C 题 电动汽车目标客户销售策略研究

第一问首先要对数据进行处理，第一问不能只看附录 1，要看看附录三中有什么，其实可以发现所有指标除了 B7 含有#NULL!外都比较正常，我们看看 B7 是什么问题，“B7 请问您有几个孩子？ ___个”，一般来讲#NULL!的出现也只能是未填写了，也就是 0，因此，第一步需要将所有#NULL!替换为 0。a1~8、B1~17 最终决定的是购买意愿，购买意愿分为 0/1，那么这道题肯定是会涉及到机器学习领域，且附件 1 是要用来做模型训练的，那么在这之前可以通过筛选数据降低数据集的离散度，这一步数据清洗也十分重要，会决定后文机器学习的训练效果，那我们怎么清洗数据集呢，首先不管 0/1 类的数据交叉部分，先分品牌类型分别对购买意愿 0/1 类数据通过 LOF 算法计算样本数据的离群度大小，将离群度值较大者排除即可，判断的临界值自行设定，程序中标红部分为可改进步骤

主程序

```
clear
clc
load x.txt;%这里的 x 是二维数据
%改进：减去均值 x=x-mean(x)
K=3;%设置第 K 距离
[n,m]=size(x);
x2=sum(x.^2,2);%每个点的横纵坐标平方相加
dist=[];
for i=1:n
    for j=1:n
        dist(i,j)=sqrt((x(i,1)-x(j,1)).^2+(x(i,2)-x(j,2)).^2);
    end
end
%repmat(x2,1,n)+repmat(x2',n,1)-2*x*x')这段程序其实就等价于二维点距离计算，没什么特别的，就是简化程序算的时间少
lof = LOF(dist,K);

figure;
clf;%clf 函数用于清除当前图像窗口
plot(x(:,1),x(:,2),'rx');
hold on
for i=1:n
    %这里绘制的圈大小按照 lof 值的十倍来，这样可视化后更明显
    plot(x(i,1),x(i,2),'bo','Markersize',lof(i)*10)
end
```

自定义函数：

```
%LOF 算法
function lof = LOF(dist,K)
%dist 为距离矩阵
m=size(dist,1);                    %m 为对象数，dist 为两两之间的距离
```

```
distance = zeros(m,m);
num = zeros(m,m);                    %distance 和 num 用来记录排序后的顺序，和对象编号顺序
kdistance = zeros(m,1);              %计算每个对象的 kdistance
count    = zeros(m,1);                %k 邻域的对象数
reachdist = zeros(m,m);              %计算两两之间的 reachable-distance
lrd = zeros(m,1);
lof = zeros(m,1);
%计算 k-距离
for i=1:m
    [distance(i,:),num(i,:)]=sort(dist(i,:),'ascend');
    kdistance(i)=distance(i,K+1);
    count(i) = -1;%自己的距离为 0，要去掉自己
    for j = 1:m
        if dist(i,j)<=kdistance(i)
            count(i) = count(i)+1;
        end
    end
end
for i = 1:m
    for j=1:i-1
        reachdist(i,j) = max(dist(i,j),kdistance(j));
        reachdist(j,i) = reachdist(i,j);
    end
end
for i = 1:m
    sum_reachdist=0;
    for j=1:count(i)
        sum_reachdist=sum_reachdist+reachdist(i,num(j+1));
    end
    %计算每个点的 lrd
    lrd(i)=count(i)/sum_reachdist;
end
% 得到 lof 值
for i=1:m
    sumlrd=0;
    for j=1:count(i)
        sumlrd=sumlrd+lrd(num(j+1))/lrd(i);
    end
    lof(i)=sumlrd/count(i);
    %改进：求方差 lof(i,1)=std(sumlrd);
end
end
```

删除离群度较大的样本数据后，0/1 类数据已有较为明显的划分，接来下可对进一步对交叉部分的数据进行剔除，这部分可选做，随便才有一个分类算法带入数据集训练，并测试分类效果，凡识别错误的样本均可排除。以上步骤会得到划分较为明显的数据集，接下来分别对各品牌不同购买意愿的群体的特征进行统计分析，例如愿意购买品牌 1 的群人有什么特征，其中 a1 的平均值是多少。。。。。。与不够买的人群有什么区别。。。。。。

第二问实则就是一个交叉验证实验，对不同品牌的销售影响，那么可以将不同品牌两种购买意愿分别看作是一个种类，品牌有三个，那么就有 6 种，在交叉实验中我们就可以看作是六种方案，这个问可以采用不均匀样本的单因素方差分析方法，详情见推文：

https://mp.weixin.qq.com/s/rdvmOjClKNUhD9_S-x_rqQ

采用该方法依次对每个指标进行分析，最后列出显著的指标

第三问，模型的优良性就自己吹牛逼了，第一问做的数据清洗，可以直接从处理的结果上处理后的数据划分明显，第二问最后得到的影响显著的指标也比较符合实际等等，接下来就是对附件 3 通过机器学习训练并识别了，但切记品牌 123 是单独的，不能一起训练，不严谨，包括前面第一问数据清洗，即使用的其他方法也得按品牌分开来做。第三问分类模型有很多，随机森林、pnn、grnn、bp、聚类、svm 等等都可以

第四问的意思就是说，选出一名没有购买电动汽车的客户，随便哪个品牌的都可以，分类模型还是用第三问的，针对 a1~8 中第二问求出的显著影响的指标，进行控制变量法分析，可按 10%、20%、等提高指标值，将其作为测试集带入模型进行预测识别，看针对该客户，主要的服务指标需要达到多少后最后识别出来才会是 1。

第五问是写建议，不用多说，但一定要注意前后问的连贯性，提高那些方面的服务性指标肯定和第二问有关，提高多少肯定和第 4 问有关。