- 监督学习中的回归对应了无监督学习中的密度估计（只从x样本本身估计分布）
- 分类对应了聚类
- 训练（得到簇时）不需要标记，评价时才需要，而且簇标记其实并无意义，簇就是意义

> 也可以针对有标记的数据展开

一、常见聚类方法

- 聚类的准则很多，没有对错，如何定义距离？

原型聚类

> 原型表示有代表性的点
> 基本有公式，可拓展性好，容易融合到其它模型中
> 只能聚类为椭球类型，其他形状不行

- 假设聚类结构能通过一组原型刻画
- 先对原型初始化，再迭代更新求解（初始化未必找中心点）
- 比如：k均值，LVQ，高斯混合聚类

密度聚类

- 假设可以通过紧密程度划分
- 用密度刻画样本间可连接性
- 比如：DBSCAN，OPTICS，DENCLUE

层次聚类

- 在不同层次对数据集进行划分。形成树形的聚类结构

> 可动态选择需要多少簇

- 可采用"自底向上"的聚合策略，也可采用"自顶向下"的分拆策略
- AGNES

二、性能度量

> 有效性指标

- 目标：簇内相似度高，簇间相似度低
- 外部指标：结果与其他参考模型比较（JAccard系数，Fm指数，Rand指数）
- 内部指标：直接考察结果（DB指数，Dunn指数）

距离度量四个性质

- 非负性：距离非负
- 同一性：只能和自己距离为0
- 对称性：x到y和y到x相同
- 直递性（三角不等式）：任意三个点满足两边之和大于等于第三变

> 向量内积不满足非负性

常用距离形式

闵可夫斯基距离Minkowski

> Lp范数
>
> 用于有序属性

- p为无穷，是切比雪夫距离
- p=2为欧式距离（严格欧式距离有根号）
- p=1为曼哈顿距离，街区距离

$$\operatorname{dist}_{\mathrm{mk}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n} |x_{iu} - x_{ju}|^p\right)^{\frac{1}{p}}$$

VDM距离

> 无序属性，如{火车，飞机}

对无序属性可采用 VDM (Value Difference Metric) [Stanfill and Waltz, 1986]. 令 $m_{u,a}$ 表示在属性 u 上取值为 a 的样本数, $m_{u,a,i}$ 表示在第 i 个样本簇中在属性 u 上取值为 a 的样本数, k 为样本簇数, 则属性 u 上两个离散值 a 与 b 之间的 VDM 距离为

$$\mathrm{VDM}_p(a, b) = \sum_{i=1}^{k} \left| \frac{m_{u,a,i}}{m_{u,a}} - \frac{m_{u,b,i}}{m_{u,b}} \right|^p . \tag{9.21}$$

可将二者混合MinkovDM

$$\mathrm{MinkovDM}_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\sum_{u=1}^{n_c} |x_{iu} - x_{ju}|^p + \sum_{u=n_c+1}^{n} \mathrm{VDM}_p(x_{iu}, x_{ju})\right)^{\frac{1}{p}}$$

三、k-means

> 贪心算法，原型聚类
> 用簇内均值代表簇
> 采用欧式距离
> k-medoids是把距离改为最近样本

输入: 样本集 $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$;
聚类簇数 k.

过程:

从 D 中随机选择 k 个样本作为初始均值向量 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$
repeat
　令 $C_i = \varnothing \ (1 \leqslant i \leqslant k)$
　for $j = 1, 2, \ldots, m$ **do**
　　计算样本 $\boldsymbol{x}_j$ 与各均值向量 $\boldsymbol{\mu}_i \ (1 \leqslant i \leqslant k)$ 的距离: $d_{ji} = ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||_2$;
　　根据距离最近的均值向量确定 $\boldsymbol{x}_j$ 的簇标记: $\lambda_j = \arg\min_{i \in \{1,2,\ldots,k\}} d_{ji}$;
　　将样本 $\boldsymbol{x}_j$ 划入相应的簇: $C_{\lambda_j} = C_{\lambda_j} \bigcup \{\boldsymbol{x}_j\}$;
　end for
　for $i = 1, 2, \ldots, k$ **do**
　　计算新均值向量: $\boldsymbol{\mu}_i' = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i} \boldsymbol{x}$;
　　if $\boldsymbol{\mu}_i' \neq \boldsymbol{\mu}_i$ **then**
　　　将当前均值向量 $\boldsymbol{\mu}_i$ 更新为 $\boldsymbol{\mu}_i'$
　　else
　　　保持当前均值向量不变
　　end if
　end for
until 当前均值向量均未更新

输出: 簇划分 $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$

图 9.2　k 均值算法

> 一般需要设置最大运行轮数或最小调整幅度阈值

1. 随机选k个作为k个中心
2. 其他样本根据距离划分给最近的簇（E）
3. 更新各簇的均值，作为新的簇中心（M）
4. 所有簇中心不再改变时停止迭代，否则重复step2

最小化均方误差和

$$E = \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} ||\boldsymbol{x} - \boldsymbol{\mu}_i||_2^2 ,$$

算法影响因素

- 样本输入顺序
- 模式相似性测度
- 初始类中心选取

四、学习向量量化LVQ

> Learning Vector Quantization
> 监督学习算法

- 假设样本有类别标记，利用样本监督信息辅助聚类

输入： 样本集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
原型向量个数 q, 各原型向量预设的类别标记 $\{t_1, t_2, \ldots, t_q\}$;
学习率 $\eta \in (0, 1)$.

过程：

```
初始化一组原型向量 {p_1, p_2, ..., p_q}
repeat
    从样本集 D 随机选取样本 (x_j, y_j);
    计算样本 x_j 与 p_i (1 ≤ i ≤ q) 的距离: d_ji = ||x_j − p_i||_2;
    找出与 x_j 距离最近的原型向量 p_i*, i* = arg min_{i∈{1,2,...,q}} d_ji;
    if y_j = t_i* then
        p' = p_i* + η · (x_j − p_i*)
    else
        p' = p_i* − η · (x_j − p_i*)
    end if
    将原型向量 p_i* 更新为 p'
until 满足停止条件
```

输出： 原型向量 $\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q\}$

图 9.4 学习向量量化算法

- 把原型向样本方向迭代更新

五、高斯混合聚类

> （Mixture-of-Gaussian）
>
> 采用概率模型表达聚类类型
>
> 生成式建模

- 假设每个簇都是高斯分布
- 极大似然估计

> 但不太好算，引入EM算法，针对未观测到的变量、隐变量

EM算法：专杀隐变量

> 隐变量：未观测变量（西瓜根蒂脱落，观测不到）

- 隐变量Z：应该属于哪个簇

> 因为有隐变量所以不能直接极大似然估计

- 先计算隐变量的后验分布
- E step根据当前参数进行估算，M step根据估算结果更新参数
- E：根据当前参数值估算隐变量期望
- M：似然估计求使E最大的参数值，更新参数
- E：计算期望：利用当前估计的参数值计算对数似然的期望值；
- M：最大化：寻找能使E步产生的似然期望最大化的参数值