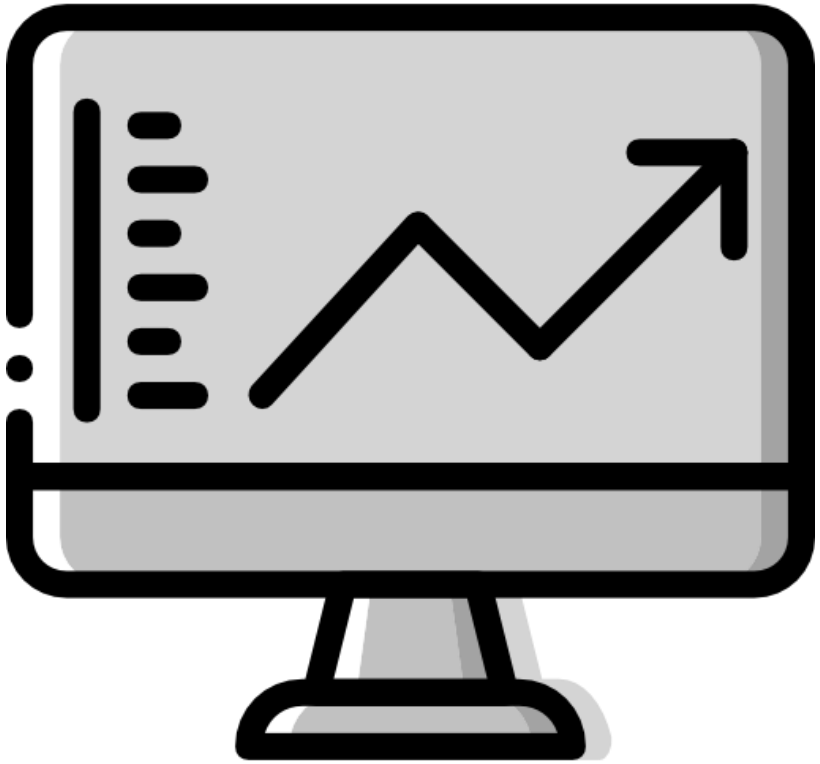
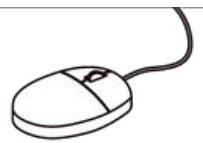
南通师范高等专科学校
Nantong Normal College



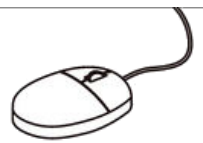
Python
数据分析与应用

执教：朱亚林

利用**NumPy**进行统计分析



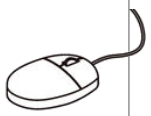
排序



排序

NumPy的排序有两种类型

- 直接排序：对数值直接进行排序
- 间接排序：根据一个或多个键对数据进行排序

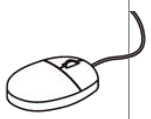


排序

sort()函数

- sort()函数是最常用的排序方法，无返回值 。
- 在函数运行之后，原始数据会被修改。
- 使用sort()函数排序时可以设置axis参数，用以指定排序的方向 。

```
>>> import numpy as np
>>> np.random.seed(50)
>>> arr = np.random.randint(1,10,10)
# 创建一个数组arr包括10个1 ~9的随机整数
>>> arr.sort()
# 对arr进行排序，并将修改arr中原有的值为新的顺序# 此时顺序为升序，思考：如何改成逆序
>>> arr = np.random.randint(1,10,size=(3,3))
>>> arr.sort(axis = 1)
>>> arr.sort(axis = 0)
# axis参数，1代表横向，0代表纵向
```



排序

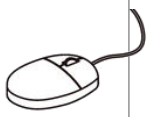
random.seed()的作用是什么？

seed() 用于指定随机数生成时所用算法开始的整数值。

1.如果使用相同的seed()值，则每次生成的随即数都相同；

2.如果不设置这个值，则系统根据时间来自己选择这个值，此时每次生成的随机数因时间差异而不同；

3.设置的seed()值仅一次有效。



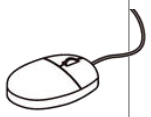
排序

sort()函数order的用法

sort函数排序时，使用order 指定排序字段

```
>>> dt = np.dtype([('name', 'S10'),('age', int)])
>>> a = np.array([("kevin", 21),("peter",25),("tony", 17), ("mike",27)], dtype = dt)
>>> print ('数组:')
>>> print (a)
>>> print ('\n')
>>> print ('按 name 排序:')
>>> print (np.sort(a, order = 'name'))
```

此例会输出什么？



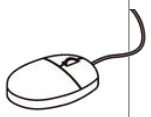
排序

argsort()函数

argsort函数对一维数组进行排序，排序时记录的是这个数组中排好序之后的索引数组。这个数组表示的是数据在新的序列中的位置。

```
>>> arr = np.array([3,2,5,7,8,0,4,1,9,6])
>>> arr.argsort()
>>> arr[arr.argsort()]
```

此例会输出什么？



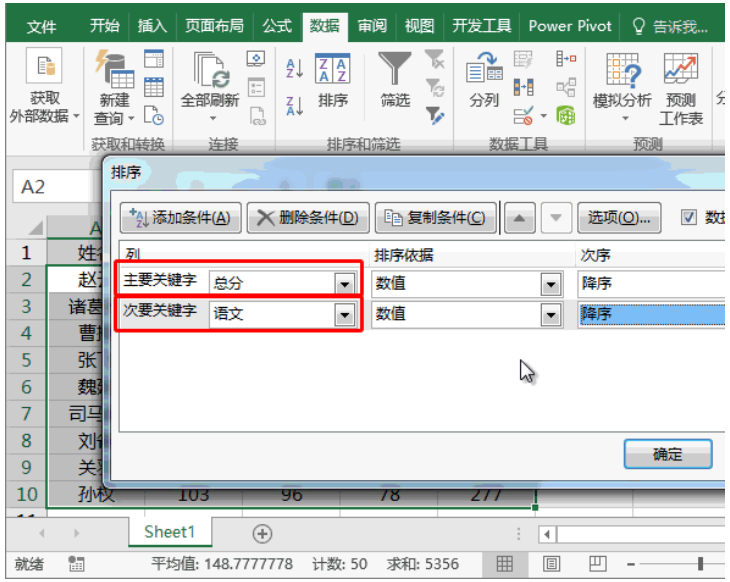
排序

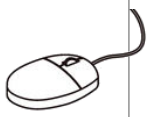
lexsort()函数

lexsort函数可以一次性对满足多个键的数组执行间接排序。其排序返回的值与argsort函数一样，都是返回的索引。对于这个函数的应用，可以把它想象成对电子表格进行排序，每一列代表一个序列，排序时优先照顾靠后的列。

一个应用场景：

小升初考试，重点班录取学生按照总成绩录取。在总成绩相同时，数学成绩高的优先录取，在总成绩和数学成绩都相同时，按照英语成绩录取…… 这里，总成绩排在电子表格的最后一列，数学成绩在倒数第二列，英语成绩在倒数第三列 。



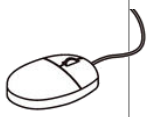


排序

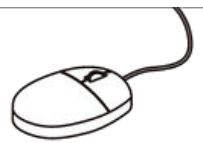
lexsort()函数

```
>>> arr1 = np.array([3,2,5,7,8])
>>> arr2 = np.array([8,42,15,71,28])
>>> arr3 = np.array([115,342,115,771,528])
>>> arr4 = np.lexsort((arr1,arr2,arr3))
>>> list(zip(arr1[arr4],arr2[arr4],arr3[arr4]))
```

此例会输出什么？



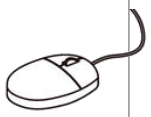
数据清洗



数据清洗

清洗数据

数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。



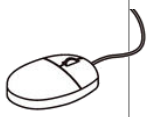
数据清洗

unique()函数

数据重复也是数据清洗过程中经常要去完成的一个方面。NumPy中提供了unique()函数来找出数组中的唯一值，并返回排序好的数组。

```
>>> vehicle = ['train','bus','ship','car','subway','ship','bicycle']
>>> np.unique(vehicle)
```

此例会输出什么？

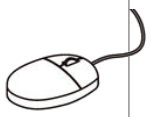


tile函数()

tile()函数用来对数据进行重复拼接。

函数格式：tile(A,reps)

- A : array_like，A的类型众多，几乎所有类型都可以：array, list, tuple, dict, matrix以及基本数据类型int, string, float以及bool类型。
- reps : array_like, reps的类型也很多，可以是tuple，list, dict, array, int, bool.但不可以是float, string, matrix类型。表示A沿各个维度重复的次数

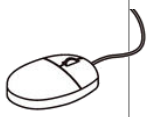


数据清洗

tile函数()

```
>>> arr = [1,2]
>>> tile(arr,2)
array([1, 2, 1, 2])
# 表示一个维度重复2次
 >>> tile(arr,(1,2))
array([[1, 2, 1, 2]])
#表示第一个维度重复两次，第二个维度重复一次
>>> tile(arr,(2,2,3))
array([[[1, 2, 1, 2, 1, 2],
        [1, 2, 1, 2, 1, 2]],

        [[1, 2, 1, 2, 1, 2],
        [1, 2, 1, 2, 1, 2]]])
# 表示第一维度重复三次，第二个维度重复两次，第三个维度重复两次
```



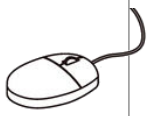
repeat()函数

repeat()函数同样是用来对数据进行重复的。

用法有两种：

- numpy.repeat(a, repeats, axis=None)
- a.repeat(repeats, axis=None)

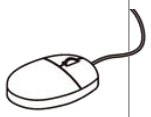
其中a为数组，repeats为重复的次数，axis表示数组维度



数据清洗

repeat()函数

```
>>> a = np.arange(10)
>>> a
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
>>> a.repeat(5)
array([0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4,
       4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9,
       9, 9, 9, 9])
>>> a np.array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
#a数组的内容没改变
>>> a=np.array([10,20])
>>>a
array([10,20])
>>> a.repeat([3,2])
array([10, 10, 10, 20, 20])
>>> repeat(a,[3,2])
# 对a数组中的对应元素进行重复复制，需要注意的是len(repeats)==a.shape[axis]
array([10, 10, 10, 20, 20])
>>> a=np.array([[10,20],[30,40]])
>>> a.repeat([3,2],axis=0)
array([[10, 20], [10, 20], [10, 20], [30, 40], [30, 40]])
>>> a.repeat([3,2],axis=1)
array([[10, 10, 10, 20, 20],
       [30, 30, 30, 40, 40]])
```



统计函数

在NumPy中，常见的统计函数有以下几种：

函数名	作用
sum(a, axis=None)	据给定轴axis计算数组a相关元素之和，axis整数或元组
mean(a, axis=None)	给定轴axis计算数组a相关元素的期望，axis整数或元组
average(a,axis=None,weights=None)	据给定轴axis计算数组a相关元素的加权平均值
std(a, axis=None)	据给定轴axis计算数组a相关元素的标准差
var(a, axis=None)	据给定轴axis计算数组a相关元素的方差
min()	求最小值
max()	求最大值
ptp()	最大值与最小的值差
np.intersect1d(arr1, arr2)	找出两个数组公共的元素

以上函数在对数据进行统计时，如不指定axis参数，则表示对整个数组进行统计。

当axis参数为1时，表示沿横轴进行计算，当axis参数为0时，表示沿纵轴进行计算。