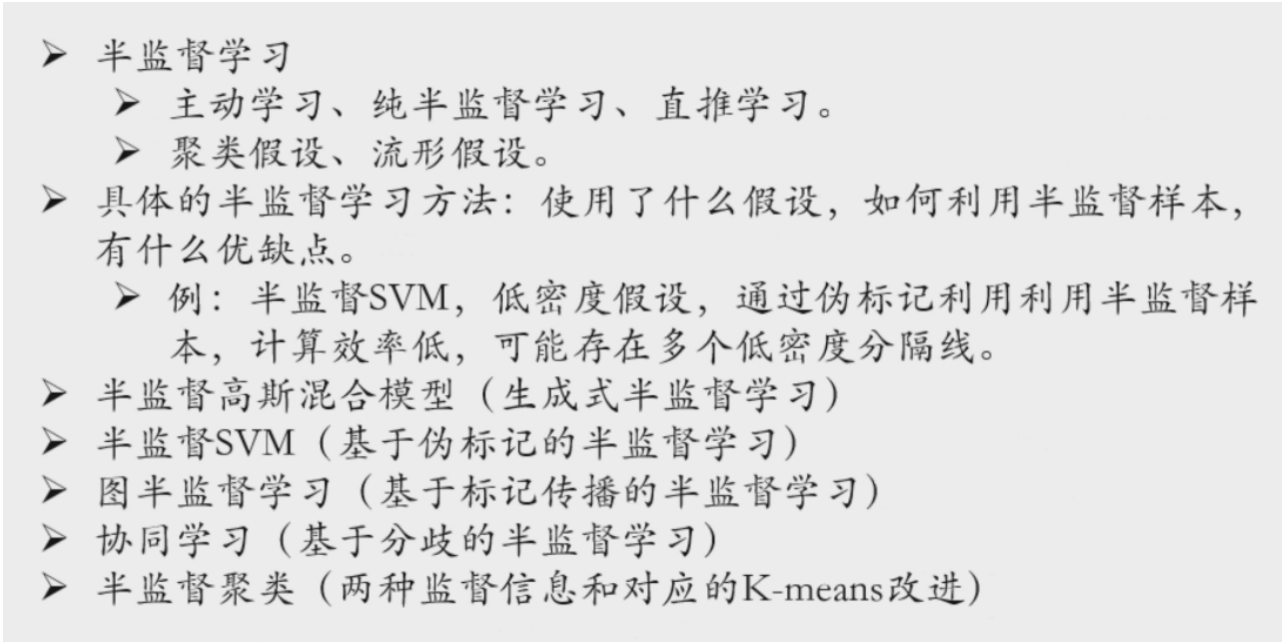


3个区分

这些算法做出什么假设，优缺点，适用于场景，如何利用无标记样本

（半监督SVM，低密度假设，伪标记）

半监督Semi-Supervised ：学习器不依赖外界交互，自动利用未标记样本来提升学习性能

- 纯半监督（开放世界）：未标记样本非待预测数据
- 直推学习（封闭世界）：未标记样本恰为待预测数据（无泛化能力）

无标记样本和测试样本都无标记，不区分

主动学习：使用尽量少的样本进行“查询”，希望每次挑出的是改善模型帮助大的样本，最大化标记无标记样本

> 主动学习引入额外专家知识，半监督不用

一、未标记样本的假设

- 假设：相似样本有相似输出

聚类假设

假设存在簇结构，同簇的同类

> 关注整体特性

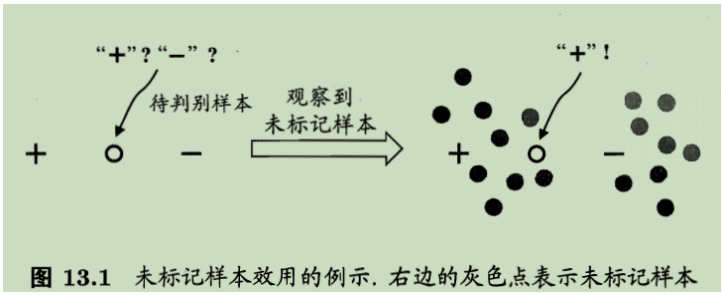


图 13.1 未标记样本效用的例示. 右边的灰色点表示未标记样本

流形假设

假设临近样本有相似输出值，分布在一个流形结构上

> 用相似程度来刻画，关注局部特性

- 是聚类假设的推广，但对输出值没有限制，适用范围更广

二、生成式方法

假设数据由同一个潜在模型生成，未标记样本看成模型的缺失参数

- 缺失参数通常用EM进行极大似然估计求解
- 关键：模型假设必须准确！

> 往往很难假设准确，除非有充分可靠的领域知识

无监督项. 显然, 高斯混合模型参数估计可用 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 $\boldsymbol{x}_j$ 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

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

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

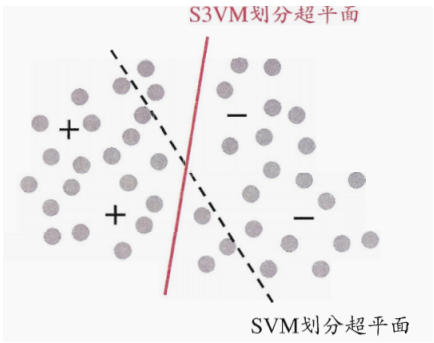
$$\alpha_i = \frac{1}{m} \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) . \tag{13.8}$$

以上过程不断迭代直至收敛, 即可获得模型参数. 然后由式(13.3)和(13.2)就能对样本进行分类.

三、半监督SVM--S3VM

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

基本假设：低密度分隔



TSVM

> Transductive

标记指派，每个未标记都分别作为正例和反例，寻找间隔最大化划分超平面

> 穷举，开销巨大

算法：先根据SVM结果进行伪标记，给真标记更大权重，再逐步微调

- 类别不平衡：计算正反例个数，乘权重

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：

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

Q：为什么TSVM开销巨大?

四、图半监督

边的强度正比于样本点的相似度，标记样本是染色点，半监督学习看成颜色扩散过程

- 图可以对应矩阵，进行矩阵运算

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：

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

缺点：

- 存储开销大，矩阵规模O（nn）
- 难以判断新样本再图中位置，只能重构图和标记传播

五、基于分歧的方法（多视图学习）

> 分歧指学习器之间的分歧，co-training协同训练，最初针对多视图数据
> 分歧对未标记数据的利用很重要

不同视图、不同算法、不同数据采样、不同参数设置等，都仅是产生差异的渠道，而非必备条件

条件

- 相容性：即其所包含的关于输出空间y的信息是一致的
- 充分：每个视图都包含足以产生最优学习器的信息
- 条件独立：在类别标记条件下两个视图独立（现实任务很难满足）

训练

每个学习器标记最有把握的未标记样本，然后提供给其他学习器进行迭代互相学习

六、半监督聚类

因为在现实聚类任务中我们往往能获得一些额外的监督信息，分两种类型：

- 必连和勿连：约束kmeans
- 少量标记：约束种子kmeans（直接作为初始化的聚类中心）

增加步骤：xi的划分是否违反了已知信息，违反则划进最近的