# 现场考核：数据统计分析

### 1. 题目

人类DNA是由4个字符A，T，C，G按一定顺序排成的长约30亿的序列。当前发现DNA序列中的一些规律性和结构：

* 在全序列中有一些是用于编码蛋白质的序列片段， 即由这4个字符组成的64种不同的3字符串，其中大多数用于编码构成蛋白质的20种氨基酸；
* 在不用于编码蛋白质的序列片段中，A和T的含量特别多些；
* 统计方法发现序列的某些片段之间具有相关性

在这项研究中最普通的思想是省略序列的某些细节，突出特征，然后将其表示成适当的数学对象。这种被称为粗粒化和模型化的方法往往有助于研究规律性和结构。作为研究 DNA序列的结构的尝试，提出以下对序列集合进行分类的问题：

附件中有20个已知类别的人工制造的序列，其中序列标号1—10为A类，11-20为B类。请从中提取统计特征，并作简单可视化，分析这两类各有何特征，将程序、结果与分析填写在第二页中。

### 2. 读取数据程序(参考)：

import numpy as np

import pandas as pd

from itertools import islice # 需先安装 itertools

def get\_frequency(filename):

str0 = ["a", "t", "c" , "g"]

df\_count = pd.DataFrame(columns=str0) # 用于记录频数

with open(path\_data + 'Art-model-data.txt','r', newline='') as filereader:

k = 0

for row in islice(filereader, 1, None): # islice() 跳过第1行

if len(row)>2: # 空行可能有换行符，占一定长度

k = k+1

k\_count = []

for item in str0:

k\_count.append(row.count(item))

df\_count.loc[k] = k\_count

df\_RowSum = df\_count.sum(axis=1) # 按行相加，用于将频数转换为频率

df\_freq = df\_count.div(df\_RowSum, axis="rows") # 每行除以对应总和， 按行计算频率

return df\_freq.astype("float").round(4) # 转换类型为浮点数后，只保留4位小数

path\_data = "D:/现场考核：数据统计分析/" # 设置 数据路径

file0 = path\_data + 'Art-model-data.txt'

df\_freq = get\_frequency(file0)

### 3. 分析程序

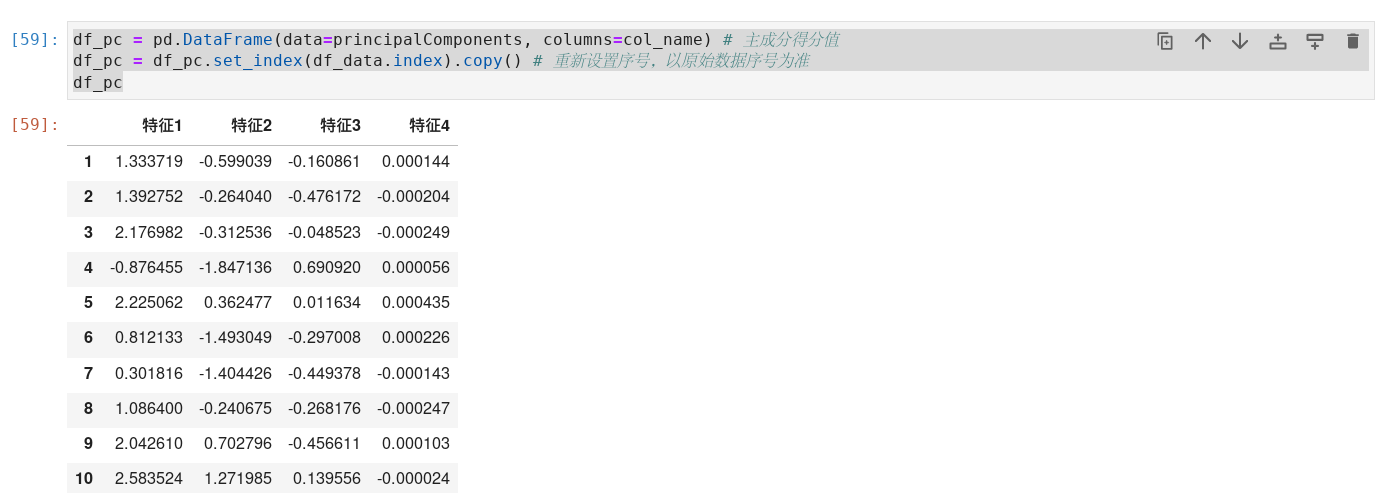
探索性数据分析



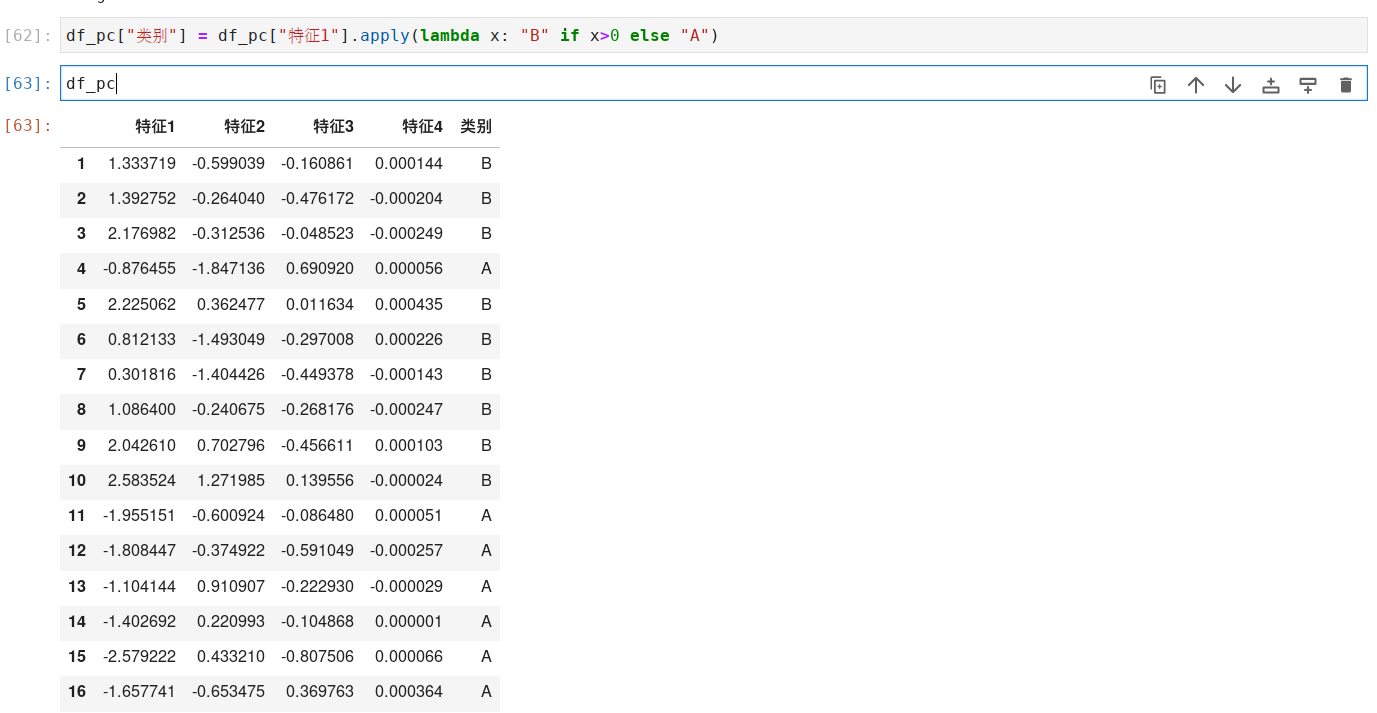
PCA主成分分析



计算样品主成分得分值

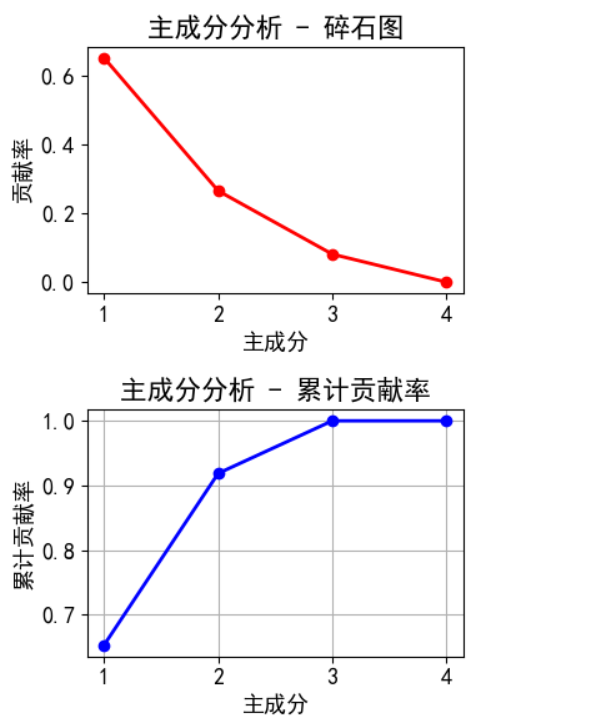


简单分类，线性分类器

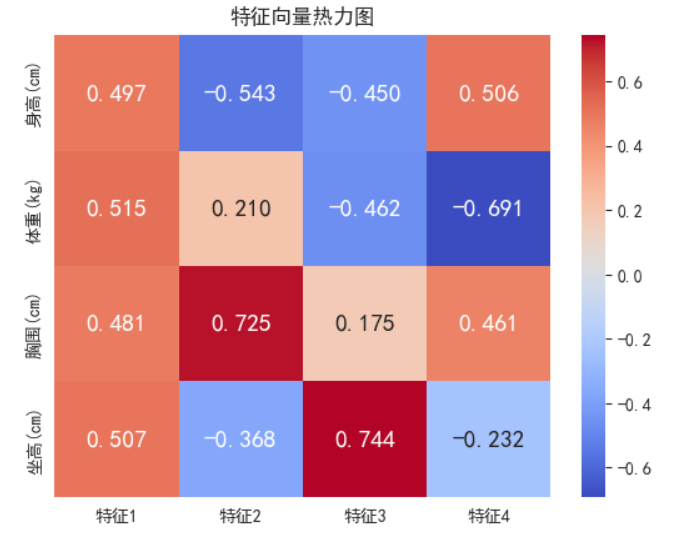


### 程序结果

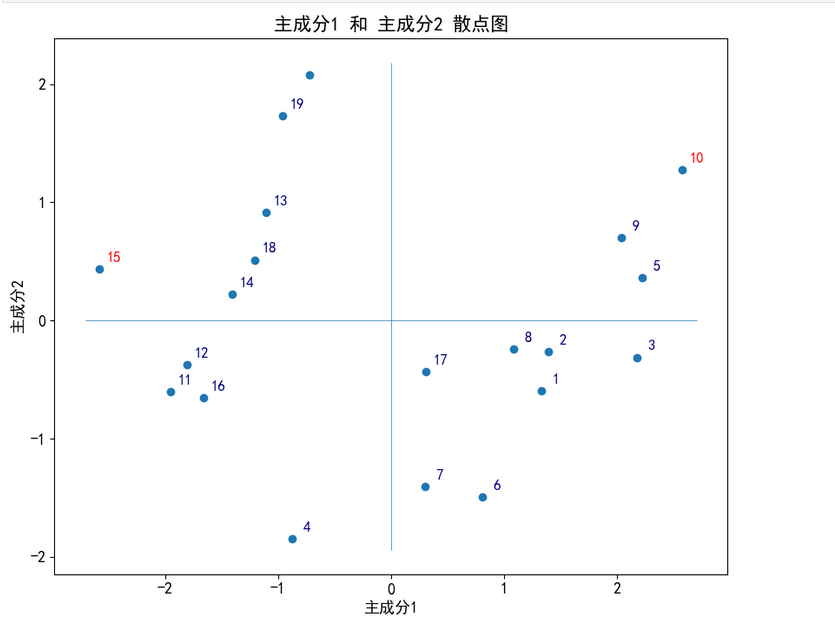
碎石图



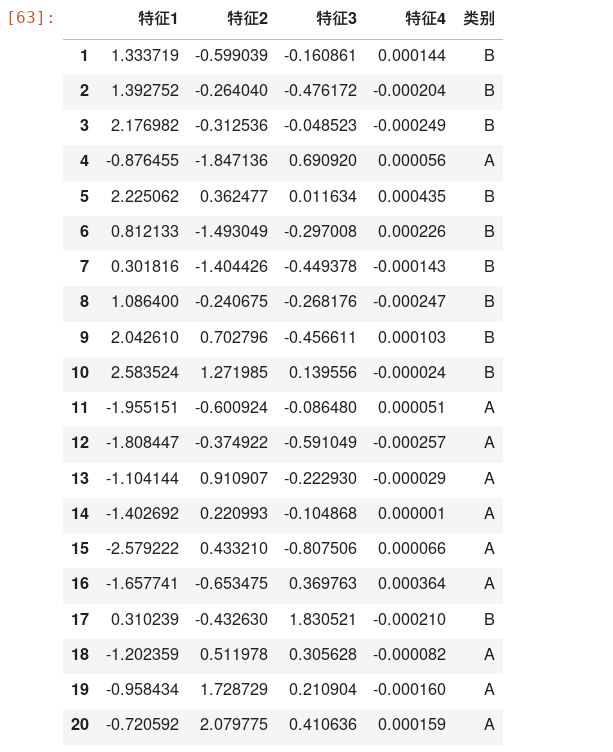
特征向量热力图

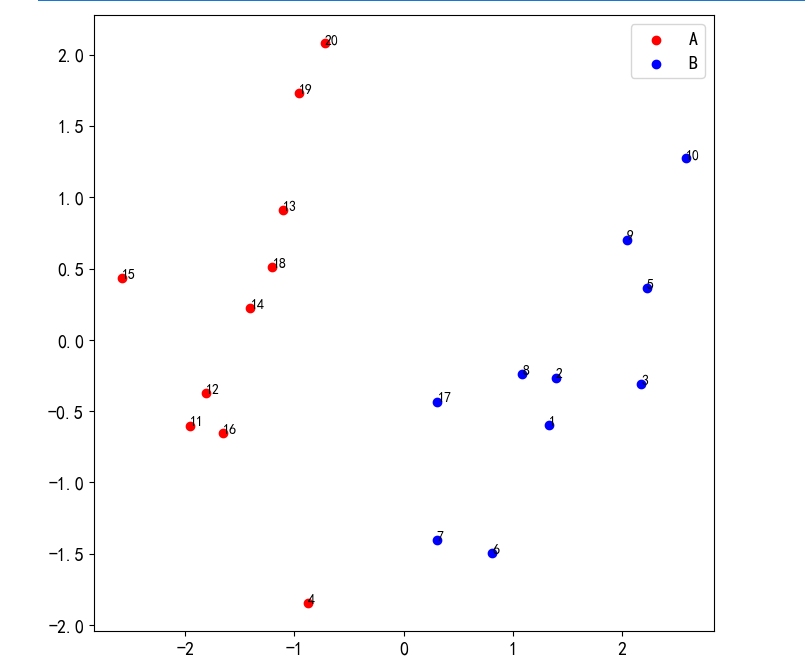


主成分1和主成分2的散点图



分类结果





### 5. 结果分析

通过主成分分析发现特征2主要代表A类，特征1主要代表B类，所以构造线性分类器进行划分，结果如上图所示 4 11 12 13 14 15 16 18 19 20为A类， 其余为B类