燕山大学

Python 机器学习实验指导书

Python machine learning Experiment Instruction Book

实验四：K-Means

教 务 处

2021 年 2 月

实验四　K-Means

一、实验目的

1.理解并掌握聚类算法模型。

2.能够基于 K-Means 聚类算法模型实现鸢尾花分类、基于经纬度的城市聚类。

3.能够举一反三，基于 K-Means 聚类算法实现果汁饮料聚类分析。

二、实验原理

聚类（Clustering）是一种无监督学习(unsupervised learning)，简单地说就是把相似的对象归到同一簇中。簇内的对象越相似，聚类的效果越好。

聚类算法是无监督学习中最常见的一种，给定一组数据，需要聚类算法去挖掘数据中的隐含信息。聚类算法的应用很广：顾客行为聚类，google 新闻聚类等。

（一）K-Means 算法

K-Means 算法是最为经典的基于划分的聚簇方法，是十大经典数据挖掘算法之一。简单的说 K-Means 就是在没有任何监督信号的情况下将数据分为 K 份的一种方法。

K-Means 算法的思想：对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。

具体的算法步骤如下：

1.随机选择 K 个中心点。

2.把每个数据点分配到离它最近的中心点。

3.重新计算每类中的点到该类中心点距离的平均值。

4.分配每个数据到它最近的中心点。

5.重复步骤 3 和 4，直到所有的观测值不再被分配或是达到最大的迭代次数（R 把 10 次作为默认迭代次数）。

K-Means 算法也可以采用 Python 里的 Scikit-learn 库来实现，在 Scikit-learn 中的 cluster（聚类）程序包可以调取 KMeans()函数直接实现对 K-Means 算法的运行，无需过多的编程或者对参数的调整。

（二）鸢尾花数据集 iris

鸢尾花数据集总共包含 150 行数据。每一行数据由 4 个特征值及一个目标值组成。

4 个特征值分别为：萼片长度（SepalLengthCm）、萼片宽度（SepalWidthCm）、花瓣长度（PetalLengthCm）、花瓣宽度（PetalWidthCm）。

目标值为三种不同类别的鸢尾花，分别为：Iris-setosa（山鸢尾），Iris-versicolor（杂色鸢尾），Iris-virginica（维吉尼亚鸢尾）。

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
```

鸢尾花分类数据集

	花萼长度	花萼宽度	花瓣长度	花瓣宽度			品种（标签）
特征	5.1	3.3	1.7	0.5	model	结果	0（山鸢尾）
	5.0	2.3	3.3	1.0			1（变色鸢尾）
	6.4	2.8	5.6	2.2			2（维吉尼亚鸢尾）

鸢尾花分类

三、实验环境

计算机；网络环境。

四、实验内容及步骤

（一）实验内容

1. 对于给定的例题，基于 K-Means 聚类算法进行鸢尾花分类、基于经纬度的城市聚类等练习。

2. 对于给定的项目，自行编写程序，使用 K-Means 聚类算法实现果汁饮料聚类分析。

（二）实验步骤

1. 进入指定实验课程，可通过阅览“知识讲解”进行实验相关知识的查缺补漏。

   ➢ 知识讲解：K-Means

2. 在熟悉了实验原理的基础上，进行“实验四：K-Means”例题部分的实操练习。

   ➢ 例题 1：鸢尾花分类（K-Means）（可观看扩展实验 K-Means 讲解视频辅助理解）

   ➢ 例题 2：基于经纬度的城市聚类

3.在步骤 2 例题练习的基础上，对给定的实操项目，自行编写程序，使用 K-Means 聚类算法实现果汁饮料的聚类分析。

   ➢ 实操项目——不同含量果汁饮料的聚类

3.完成实验报告。

请严格基于实验报告的模板撰写实验报告。

本课程所有实验全部结束后再统一打印。

附：实操项目要求

➢ 实操项目——不同含量果汁饮料的聚类

某企业通过采集企业自身流水线生产的一种果汁饮料含量的数据集，来实现K-Means算法。通过聚类以判断该果汁饮料在一定标准含量偏差下的生产质量状况，对该饮料进行类别判定。

【数据集】

该数据集共有样本59个，变量2个，包括juice（该饮料的果汁含量偏差）、sweet（该饮料的糖分含量偏差），单位均为mg/ml。

所有特征变量都为与标准含量相比的偏差，该数据集没有目标类别标签变量。

【实验要求】

1.加载数据集，读取数据，探索数据。

（数据集路径：data/data76878/4_beverage.csv）

2.样本数据转化（可将pandasframe格式的数据转化为数组形式），并进行可视化（绘制散点图），观察数据的分布情况，从而可以得出k的几种可能取值。

3.针对每一种k的取值，进行如下操作：

（1）进行K-Means算法模型的配置、训练。

（2）输出相关聚类结果，并评估聚类效果。

这里可采用CH指标来对聚类有效性进行评估。在最后用每个k取值时评估的CH值进行对比，可得出k取什么值时，聚类效果更优。

注：这里缺乏外部类别信息，故采用内部准则评价指标（CH）来评估。(metrics.calinski_harabaz_score())

（3）输出各类簇标签值、各类簇中心，从而判断每类的果汁含量与糖分含量情况。

（4）聚类结果及其各类簇中心点的可视化（散点图），从而观察各类簇分布情况。（不同的类表明不同果汁饮料的果汁、糖分含量的偏差情况。）

4.【扩展】（选做）：设置 k 一定的取值范围，进行聚类并评价不同的聚类结果。

参考思路：设置 k 的取值范围；对不同取值 k 进行训；计算各对象离各类簇中心的欧氏距离，生成距离表；提取每个对象到其类簇中心的距离，并相加；依次存入距离结果；绘制不同 K 值对应的总距离值折线图。

封面设计： 贾丽

地 址：中国河北省秦皇岛市河北大街 438 号
邮 编：066004
电 话：0335-8057068
传 真：0335-8057068
网 址：http://jwc.ysu.edu.cn