法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

第二讲

科学计算及数据可视化入门

--梁斌

目录

- NumPy
- SciPy
- Matplotlib入门
- 实战案例：2016 Election Pools

目录

- NumPy
- SciPy
- Matplotlib入门
- 实战案例：2016 Election Pools

NumPy

NumPy, Numerical Python

- 高性能科学计算和数据分析的基础包
- ndarray，多维数组（矩阵），具有矢量运算能力，快速、节省空间
- 矩阵运算，无需循环，可完成类似Matlab中的矢量运算
- 线性代数、随机数生成
- `import numpy as np`

NumPy ndarray

ndarray，N维数组对象（矩阵）

- 所有元素必须是相同类型
- ndim属性，维度个数
- shape属性，各维度大小
- dtype属性，数据类型

示例代码： numpy_codes.ipynb

创建ndarray

- np.array(collection)，collection为序列型对象(list)，嵌套序列(list of list)
- np.zeros, np.ones，np.empty 指定大小的全0或全1数组
  - 注意：第一个参数是元组，用来指定大小，如(3,4)
  - empty不是总是返回全0，有时返回的是未初始的随机值

创建ndarray（续）

- np.arange()类似range() 注意是arange，不是英文arrange

ndarray数据类型

- dtype，类型名+位数，如float64，int32
- 转换数组类型
  - astype

示例代码： numpy_codes.ipynb

NumPy

多维数组运算

矢量化（vectorization）

- 矢量运算，相同大小的数组键间的运算应用在元素上
- 矢量和标量运算，“广播” – 将标量“广播”到各个元素

索引与切片

- 一维数组的索引与Python的列表索引功能相似
- 多维数组的索引

示例代码： numpy_codes.ipynb

NumPy

索引与切片（续）

- 多维数组的索引
  - `arr[r1:r2, c1:c2]`
  - `arr[1,1]` 等价 `arr[1][1]`
  - `[:]` 代表某个维度的数据

0,0	**0,1**	**0,2**
1,0	1,1	1,2
2,0	2,1	2,2

示例代码： numpy_codes.ipynb

NumPy

索引与切片（续）

- 条件索引
  - 布尔值多维数组　arr[condition] condition可以是多个条件组合
  - 注意，多个条件组合要使用 & |，而不是and or

0	1	2
3	4	5
6	7	8

T	F	F
F	T	F
F	F	T

0

4

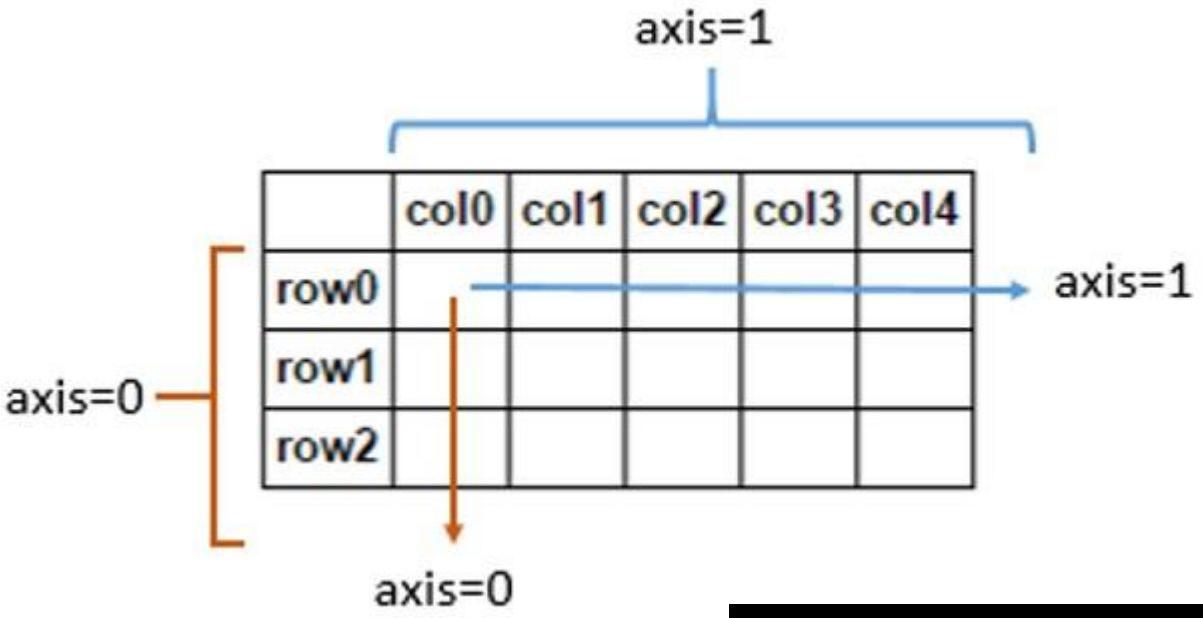
8

示例代码： numpy_codes.ipynb

NumPy

维数转换

- 转置 transpose
- 高维数组转置要指定维度编号 (0,1,2,…)



示例代码： numpy_codes.ipynb

NumPy

通用函数（ufunc）

- 元素级运算

常用的通用函数

- ceil，向上最接近的整数
- floor，向下最接近的整数
- rint，四舍五入
- isnan，判断元素是否为 NaN(Not a Number)
- multiply，元素相乘
- divide，元素相除

示例代码： numpy_codes.ipynb

NumPy

np.where

- 矢量版本的三元表达式 `x if condition else y`
- `np.where(condition, x, y)`

常用的统计方法

- `np.mean, np.sum,`
- `np.max, np.min`
- `np.std, np.var`
- `np.argmax, np.argmin`
- `np.cumsum, np.cumprod`
- 注意多维的话要指定统计的维度，否则默认是全部维度上做统计。

示例代码： `numpy_codes.ipynb`

NumPy

np.all和np.any

- all，全部满足条件
- any，至少有一个元素满足条件

np.unique

- 找到唯一值并返回排序结果

操作文本文件

- 读取
  - np.loadtxt

示例代码： numpy_codes.ipynb

目录

- NumPy
- SciPy
- Matplotlib入门
- 实战案例：2016 Election Pools

SciPy

Scipy

- 在NumPy库的基础上增加了众多的数学、科学及工程常用的库函数
- 线性代数、常微分方程求解、信号处理、图像处理、稀疏矩阵等
- `import scipy as sp`
- 一般的数据处理numpy已经够用
- 常用的统计函数在matplotlib部分讲解

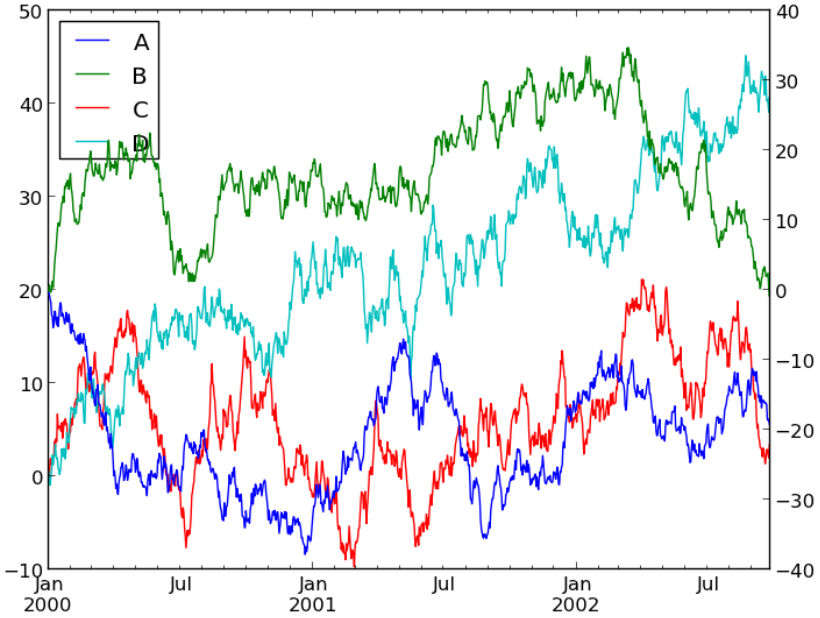
目录

- NumPy
- SciPy
- Matplotlib入门
- 实战案例：2016 Election Pools

Matplotlib

Matplotlib

- 用于创建出版质量图表的绘图工具库
- 目的是为Python构建一个Matlab式的绘图接口
- import matplotlib.pyplot as plt
  - pyploy模块包含了常用的matplotlib API函数



Matplotlib

figure

- Matplotlib的图像均位于figure对象中
- 创建figure
  - plt.figure()

示例代码： matplotlib_codes.ipynb

Subplot

- fig.add_subplot(a, b, c)
  - a,b 表示讲fig分割成axb的区域
  - c 表示当前选中要操作的区域，
    - 注意：从1开始编号

Matplotlib

Subplot (续)

- fig.add_subplot(a, b, c)
  - 返回的是AxesSubplot对象
  - plot 绘图的区域是最后一次指定subplot的位置 (jupyter里不能正确显示)

- 在指定subplot里结合scipy绘制统计图
  - 正态分布 `sp.stats.norm.pdf`
  - 正态直方图 `sp.stats.norm.rvs`

示例代码：

matplotlib_codes.ipynb,

matplotlib_codes.py

Matplotlib

Subplot (续)

- 直方图 hist
- 散点图 scatter
- 柱状图 bar
- 矩阵绘图 plt.imshow()
  - 混淆矩阵，三个维度的关系

示例代码：

matplotlib_codes.ipynb,

matplotlib_codes.py

Matplotlib

plt.subplots()

- 同时返回新创建的figure和subplot对象数组
- fig, subplot_arr = plt.subplots(2,2)
- 在jupyter里可以正常显示，推荐使用这种方式创建多个图表

示例代码：

matplotlib_codes.ipynb,

matplotlib_codes.py

目录

- NumPy
- SciPy
- Matplotlib入门
- 实战案例：2016 Election Pools

2016 美国大选

示例代码：

lecture02_project.ipynb,

lecture02_project.py

项目介绍

- 项目地址：https://www.kaggle.com/fivethirtyeight/2016-election-polls
- 该数据集包含了2015年11月至2016年11月期间对于2016美国大选的选票数据
- 27列数据

2016 美国大选

项目目的

- 分析每个月的民意调查统计趋势

涉及知识点

- 高阶函数：filter
- Numpy读取文本文件
- 处理日期格式数据
- Numpy的切片与索引
- Numpy的统计方法
- 高级特性：列表推导式
- 数据结构zip
- Matplotlib进行简单的数据可视化

```
示例代码：
lecture02_project.ipynb,
lecture02_project.py
```

参考

- 快速入门numpy、scipy

  https://docs.scipy.org/doc/numpy-dev/user/quickstart.html
- numpy教程

  http://cs231n.github.io/python-numpy-tutorial/
- numpy scipy介绍

  https://engineering.ucsb.edu/~shell/che210d/numpy.pdf
- 13个numpy scipy教程

  http://www.erzama.com/scipy-numpy-tutorials-w-12023/
- 《Python数据分析基础教程：NumPy学习指南》
- Matplotlib示例库

  http://matplotlib.org/gallery.html

参考

- 《Python for Data Analysis》

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答 @Robin_TY

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop