
● 典型代表-最近邻自学习算法

- 用已标记样本 D_l 生成分类器 f
- 选择 $x = \operatorname{argmin} d(x, x_0)$, $x \in D_u$, $x_0 \in D_l$, 也就是选择离已标记样本最近的无标记样本
- 用 f 给 x 确定一个类别 $f(x)$, 并将 $(x, f(x))$ 加入 D_l 中
- 重复上述步骤, 直到 D_u 为空集

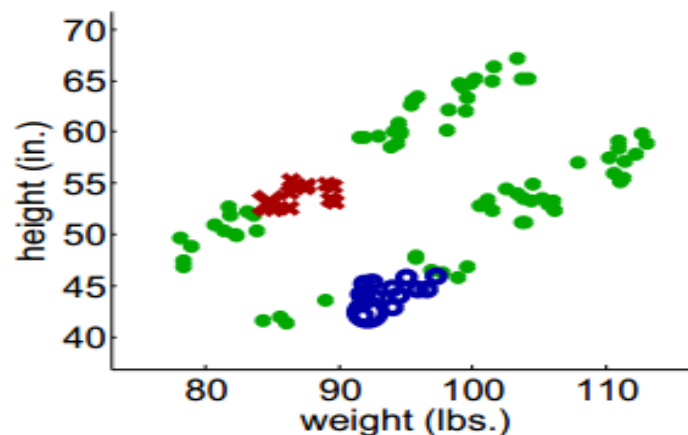
其中 $d(x_1, x_2)$ 为两个样本的欧式距离

自学习方法 (Self-Training Methods)

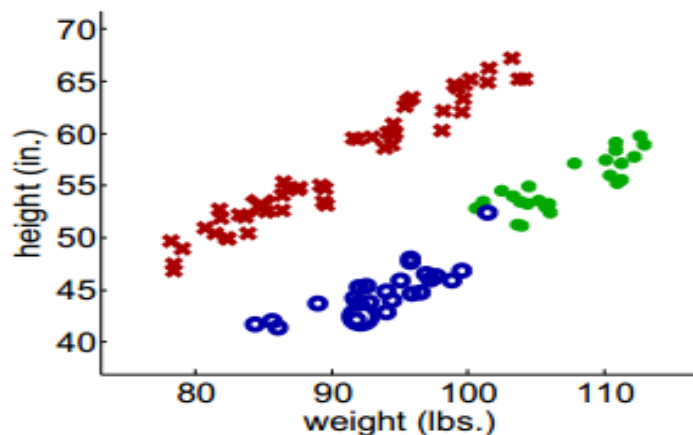
● 典型代表-最近邻自学习算法



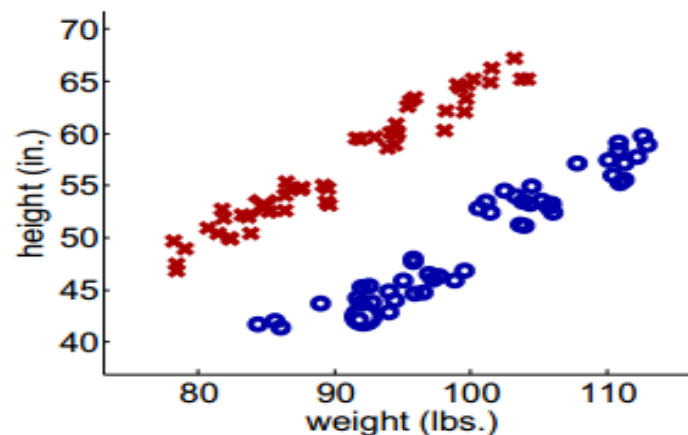
初始情况



第25轮迭代后



第74轮迭代后



最终结果

半监督SVM

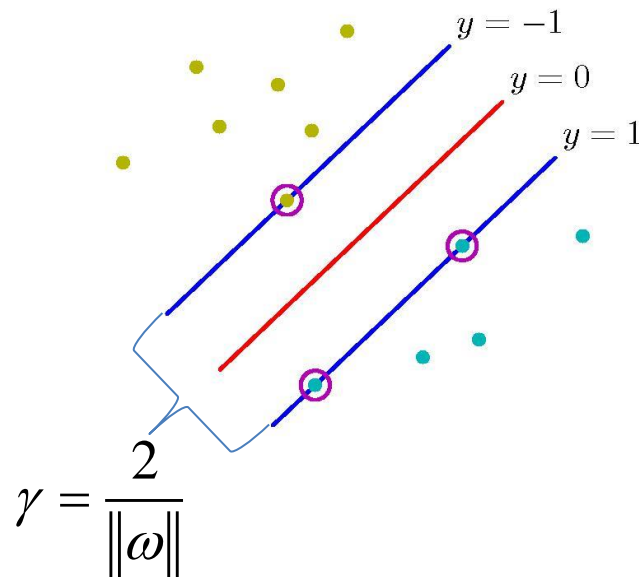
(Semi-Supervised SVM, S3VM)

● 标准SVM

针对二分类问题，利用已标记数据，在样本空间中寻找一个最优超平面使两类样本间**分类间隔**最大

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2$$

$$s.t. \ y_i (\omega^T x_i + b) \geq 1, i = 1, 2, \dots, N$$



半监督SVM

(Semi-Supervised SVM, S3VM)

● 标准SVM

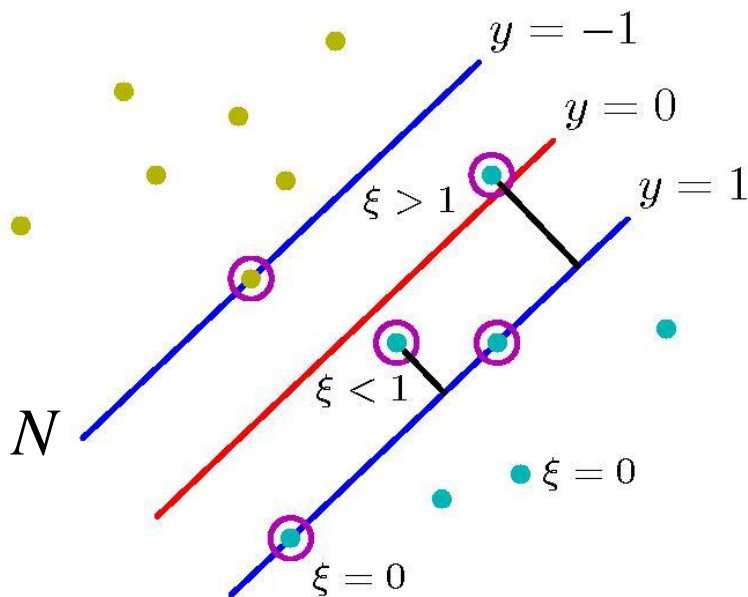
● 软间隔

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_i^N \xi_i$$

$$s.t. \ y_i (\omega^T x_i + b) \geq 1 - \xi_i, i = 1, 2, \dots, N$$

$$\xi_i \geq 0$$

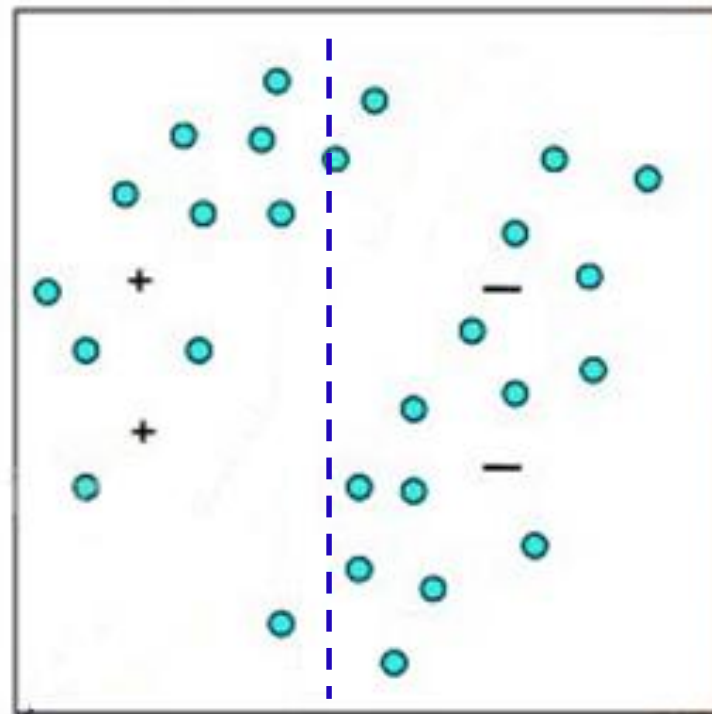
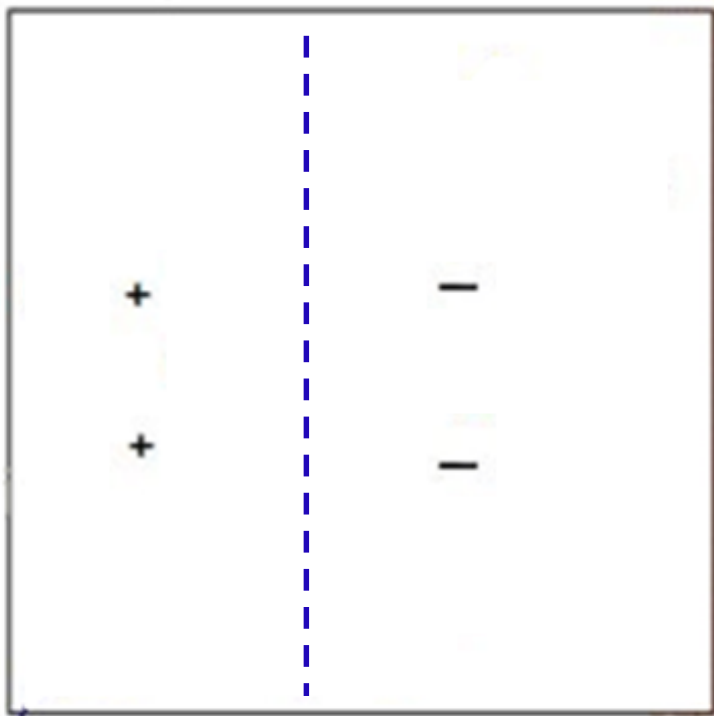
其中 ξ_i 为松弛向量



半监督SVM

(Semi-Supervised SVM, S3VM)

- 标准SVM



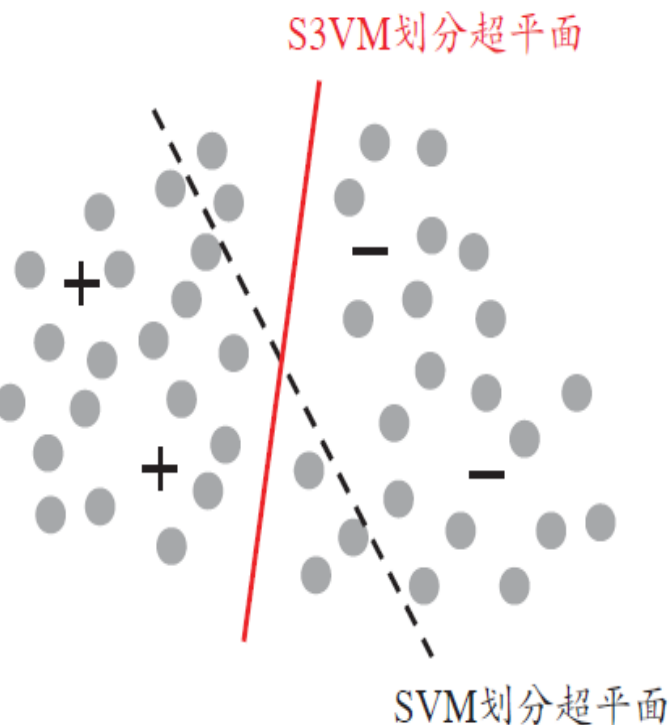
半监督SVM

(Semi-Supervised SVM, S3VM)

- S3VM是标准SVM方法在未标记样本上的一种推广，其中**直推式支持向量机T-SVM (Transductive SVM)**是最典型的算法

- 基本思想

- 针对二分类问题，同时利用标记和未标记样本，**通过尝试将每个未标记样本分别作为正例和反例来寻找最优分类边界**，来得到原始数据中两类样本的**最大分类间隔**（最小错误率），即寻找穿过数据低密度分布区域的分类面



半监督SVM

(Semi-Supervised SVM, S3VM)

● 直推式支持向量机T-SVM

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

$$\text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad i = 1, \dots, l,$$

$$\hat{y}_i(w^\top x_i + b) \geq 1 - \xi_i, \quad i = l + 1, \dots, m,$$

$$\xi_i \geq 0, \quad i = 1, \dots, m,$$

其中， C_l 和 C_u 分别表示已标记样本和未标记样本的**惩罚因子**，用于调整不同样本的权重， ξ 为**松弛向量/因子**，用于调整对错分样本的容忍程度

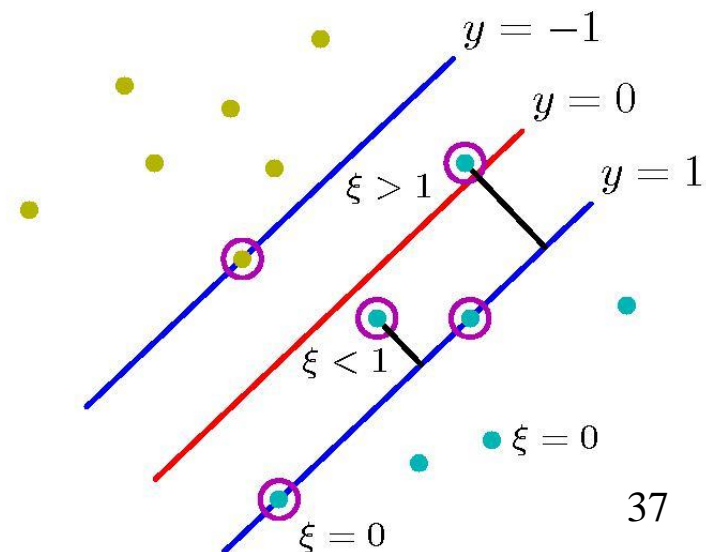
半监督SVM

(Semi-Supervised SVM, S3VM)

● T-SVM

- 尝试未标记样本的各种标记指派是一个穷举遍历过程，未标记样本数太大时不能直接求解
- 采用局部搜索迭代求近似解，通过局部搜索和调整指派为异类且可能错误的标记指派，使目标函数值不断下降

- 1局部搜索: x_i, x_j
- 2生成伪标记: \hat{y}_i, \hat{y}_j
- 3问题求解: 得到超平面和 ξ_i, ξ_j
- 4异类: $\hat{y}_i \hat{y}_j < 0$
- 5可能错误: $\xi_i + \xi_j > 2$
- 6调整: 互换标记
- 7循环: 3-7



半监督SVM

(Semi-Supervised SVM, S3VM)

● T-SVM算法

输入: 有标记样本集 $D_l = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$;
未标记样本集 $D_u = \{x_{l+1}, x_{l+2}, \dots, x_{l+u}\}$;
折中参数 C_l, C_u .

过程:

- 1: 用 D_l 训练一个 SVM_l ;
- 2: 用 SVM_l 对 D_u 中样本进行预测, 得到 $\hat{y} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \dots, \hat{y}_{l+u})$;
- 3: 初始化 $C_u \ll C_l$;
- 4: **while** $C_u < C_l$ **do**
- 5: 基于 $D_l, D_u, \hat{y}, C_l, C_u$ 求解式(13.9), 得到 $(w, b), \xi$;
- 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 $D_l, D_u, \hat{y}, C_l, C_u$ 重新求解式(13.9), 得到 $(w, b), \xi$
- 10: **end while**
- 11: $C_u = \min\{2C_u, C_l\}$
- 12: **end while**

未标记样本的伪标记不准确,
对应的系数要小

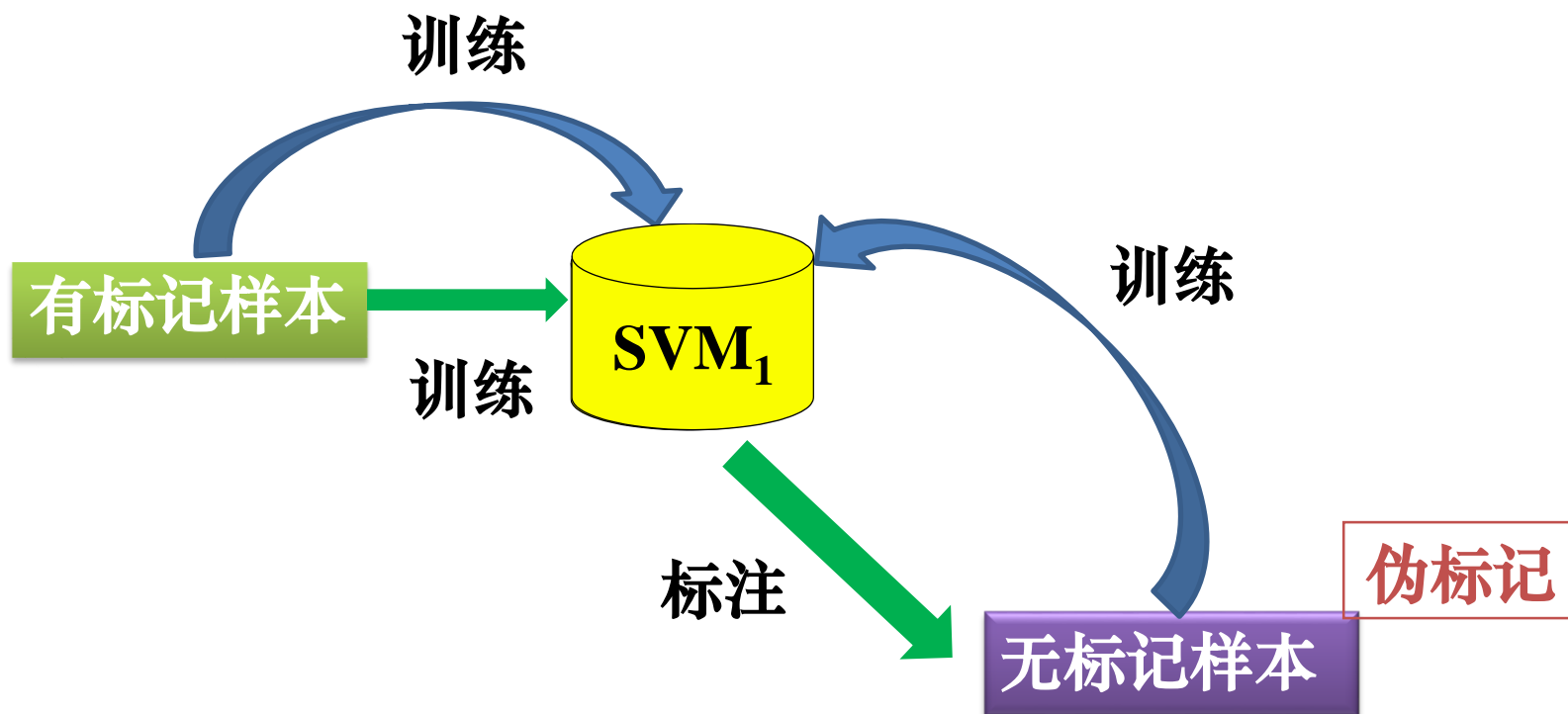
逐渐增大 C_u 提高未标记样本贡献

输出: 未标记样本的预测结果: $\hat{y} = (\hat{y}_{l+1}, \hat{y}_{l+2}, \dots, \hat{y}_{l+u})$

半监督SVM

(Semi-Supervised SVM, S3VM)

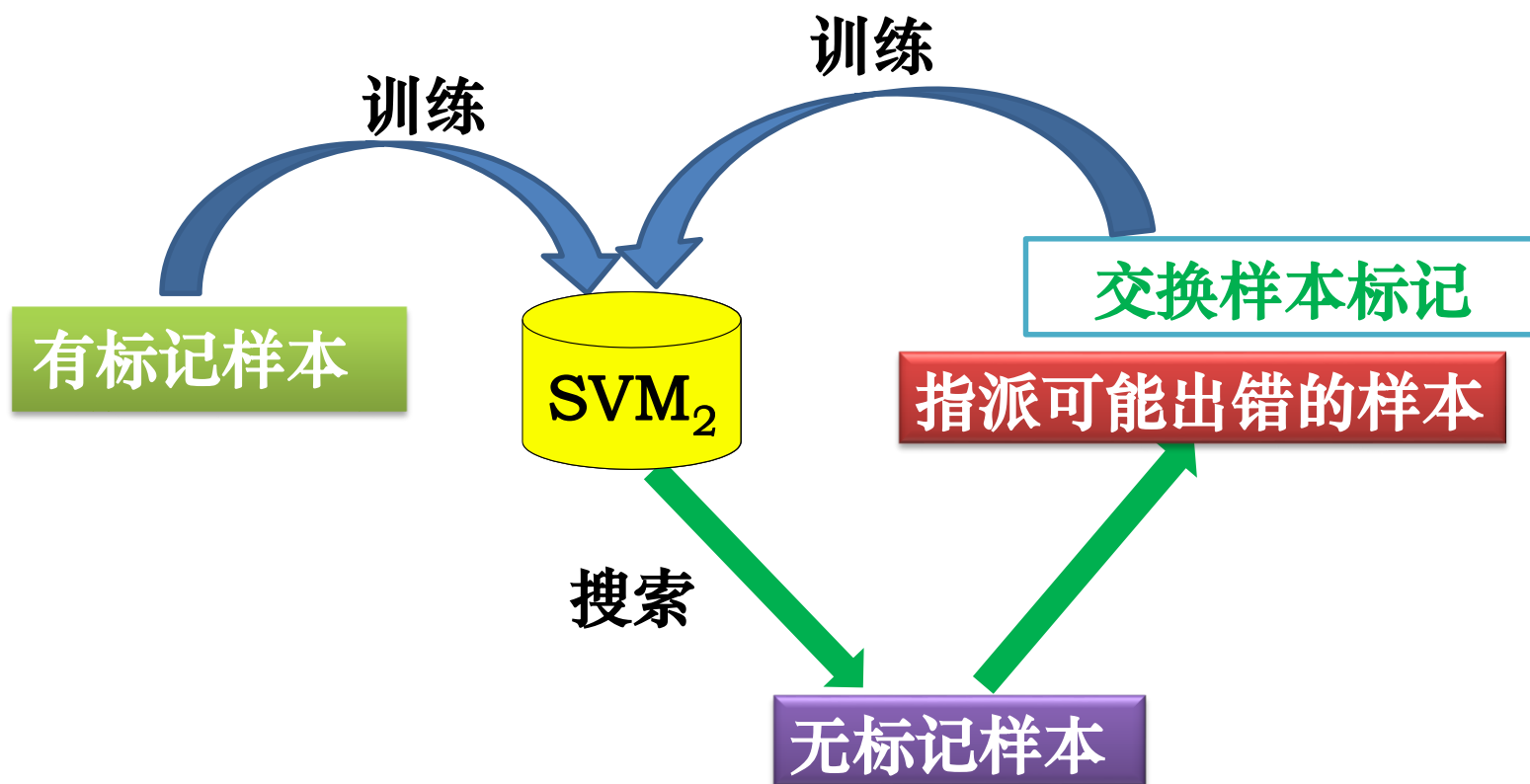
- T-SVM: 基于局部搜索的迭代近似优化求解



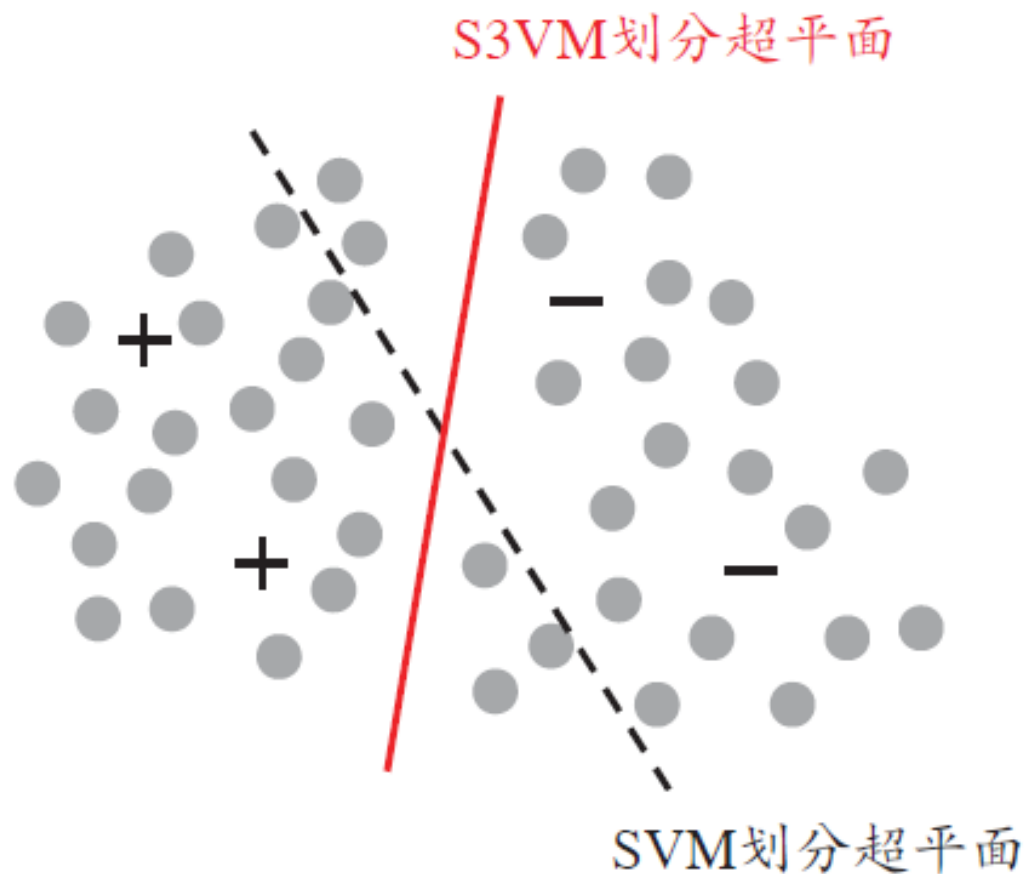
半监督SVM

(Semi-Supervised SVM, S3VM)

- T-SVM: 基于局部搜索的迭代近似优化求解



T-SVM和SVM算法对比



T-SVM的问题

- 搜寻标记指派出错的每一对未标记样本并进行调整，是一个复杂的大规模优化问题

高效优化求解算法

- 基于图和梯度下降的LDS算法 [Chapelle et al., 2005]
- 基于标记均值估计的MeanS3VM [Li et al., 2009]
- [Chapelle et al., 2005] Semi-Supervised Classification by Low Density Separation, AISTATS 2005.
- [Li et al., 2009] Semi-supervised Learning using Label Mean, ICML 2009.

半监督聚类

(Semi-Supervised Clustering)

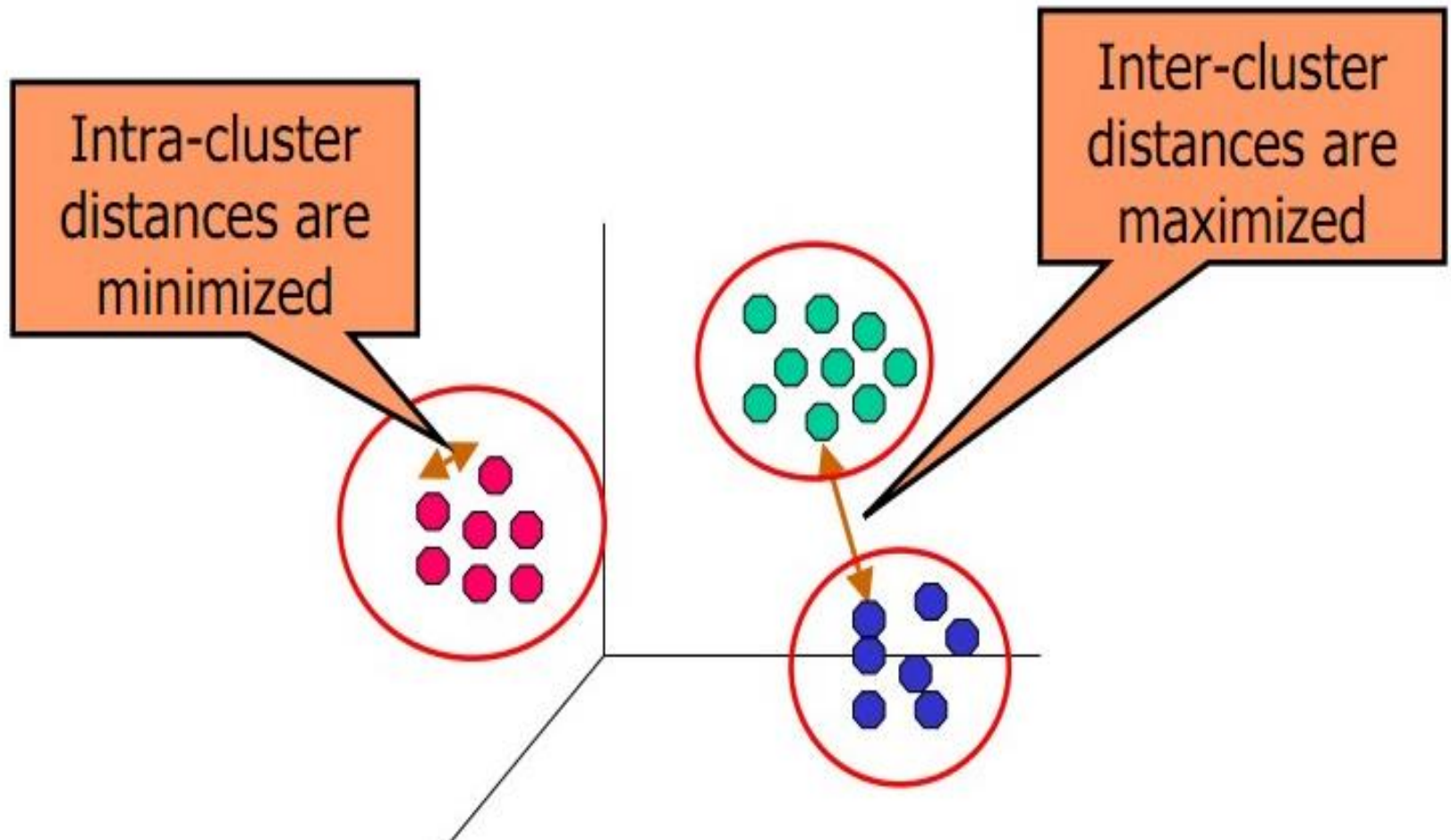
● 聚类

- 原理思路：“物以类聚，人以群分”，将数据样本根据其特定属性聚集几类（如K-Means），或者给出一个样本属于每一类的概率
- 同一个簇中的对象具有相似的属性，不同簇中的对象具有不同的属性
- 不需要事先根据训练数据去训练聚类器，是一种典型的无监督学习任务

半监督聚类

(Semi-Supervised Clustering)

- 聚类



半监督聚类

(Semi-Supervised Clustering)

● 聚类-K均值聚类

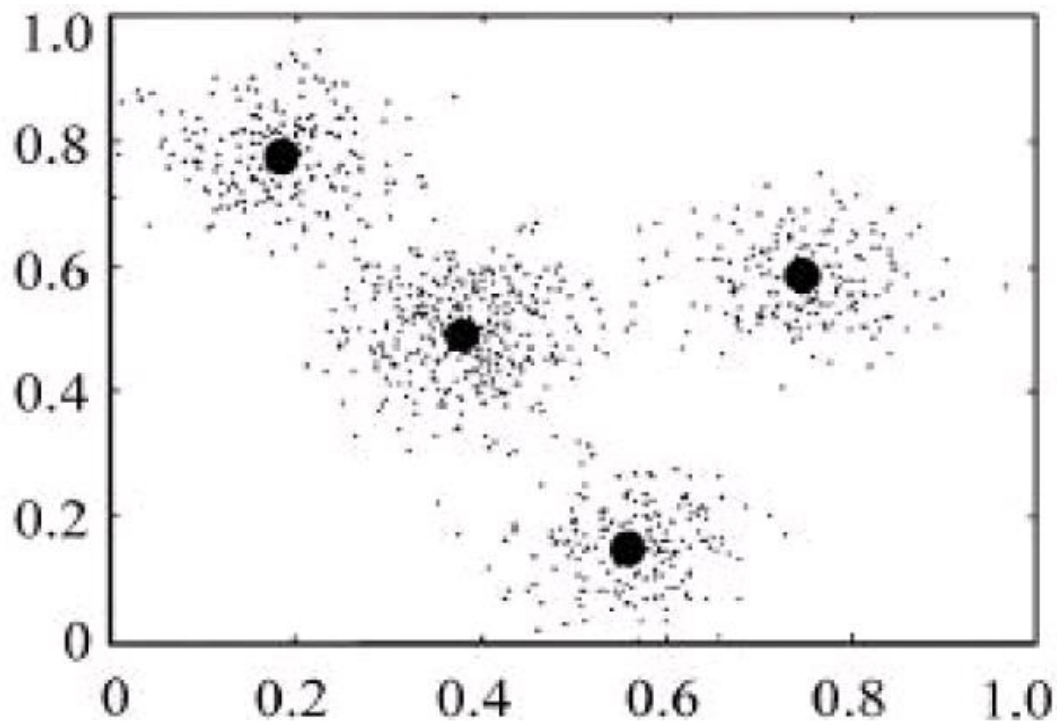
- 给定 D 维空间的数据集 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ，不知道这些数据所对应标签，通过聚类方法将这些数据集划分成 K 类
- 对于 K 个聚类中的每一类，分别建立一个代表点 μ_k ，将每一个样本划归到离该样本最近的 μ_k 所代表的聚类
- 目标：最小化准则函数

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

半监督聚类

(Semi-Supervised Clustering)

- 聚类-K均值聚类



半监督聚类

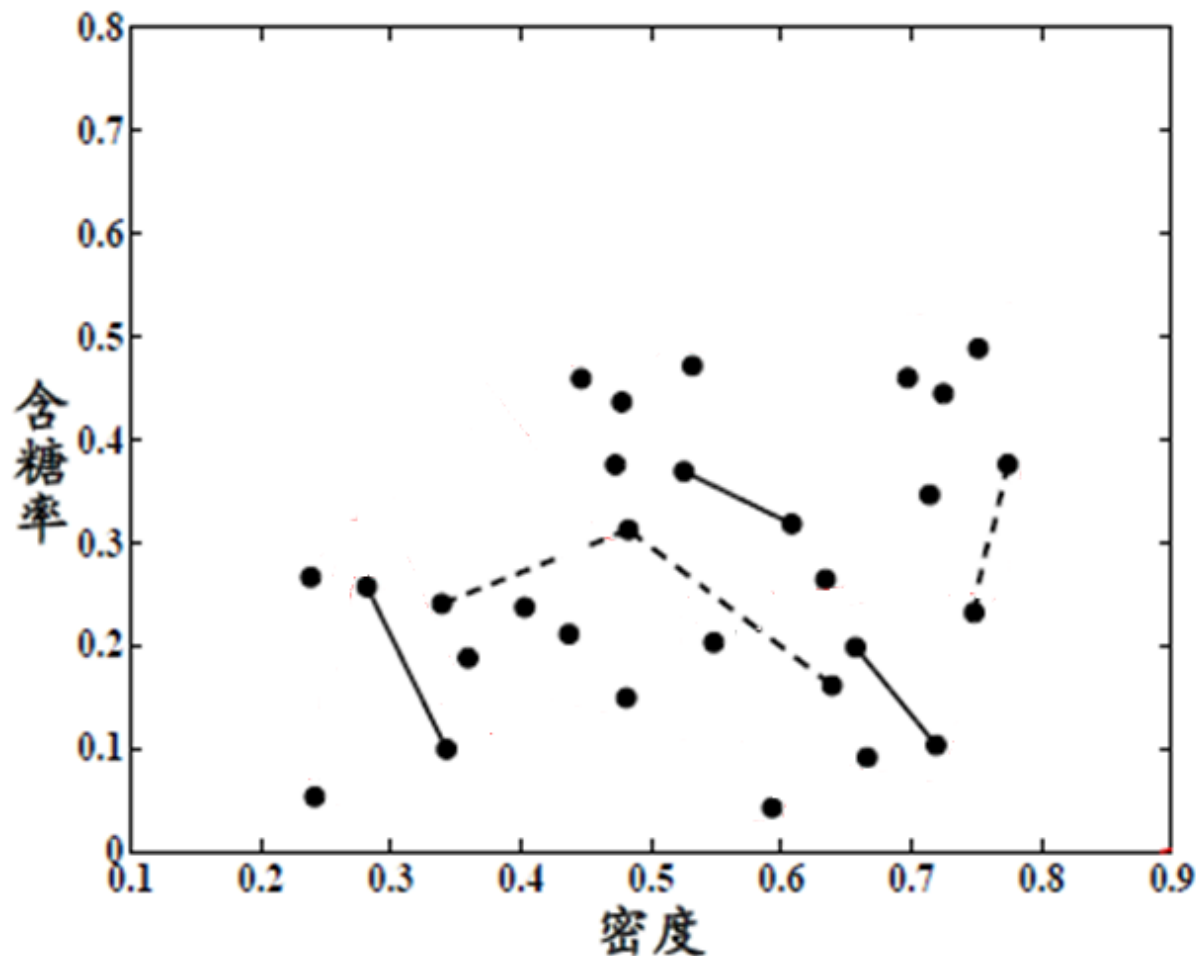
(Semi-Supervised Clustering)

- **实际聚类任务**：除了大量的未标记样本数据外，往往能获得一些额外的监督信息
 - 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**：利用少量有标记数据的监督信息获得更好的聚类效果

半监督聚类

(Semi-Supervised Clustering)

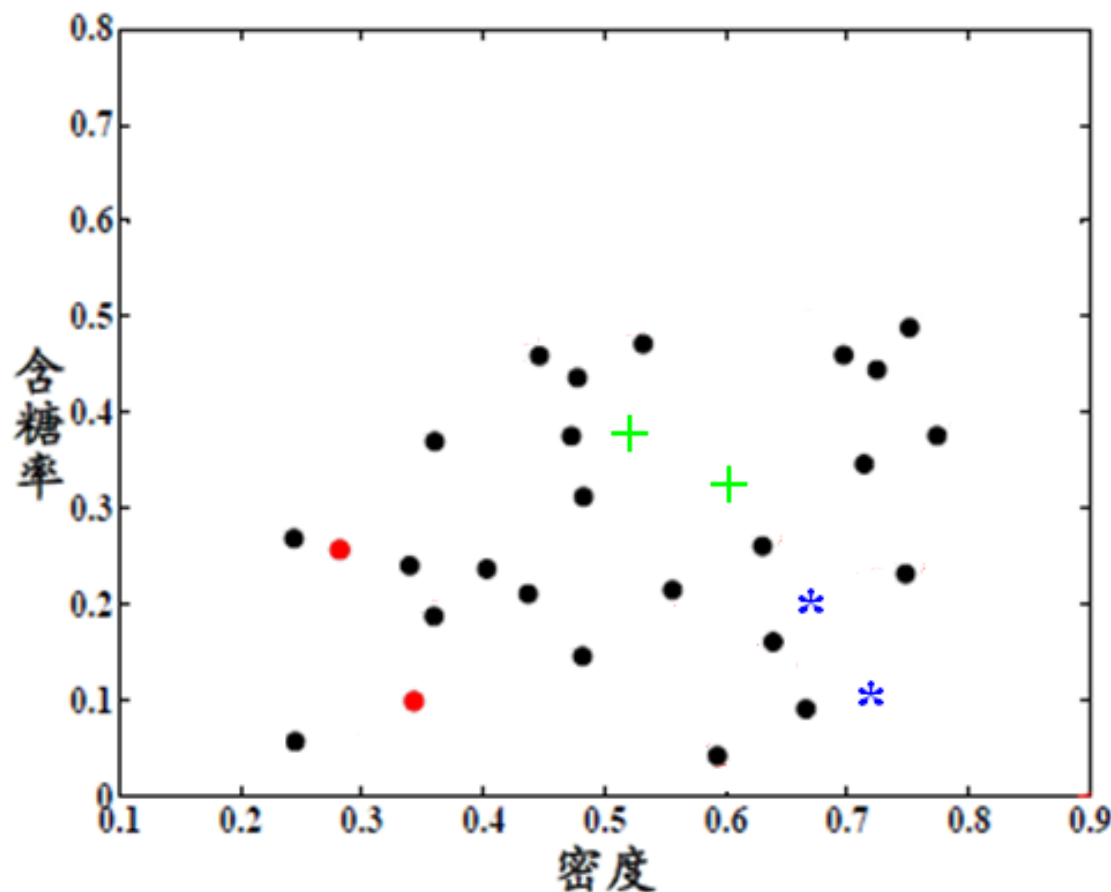
- 第一类监督信息：“必连”与“勿连”关系约束



半监督聚类

(Semi-Supervised Clustering)

- 第二种监督信息：具有少量有标记样本



半监督聚类-约束K均值

- 利用第一类监督信息的典型算法

约束K均值(Constrained K-means)算法

[Wagstaff et al., 2001]

- 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

- [Wagstaff et al., 2001] Wagstaff K, Cardie C, Rogers S, et al. Constrained k-means clustering with background knowledge, ICML 2001.

半监督聚类-约束K均值

```
8:   while  $\neg$  is_merged do
9:       基于  $\mathcal{K}$  找出与样本  $x_i$  距离最近的簇:  $r = \arg \min_{j \in \mathcal{K}} d_{ij}$ ;
10:      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;
11:      if  $\neg$  is_violated then
12:           $C_r = C_r \cup \{x_i\}$ ;
13:          is_merged=true
14:      else
15:           $\mathcal{K} = \mathcal{K} \setminus \{r\}$ ;
16:          if  $\mathcal{K} = \emptyset$  then
17:              break并返回错误提示
18:          end if
19:      end if
20:  end while
```

不冲突，选择最近的簇

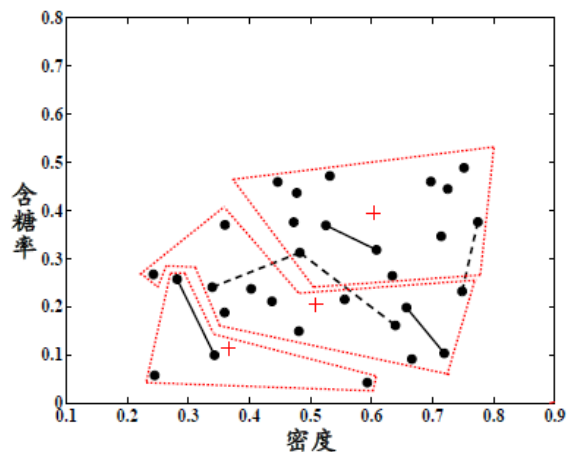
冲突，尝试次近的簇

更新均值向量.

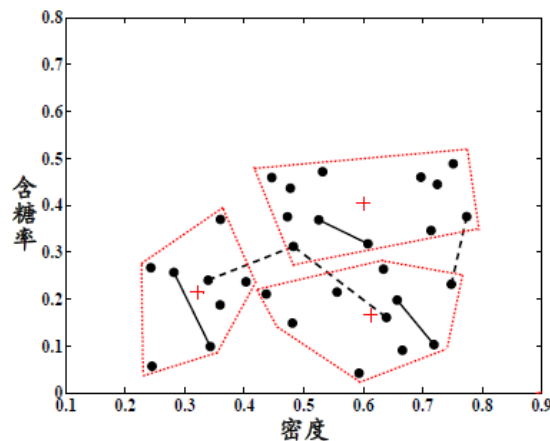
```
20:   end while
21: end for
22: for  $j = 1, 2, \dots, k$  do
23:      $\mu_j = \frac{1}{|C_j|} \sum_{x \in C_j} x$ ;
24: end for
25: until 均值向量均未更新
输出: 簇划分  $\{C_1, C_2, \dots, C_k\}$ 
```

图 13.7 约束 k 均值算法

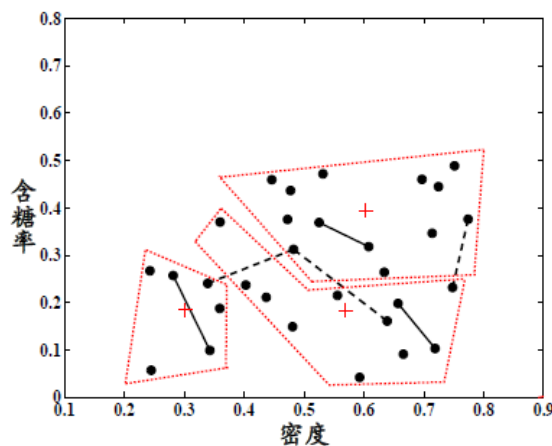
半监督聚类-约束K均值



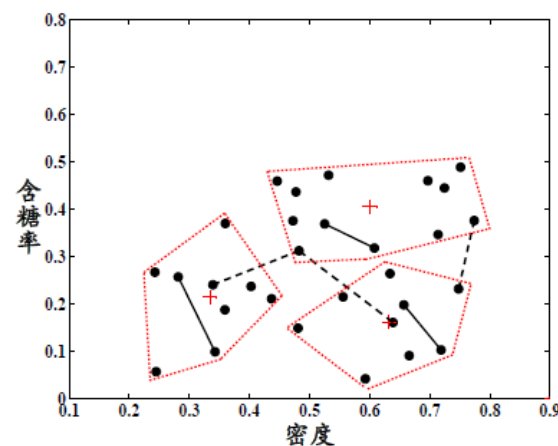
(a) 第 1 轮迭代后



(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-约束种子K均值

- 利用第二类监督信息的典型算法

约束种子K均值(Constrained Seed K-means)

[Basu et al., 2002]

监督信息：具有少量有标记样本，即假设少量有标记样本属于K个聚类簇

监督信息利用：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

- [Basu et al., 2002] S. Basu, A. Banerjee, R. Mooney, Semi-supervised Clustering by Seeding, ICML 2002.

半监督聚类-约束种子K均值

输入: 样本集 $D = \{x_1, x_2, \dots, x_m\}$;

用有标记样本初始化簇
中心.

```
1: for  $j = 1, 2, \dots, k$  do  
2:    $\mu_j = \frac{1}{|S_j|} \sum_{x \in S_j} x$   
3: end for
```

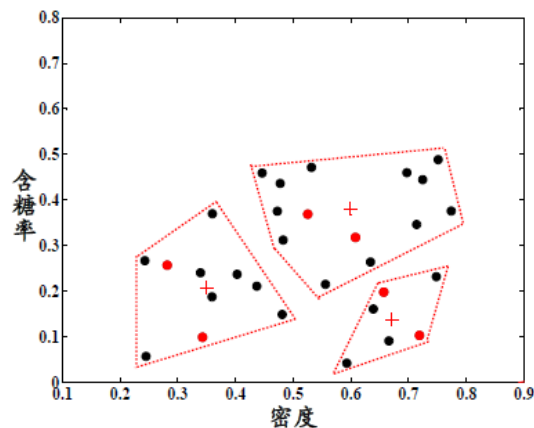
用有标记样本初始化 k
个簇.

```
6:   for  $j = 1, 2, \dots, k$  do  
7:     for all  $x \in S_j$  do  
8:        $C_j = C_j \cup \{x\}$   
9:     end for
```

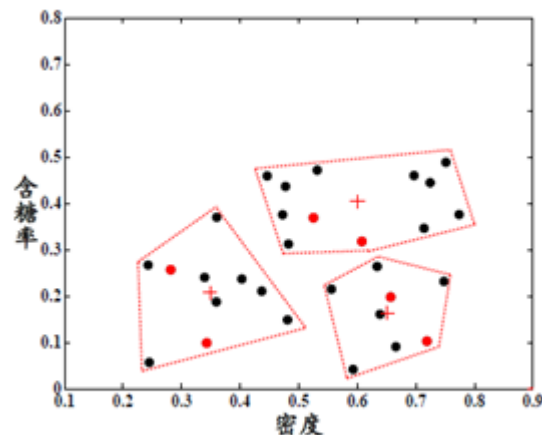
```
10:   找出与样本  $x_i$  距离最近的簇:  $r = \arg \min_{j \in \{1, 2, \dots, k\}} d_{ij}$ ;  
11:   for all  $x_i \in D \setminus S$  do  
12:     计算样本  $x_i$  与各均值向量  $\mu_j$  ( $1 \leq j \leq k$ ) 的距离:  $d_{ij} = \|x_i - \mu_j\|_2$ ;  
13:     找出与样本  $x_i$  距离最近的簇:  $r = \arg \min_{j \in \{1, 2, \dots, k\}} d_{ij}$ ;  
14:     将样本  $x_i$  划入相应的簇:  $C_r = C_r \cup \{x_i\}$   
15:   end for
```

图 13.9 约束种子 k 均值算法

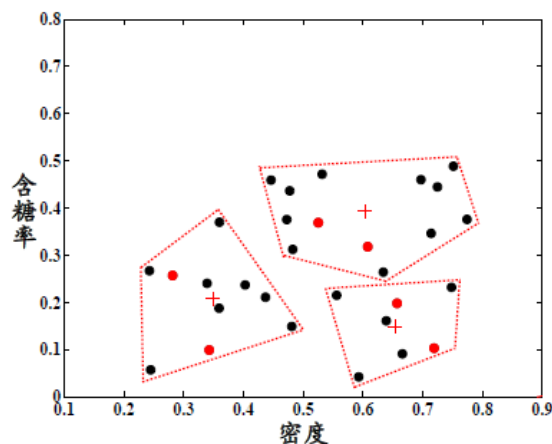
半监督聚类-约束种子K均值



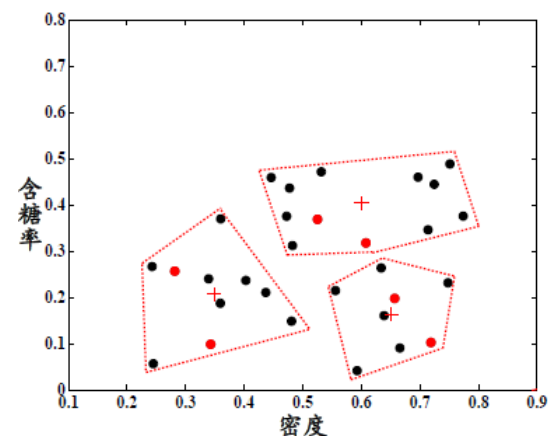
(a) 第 1 轮迭代后



(d) 第 3 轮迭代后



(b) 第 2 轮迭代后



(d) 第 4 轮迭代后

存在问题及发展趋势

- 新假设的提出

- 现有算法仅基于聚类和流形假设，期待新的模型假设来扩充和改善半监督学习算法

- 半监督学习的抗干扰性比较弱

- 现有算法多针对无噪声干扰样本数据，而实际数据通常存在噪声干扰
- 需综合考虑噪声干扰下未标记样本数据分布的不确定性及复杂性

延伸阅读

- 其他半监督学习算法

- 基于分歧的方法-协同训练(Co-training)[1]

- 基于图的方法(Graph-based methods)[2]

- Blum, A., & Mitchell, T. (1998). **Combining labeled and unlabeled data with co-training**. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, 92-100.

- Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). **Semi-supervised learning using Gaussian fields and harmonic functions**. *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 912-919.

本节课程结束!

下节课再见!