一、k-means聚类算法原理与一些性质

聚类问题从本质上来说是将包含若干元素的集合按某一准则划分成若干个互不相交的集合的并集，该准则常常用一个函数来定义，称为目标函数。我们优化（聚类）的目标往往是极大化或者极小化该目标函数。我们将它进行数学表述：

给定一个集合：

$$J_M = \{x_1, x_2 \dots x_M\}$$

我们在 J_M 的非空子集定义一个集函数 $\rho(\omega)$ ，并通过该函数构造一个目标函数，通过优化该目标函数实现聚类。首先需要确定一个集合的划分方式：即将集合 J_M 分解为 k 个互不相交的非空集合 ω_i 的并集，使得

$$\bigcup_{i=1}^{k} \omega_i = J_M$$

$$\omega_i \bigcap \omega_j = \phi \ \ (i \neq j)$$

使得目标函数最大（或最小），即

$$minimize(or\ maximize) \sum_{i=1}^{k} J(\omega_i)$$

其中 k 是一个超参数，需要人工设定。

考虑向量聚类问题：

在该问题中， J_M 中的每个元素都是一个 n 维向量，将第 t 个元素记为 x_t ：$x_t = (x_1(t), x_2(t) \dots x_n(t))$ 。对于集合 J_M 的一个非空子集 ω_i ，定义：

$$\rho(\omega_i) = \sum_{x_t \in \omega_i} (x_t - e_{w_i})'(x_t - e_{w_i})$$

该求和的意义是对 ω_i 中的所有元素进行求和， e_{ω_i} 是 ω_i 中的所有元素的中心，即

$$e_{\omega_i} = \frac{1}{|w_i|} \sum_{x_t \in \omega_i} x_t$$

其中 $|\omega_i|$ 表示集合 ω_i 中的元素个数。

对于集合 J_M 的一个给定划分，可以得到

$$\sum_{i=1}^{k} \rho(\omega_i)$$

我们将它定义为向量聚类的目标函数，对它进行极小化。不难得到该目标函数的等价形式：

$$\sum_{i=1}^{k} \rho(\omega_i) = \sum_{i=1}^{n} (x_i - x^*)'(x_i - x^*)$$

其中 x^* 是 x_i 所在类的中心。

极小化该目标函数的意义是使得同一类向量到它的中心位置的距离最近。该问题由于复杂度较高，无法直接求出目标函数的最小值，所以只能从某一种分类出发，逐步进行迭代，使得目标函数朝着改善（即目标函数值减小）的方向前进，这就是著名的 $k-means$ 聚类算法。采用以下步骤进行：

1、任选 k 个数据点 $c_1, c_2 \dots c_k$ 作为初始聚类中心（ k 是一个超参数，是最终聚成的类别数，需要人为设定）。

2、计算每个数据到 k 个中心的距离，并将数据划分到最近距离所在的类。这样就得到了集合 的一种划分方式，重新计算每一类的中心位置 $c_1^{'}, c_2^{'} \dots c_k^{'}$ ，这样就得到了目标函数的一个取值，即

$$\sum_{i=1}^{k} \rho(\omega_i) = \sum_{i=1}^{n} (x_i - x^*)'(x_i - x^*)$$

x^* 是元素 x_i 所在的类的中心，即 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 中的某一个。

3、计算每个数据到 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 这 k 个中心的距离，并将数据划分到最近距离所在的类。这样就得到了集合 J_M 的又一种划分方式，这样可以计算

$$\sum_{i=1}^{n} (x_i - x^{'})'(x_i - x^{'})$$

的值，其中 x_i 是集合 J_M 中的元素， x' 仍然是 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 中的某一个，与 x_i 运算的 x' 是根据步骤3中的新划分方式确定的，是 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 距离 x_i 最近的一个。由于 x^* 是 k 个中心的某一个，而 x' 是 k 个中心中距 x_i 最近的一个，所以

$$(x_i - x^{'})'(x_i - x^{'}) \leq (x_i - x^*)'(x_i - x^*)$$

即

$$\sum_{i=1}^{n} (x_i - x^{'})'(x_i - x^{'}) \leq \sum_{i=1}^{n} (x_i - x^*)'(x_i - x^*) = \sum_{i=1}^{k} \rho(\omega_i)$$

但此时 $\sum_{i=1}^{n} (x_i - x^{'})'(x_i - x^{'})$ 不是该新划分下的目标函数，因为 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 不是新划分下的中心位置。

4、计算3中的新划分得到的每一类的中心位置 $c_1^{**}, c_2^{**} \dots c_k^{**}$ ，从而得到目标函数的一个新取值。我们下面说明，经过这样的一步操作之后，目标函数值不增（一般会较少，除非 $c_1^{**}, c_2^{**} \dots c_k^{**}$ 与 $c_1^{'}, c_2^{'} \dots c_k^{'}$ 完全重合）。由于

$$\sum_{i=1}^{n} (x_i - x^{'})'(x_i - x^{'}) \leq \sum_{i=1}^{n} (x_i - x^*)'(x_i - x^*)$$

现只需证明

$$\sum_{i=1}^{n} (x_i - x^{**})'(x_i - x^{**}) \leq \sum_{i=1}^{n} (x_i - x^{'})'(x_i - x^{'})$$

x^{**} 是新划分的每一类的中心位置。由于左右两边对集合 J_M 的划分方式是一样的，只是左边的 x^{**} 是每类的中心，而右边不是。证明该不等式，就转化为了证明如下定理：

定理一、 设 $x_1, x_2 \dots x_n$ 是欧式空间的 n 个向量，则 $\sum_{i=1}^{n} (x_i - x)'(x_i - x)$ 取到最小值当且仅当 x 是这 n 个向量的中心位置，即 $x = \frac{1}{n} \sum_{i=1}^{n} x_i$

证明、

$$\frac{\partial \sum_{i=1}^{n}(x_i-x)'(x_i-x)}{\partial x}=-2(x_i-x)=0$$

$x=\frac{1}{n}\sum_{i=1}^{n}x_i$ 是该函数的一个驻点。显然该目标函数是严格凸的，所以 $x=\frac{1}{n}\sum_{i=1}^{n}x_i$ 是目标函数的唯一最小值点。这就说明，只要步骤4中有某个 中心位置发生了变化，那么目标函数值严格较小。综上所述，从步骤2到步骤4，经过一次迭代后，目标函数不增，即

$$\rho(\omega_i')=\sum_{i=1}^{n}(x_i-x^{**})'(x_i-x^{**})\le\sum_{i=1}^{n}(x_i-x')'(x_i-x')\le\sum_{i=1}^{n}(x_i-x^{*})'(x_i-x^{*})=\rho(\omega_i)$$

经过这样一次处理后，目标函数的值是递减的，且只要在该过程中中心位置发生了变化，则

$$\rho(\omega_i')<\rho(\omega_i)$$

即不等号严格成立,目标函数严格减小。

5、对于得到的中心点，计算每个数据，到 k 个中心的距离，并将数据划分到最近距离所在的类。这样就得到了集合 J_M 的又一种划分方式。在该划分方式下，再计算每一类的中心位置；对于该中心位置，再通过距离进行重新划分，一直循环下去，这样的算法称为 $k-means$ 算法。

二、k-means聚类算法的收敛性证明

定理二、对于任意给定的迭代聚类中心初值（或者任意给定的一种划分方式）， $k-means$ 算法的目标函数一定会收敛。

证明、将目标函数记为 $f(T)$ ，其中 T 是对给定数据集的一种划分方式，例如划分 T_1 是将数据集划分成 $\{\omega_1,\omega_2\dots\omega_k\}$ 这 k 个互不相交的集合，则

$$f(T_1)=\sum_{i=1}^{k}\rho(\omega_i)=\sum_{j=1}^{n}(x-x^*)'(x-x^*)$$

显然对于任何划分方式 T ，$f(T)\ge 0$ 。对于从任意一个初始方式开始，不断进行迭代，就会得到对数据集的一列划分： $T_1,T_2,T_3\dots$ ，同时对应地得到一列目标函数值 $f(T_1),f(T_2),f(T_3)\dots$ 。由前文所述，该数列 $\{f(T_n)\}$ 单调递减且有下界 0 ，由单调有界数列的收敛定理知， $\{f(T_n)\}$ 收敛，即

$$\lim_{n\to\infty}f(T_n)$$

存在，这就意味着，随着算法迭代次数增加，目标函数一定会收敛。

定理三、随着迭代次数趋向无穷， k 个中心点必然会收敛，即对于每种确定的初值选取方式，聚类结果是唯一确定的。

若将第 k 次迭代得到的中心点的位置记为 $c_n^1,c_n^2\dots c_n^k$ 。现要证明对于任意给定的迭代聚类中心初值（或者任意给定的一种初始划分方式），对任意固定的 p $(p\in N^+,1\le p\le k)$

$\lim_{n\to\infty}c_n^p$ 存在。

证明、对于一个给定的数据集来说，元素个数是有限的，所以总的划分方式是有限的，所以 $f(T_n)$ 只有有限个不同的取值。所以，随着 n 的增大， $f(T_n)$ 不可能一直严格递减，即严格递减过程必然要中止，即存在 N ，当 $n>N$ 时， $f(T_n)$ 恒为常数。由定理一可知，当 $n>N$ 时， $c_n^1,c_n^2\dots c_n^k$ 恒为常数，不会再改变（因为中心点只要有一个改变，那么 $f(T_n)$ 必然严格递减，与 $f(T_n)$ 恒为常数矛盾）。所以对任意固定的 p $(p\in N^+,1\le p\le k)$ ，$\lim_{n\to\infty}c_n^p$ 存在。证毕。

推论、从证明过程可以看出，不仅仅可以证明对于任意一个初值给定方式，目标函数 $f(T_n)$ 与中心点 $c_n^1,c_n^2\dots c_n^k$ 都会收敛，而且可以得到更强的结论：经过有限次迭代后，目标函数值与中心点位置都会恒为常量，即该算法迭代过程不是一直无限逼近极小值点的过程，而是经过有限步逼近后，必然会严格等于极小值点，此后再进行迭代，划分方法、中心点、目标函数等都不会再改变。由于对于给定的初值，每一步的过程是完全确定的，不含随机因素。所以对于给定初值，聚类结果是唯一确定的。

注、虽然对于给定的初值，算法可以保证收敛，但是对于不同的初值选取情况，算法收敛到的结果可能是不一样的。显然，对于不同的类别数 k ，聚类结果必然不同。所以初始中心位置（或初始划分方式）与类别数 k 是该算法需要调节的超参数。