# means算法原理

K-means**算法介绍**：K-means算法是典型的基于距离的非层次聚类算法，在最**小化误差函数的基础上**将数据划分为预定的类数K，采用**距离作为相似性的评价标准**，即认为两个对象的距离越近，其相似度越大。

**算法思想**：首先，确定 k 个初始点作为分簇的中心；接着，将数据集中的每个点分配到一个簇中，即为每个点找到距离其最近的质心，并将其分配给该质心所对应的簇；然后，每个簇的中心更新为该簇所有点的平均值。再次重新分配数据集中所有的点，如果所有的点被分配的簇和之前一样，即簇的中心不会再改变，则此时的 k 个簇就是我们所需要的；如果某个点被分配的簇改变了，则分配完所有的点之后重新更新每个簇的质心，重复分配、更新操作直到所有簇的中心不再改变。

**K-means实现**：

使用的数据集为uci上的聚类算法数据集[3D Road Network](http://archive.ics.uci.edu/ml/datasets/3D+Road+Network+%28North+Jutland%2C+Denmark%29)。

1. 从n个样本数据中随机选取k个对象作为初始的聚类中心。
2. 分别计算每个样本到各个聚类中心的距离，将对象分配到距离最近的聚类中。
3. 所有对象分配完成后，重新计算k个聚类的中心。
4. 与前一次计算得到的k个聚类中心比较，如果聚类中心发生变化，转至步骤2)，否则转置步骤5)。
5. 当质心不再发生变化时，停止并输出聚类结果。