

模糊聚类 —— FCM
FC (模糊 C 均值聚类)

硬聚类 HC : 将数据点准确地划分到某一聚类中.

e.g. K-means

模糊聚类 : 数据点可能归属不止一个聚类.

通过一个 成员 将数据点与聚类水平联系起来.

K-means 聚类分析

(即 K 均值聚类分析)

思想：利用距离远近的思想将目标数据聚为指定的 K 个簇，进而使得样本的簇内差异小，簇间差异大。

聚类模型特征：无监督模型。

定义：有、无监督模型（算法）取决于数据集中是否包含 因变量 y

又称标签变量

Steps：此时已选好 k 值

选 k ① 从数据集中随机选 k 个点作为原始簇中心 \Rightarrow prepare for the E-step

E-step：聚类 ② 计算其余点与簇中心距离，并将每个样本归到距离最近簇中心的类别

M-step：均值 ③ 计算各簇中心所包含同类样本点的均值，以此为新的中心，回到②直到簇中心变化趋于稳定。

④ 得到最终 k 个簇

Why the steps is correct \Rightarrow 原理

聚类目的为使得样本的簇内差异小，簇间差异大，因此为满足①：定义

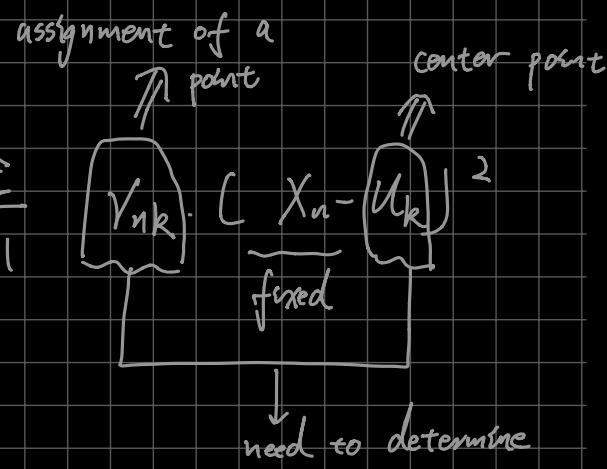
$\sum_{j=1}^k \sum_{i=1}^{n_j} r_{ji} (x_i - c_j)^2$ 为簇内离差平方和：
K-means 聚类的目标函数为簇内离差平方和：

$$\sum_{j=1}^k \sum_{i=1}^{n_j} r_{ji} (x_i - c_j)^2$$

第 j 簇
 n_j 为第 j 簇样本数
 c_j 为第 j 簇中心
簇内离差平方和

目标: 使 $J(c_1, c_2 \dots c_k)$ 最小 \Rightarrow

$$\text{minimize } J = \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \cdot \|x_n - u_k\|^2$$



How to minimize J ? \Rightarrow EM algorithm

E step: known $u_1, u_2 \dots u_k$, need to calculate

γ_{nk} (this step is to assign points to clusters) to minimize J

$$\frac{(x_n - u_i)^2 < (x_n - u_j)^2}{\dots} \Rightarrow \gamma_{nk} = \begin{cases} 1, & k = \arg \min_j \cdot (x_n - u_j)^2 \\ 0, & \text{otherwise} \end{cases}$$

M step: known γ_{nk} , need to calculate $u_1, u_2 \dots u_k$

$$\text{to minimize } J \quad \frac{(u - x_1)^2 + (u - x_2)^2 + (u - x_3)^2}{\text{最小值当且仅当 } u = \bar{x}} \Rightarrow u_k = \frac{\sum_{n=1}^N x_n \cdot \gamma_{nk}}{\sum_{n=1}^N \gamma_{nk}}$$

为何EM algorithm是
对的此后果证明↑

因为 step ②, ③ 为 E-step, M-step, 根据 EM algorithm,

重复 E-step, M-step 定会收敛, 因此 steps ②, ③, ④ are correct.

△ 步骤的背后的数学原理, 通过严谨的数学推导可知为什么要这样做, 为什么这样做是对的。
知其然也要之其所以然。

K值选择：

Method:

A: 捂点法 (考虑了①) Elbow Method

思想：在不同k值下计算簇内~~簇内~~离差平方和，通过可视化
即目标函数
方法找拐点对应k值。

原理：拐点为突然变化点，可认为此后随k增加，聚类
效果不会有大变化。

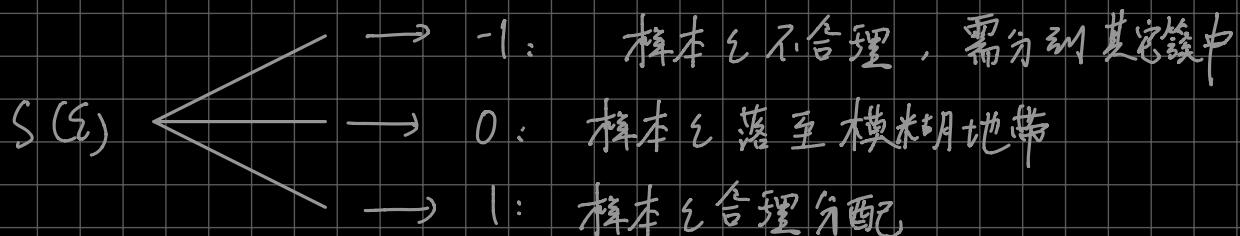
J的值，J随k↑一定是↓。

B: 轮廓系数法 (考虑了①. ②)

轮廓系数： $S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$

$a(i)$: 簇内均值 \Rightarrow 考虑了①

$b(i)$: 簇间均值最小值 \Rightarrow 考虑了②



初始值的选择：Genetic Algorithm

层次聚类 (Hierarchical clustering)

Agglomerative Nesting clustering (自底向上)

Steps:

(i) Assign each point into its own group

(ii) Merge two closest groups into one group

(iii) Repeat ii until reach stop condition

Dissolve Analysis clustering (自顶向下)

Steps:

(i) Assign each point into one group

(ii) Divide the loosest group into two group.

(iii) Repeat ii until reach stop condition

How to calculate the distance between group of points?

— Single linkage: $d(A, B) = \min_{a \in A, b \in B} d(a, b)$

— Complete linkage $d(A, B) = \max_{a \in A, b \in B} d(a, b)$

— Average linkage $d(A, B) = \frac{\sum_{a \in A, b \in B} d(a, b)}{|A| |B|}$

\Rightarrow use linkage to transform: calculate $d(A, B)$

into calculate $d(a, b)$, $a \in A, b \in B$

distance function, e.g. Euclidean · Manhattan ·

△ 终止条件由用户提供。(聚类单个数量 阈值)

△ Hierarchical clustering has a high time & space complexity -
so when the data is big we don't use it

△ Hierarchical clustering 忽略 { 不同簇间信息 簇间互连性 } 因此只能处理非静态模型

△ 优缺点分析：

优：没有使用准则函数，对数据结构假设更少 \Rightarrow 通用性更强

缺：无法对已做的分解合并进行调整，因此当分裂、合并点抉择困难时。
抉择的方法十分关键。

GMM (Gaussian Mixed Model)

K-means 缺点：

- 对数据结构做假设，要求簇的形状必须为圆形。若实际数据分布不为圆形（椭圆·月牙型）则结果误差大
- 样本属于某个簇的概率是定性的，不输出概率值，因此鲁棒性差

Solve \Rightarrow GMM，为 K-means 模型的一个优化。它并不使用硬截断进行类别分离，而是使用高斯平滑模型估计 (solve 1)，而且输出概率值 (solve 2)

原理：EM algorithm

E step：为每个点计算由 mode(内每个分量生成的概率，即已知 \vec{u} , Σ , π)
 ↓
 贝叶斯定理
 后验概率

计算 assignment $r(z_{nk})$ ($n=1, 2 \dots N, k=1, 2 \dots K$)

$$\text{soft assignment } r(z_{nk}) = \frac{\pi_k \cdot N(x_n | u_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_n | u_j, \Sigma_j)}$$

mixed coefficient

M step：调整 model params，最大化 model 生成这些 params 可能性 \Rightarrow 对数似然函数

$$L = \log \prod_{j=1}^m p(x) = \sum_{j=1}^m \log p(x)$$

即已知 assignments, 更新 π , \vec{u} , Σ

$$\pi_k = \frac{N_k}{N};$$

$$u_k = \frac{1}{N_k} \cdot \sum_{n=1}^N r(z_{nk}) \cdot x_n$$

$$\Sigma_k = \frac{1}{N_k} \cdot \sum_{n=1}^N r(z_{nk}) \cdot (x_n - u_k) \cdot (x_n - u_k)^T$$

△ EM algorithm 保证该过程中参数 总会收敛到局部最优解

优势: ① More general \Rightarrow

若允许使用全部协方差类型, GMM 理论上 可拟合任意分布
实际上, $k \geq 10$ 即可

② More Robust \Rightarrow

使用 Soft assignment 而非 Hard assignment, 可得样本的概率值.

缺点: ① 计算量大 (相比 K-means)

② 基于 EM algorithm, 容易入局部最优 (与初始化有关)

③ 当每个高斯混合 model 无足够多 samples 时, 估算协方差很困难
不适用于样本数少的情况.

④ 无法自动确定 K 值.

模糊 C 均值 (FCM) 聚类：

聚类中心计算：

$$C_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$$

w_k : membership values (隶属度)

m : fuzzier values (模糊值)

$m \uparrow$, 聚类越模糊

目标函数：

$$\arg \min_C$$

$$\sum_{i=1}^n \sum_{j=1}^c w_{ij}^m \|x_i - C_j\|^2$$

与 k-means 的区别

结果： U_{cxn} 和 C ，根据 U_{cxn} 可确定
隶属度矩阵 c 个聚类中心矩阵

每个样本应被归入什么类别 (对每列向量找

$C_j = \max_{i=1}^c (u_{ij})$ 作为样本 j 的类别。

优点：无多重迭代反复计算，计算量少，效率↑↑.

缺点：① m 值选取只能靠经验（通常取 2 ）或实验证得

② 初值矩阵 U_{cxn} 设置很重要，对初始化敏感

③ 容易陷入局部极小值。