第九章 聚类算法

By Xian2207, 13689903575, wszhangxian@126.com

1 聚类任务

聚类：试图将数据集中的样本划分为若干个通常互不相交的子集。每个子集称为一个样本簇（cluster）。同一簇样本尽可能相似，不同簇样本应尽可能不同，即“簇内相似度”（intra-cluster similarity）高于“簇间相似度”（inter-cluster similarity）。
无监督学习：训练样本的标记信息是未知的，即只有样本x_i，没有对应的类标y_i。目标是通过对无标记样本的学习揭示数据内在性质，为进一步数学分析提供基础。

2 性能度量

聚类性能度量也称聚类“有效性指标”（validity index），即通过性能度量来评估聚类结果的好坏。性能度量分两大类：一类是将聚类结果与某个“参考模型”（reference model）比较，这种参考模型也被称为“外部指标”（external index）；另一类是直接考察聚类结果而不利用任何参考模型，称为“内部指标”（internal index）。
对数据集D定义聚类给出的簇为$C_1, C_2, \dots, C_k$，参考模型给出的簇为C^*，将样本两两组合，会有四个集合即：
a: 在C中同簇，在C^*中也同簇；
b: 在C中同簇，在C^*中不同簇；
c: 在C中不同簇，在C^*中同簇；
d: 在C中不同簇，在C^*中也不同簇；
基于上述集合导出一些常用的聚类性能外部指标：
A: Jaccard 系数（Coefficient, JC）

$$JC = \frac{a}{a+b+c}$$

B: FM 指数（Foolkes and Mallows Index, FMI）

$$JC = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

C: Rand 指数（Rand Index, RI）

$$\frac{2(a+d)}{m(m-1)}$$

上述指标的值均在[0,1]之间，值越大越好。对于聚类结果的簇划分，可有
avg(C):　　簇 C 内样本的平均距离；
diam(C):　　簇 C 内样本的最远距离；
$d_{min}(C_i, C_j)$:　两个簇最近的样本间的距离；
$d_{cen}(C_i, C_j)$:　两个簇中心点间的距离；
基于上述距离，又可导出聚类内性能度量的内部指标

（1）DB 指数（Davies-Bouldin Index, DBI）

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i} \frac{avg(C_i)+avg(C_j)}{d_{cen}(\mu_i,\mu_j)}$$

（2）Dunn 指数（Dunn Index, DI）

$$DI = \min_{1\leq i\leq k}\{\min_{j\neq i}(\frac{d_{min}(C_i,C_j)}{\max_{1\leq l\leq k} diam(C_l)})\}$$

显然，DBI 的值越小，即簇内样本的平均距离越小，说明样本同在一个簇内；DI 越大，即两个簇最近的样本距离越大，说明两个簇彼此距离越远，属于不同簇。

3 距离计算

距离具有四种特性，非负性（$d\geq 0$），对称性（$d(x_i,x_j)=d(x_j,x_i)$），直递性（$d(x_i,x_j)\leq d(x_i,x_k)+d(x_k,x_j)$）。最常用的是闵可夫斯基距离（Minkovski Distance, mk），

$$d_{mk}(x_i,x_j)=(\sum_{u=1}^{n}|x_{i,u}-x_{j,u}|^p)^{\frac{1}{p}}$$

当$p=1$即曼哈顿距离（Manhattan Distance）；当$p=2$即欧拉距离（Euclidean Distance）。属性划分一般为连续属性（continuous attribute）和离散属性（categorical attribute），前者可能有无穷多个取值，后者为有限个取值。然而在讨论距离计算时，我们关注属性定义是否“有序”。例如离散属性{1,2,3}的定义与连续属性的性质更接近一些，能直接在属性值上计算距离，如 1 到 2 的距离比 1 到 3 的距离更近。类似这样的属性，我们称为“有序属性”（ordinal attribute）；而对于{airplane, horse, person}这样的属性，不能计算出距离，称为“无序属性”（non-ordinal attribute）。故闵可夫斯基距离主要适用于有序属性，且根据属性重要性，衍生了加权的闵可夫斯基距离。对无序属性，我们采用 VDM（Value Difference Metric）距离。将闵可夫斯基距离和 VDM 结合起来，即可处理混合属性的样本距离。但是，有些距离计算不满足距离的直递性，如“人”，“马”分别与“人马”距离相同，但人马距离（人马差别很大，可直观想象）大于人到人马+人马到马之间的距离，如图示。

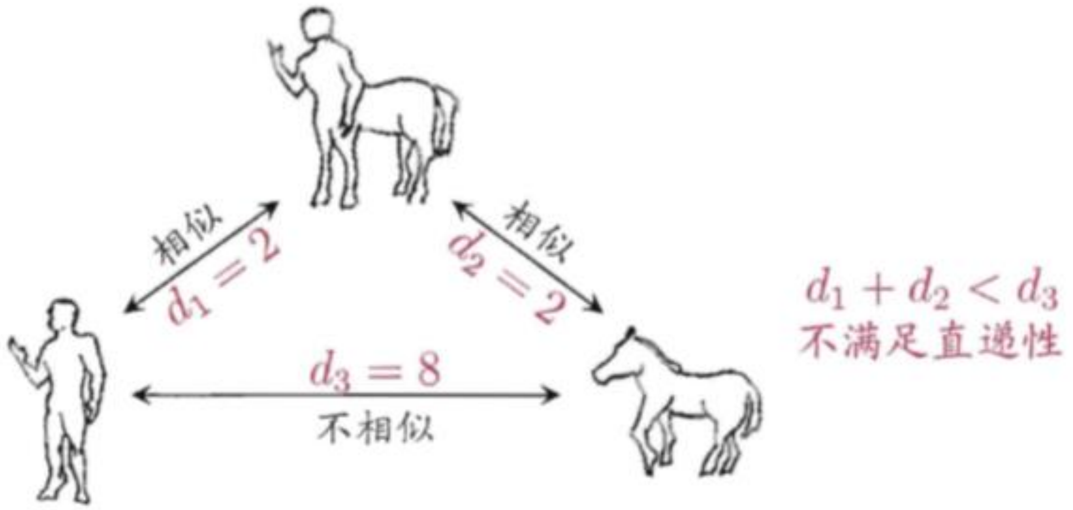


图 9.1 非度量距离的一个例子

此类距离称为“非度量距离”（non-metric distance），要解决此类距离的计算，须借助“距离

度量学习”（distance metric learning）来实现。

4 原型聚类

原型聚类定义为“基于原型的聚类”（prototype-based clustering），原型即指样本空间具有代表性的点。机制是假设聚类结构能通过一组原型刻画，算法先对原型初始化，然后对原型进行迭代更新求解。采用不同的原型表示，不同的求解方式，可产生不同的算法。主要的算法有：

4.1 k-均值（k-means）算法

给定样本集$D = \{x_1, x_2, \ldots, x_m\}$，“k-均值”（k-means）算法针对聚类所得的簇划分$C = \{C_1, C_2, \ldots, C_k\}$，则C最小化平方误差为

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - u_i)^2$$

输入：样本集D，聚类簇数k

输出：k个簇的样本划分结果，

过程：

（1） 从 D（样本空间中样本总数等于 m）中选出 k 个样本（每个样本总数等于 m）；

（2） 对每个样本，计算样本与自身，样本与其他样本之间的欧式距离，得到 k 个簇数；

（3） 计算 k 个 cluster 的中心点即均值，依次用这 k 个中心点遍历样本，计算距离，重新划分样本到 k 个 cluster 里；

（4） 重复 2 直到这 k 个中心点的均值不再更新。

4.2 k-近邻（k-Nearest Neighbor）算法

KNN 算法思想是“近朱者赤，近墨者黑”，由邻居判断类别。

计算步骤如下：

A: 算距离

给定测试对象，计算它与训练集中每个样本之间的距离；

B: 找邻居

圈定距离最近的 k 个训练对象，作为分类依据；

优点：

无需参数估计，无需训练，实现简单，适合多分类，可能比 SVM 还好。

缺点：

懒作算法，对测试样本分类时的计算量答，内存开销大，评分慢，可解释性没决策树好。

常见问题：

K 值设定多大？太大，过拟合；太小，增加过多的分类，计算开销大，且容易受噪音影响。

4.3 学习向量量化

学习向量量化（Learning Vector Quantization, LVQ）原理同 k-means 类似，也试图找一组原型向量来刻画聚类结构，但与 k-means 不同的是，它假设样本带有类别标记，利用这些样本辅助聚类划分。详情参考 P204 页。

4.4 高斯混合聚类

与 k-means, LVQ 不同，高斯混合聚类（Gaussian Mixture Clustering）采用概率模型来表达聚类原型。参考 P205 页。

4.5 密度聚类

密度聚类也被称为“基于密度的聚类”（density-based clustering），机制是通过样本分布的密度来确定。一般，该算法从样本密度角度来考察样本之间的可连续性，并基于可连续样本不断拓展聚类簇，以获得最终划分结果。知名算法是 DBSCAN，详情参考 P211。

4.6 层次聚类

顾名思义，层次聚类（hierarchical clustering）试图在不同层次对数据进行划分，从而形成树形的聚类结构。数据集的划分可采用“自底而上”的聚合策略，也可采用“自顶而下”的分拆策略。参考 P214 页。

5 k-means 实例

参考链接：网上答案有误，这里重新做一遍。
https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput695/F07/exercises/Exercises695Clus-solution.pdf
假设数据集空间$\mathrm{D}=\{A_1(2,10),A_2(2,5),A_3(8,4),A_4(5,8),A_5(7,5),A_6(6,4),A_7(1,2),A_8(4,9)\}$，要求是将其使用 k-means 算法和欧拉距离划分为 3 个簇（cluster），即$\mathrm{k}=3$。
答：
欧拉距离：

$$d_{mk}(x_i,x_j)=(\sum_{u=1}^{n}|x_{i,u}-x_{j,u}|^p)^{\frac{1}{p}},\ p=2$$

则

$$d_{\text{Eulidean}}(x_i,x_j)=\sqrt{(x_i-x_j)^2}$$

问题分析：
由于划分 3 个 cluster，故须从样本空间D中任选 3 个样本作为 3 个 cluster 的初始化样本依据。这里选择$\mathrm{D}'=\{A_1(2,10),\ A_4(5,8),\ A_7(1,2)\}$三个点，代表三个簇的均值，即$A_1\in$ cluster 1；$A_4\in$ cluster 2；$A_7\in$ cluster 3

第一次迭代：
遍历样本
依次计算样本与这三个簇均值的距离。三个 cluster 都只有一个点，故均值就是点本身
$d_{1,1}(x_1,x_1)=d_{1,1}(A_1,A_1)=\sqrt{(2-2)^2+(10-10)^2}=0;$
$d_{1,4}(x_1,x_4)=d_{1,4}(A_1,A_4)=\sqrt{(2-5)^2+(10-8)^2}=\sqrt{13};$
$d_{1,7}(x_1,x_7)=d_{1,7}(A_1,A_7)=\sqrt{(2-1)^2+(10-2)^2}=\sqrt{65};$
故$d_{1,1}$最小，即A_1与 cluster 1 最近，所以$A_1\in$ cluster 1;
$d_{2,1}(x_2,x_1)=d_{2,1}(A_2,A_1)=\sqrt{(2-2)^2+(5-10)^2}=5;$
$d_{2,4}(x_2,x_4)=d_{2,4}(A_2,A_4)=\sqrt{(2-5)^2+(5-8)^2}=\sqrt{18};$
$d_{2,7}(x_2,x_7)=d_{2,7}(A_2,A_7)=\sqrt{(2-1)^2+(5-2)^2}=\sqrt{10};$
故$d_{2,7}$最小，即A_2与 cluster 3 最近，所以$A_2\in$ cluster 3;
$d_{3,1}(x_3,x_1)=d_{3,1}(A_3,A_1)=\sqrt{(8-2)^2+(4-10)^2}=\sqrt{72};$
$d_{3,4}(x_3,x_4)=d_{3,4}(A_3,A_4)=\sqrt{(8-5)^2+(4-8)^2}=\sqrt{25};$
$d_{3,7}(x_3,x_7)=d_{3,7}(A_3,A_7)=\sqrt{(8-1)^2+(4-2)^2}=\sqrt{53};$
故$d_{3,4}$最小，即A_3与 cluster 2 最近，所以$A_3\in$ cluster 2;
$d_{4,1}(x_4,x_1)=d_{4,1}(A_3,A_1)=\sqrt{(5-2)^2+(8-10)^2}=\sqrt{13};$
$d_{4,4}(x_4,x_4)=d_{4,4}(A_3,A_4)=\sqrt{(5-5)^2+(8-8)^2}=0;$
$d_{4,7}(x_4,x_7)=d_{4,7}(A_3,A_7)=\sqrt{(5-1)^2+(8-2)^2}=\sqrt{52};$
故$d_{4,4}$最小，即A_4与 cluster 2 最近，所以$A_4\in$ cluster 2;
$d_{5,1}(x_5,x_1)=d_{5,1}(A_5,A_1)=\sqrt{(7-2)^2+(5-10)^2}=\sqrt{50};$
$d_{5,4}(x_5,x_4)=d_{5,4}(A_5,A_4)=\sqrt{(7-5)^2+(5-8)^2}=\sqrt{13};$
$d_{5,7}(x_5,x_7)=d_{5,7}(A_5,A_7)=\sqrt{(7-1)^2+(5-2)^2}=\sqrt{45};$
故$d_{5,4}$最小，即A_5与 cluster 2 最近，所以$A_5\in$ cluster 2;
$d_{6,1}(x_6,x_1)=d_{6,1}(A_6,A_1)=\sqrt{(6-2)^2+(4-10)^2}=\sqrt{52};$
$d_{6,4}(x_6,x_4)=d_{6,4}(A_6,A_4)=\sqrt{(6-5)^2+(4-8)^2}=\sqrt{17};$
$d_{6,7}(x_6,x_7)=d_{6,7}(A_6,A_7)=\sqrt{(6-1)^2+(4-2)^2}=\sqrt{29};$
故$d_{6,4}$最小，即A_6与 cluster 2 最近，所以$A_6\in$ cluster 2;
$d_{7,1}(x_7,x_1)=d_{7,1}(A_7,A_1)=\sqrt{(1-2)^2+(2-10)^2}=\sqrt{65};$
$d_{7,4}(x_7,x_4)=d_{7,4}(A_7,A_4)=\sqrt{(1-5)^2+(2-8)^2}=\sqrt{52};$
$d_{7,7}(x_7,x_7)=d_{7,7}(A_7,A_7)=\sqrt{(1-1)^2+(2-2)^2}=0;$
故$d_{7,7}$最小，即A_7与 cluster 3 最近，所以$A_7\in$ cluster 3;
$d_{8,1}(x_8,x_1)=d_{8,1}(A_8,A_1)=\sqrt{(4-2)^2+(9-10)^2}=\sqrt{5};$
$d_{8,4}(x_8,x_4)=d_{8,4}(A_8,A_4)=\sqrt{(4-5)^2+(9-8)^2}=\sqrt{2};$
$d_{8,7}(x_8,x_7)=d_{8,7}(A_8,A_7)=\sqrt{(4-1)^2+(9-2)^2}=\sqrt{58};$
故$d_{8,4}$最小，即A_8与 cluster 2 最近，所以$A_8\in$ cluster 2;

Clustering： cluster 1 =$\{A_1\}$; cluster 2 =$\{A_3,A_4,A_5,A_6,A_8\}$; cluster 3 =$\{A_2,A_7\}$

更新均值：
$\alpha_1=(x \text{ of } A_1, y \text{ of } A_1)$

$\alpha_2=(\frac{x \text{ of } A_3+A_4+A_5+A_6+A_8}{5},\frac{y \text{ of } A_3+A_4+A_5+A_6+A_8)}{5}$

$\alpha_3=(\frac{x \text{ of } A_2+A_7}{2},\frac{y \text{ of } A_2+A_7)}{2}$

故

$\alpha_1 = (2,10)$

$\alpha_2 = (\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}) = (6,6)$

$\alpha_3 = (\frac{2+1}{2}, \frac{5+2)}{2} = (1.5,3.5)$

第二次迭代：

$d_{A_1,\alpha_1}(A_1, \alpha_1) = \sqrt{(2-2)^2 + (10-10)^2} = 0;$

$d_{A_1,\alpha_2}(A_1, \alpha_2) = \sqrt{(2-6)^2 + (10-6)^2} = \sqrt{36};$

$d_{A_1,\alpha_3}(A_1, \alpha_3) = \sqrt{(2-1.5)^2 + (10-3.5)^2} = \sqrt{42.5};$

d_{A_1,α_1}最小，即$A_1 \in$ cluster 1;

$d_{A_2,\alpha_1}(A_2, \alpha_1) = \sqrt{(2-2)^2 + (5-10)^2} = \sqrt{25};$

$d_{A_2,\alpha_2}(A_2, \alpha_2) = \sqrt{(2-6)^2 + (5-6)^2} = \sqrt{17};$

$d_{A_2,\alpha_3}(A_2, \alpha_3) = \sqrt{(2-1.5)^2 + (5-3.5)^2} = \sqrt{2.5};$

d_{A_2,α_3}最小，即$A_2 \in$ cluster 3;

$d_{A_3,\alpha_1}(A_3, \alpha_1) = \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{72};$

$d_{A_3,\alpha_2}(A_3, \alpha_2) = \sqrt{(8-6)^2 + (4-6)^2} = \sqrt{8};$

$d_{A_3,\alpha_3}(A_3, \alpha_3) = \sqrt{(8-1.5)^2 + (4-3.5)^2} = \sqrt{42.5};$

d_{A_3,α_2}最小，即$A_3 \in$ cluster 2;

$d_{A_4,\alpha_1}(A_4, \alpha_1) = \sqrt{(5-2)^2 + (8-10)^2} = \sqrt{13};$

$d_{A_4,\alpha_2}(A_4, \alpha_2) = \sqrt{(5-6)^2 + (8-6)^2} = \sqrt{5};$

$d_{A_4,\alpha_3}(A_4, \alpha_3) = \sqrt{(5-1.5)^2 + (8-3.5)^2} = \sqrt{32.5};$

d_{A_4,α_2}最小，即$A_4 \in$ cluster 2;

$d_{A_5,\alpha_1}(A_5, \alpha_1) = \sqrt{(7-2)^2 + (5-10)^2} = \sqrt{50};$

$d_{A_5,\alpha_2}(A_5, \alpha_2) = \sqrt{(7-6)^2 + (5-6)^2} = \sqrt{2};$

$d_{A_5,\alpha_3}(A_5, \alpha_3) = \sqrt{(7-1.5)^2 + (5-3.5)^2} = \sqrt{32.5};$

d_{A_5,α_2}最小，即$A_5 \in$ cluster 2;

$d_{A_6,\alpha_1}(A_6, \alpha_1) = \sqrt{(6-2)^2 + (4-10)^2} = \sqrt{52};$

$d_{A_6,\alpha_2}(A_6,\alpha_2) = \sqrt{(6-6)^2+(4-6)^2} = \sqrt{4};$

$d_{A_6,\alpha_3}(A_6,\alpha_3) = \sqrt{(6-1.5)^2+(4-3.5)^2} = \sqrt{20.5};$

d_{A_6,α_3}最小，即$A_6 \in$ cluster 2;

$d_{A_7,\alpha_1}(A_7,\alpha_1) = \sqrt{(1-2)^2+(2-10)^2} = \sqrt{65};$

$d_{A_7,\alpha_2}(A_7,\alpha_2) = \sqrt{(1-6)^2+(2-6)^2} = \sqrt{41};$

$d_{A_7,\alpha_3}(A_7,\alpha_3) = \sqrt{(1-1.5)^2+(2-3.5)^2} = \sqrt{2.5};$

d_{A_7,α_3}最小，即$A_7 \in$ cluster 3;

$d_{A_8,\alpha_1}(A_8,\alpha_1) = \sqrt{(4-2)^2+(9-10)^2} = \sqrt{5};$

$d_{A_8,\alpha_2}(A_8,\alpha_2) = \sqrt{(4-6)^2+(9-6)^2} = \sqrt{13};$

$d_{A_8,\alpha_3}(A_8,\alpha_3) = \sqrt{(4-1.5)^2+(9-3.5)^2} = \sqrt{36.5};$

d_{A_8,α_3}最小，即$A_8 \in$ cluster 1;

Clustering：cluster 1 =$\{A_1, A_8\}$; cluster 2 =$\{A_3, A_4, A_5, A_6\}$; cluster 3 =$\{A_2, A_7\}$

更新权值：

$\alpha_1 = \left(\frac{2+4}{2}, \frac{10+9}{2}\right) = (3,9.5)$

$\alpha_2 = (\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}) = (6.5,5.25)$

$\alpha_3 = (\frac{2+1}{2}, \frac{5+2)}{2} = (1.5,3.5)$

第三次迭代：

$d_{A_1,\alpha_1}(A_1,\alpha_1) = \sqrt{(2-3)^2+(10-9.5)^2} = \sqrt{1.25};$

$d_{A_1,\alpha_2}(A_1,\alpha_2) = \sqrt{(2-6.5)^2+(10-5.25)^2} = \sqrt{42.812};$

$d_{A_1,\alpha_3}(A_1,\alpha_3) = \sqrt{(2-1.5)^2+(10-3.5)^2} = \sqrt{42.5};$

d_{A_1,α_1}最小，即$A_1 \in$ cluster 1;

$d_{A_2,\alpha_1}(A_2,\alpha_1) = \sqrt{(2-3)^2+(5-9.5)^2} = \sqrt{21.25};$

$d_{A_2,\alpha_2}(A_2,\alpha_2) = \sqrt{(2-6.5)^2+(5-5.25)^2} = \sqrt{20.312};$

$d_{A_2,\alpha_3}(A_2,\alpha_3) = \sqrt{(2-1.5)^2+(5-3.5)^2} = \sqrt{2.5};$

d_{A_2,α_3}最小，即$A_2 \in$ cluster 3;

$d_{A_3,\alpha_1}(A_3,\alpha_1)=\sqrt{(8-3)^2+(4-9.5)^2}=\sqrt{55.25};$

$d_{A_3,\alpha_2}(A_3,\alpha_2)=\sqrt{(8-6.5)^2+(4-5.25)^2}=\sqrt{3.8125};$

$d_{A_3,\alpha_3}(A_3,\alpha_3)=\sqrt{(8-1.5)^2+(4-3.5)^2}=\sqrt{42.5};$

d_{A_3,α_2}最小，即$A_3\in$ cluster 2;

$d_{A_4,\alpha_1}(A_4,\alpha_1)=\sqrt{(5-3)^2+(8-9.5)^2}=\sqrt{6.25};$

$d_{A_4,\alpha_2}(A_4,\alpha_2)=\sqrt{(5-6.5)^2+(8-5.25)^2}=\sqrt{9.8125};$

$d_{A_4,\alpha_3}(A_4,\alpha_3)=\sqrt{(5-1.5)^2+(8-3.5)^2}=\sqrt{32.5};$

d_{A_4,α_2}最小，即$A_4\in$ cluster 1;

$d_{A_5,\alpha_1}(A_5,\alpha_1)=\sqrt{(7-3)^2+(5-9.5)^2}=\sqrt{36.25};$

$d_{A_5,\alpha_2}(A_5,\alpha_2)=\sqrt{(7-6.5)^2+(5-5.25)^2}=\sqrt{0.3125};$

$d_{A_5,\alpha_3}(A_5,\alpha_3)=\sqrt{(7-1.5)^2+(5-3.5)^2}=\sqrt{32.5};$

d_{A_5,α_2}最小，即$A_5\in$ cluster 2;

$d_{A_6,\alpha_1}(A_6,\alpha_1)=\sqrt{(6-3)^2+(4-9.5)^2}=\sqrt{39.25};$

$d_{A_6,\alpha_2}(A_6,\alpha_2)=\sqrt{(6-6.5)^2+(4-5.25)^2}=\sqrt{1.8125};$

$d_{A_6,\alpha_3}(A_6,\alpha_3)=\sqrt{(6-1.5)^2+(4-3.5)^2}=\sqrt{20.5};$

d_{A_6,α_3}最小，即$A_6\in$ cluster 2;

$d_{A_7,\alpha_1}(A_7,\alpha_1)=\sqrt{(1-3)^2+(2-9.5)^2}=\sqrt{60.25};$

$d_{A_7,\alpha_2}(A_7,\alpha_2)=\sqrt{(1-6.5)^2+(2-5.25)^2}=\sqrt{40.812};$

$d_{A_7,\alpha_3}(A_7,\alpha_3)=\sqrt{(1-1.5)^2+(2-3.5)^2}=\sqrt{2.5};$

d_{A_7,α_3}最小，即$A_7\in$ cluster 3;

$d_{A_8,\alpha_1}(A_8,\alpha_1)=\sqrt{(4-2)^2+(9-10)^2}=\sqrt{5};$

$d_{A_8,\alpha_2}(A_8,\alpha_2)=\sqrt{(4-6)^2+(9-6)^2}=\sqrt{13};$

$d_{A_8,\alpha_3}(A_8,\alpha_3)=\sqrt{(4-1.5)^2+(9-3.5)^2}=\sqrt{36.5};$

d_{A_8,α_3}最小，即$A_8\in$ cluster 1;

Clustering：cluster 1 $=\{A_1,A_4,A_8\}$; cluster 2 $=\{A_3,A_5,A_6\}$; cluster 3 $=\{A_2,A_7\}$

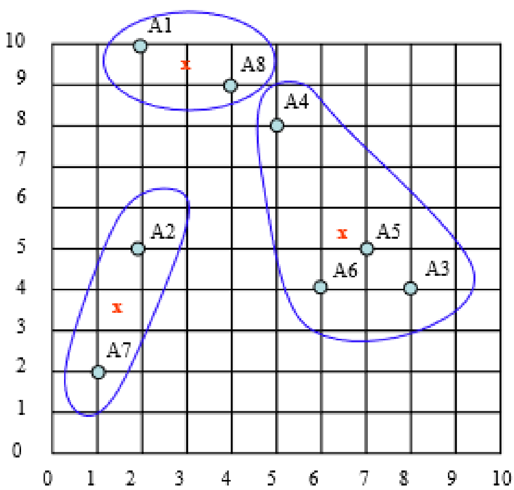
此处可不更新权值，为方便得出结论，供规律发现，这里更新一下：

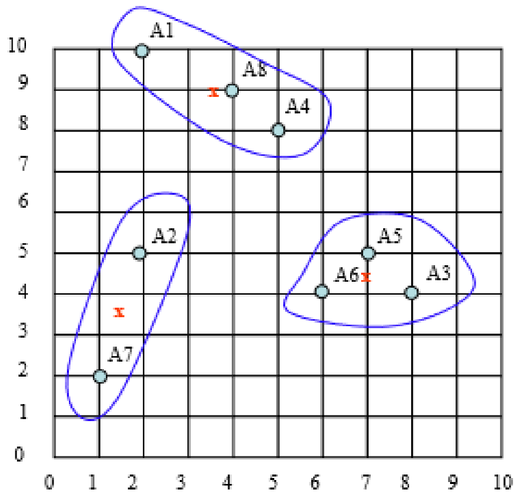
$\alpha_1 = \left(\frac{2+5+4}{3}, \frac{10+8+9}{3}\right) = (3.667, 9)$

$\alpha_2 = (\frac{8+7+6}{3}, \frac{4+5+4}{3}) = (7, 4.333)$

$\alpha_3 = (\frac{2+1}{2}, \frac{5+2)}{2} = (1.5, 3.5)$

最终，分类如下，左图为第二次迭代结果，右图为第三次迭代结果，红叉代表中心点即第三次迭代产生的$\alpha_1, \alpha_2, \alpha_3$





6 KNN 实例

假设数据集空间$D = \{A_1(2,10), A_2(2,5), A_3(8,4), A_4(5,8), A_5(7,5), A_6(6,4), A_7(1,2), A_8(4,9)\}$
请用 KNN 算法分类。

初始化

任意选取一个点，即 k=1，则 cluster 1 = A_1；距离阈值为 t = 4，即小于距离 4 则在簇附近;

遍历样本

计算A_1到A_1的距离

$d_{A_1,A_1}(A_1, A_1) = \sqrt{(2-2)^2 + (10-10)^2} = \sqrt{0};$

$d_{A_1,A_1} <$ t; 则$A_1 \in$ cluster 1;

计算A_2到A_1的距离

$d_{A_2,A_1}(A_2, A_1) = \sqrt{(2-2)^2 + (5-10)^2} = \sqrt{25};$

$d_{A_2,A_1} >$ t; 则$A_2 \in$ cluster 2;

计算A_3到A_1, A_2的距离

$d_{A_3,A_1}(A_3, A_1) = \sqrt{(8-2)^2 + (4-10)^2} = \sqrt{72};$

$d_{A_3,A_2}(A_3, A_2) = \sqrt{(8-2)^2 + (4-5)^2} = \sqrt{37};$

$d_{A_3,A_2} > \mathrm{t}$; 则$A_3 \in$ cluster 3;
计算A_4到A_1, A_2, A_3的距离

$$d_{A_4,A_1}(A_4, A_1) = \sqrt{(5-2)^2 + (8-10)^2} = \sqrt{13} = 3.6056;$$

$$d_{A_4,A_2}(A_4, A_2) = \sqrt{(5-2)^2 + (8-5)^2} = \sqrt{18};$$

$$d_{A_4,A_3}(A_4, A_3) = \sqrt{(5-8)^2 + (8-4)^2} = \sqrt{25};$$

$d_{A_4,A_1} < \mathrm{t}$; 则$A_4 \in$ cluster 1;
计算A_5到A_1, A_2, A_3, A_4的距离

$$d_{A_5,A_1}(A_5, A_1) = \sqrt{(7-2)^2 + (5-10)^2} = \sqrt{50};$$

$$d_{A_5,A_2}(A_5, A_2) = \sqrt{(7-2)^2 + (5-5)^2} = \sqrt{25};$$

$$d_{A_5,A_3}(A_5, A_3) = \sqrt{(7-8)^2 + (5-4)^2} = \sqrt{2};$$

$$d_{A_5,A_4}(A_5, A_4) = \sqrt{(7-5)^2 + (5-8)^2} = \sqrt{13};$$

$d_{A_5,A_3} < \mathrm{t}$；即$A_5 \in$ cluster 3（因A_5距离A_3最近）；
计算A_6到A_1, A_2, A_3, A_4, A_5的距离

$$d_{A_6,A_1}(A_6, A_1) = \sqrt{(6-2)^2 + (4-10)^2} = \sqrt{52};$$

$$d_{A_6,A_2}(A_6, A_2) = \sqrt{(6-2)^2 + (4-5)^2} = \sqrt{17};$$

$$d_{A_6,A_3}(A_6, A_3) = \sqrt{(6-8)^2 + (4-4)^2} = \sqrt{4};$$

$$d_{A_6,A_4}(A_6, A_4) = \sqrt{(6-5)^2 + (4-8)^2} = \sqrt{17};$$

$$d_{A_6,A_5}(A_6, A_5) = \sqrt{(6-7)^2 + (4-5)^2} = \sqrt{2};$$

$d_{A_6,A_5} < \mathrm{t}$；即$A_6 \in$ cluster 3（因A_6距离A_5最近）；
计算A_7到$A_1, A_2, A_3, A_4, A_5, A_6$的距离

$$d_{A_7,A_1}(A_7, A_1) = \sqrt{(1-2)^2 + (2-10)^2} = \sqrt{65};$$

$$d_{A_7,A_2}(A_7, A_2) = \sqrt{(1-2)^2 + (2-5)^2} = \sqrt{10};$$

$$d_{A_7,A_3}(A_7, A_3) = \sqrt{(1-8)^2 + (2-4)^2} = \sqrt{53};$$

$$d_{A_7,A_4}(A_7, A_4) = \sqrt{(1-5)^2 + (2-8)^2} = \sqrt{52};$$

$$d_{A_7,A_5}(A_7, A_5) = \sqrt{(1-7)^2 + (2-5)^2} = \sqrt{45};$$

$$d_{A_7,A_6}(A_7, A_6) = \sqrt{(1-6)^2 + (2-4)^2} = \sqrt{29};$$

$d_{A_7,A_2} < \text{t}$；即$A_7 \in$ cluster 2（因A_7距离A_2最近）；
计算A_8到$A_1, A_2, A_3, A_4, A_5, A_6, A_7$的距离

$$d_{A_8,A_1}(A_8, A_1) = \sqrt{(4-2)^2 + (9-10)^2} = \sqrt{5};$$

$$d_{A_8,A_2}(A_8, A_2) = \sqrt{(4-2)^2 + (9-5)^2} = \sqrt{20};$$

$$d_{A_8,A_3}(A_8, A_3) = \sqrt{(4-8)^2 + (9-4)^2} = \sqrt{41};$$

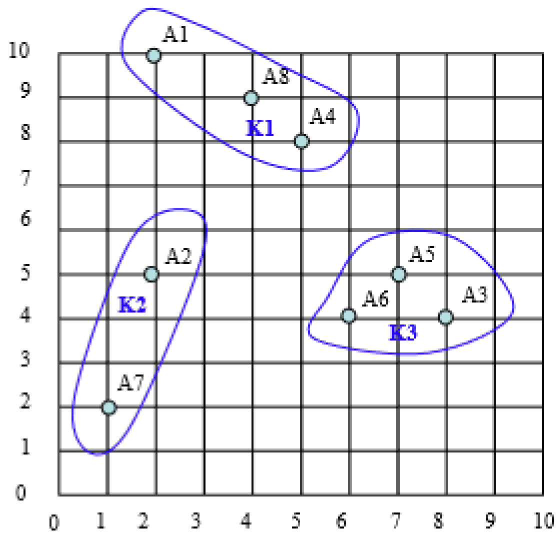
$$d_{A_8,A_4}(A_8, A_4) = \sqrt{(4-5)^2 + (9-8)^2} = \sqrt{2};$$

$$d_{A_8,A_5}(A_8, A_5) = \sqrt{(4-7)^2 + (9-5)^2} = \sqrt{25};$$

$$d_{A_8,A_6}(A_8, A_6) = \sqrt{(4-6)^2 + (9-4)^2} = \sqrt{29};$$

$$d_{A_8,A_7}(A_8, A_7) = \sqrt{(4-1)^2 + (9-2)^2} = \sqrt{58};$$

$d_{A_8,A_4} < \text{t}$；即$A_8 \in$ cluster 1（因A_8距离A_4最近）；
Clustering：cluster 1 =$\{A_1, A_4, A_8\}$; cluster 2 =$\{A_3, A_5, A_6\}$; cluster 3 =$\{A_2, A_7\}$



注意，分类结果同 k-means 的一样，且 k=3。上面这个阈值 t 很关键，一下解决了 KNN 的分类。根据西瓜书和各大 CSDN 博主资料显示，阈值的确定可以用 k-折交叉验证的方法确定。后续要查看下，网上多为调库，黑箱操作，看不到计算流程。

总结

聚类是机器学习中“新算法”出现最多，最快的领域，且常被用来异常检测。原因在于不存在客观标准，给定数据集，总能从某个角度找到以往算法未覆盖的某种标准，如密度，概率，距离等设计出新算法。相对于机器学习其他分类算法，聚类算法知识体系不够系统化，有待进一步的发展和总结。