# 实验报告一

学号：20191060239

姓名：李品鑫

实验名称：数据预处理与探索

实验内容：

1. 原始数据的描述：样本数、属性数、属性类型（离散、连续）等。

2. 数据预处理

1）缺失值处理：采用合适的方式对数据中的缺失值进行处理。

2）数据规范化：采用合适的方法对连续属性进行规范化处理。

3. 采用合适的度量对数据进行汇总统计（如，百分位、均值、中位数、方差等）。

4. 采用合适的工具对数据进行可视化（如直方图、盒状图、散点图等）。

5. 其他有价值的探索（自行发现，鼓励思考）。

实验要求：

独立完成实验，拒绝抄袭。撰写实验报告，需包含实验目的（对数据预处理与探索的理解，数据预处理与探索在数据挖掘中的作用），实验过程及原理（所采用方法的简要介绍），实验结果展示等。

可根据自己能力采用合适的工具完成实验，鼓励自己编程完成，使用的编程语言不限，但需要附上自己编写的代码以及必要的注释。

实验结果：

1. 原始数据的描述：样本数、属性数、属性类型（离散、连续）等。

原始数据样本数有150个，属性数有4种，属性类型为离散型。

2. 数据预处理

对属性的缺失值进行平均值替换，对分类缺失采用贝叶斯分类方法确认其类别。 数据规范化处理采用最大最小值方法：

代码如下：

close all;

clear all;

clc;

%读取数据

filename = 'iris\iris.data';

fileID = fopen(filename);

data = textscan(fileID,'%f%f%f%f%s','Delimiter',',','TreatAsEmpty',{'Null'});

fclose(fileID);

attrib = [data{1,1}, data{1,2}, data{1,3}, data{1,4}];

class = data{1,5};

setosa = attrib(1:50,1:4);

versicolor = attrib(51:100,1:4);

virginica = attrib(101:149,1:4);

u1=nanmean(setosa);u2=nanmean(versicolor);u3=nanmean(virginica);

%填充空值

for i=1:50

for j = 1:4

if isnan(setosa(i,j))

setosa(i,j) = u1(1,j);

end

if isnan(versicolor(i,j))

versicolor(i,j) = u2(1,j);

end

if i<=49

if isnan(virginica(i,j))

virginica(i,j) = u3(1,j);

end

end

end

end

%规范化

setosa=mapminmax(setosa',0,1)';

versicolor=mapminmax(versicolor',0,1)';

virginica=mapminmax(virginica',0,1)';

%贝叶斯分类器

c1=cov(setosa);c2=cov(versicolor);c3=cov(virginica);

t1=diag(c1);t2=diag(c2);t3=diag(c3);

c1=diag(t1);c2=diag(t2);c3=diag(t3);

inv\_c1=inv(c1);inv\_c2=inv(c2);inv\_c3=inv(c3);

d1=det(c1);d2=det(c2);d3=det(c3);

x\_test=[attrib(150,1), attrib(150,2), attrib(150,3), attrib(150,4)];

p1=-0.5\*(x\_test-u1)\*inv\_c1\*(x\_test-u1)'-0.5\*log(d1);

p2=-0.5\*(x\_test-u2)\*inv\_c2\*(x\_test-u2)'-0.5\*log(d2);

p3=-0.5\*(x\_test-u3)\*inv\_c3\*(x\_test-u3)'-0.5\*log(d3);

[~,max\_id]=max([p1,p2,p3]);

if max\_id == 1

fprintf('x\_test属于iris-setosa\n ');

class{150}='iris-setosa';

elseif max\_id == 2

fprintf('x\_test属于iris-versicolor\n');

class{150}='iris-versicolor';

elseif max\_id == 3

fprintf('x\_test属于iris-virginica\n');

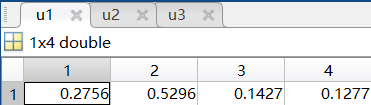
class{150}='iris-virginica';

end

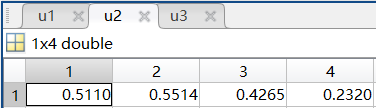
3. 采用合适的度量对数据进行汇总统计

植物外观一般不会有十分极端的差异，故采用平均值对数据进行汇总与可视化。三类鸢尾花的平均值在前面的代码中已求出，标准化后结果如下：

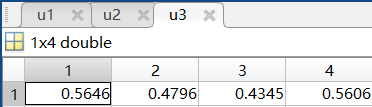
第一类（iris-setosa）



第二类（iris-versicolor）



第三类（iris-virginica）



4.数据可视化：

采用直方图统计各鸢尾花品种四个属性的分布，结果如下：

