廣東工業大學

QG 中期考核详细报告书

题　　目　数据挖掘中期考核

学　　院　计算机学院

专　　业　信息安全

年级班别　19 级（1）班

学　　号　3119005436

学生姓名　徐国涛

2020 年 04 月

一、K-means 算法

1、数据集的处理

数据间是以逗号隔开，所以用 pd.read_csv 进行读取，因为数据集中没有作为 column 的行，所以参数 header=None，进行 kmeans 算法时将 kmeans 所需的容器 clusterAssment 拼接到数据集之后，作为结果返回，便于观察结果和计算。

2、算法步骤和思想

思想：

对于给定的数据集和簇个数 k，先随机生成 k 个质心，计算每个点与各质心的距离，通过距离的最小值觉得该点归属于哪个簇，随后计算各簇内的所有点的均值作为新的质心，再与各点进行比较，直到簇不再发生变化或者达到最大所要求的最大迭代次数。

步骤（伪代码）：

创建 k 个点作为初始质心

    当任意一个点的簇分配结果发生改变时：

        对于数据集中的每个点：

            对每个质心：

                计算质心与数据点之间的距离

            将数据点分配到据其最近的簇

        对每个簇，计算簇中所有点的均值并将均值作为新的质心

直到簇不再发生变化或者达到最大迭代次数（本算法选取前者）

具体步骤：

随机生成 k 个质心

创建容器 clusterAssment：第一列存放点到质心最小距离的平方(SSE)，第二列存放本次计算后所归属的簇，第三列存放上一次计算后所归属的簇

初始化容器：第一列为无穷，第二三列为-1

将容器拼接到所给数据集之后得到 result_set

接下来需根据后两列数据是否完全相同（即簇不再发生变化）判断是否进行迭代，初始化 clusterChanged 为 false

进入循环部分

对每一行进行计算：

计算所定义的距离

找到最小值并存入容器第一列

将最近的质心的编号存入容器第二列

        判断最后两列是否完全相等

        如果不完全相等：

            对本次计算后的簇分组后求均值

            更新质心

            将新的质心更新到最后一列

返回质心和数据集

3、算法实现结果评估
对不同 k 值进行测试，得到 SSE 后绘图与 sklearn 库中的模型对比，曲线走向大致相同，但拐点不能明显看出来。

4、不足和优化之处
测试多个 k 值时拐点不明显。
优化：二分 kmeans

二、K 近邻算法
1、数据集的处理
数据间是以逗号隔开，所以用 pd.read_csv 进行读取，因为数据集中没有作为 column 的行，所以参数 header=None。从数据集中根据比例随机抽取数据集得到新数据集作为训练集和测试集，测试时选择比例为 8 : 2。优化：对训练集和测试集进行归一化

2、算法步骤和思想
思想：
对于给定的测试集、训练集和 k 值，计算测试集中每个点到训练集中所有点的距离，选取距离最近的 k 个点，以这 k 个点中标签出现频率最高的标签作为预测结果返回

步骤（伪代码）：
对未知类别属性的数据集中的每个点执行以下操作：
1) 计算已知类别数据集中的点与当前点之间的距离
2) 按照距离递增次序排序
3) 选取与当前点距离最小的 k 个点
4) 确定当前 k 个点所在类别的出现频率
5) 返回前 k 个点出现频率最高的类别作为当前点的预测分类

具体步骤：
1) 获取基本信息：标签、测试集和训练集的行数列数
2) 初始化存放距离的容器 dist
3) 计算测试集中各点到训练集中所有点的距离
4) 给行贴上标签
5) 对 dist 的每一行：
   进行升序排序后，选取距离最小的 k 个点
   对标签进行统计，统计频率最高的标签作为预测结果存入列表 result 中
6) 返回结果

3、算法实现结果评估
所写算法能够实现对测试集教好的预测，对不同 k 值进行比较，结果都比较稳定，准确率较高。

4、不足和优化之处

第一次实现的 KNN 算法中对数据集中各列所占权重不同，对结果有影响。

优化：对训练集和测试集分别进行归一化处理，将所有数据映射到 0-1 之间