九、聚类

1、K-means

k-means评价的标准，同一类别内的距离小，不同类别间的距离大

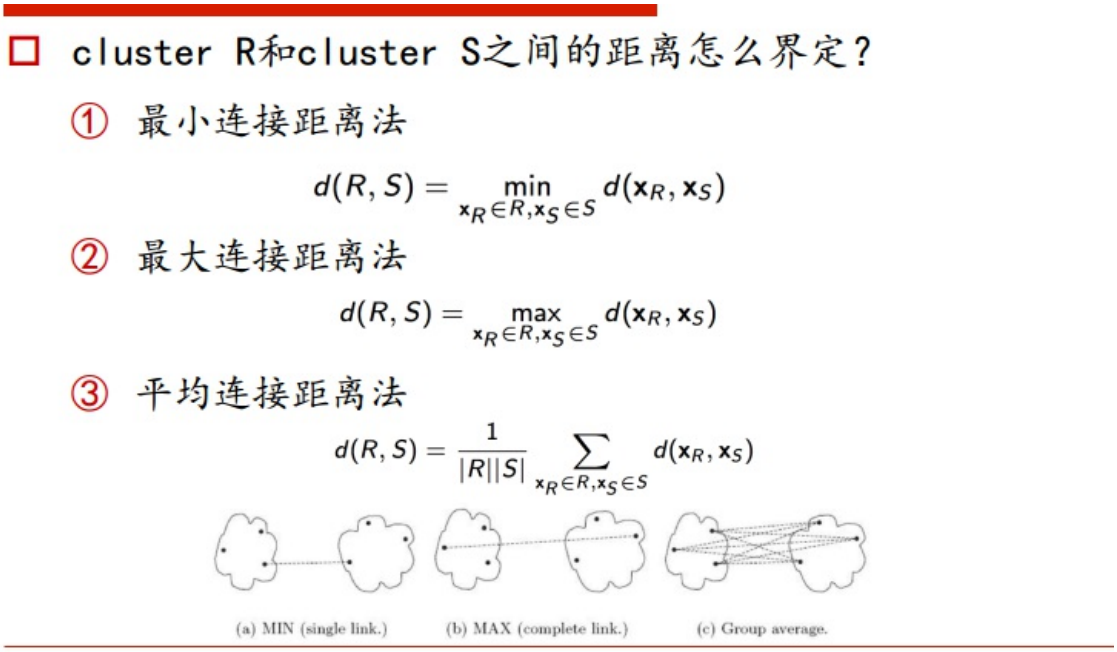
k-means收敛的标准：（1）聚类中心不再变化（2）内部点距离聚类中心的距离和不再变化

k-means和k-means++区别：初始化不同，k-means随机选，k-means++选取离得远的

k-means的缺点：对于异常点敏感，不容易收敛

2、层次聚类

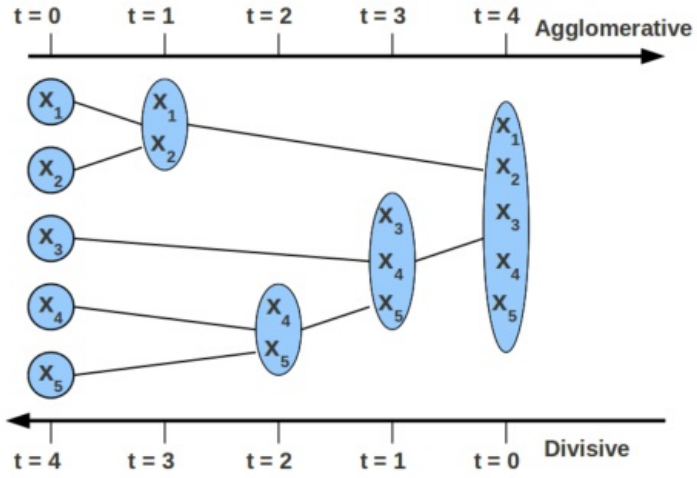
层次聚类采用的是簇中的距离



层次聚类的流程（有底到顶）

1、每次计算出距离最近的两个簇，融合一起，形成新的簇

2、repeat 1,直到不能再分



3、高斯混合模型

多个高斯分布以合适的概率分布，每个高斯模型都有自己的参数，均值，高斯混合模型需要得到每个样本点属于所有高斯分布的概率值。高斯混合模型采用EM算法求解

高斯混合模型

GMM+EM = "Soft K-means"

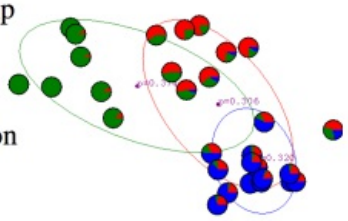
- Decide the number of clusters, K
- Initialize parameters (randomly)
- E-step: assign *probabilistic* membership

$$p_{ij} = P(C=i|\mathbf{x}_j) = \alpha P(\mathbf{x}_j|C=i)P(C=i)$$
$$p_i = \textstyle\sum_j p_{ij}.$$

- M-step: re-estimate parameters based on *probabilistic* membership

$$\boldsymbol{\mu}_i \leftarrow \sum_j p_{ij}\mathbf{x}_j / p_i$$
$$\boldsymbol{\Sigma}_i \leftarrow \sum_j p_{ij}\mathbf{x}_j\mathbf{x}_j^\top / p_i$$
$$w_i \leftarrow p_i\ .$$

- Repeat until change in parameters are smaller than a threshold



See R&N for details