

人工智能实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 题目 | 实验六：无监督机器学习算法 | | |
|  | Kmeans算法实现 | | |
|  |  |  |  |
| 学院 |  | 年级 |  |
| 专业 |  | 学号 |  |
| 姓名 |  | | |
| 完成时间 |  | | |
| 指导老师 | 刘坤 | | |

2025 年 月 日实验题目

**1. 实验目的：**

k-均值算法（英文：k-means clustering）源于信号处理中的一种向量量化方法，现在则更多地作为一种聚类分析方法流行于数据挖掘领域。k-均值聚类的目的是：把n个点（可以是样本的一次观察或一个实例）划分到k个聚类中，使得每个点都属于离他最近的均值（此即聚类中心）对应的聚类，以之作为聚类的标准。

**2. 实验原理**

K-means聚类是典型的基于距离的聚类算法，采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。K-means聚类的基本思想是，在指定聚类个数K的情况下，从数据集中随机化选取K个个案作为起始的聚类中心点，计算其他个案所代表的点与初始聚类中心点的欧式距离，将个案分到距离聚类中心最近的那个类，所有数据个案划分类别后，形成了K个数据集(K个簇), 重新计算每个簇中数据个案的均值，将均值作为新的聚类中心。因此聚类中心处于变化中，这个过程不断重复，直到聚类中心点不再变化为止。

有两点需要注意，第一是K个初始聚类中心的选择具有随机性；第二是计算距离通常使用标准化欧式距离，不同量纲单位的聚类数据应提前进行数据标准化处理。

已知观测集，其中每个观测都是一个 d-维实向量，k-平均聚类要把这 n个观测划分到k个集合中(k≤n),使得组内平方和最小。换句话说，它的目标是找到使得下式满足的聚类，其中是的中所有点的均值。



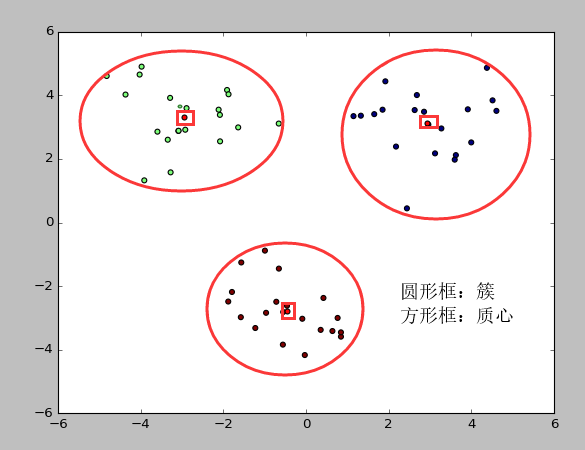
k-means术语：

簇: 所有数据的点集合，簇中的对象是相似的。

质心: 簇中所有点的中心（计算所有点的均值而来）.

SSE: Sum of Sqared Error（误差平方和）, 它被用来评估模型的好坏，SSE 值越小，表示越接近它们的质心. 聚类效果越 好。由于对误差取了平方，因此更加注重那些远离中心的点（一般为边界点或离群点）。详情见kmeans的评价标准。

有关 簇 和 质心 术语更形象的介绍, 请参考下图:



k-means应用场景

kmeans，用于数据集内种类属性不明晰，希望能够通过数据挖掘出或自动归类出有相似特点的对象的场景。其商业界的应用场景一般为挖掘出具有相似特点的潜在客户群体以便公司能够重点研究、对症下药。

例如，在2000年和2004年的美国总统大选中，候选人的得票数比较接近或者说非常接近。任一候选人得到的普选票数的最大百分比为50.7%而最小百分比为47.9% 如果1%的选民将手中的选票投向另外的候选人，那么选举结果就会截然不同。 实际上，如果妥善加以引导与吸引，少部分选民就会转换立场。尽管这类选举者占的比例较低，但当候选人的选票接近时，这些人的立场无疑会对选举结果产生非常大的影响。如何找出这类选民，以及如何在有限的预算下采取措施来吸引他们？ 答案就是聚类（Clustering)。

那么，具体如何实施呢？首先，收集用户的信息，可以同时收集用户满意或不满意的信息，这是因为任何对用户重要的内容都可能影响用户的投票结果。然后，将这些信息输入到某个聚类算法中。接着，对聚类结果中的每一个簇（最好选择最大簇 ）， 精心构造能够吸引该簇选民的消息。最后， 开展竞选活动并观察上述做法是否有效。

另一个例子就是产品部门的市场调研了。为了更好的了解自己的用户，产品部门可以采用聚类的方法得到不同特征的用户群体，然后针对不同的用户群体可以对症下药，为他们提供更加精准有效的服务。

**3. 实验步骤**

需求：Python语言，依赖库sklearn, numpy.

先随机选取K个对象作为初始的聚类中心。然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。一旦全部对象都被分配了，每个聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是以下任何一个：

没有（或最小数目）对象被重新分配给不同的聚类。

没有（或最小数目）聚类中心再发生变化。

误差平方和局部最小。

得到相互分离的球状聚类，在这些聚类中，均值点趋向收敛于聚类中心。 一般会希望得到的聚类大小大致相当，这样把每个观测都分配到离它最近的聚类中心（即均值点）就是比较正确的分配方案。

k-means工作流程：

创建 k 个点作为起始质心（通常是随机选择）

当任意一个点的簇分配结果发生改变时（不改变时算法结束）

对数据集中的每个数据点

对每个质心

计算质心与数据点之间的距离

将数据点分配到距其最近的簇

对每一个簇, 计算簇中所有点的均值并将均值作为质心

k-means开发流程

收集数据：使用任意方法

准备数据：需要数值型数据类计算距离, 也可以将标称型数据映射为二值型数据再用于距离计算

分析数据：使用任意方法

训练算法：不适用于无监督学习，即无监督学习不需要训练步骤

测试算法：应用聚类算法、观察结果.可以使用量化的误差指标如误差平方和（后面会介绍）来评价算法的结果.

使用算法：可以用于所希望的任何应用.通常情况下, 簇质心可以代表整个簇的数据来做出决策.

k-means优缺点

优点:

属于无监督学习，无须准备训练集  
原理简单，实现起来较为容易  
结果可解释性较好

缺点:

聚类数目k是一个输入参数。选择不恰当的k值可能会导致糟糕的聚类结果。这也是为什么要进行特征检查来决定数据集的聚类数目了。  
可能收敛到局部最小值, 在大规模数据集上收敛较慢  
对于异常点、离群点敏感

使用数据类型 : 数值型数据

**4. 实验算法和代码**

问题：使用K-means2算法将下列数据点进行聚类（这里使用欧式距离

作为度量，K取值为2.

数据点：P1(1,2), P2(2,3), P3(5,6), P4(7,8), P5(9,10), P6(15,14), P7(16,13), P8(18,17), P9(20,19)

解题步骤：

1.选择聚类的个数K=2。

2.任意选择两个数据点作为初始聚类中心，例如选择P1和P6。

3.计算每个数据点到两个聚类中心的距离，并将每个数据点分配到距

离最近的聚类中。

4.重新计算每个聚类的中心点，即将该聚类中所有数据点的坐标求平

均值得到新的聚类中心。

5.重复步骤3和4，直到聚类中心不再发生变化或达到最大迭代次数。

6.输出最终的聚类结果。

根据以上步骤，可以得到该数据集的聚类结果为：{P1,P2,P3,P4,P5}和{P6,P7,P8,P9}。其中，第一个聚类中心为(3,4)，第二个聚类中心为(17,16)。

**代码：**

**5. 实验结果**

**6. 实验总结**