# 实验报告五

学号：20191060239

姓名：李品鑫

实验名称：异常检测算法实现及应用

【实验目标】：

实现基于距离的嵌套循环异常检测算法，并在iris数据集上进行应用。

【实验任务】：

1. 阐述基于距离的嵌套循环算法检测异常的原理及过程。

2. 编程实现基于距离的嵌套循环异常检测算法。

3. 在iris数据集上使用基于距离的嵌套循环算法检测异常，并对检测结果进行分析，相关数据集在文件夹中给出。

4. 实验总结与思考。

【实验要求】：

（1）独立完成实验，拒绝抄袭。撰写实验报告，需包含实验目的，实验原理及过程，实验结果展示等。

（2）可根据自己能力借助合适的工具（如sklearn库）完成实验，但需清楚算法原理。鼓励自己动手编程实现算法，评分时，自己动手编程实现借助工具。

（3）附代码（使用工具完成的附操作过程）。

【实验过程】：

1. 阐述基于距离的嵌套循环算法检测异常的原理及过程。

基于距离的异常值检测方法原理是利用计算数据对象相互之间的距离，并设定一个距离阈值与支持度阈值，当某一数据点与其它数据对象的距离超出距离阈值的支持度大于阈值时，则将该点作为离群点。算法流程约如下伪码所示：

dis=X;sup=Y; //设定距离阈值dis与支持度阈值sup。X、Y为一常数

for i in data //遍历数据对象

count=0; //支持度计数

for j in data

if distance(x(i),x(j))>dis //当距离大于阈值时，支持度计数+1

count++;

if count>sup //累计支持度大于阈值时

print(“该点x(i)为离群点”);

2. 编程实现基于距离的嵌套循环异常检测算法。

针对iris数据集编程进行的异常检测算法如下：

close all;

clear all;

clc;

%读取数据，该脚本由matlab自动生成。

importdata;

n=size(iris,1);

sum=0;

%计算总距离并由此计算距离阈值。

for i = 1:n

for j = 1:n

sum=sum+dist(iris(i,2:5),iris(j,2:5)');

end

end

r = sum/(n\*(n-1));

%设定支持度阈值。

sup = 0.02;

err = [];

for i = 1:n

count=0;

for j =1:n

if and(i~=j,dist(iris(i,2:5),iris(j,2:5)') <= r)

count=count+1;

end

end

if count < sup\*n

err=cat(1,err,iris(i,1));

end

end

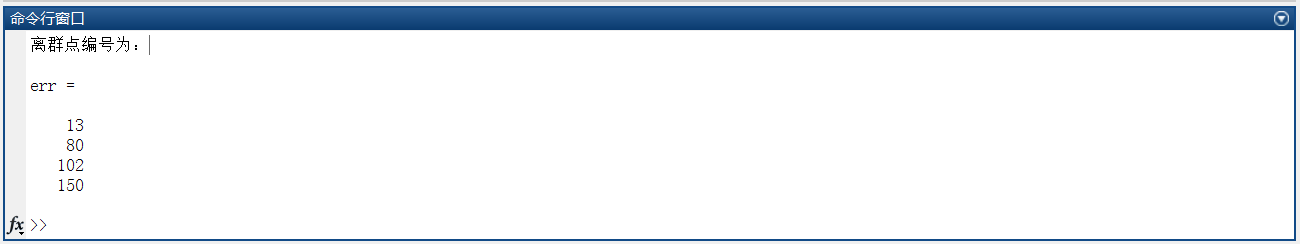
err

disp('离群点编号为：');

err

3. 在iris数据集上使用基于距离的嵌套循环算法检测异常，并对检测结果进行分析

以上代码采用了欧氏距离作为度量，运行结果如下：



为验证结果是否准确，排除欧式距离中各个维度尺度不一致且相关的影响，故对代码进行修改，在保持支持度不变、距离阈值分数不变的情况下使用matlab中自带的马氏距离函数 pdist2 进行计算。运行结果如下：



两者比对，发现结果相同。说明在iris数据集中各维度尺度对结果影响不大。

根据算法的原理，随着sup值的增大，选择出来的离群点err也就越多，为验证此想法，分别设置sup为不同的值进行多次计算，部分结果如下表：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| sup | 0.01 | 0.02 | 0.2 | 0.35 | 0.4 | 0.8 |
| err | 13, 80,  102, 150 | 13, 80,  102, 150 | 13, 80,  102, 150 | 13, 14,  15, 16,  17, 23,  33, 34,  39, 43,  80, 102,  119, 150, | 1, 2  3, 4  5, 6  7, 8  9, 10  11, 12  13 ,14  …等47个数据 | 1, 2  3, 4  5, 6  7, 8  9, 10  11, 12  13 ,14  …等149个数据 |

可以发现，除了0.2之前sup变化幅度过小导致err无变化之外，其余随着sup的增大，err所涵盖的数据点就更多。

同时发现了一个离群点检测算法的特点，就是当sup==0.4时，所选择出来的“离群点”几乎都是属于同一类鸢尾花的数据，这表明在离群点检测算法可以运用在寻找新类别中。而当sup过大时，则几乎涵盖了整个数据集，说明检测失效。

【实验总结】：

通过该次实验对基于距离的嵌套循环异常检测算法的原理与过程分析以及使用Matlab进行编写代码实操，很好地加强了对代码的理解与运用。并且在结果分析中采用了两个不同的距离函数（欧氏距离、马氏距离）进行比较，理解了欧氏距离与马氏距离他们各自的数学意义与特点。