专题：鸢尾花智能分类

【背景说明】

自然界中鸢尾花（iris）主要有两个品种：山鸢尾和变色鸢尾，如图所示，两个品种的分辨特征是花瓣的长度和宽度，测量并收集样本的鸢尾花数据，得到样本数据表（参见下表）。本专题使用“人工智能”的方法，使程序能够“学习”所收集的数据，可以对样本自动分类，也可以判定新鸢尾花的类别。

山 鸢（yuān） 尾　　变 色 鸢 尾　　花 瓣 特 征 值 测 量

图1. 鸢尾花的信息采集

样本序号	花瓣长度	花瓣宽度	类别/品种
1	1.1	0.1	1/山鸢尾
2	1.7	0.5	1/山鸢尾
3	1.4	0.3	1/山鸢尾
4	1.6	0.6	1/山鸢尾

样本序号	花瓣长度	花瓣宽度	类别/品种
5	5.0	1.7	2/变色鸢尾
6	4.0	1.0	2/变色鸢尾
7	4.5	1.5	2/变色鸢尾
8	3.0	1.1	2/变色鸢尾

表1. 鸢尾花的样本数据示例（单位：厘米）

各问摘要：

第1问，基本处理，输入4组长度宽度，计算平均值
第2问，文件访问，从文件读取8组数据，按类别计算中心点坐标
第3问，就近判断，文件数据可变长，按距离各中心点的远近判断类别
第4问，学习型判断，初始分界线并依此判断类别，失败时调整参数
第5问，自动聚类，一批未分类数据自动划分为两类
第6问，应用提升，先自动聚类，按最宽分界面生成分界线，再判断类别

【第 1 问，样本数据计算均值】

专题第一步，能够对样本作基本数据处理。

程序功能：输入4朵鸢尾花的花瓣长度及宽度（单位：厘米），计算并输出花瓣长度及宽度的平均值（保留3位小数）。程序保存为C:\KS\iris01.c。

运行示例：（前4行为输入）

```
1.1  0.1
1.7  0.5
1.4  0.3
1.6  0.6
1.450 0.375
```

【第 2 问，从文件读取样本】

专题第二步，能够从文件中读取样本数据，并分类处理。

程序功能：从文件iris02.txt中读取8朵鸢尾花的花瓣长度、宽度及类别，按照类别统计数量，并计算其中心点坐标。程序保存为C:\KS\iris02.c。

人工智能：参考下图，以样本花瓣的长度（对应横轴x）和宽度（对应纵轴y）绘制坐标系，将样本数据绘制在坐标系上，以更直观展示数据。图中类别1有4个样本，中心点为A点，类别2也有4个样本，其中心点为B。

细则要求：

（1）如果文件打开失败，输出“文件xxx打开失败”（其中xxx为文件名）。

（2）鸢尾花类别为1或2，其他类别码作为0处理（未分类）。

（3）输出样式参考运行示例，中心点坐标保留3位小数。

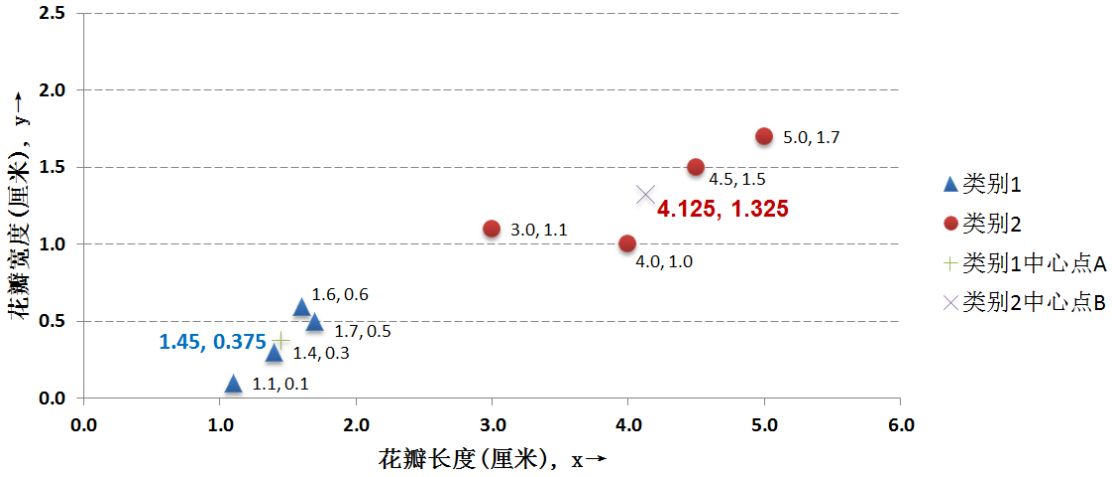
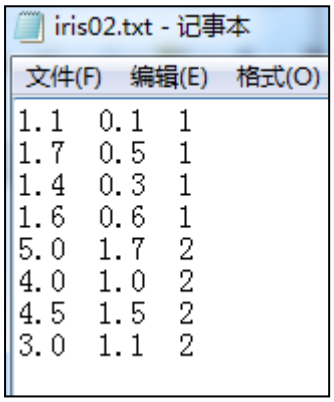


图2. 鸢尾花样本数据在坐标平面上的位置

运行示例：（含文件内容和屏幕输出）




```
类别1：4个样本，中心点A(1.450,0.375)
类别2：4个样本，中心点B(4.125,1.325)
```

【第 3 问，变长文件+就近判断】

专题第三步，从文件中读取数据并处理，按就近原则判断新鸢尾花的类别。

程序功能：从文件iris03.txt中连续读取各鸢尾花的花瓣长度、宽度及类别（保存于结构体数组中）。再从键盘输入一朵新鸢尾花的长度和宽度，计算与类别中心点A和B的距离，按就近原则判断其类别。程序保存为C:\KS\iris03.c。

人工智能：参考下图，各样本及中心点绘制在坐标系上，待判断类别的新鸢尾花标为C点，分别计算C与中心点A、B的距离，如果AC距离更短，判断为类别1，否则判断为类别2。

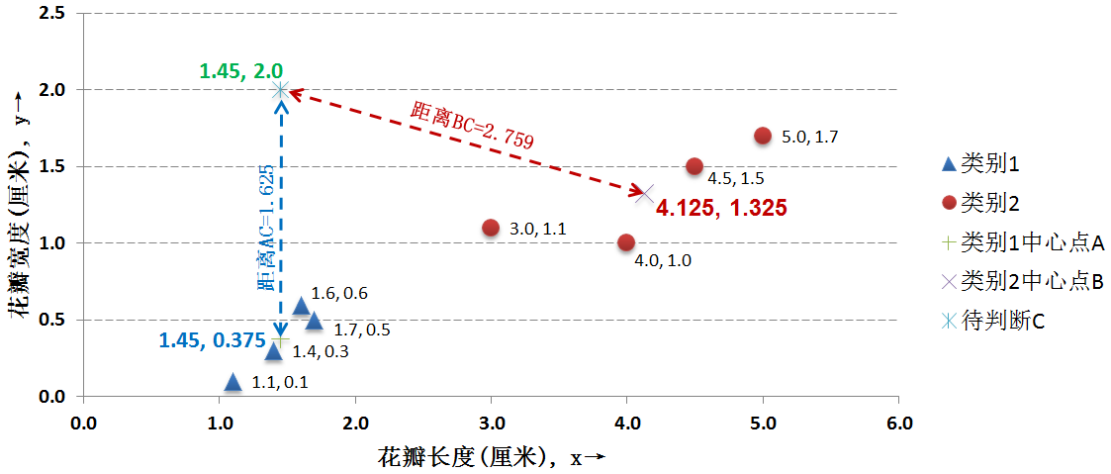


图3. 计算新鸢尾花与2个中心点的距离，按就近原则判断其类别

细则要求：

（1）当文件读取中遇到负数、无效符号或文件结束，或者样本数量超过100个，自动停止读取。

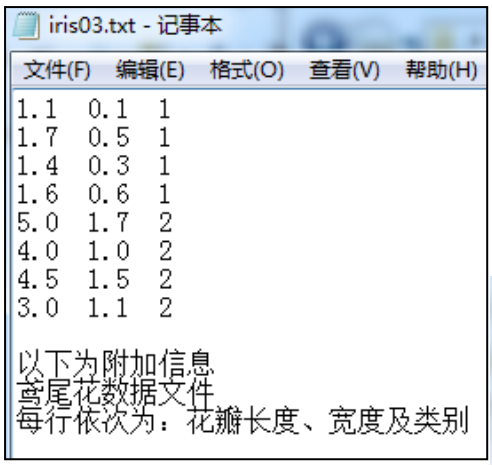
（2）使用结构体存放各样本的数据，结构体定义如下：

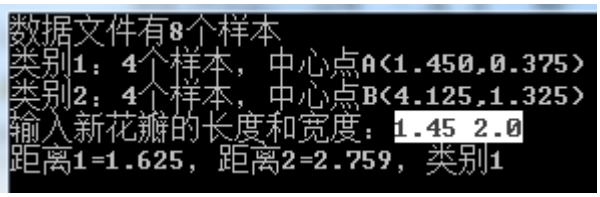
```
struct feature
{
    double x, y;     //鸢尾花的长度(x)和宽度(y)
    int type;        //鸢尾花类别，类别1或2，0表示未分类
};
```

（3）坐标系中(x1,y1)到(x2,y2)的距离公式：sqrt((x2-x1)^2+(y2-y1)^2)

（4）输出样式及小数保留位数参考运行示例。

运行示例：（含文件内容和输入输出）





【第 4 问，学习型类别判断】

专题第四步，在读取样本的过程中不断修正判断参数，通过学习使判断更有效。

人工智能：参考下图，为更直观对鸢尾花进行分类，在坐标系中绘制一条分界线（如图中黄色实线），直线型分界线形如“ax+by+c=0”，对于一个花瓣数据(x,y)，计算f(x,y)=ax+by+c，由结果的正负可以判断其位于分界线的两侧。学习的策略是不断“试错+调整”，具体是先画一条初始分界线，对每个样本数据进行试判，如果发现误判，则修改分界线参数，并继续

下一样本的检测，重复上述过程，直至所有样本检测成功（可能需要好几轮循环学习）。

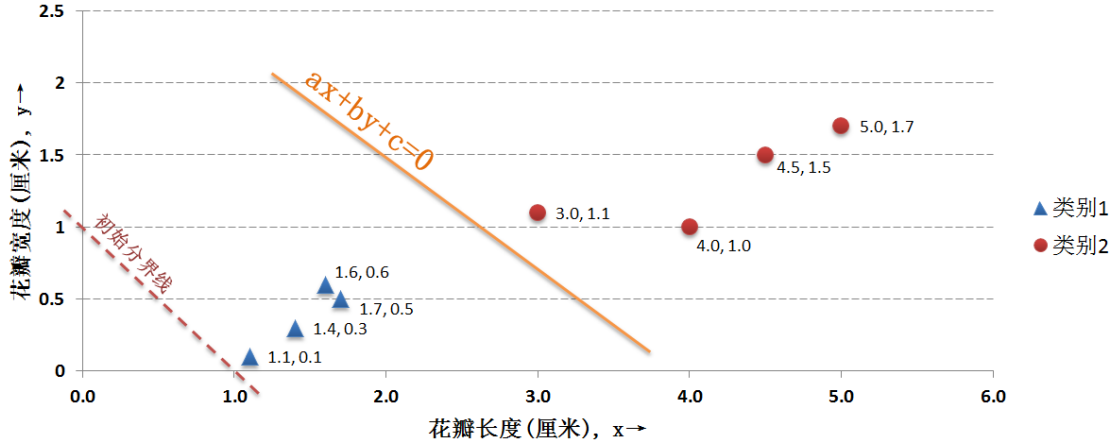


图4. 使用分界线公式判断样本鸢尾花的类别，误判时修正公式参数

学习算法：

（1）设置初始分界线经过(0,1)和(1,0)两点，公式为“x+y-1=0”（a=1,b=1,c=-1）

（2）使用分界线公式试判各个样本，误判时调整公式参数如下：

```
a ← a + rate*t*x          （其中x,y为样本数据，rate为修正学习率）
b ← b + rate*t*y          （样本为类别1时，t取-1，类别2时t取1）
c ← c + rate*t
```

（3）参数调整后，继续下一样本的检测，本轮所有样本检测后，需要进行下一轮检测。

（4）持续上述的“检测+调整”的过程直至所有样本检测成功。

编程要求：从键盘输入学习修正率，然后从文件iris03.txt中读取各样本数据。分界线参数学习修正后，再从键盘输入一朵新鸢尾花的长度和宽度，判断其类别。程序保存为C:\KS\iris04.c。

细则要求：

（1）对花瓣数据(x,y)计算f=ax+by+c，结果为负判断为类别1，否则判为类别2。

（2）为避免死循环，限定最多检测10轮。

（3）文件及结构体数组参考“第2问”。

（4）输出样式参考运行示例，其中a,b,c,f保留两位小数。

运行示例：（文件iris03.txt内容见第3问，兰色框线内为输入）

```
数据文件有8个样本
输入学习率：0.3
分界线：a=0.52，b=0.94，c=-1.90
输入新花瓣的长度和宽度：1.5 0.5
f=-0.65，类别1
```

```
数据文件有8个样本
输入学习率：1.0
分界线：a=0.90，b=2.20，c=-5.00
输入新花瓣的长度和宽度：3.5 1.5
f=1.45，类别2
```

【第 5 问，自动聚类】

专题第五步，从文件中读取未分类样本，学习后划分为2类。再对新数据判断其类别。

人工智能：参考下图，收集未分类的样本数据，通过学习将所有样本划分成2类，策略是先选择初期划分，然后不断按距离中心点的远近重新划分。

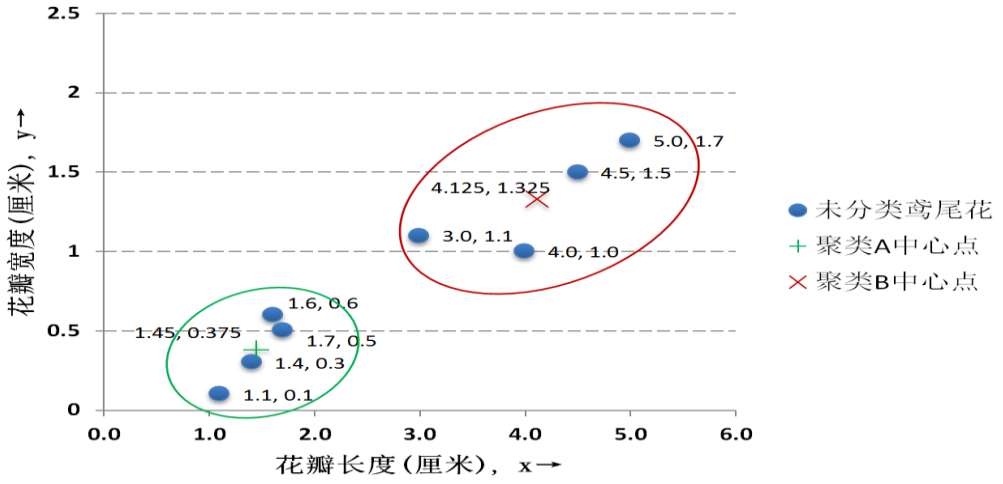


图5. 初始聚类划分后计算中心点AB，以AB为中心重新分类，直至划分稳定

关键算法：将所有样本划分成2类的算法过程如下。

（1）从所有样本中选取2个样本，作为每一个类别的初始聚类中心。

（2）计算每个样本到2个聚类中心的距离，将样本划分到距离最近的类别中。

（3）按新的类别划分，计算2个类别的中心点（作为新的聚类中心）。

（4）重复步骤（2）至（3），直到聚类中心与划分不再发生变化。

编程要求：从文件iris05.txt中读取未分类样本数据，从键盘输入2个类别的初始样本序号，聚类划分后输出各样本的分类结果。再从键盘输入一朵新鸢尾花的长度和宽度，按就近原则判断其类别。程序保存为C:\KS\iris05.c。

细则要求：

（1）为避免死循环，限定聚类重新划分最多100次。

（2）文件及结构体数组参考“第2问”，变长文件读取参见“第3问”。

（3）屏幕输出的内容、样式及精度均参考运行示例。

运行示例：（含输入+输出，文件iris05.txt略，可参考输出信息）

【第 6 问，自动聚类+最宽分界面】

专题第六步，自动聚类+分界面判断，分界线采用宽分界面方法。

人工智能：参考下图，聚类划分后，两个聚类之间，存在宽范围的分界面，设法得到最宽的分界面，选取宽分界面的中间为分界线，该分界线具有更高的可靠性。

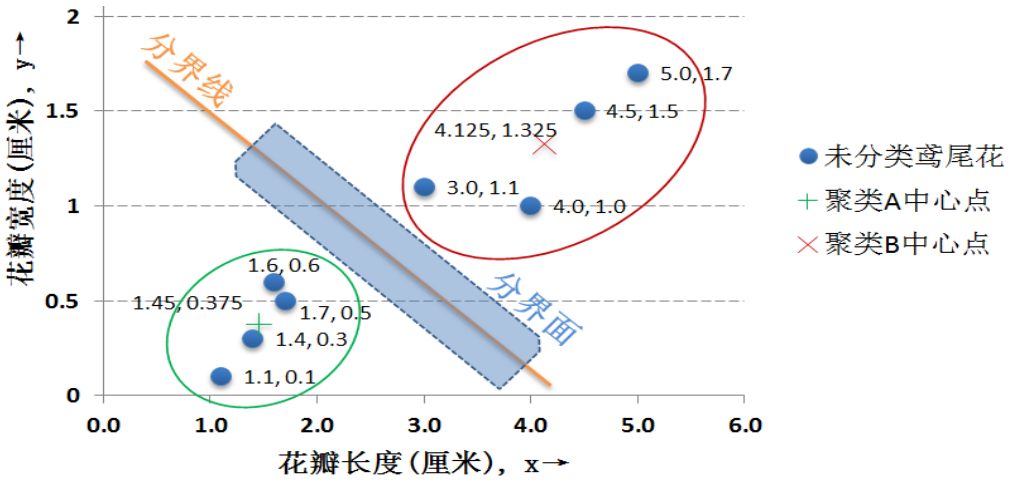


图6. 自动聚类后，找出两类中距离最短的2个样本，其垂直平分线为分界线

编程要求：从文件iris06.txt中读取未分类样本数据，从键盘输入2个类别的初始样本序号，聚类划分后输出各样本的分类结果。依据分类结果设法计算出“最宽”分界面，取分界面的中间为分界线，输出分界线参数。验证该分界线是否正确。最后输入新鸢尾花的长度和宽度，判断其类别。程序保存为C:\KS\iris06.c。

关键算法：（宽分界面算法）

（1）分别从2个类别中选取一个样本，使它们之间的距离最短。

（2）计算这两点之间的垂直平分线，作为初始的分界线（ax+by+c=0）。

坐标点(x_1, y_1)和(x_2, y_2)之间的垂直平分线，参考计算公式：

$a = (x_2-x_1)$

$b = (y_2-y_1)$

$c = (x_1^2-x_2^2+y_1^2-y_2^2)/2$

（3）对每个样本验算f=ax+by+c。类别1的样本，得到的f应该为负数，求解其最大值；类别2的f应该为正数，求解其最小值。

细则要求：

（1）聚类算法参考“第5问”。

（2）分界线判断方法参考“第4问”。

（3）文件及结构体数组参考“第2问”，变长文件读取参见“第3问”。

（4）屏幕输出的内容、样式及精度均参考运行示例。

运行示例：（含输入+输出，文件iris06.txt略，可参考输出信息）

```
数据文件有8个样本
输入2个初始样本：0 7
样本0(1.1,0.1)→类别1
样本1(1.7,0.5)→类别1
样本2(1.4,0.3)→类别1
样本3(1.6,0.6)→类别1
样本4(5.0,1.7) →类别2
样本5(4.0,1.0) →类别2
样本6(4.5,1.5) →类别2
样本7(3.0,1.1) →类别2
距离最近的2个样本是：1 7，距离=2.050
初始分界线：a=1.30，b=0.60，c=-3.54
分界线误判断数量=0
类别1最大f=-1.025 类别2最小f=1.025
输入新花瓣的长度和宽度：1.45 2
类别1
```

```
数据文件有8个样本
输入2个初始样本：0 4
样本0(1.1,0.1)→类别1
样本1(1.7,0.5)→类别1
样本2(1.4,0.3)→类别1
样本3(1.6,0.6)→类别1
样本4(5.0,1.7) →类别2
样本5(4.0,1.0) →类别2
样本6(4.5,1.5) →类别2
样本7(3.0,1.1) →类别2
距离最近的2个样本是：1 7，距离=2.050
初始分界线：a=1.30，b=0.60，c=-3.54
分界线误判断数量=0
类别1最大f=-1.025 类别2最小f=1.025
输入新花瓣的长度和宽度：3.5 1.5
类别2
```