聚类分析

相似性测度

1. 欧式距离

$$D(X_i,X_j)=|X_i-X_j|=\sqrt{(X_i-X_j)^T(X_i-X_j)}=\sqrt{(x_{i1}-x_{j1})^2\cdots(x_{in}-x_{jn})^2}$$

2. 马氏距离

设X为模式向量，M为某类模式的均值向量，C为该类模式总体的协方差矩阵。
$D^2=(X-M)^TC^{-1}(X-M)$

> 欧式距离类似于两个点之间的直线距离，而马氏距离更类似于等高线；欧式距离相等代表集合距离上的相等，马氏距离的相等代表处于同一等高线，也就是同一分布水平上。

3. 明氏距离

Xi，Xj是n维向量，明氏距离Dm表示为：
$D_m(X_i,X_j)=[\sum_{k=1}^{n}|x_{ik}-x_{jk}|^m]^{\frac{1}{m}}$

m=2时，明氏距离即为欧式距离；
m=1时，"街坊距离"

4. 汉明距离

如果模式向量各分量的值仅取1或-1，即为二值模式，可用汉明距离衡量模式间的相似性。
设Xi,Xj为n维二值模式向量，Xi，Xj之间的汉明距离定义为：
$D_h(X_i,X_j)=\frac{1}{2}(n-\sum_{k=1}^{n}x_{ik}\cdot x_{jk})$

两个模式向量取值不同的分量数即为汉明距离。

5. 角度相似性函数

模式向量Xi,Xj之间夹角的余弦，也是Xi的单位向量与Xj的单位向量之间的点积。
$S(X_i,X_j)=\frac{X_i^TX_j}{|X_i|\cdot|X_j|}$

角度相似性函数反映了几何上相似性的特征，他对于坐标系的旋转及放大缩小是不变的，但对位移和一般的线性变换不具有不变性的
在0，1的二值情况下，S等于Xi，Xj两向量共有的特征数目。

聚类准则

1.
阈值准则：根据规定的距离阈值进行分类的准则。

2.
函数准则：模式类别之间的相似性或差一行可用一个函数来表示。在聚类分析，表示模式类间的相似性或差异性的函数成为聚类准则函数。

一个常见的指标是无差评纺织和，适用于各类样本密集且数目相差不多，而不同类间的样本有明显分开的情况。

基于距离阈值的聚类算法

近邻聚类法

对于每个点，与每一个聚类中心计算欧式距离，若大于距离阈值，则作为新的聚类中心，否则，选择欧式距离最短的聚类，加入其中。

> 1. 近邻聚类法的聚类结果很大程度上依赖于第一个聚类中心的位置选择，待分类模式样本的排列次序，距离阈值的大小以及样本分布的几何性质等
> 2. 需要用先验知识知道阈值T和起始点Z1的选择。

最大最小距离算法

任选一个样本作为第一个聚类中心Z1，选择距离Z1最远的样本作为第二聚类中心Z2，诸逐个计算每个模式样本与以确定的所有聚类中心之间的距离，并选出其中的最小距离。在所有最小距离中选出最大距离，若改制达到Z1，Z2之间距离的一定壁纸以上，则将该样本定义为新的聚类中心，继续计算其他点直至确定所有的聚类中心。确定所有聚类中心之后，将其他店按照最短距离进行分配到其他聚类。

层次聚类法

N个样本初始自成一类，计算出各类之间的距离，得到NxN维的距离矩阵D。找出其中最小元素，将对应的两类合并成一类。重新计算新的距离矩阵D。当所有距离都笑大于阈值时，停止聚类。

不同类间距离计算准则

1. 最短距离法：两个聚类间所有样本的最短距离
2. 最长距离发：两个聚类间所有样本的最短距离
3. 中间距离法：若K类是由I类和J类合并而成，则H类与K类之间的距离为公式一
4. 重心法：考虑到每一类包含样本数目，改进的中间距离法为公式二
5. 类平均法：H类和K类间的距离公式定义为公式三，若K类有I类和J类合并而产生，则可以产生H和K类之间距离的递推式，为公式

公式一：$$D_{HK}=\sqrt{\frac{1}{2}D_{HI}^2\frac{1}{2}D_{HJ}^2-\frac{1}{4}D_{IJ}^2}$$

公式二：$$D_{HK}=\sqrt{\frac{n_I}{n_In_J}D_{HI}^2\frac{n_J}{n_In_J}D_{HJ}^2-\frac{n_In_J}{n_In_J}D_{IJ}^2}$$

公式三：$$D_{HK}=\sqrt{\frac{1}{n_Hn_K}\sum_{i\in H,j\in K}d_{ij}^2}$$ 公式四：$$D_{HK}=\sqrt{\frac{n_I}{n_In_J}D_{HI}^2\frac{n_J}{n_In_J}D_{HJ}^2}$$