1、k-means 算法步骤

（1）初始划分 k 个聚类，N_i为第 i 个聚类 Γ_i 中的样本数目,i=1,2,...,k,利用

$$m_i = \frac{1}{N_i}\sum_{y\in\Gamma_i} y$$

与

$$J_e = \sum_{i=1}^{k}\sum_{y\in\Gamma_i} \| y - m_i \|^2$$

计算出每类的均值和总的误差平方和，其中 y，m 均为向量

（2）任取一个样本 y，设 $y\in\Gamma_i$

（3）若 N_i=1，则转（2）；否则继续；

（4）计算（注：ρ_i为将 y 从 Γ_i移到 Γ_j后，Γ_i的误差平方和减小量，相应的，ρ_j为 Γ_j 误差平方和增量）

$$\rho_j = \frac{N_j}{N_j+1}\| y - m_j \|^2, j \neq i$$

$$\rho_i = \frac{N_i}{N_i-1}\| y - m_i \|^2$$

（5）考查 ρ_j中的最小者 ρ_s，若 $\rho_s < \rho_i$,，则把 y 从 Γ_i移到 Γ_s中

（6）重新计算 m_i，i=1,2,...,k 与 J_e

（7）若连续迭代 N 次（N 为总样本数），J_e不改变，则停止；否则转（2）

2、微信文章中基于业绩持续性的基金聚类分析步骤（H-L 方法）

目标：为了识别前期业绩较好且具有业绩持续性的基金

（1）选取 N 支目标基金在一段时间内的累计净值，离散为 t 个评价期间

（2）计算基金 i 在第 j 个评价期间的平均收益率（%）

（3）计算在每个评价期间所有基金收益率的平均值（%）**（该文中的实证例子算的是每支基金在 t 个评价期间的收益率平均值，按照《基于业绩持续性的证券投资基金聚类与实证研究》这篇参考文献，应该是作者理解错误了）**

（4）比较每支基金在每个评价期间收益率与该期间平均收益率的大小，若大于平均收益率则定义为 H，反之定义为 L，从而确定基金持续性标度

（5）定义基金初始业绩持续指数=H 的频数/t

（6）将初始业绩指数赋予不同的权重，权重取为连续出现 H 的个数。（因为显然当基金业绩连续的高于平均时，它的持续性应当更好）

（7）将基金初始业绩持续指数乘以权重得到最终业绩持续指数

（8）对最终业绩持续指数进行聚类（该文未说明聚类方法，默认 k-means）

（9）根据聚类结果可以将目标基金划分为若干等分，可分别按低持续性、中持续性、高持续性等分类。

3、参考文献中的拓展延伸

微信文章中只分析了一个指标,即收益率,而一般的基金绩效聚类分析需要考虑到更多指标。因此指标的选取成为聚类分析结果是否可靠的决定因素。

《聚类分析在开放式基金绩效研究中的应用》一文选取了：上市时间、夏普指数、詹森指

数、特雷诺指数、单位净值、基金变动率共 6 个指标。为消除量纲影响，数据应先做标准化处理。此文应用的聚类分析方法是**系统聚类**（假设有 n 个样本，第一步将每个样本聚成一类，共 n 类；第二步将最近的两类（欧氏距离）进一步聚成一类，共 n-1 类，如此下去直到只有一类，之后通过谱系图直观反映出聚类过程），利用 SPSS 实现。

《开放式基金投资评判中的聚类分析》一文选取了：累计净值、基金规模、收益率、詹森指数、特雷诺指数、夏普比率共 6 个指标，同样利用 $Y' = \frac{Y - \overline{Y}}{S.D.}$ 进行标准化处理消除量纲影响。再利用 SPSS 进行**系统聚类**得到聚类谱系图，之后人工分为若干类（文中给出 3 类），再计算每类的六项指标，分析每类的特点，判断分类是否具有基金评价指导作用。

《聚类分析在证券投资中的应用》一文给出了利用欧氏距离定义系统聚类中任意两类距离的定义（类平均法）：给出两个类 G_p、G_q 之间距离的平方如下

$$D^2(p,q) = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}^2$$

其中 d_{ij} 取欧氏距离，n_p、n_q 为两类样本个数。

该文主要讨论对**股票**的综合评价，选取了 17 个具有代表性的指标体系：行业每股收益、行业净资产收益率、行业主营收入增长率、总资产利润率、净资产利润率、主营业务收益率、每股收益、资产负债率、流动比率、速动比率、总资产周转率、主营收入增长率、净利润增长率、流通股股本、每股净资产、每股公积金、每股未分配利润

实证：随机选取 A 股市场非 ST、PT 股 50 支。同样先进行标准差标准化，再进行指标同趋化处理（逆指标正向化、适度指标正向化），再利用 SPSS 进行聚类分析