**实验4 Python数值计算**

**学号：** **姓名： 日期：2021-11-08**

* **实验正式报告请交纸质版。截止日期：2021年11月15日**
* **请注意报告版面整洁，截图清晰，内容组织逻辑清晰，易于阅读**
* **代码请附在报告中，并打包发邮件。代码要能复现报告中的结果。**

**目的和要求：**

（1）掌握基于Numpy数值分析的基本用法

（2）运用Numpy数值分析，解决实际问题

**一、基本题**

**使用Python Numpy回答下面的问题：**

1. 矩阵，计算a的行列式和逆矩阵。

2.a = [1 2 3; 4 5 6; 7 8 9], b = a;

用np.rot90 去旋转矩阵a，要旋转几次，才能和 np.flipud(np.fliplr(b)) 相等

3. 使用np.random.rand函数，建立一个5x5的随机数矩阵，求矩阵的行列式的值，秩和范数。

4. 方程组,

1. 依线性代数，写成Ax = b的形式，那么，求得x
2. 使用np.linalg.solve 求得x
3. 两者是否相等？

5. 写出代码，给出例子,验证两个方阵A，B满足|AB| =|A||B|

**二、应用题**

1948年起奥林匹克运动会女子铅球记录如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 年份 | 1948 | 1952 | 1956 | 1960 | 1964 | 1968 | 1972 | 1976 | 1980 | 1984 |
| 距离(米) | 13.75 | 15.28 | 16.59 | 17.32 | 18.14 | 19.61 | 21.03 | 21.16 | 22.41 | 23.57 |
| 平均年龄 | 28 | 28 | 27.5 | 29 | 30 | 26 | 26.5 | 28 | 27 | 31 |

a) 使用数据插值函数，从这些数据中预测1970年的奥运会女子铅球的最佳成绩？

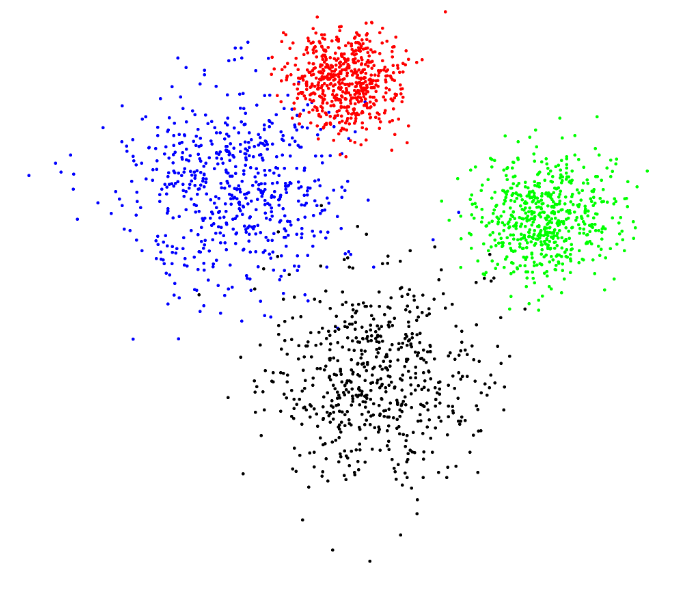
b) 表中的距离逐年呈现单调递增的趋势，请用一个一次函数y = kx + b 拟合这些数据。并回答依照拟合的函数，在2000年奥运会女子铅球的最佳成绩会是多少？

**三、综合题**

**编写函数实现Kmeans 算法聚类算法。**

使用Numpy编写代码，不要使用其他工具包，勿直接抄袭网上代码！

聚类就是根据数据之间的相似度将数据集划分为多个类别或组，使类别内的数据相似度较大而类别间的数据相似度较小。如下图所示，左边是原始数据，右边是聚类之后的效果，不同的颜色代表不同的类别。



**实验数据：**文件Lab4.dat 中含有2400个二维空间的点坐标XY。

**Kmeans 算法简要**

请参考相关书籍或网络，了解kmeans 算法。大致的步骤如下：

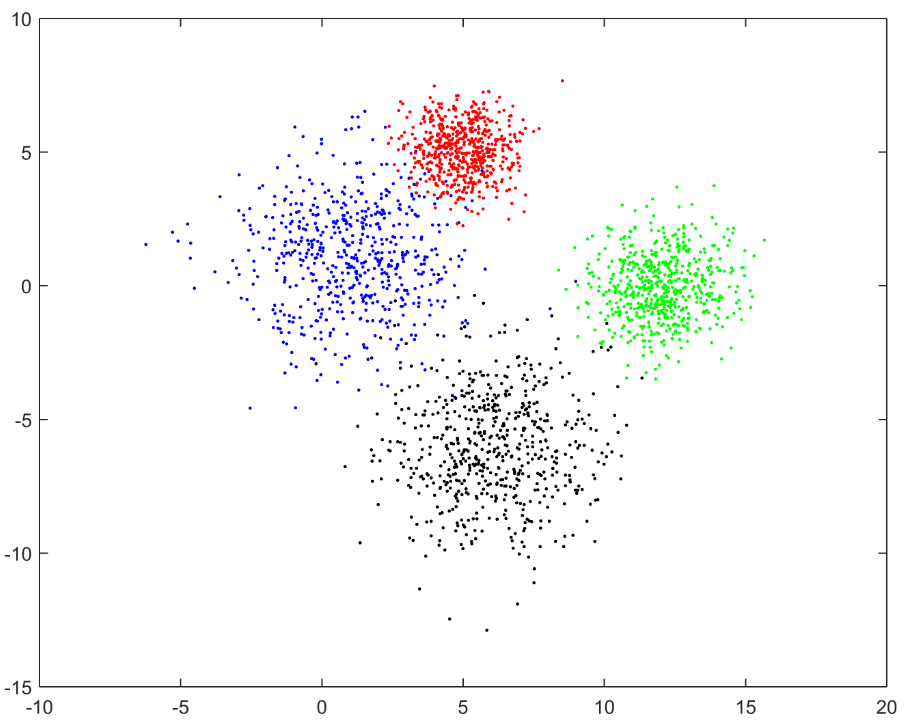
1. 设置初始类别中心和类别数
2. 根据类别中心对全部数据进行类别划分：每个点分到离自己距离最小的那个类
3. 重新计算当前类别划分下每个类的中心：例如可以取每个类别里所有的点的平均值作为新的中心。如何求多个点的平均值？ 分别计算X坐标的平均值，y坐标的平均值，从而得到新的点。**注意：**类的中心可以不是真实的点，虚拟的点也不影响。
4. 在新的类别中心下继续进行类别划分;
5. 如果连续两次的类别划分结果不变则停止算法; 否则循环2～5。例如当类的中心不再变化时，跳出循环。

**影响Kmeans算法的可能因素：**

* **如何选择距离的定义：**对数据点进行类别划分时，需要计算点到点之间的距离。距离有很多种。 例如欧式距离，, 其中和是两个不同的点; L1范数距离。
* **如何确定合理的K值？**
* **如何选择K个初始类簇的中心点？**

**要求：**

**1. 实现函数 cidx, ctrs = kmeans(X, K),** 其中

* 输入X是的矩阵，每行一个点，K是类别的个数，
* 输出ctrs 是类的中心坐标，对应的size应该为：。cidx 是各个点的类别信息，表示每个点属于哪一类，其size 为, 例如cidx(0) = 2，表示第一个点属于第二个类。
* 函数写好后，测试当K = 2, 3, 4的效果，并用散点图画出分类的效果。例如，下图显示了当K=4的效果。

请查阅matplotlib.pyplot.scatter散点图的画法。

**2. 确定最优的参数K：手肘法**

参考：<https://blog.csdn.net/qq_15738501/article/details/79036255>

对每一个K值，计算分类的SSE(sum of the squared errors，误差平方和)。

其中N是点的个数，是第个点，是对应的中心。

编写函数函数y = calSSE(X, cidx) 计算聚类效果的SSE，其中X是待聚类的数据，cidx 是kmeans 函数返回的聚类结果， 画出当类别数K分别为2,3,4,5,6等值时的SSE， 肉眼能否看出最佳的K值？类似下图的效果？

