cutd

《利用Python进行数据分析： Python for Data Analysis 》学习随笔

NoteBook of 《Data Analysis with Python》

3.IPython基础

Tab自动补齐

- 变量名
- 变量方法
- 路径

解释

- ？解释，
- ？？显示函数源码
- ？搜索命名空间

%run命令

- %run 执行所有文件
- %run -i 访问变量

- Ctrl-C中断执行
- %paste可以粘贴剪切板的一切文本
- 一般使用%cpaste因为可以改

键盘快捷键

魔术命令

- %timeit 检测任意语句的执行时间
- %magic显示魔术命令的详细文档
- %xdel v 删除变量，并清除其一切引用
- 注册超能云（SuperVessel Cloud）（注册网址： http://www.ptopenlab.com ）

4.Numpy基础：数组和矢量计算

ndarray：多维数组对象

创建ndarray

- np.array() #将输入数据转换成ndarray
- np.asarray() #将输入转换成ndarray
- np.zeros() zeros_like #全零数组
- np.ones()ones_like #全一数组
- np.empty()empty_like #创建数组只分配内存不赋值
- np.arange() #返回 range版的ndarray
- eye、identify #N*N单位矩阵

ndarray类型

- int、uint8 16 32 64
- float 16 32 64 128
- complex 64 128 256 浮点数表示的复数
- bool
- object python对象类型
- string_ 固定长度的字符串
- unicode_ 固定长度的Unicode类型
- astype可以显示转换成其他类型

常用链接

我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签

python (5) 机器学习 (4)
数据挖掘 (3) Storm (3)
kaggle (3) Linux (2)
Machine Learning (2)
mysql (2) Apache Storm (2)
elasticsearch (2) 更多

随笔分类

Android开发(1)
Python&Machine Learning(14)
sql(1)
Thinking
遇到的一些问题以及解决办法...

随笔档案

2017年4月 (1)
2017年3月 (2)
2017年2月 (1)
2017年1月 (3)
2016年12月 (2)
2016年10月 (3)
2016年8月 (2)
2016年7月 (4)

数组和标量之间的计算

批量计算

基本的索引和切片

- 数组切片是原始数据，对切片的任何修改都会直接修改原始数据
- 若需要复制一个副本就需要显示的复制.copy()
- 访问单个元素arr[0][2]=arr[0,2]
- 注意区分1和:1 #前者表示第二行，后者表示到第一行
- and or 在布尔型数据中无效

花式索引

- 花式索引就是将数据复制到新数组中
- 转置 .T transpose swapaxes

通用函数ufunc：元素级数组函数

- 一元ufunc
  - abs、fabs
  - sqrt平方根
  - square平方
  - exp指数
  - log、log10、log2、log1p对数
  - sign正负号
  - ceil大于等于该值的最小整数
  - floor小于等于该值的最大整数
  - rint四舍五入到最接近的整数
  - modf小数和整数部分分别返回
  - isnan布尔型数组返回NaN非数字
  - isfinite、isinf布尔型返回有穷和无穷
  - cos、sosh、sin、sinh、tan、tanh双曲线三角函数
  - arc cos、sosh、sin、sinh、tan、tanh反三角
  - logical_not 计算not x的真值
- 二元ufunc
  - add相加

2016年5月 (1)
2016年4月 (4)
2016年3月 (1)
2016年2月 (6)

AI

深度学习

最新评论

1. Re:Storm Windowing storm..
我在评论区说一下滚窗的问题吧。
最明显的一个问题是，滚窗比较大的时候，机器处理不过会导致很多数据延时到达，所以当滚窗窗口大小比较大的时候最好要加上延时的处理。
--cutd

2. Re:Kaggle入门教程
您好，经试验，您的这条语句：accuracy = sum(predictions[predictions==titanic['Survived']])/(titanic['Survived'].cou.....
--贝叶斯的朴素

3. Re:Kaggle入门教程
您好，使用您的代码最后提交的csv文件准确率确实为75%，但是训练数据使用进行交叉验证时，得到的accuracy为26%，是不是accuracy计算出现了问题？
--贝叶斯的朴素

4. Re:Kaggle入门教程
5 转换 sex性别列的问题 titanic.loc[titanic['Sex']=='male','Sex'] = 0titanic.loc[titanic['Sex']=='female','Sex.....

- subtract数组一减二
- multiply相乘
- divide、floor_divide除法，向下取整
- power A B 计算A^B
- maximum、fmax fmax会忽略NaN
- minimum、fmin
- mod求模
- copysign将二数组的值的符号复制给第一个数组的值
- greater、greater_equal、less、less_equal、equal、not_equal比较运算，产生布尔型，相当于>,>=
- logical_and,logical_or,logical_xor & | ^

利用数组进行数据处理

- 用数组表达式代替循环叫矢量化
- numpy.where=x if condition else y
- [(x if c else y)for x,y,z in zip(x,y,c)]=np.where(c,x,y)

数学和统计方法

- 基本数组统计方法
  - sum 对数组中全部元素或某轴向的元素求和
  - mean 算术平均值
  - std、var 标准差和方差
  - min、max最大值和最小值
  - argmin、argmax最大最小元素的索引
  - cumsum所有元素的累计和
  - cumprod所有元素的累计积
- 用于布尔型数组的方法
  - sum计算true的个数
  - any测试是否存在一个或多个true
  - all 检测数组中是否所有值都是true
- 排序
  - sort
  - np.sort返回是已排序的副本
- 数组的结合运算
  - np.unique找出数组中的唯一值并返回已排序的结果
  - intersect1d(x,y)返回有序的公共元素结果

- union1d(x,y)返回并集的有序结果
- in1d(x,y)返回x的元素是否包含于y的布尔型数组
- setdiff1d(x,y)返回集合的差，x中不在y中
- setxor1d(x,y)异或存在某一但是不同时存在2者

数组文件的输入和输出

- 二进制读写
  - np.save
  - np.load
  - np.savez将多个数组保存到一个压缩文件中
- 读取文本文件
  - np.loadtxt
  - np.genfromtxt
  - np.savetxt

线性代数

- 矩阵乘法
  - np.dot(x,y)=x.dot(y)
- 常用函数
  - diag 返回矩阵的对角线元素，或将一维数组转换成矩阵
  - cot矩阵乘法
  - trace对角线元素和
  - det 矩阵行列式值
  - eig特征值和特征向量
  - inv求逆
  - pinv计算矩阵的伪逆
  - qrQR分解
  - svd奇异值分解
  - solve Ax=b的解
  - lstsq Ax=b的最小二乘解
- 随机数生成
  - numpy.random
  - seed确定随记生成数的种子
  - permutation返回一个序列的随记排列或一个随记排列的范围

- shuffle对一个序列直接随记排列
- rand产生均匀分布的样本值
- randint从给定的范围内随机选取整数
- randn产生标准状态分布的随机值
- binomial产生二项分布的样本值
- normal产生高斯分布的样本值
- beta产生B分布的样本值
- chisquare产生卡方分布的样本值
- gamma产生gamma分布的样本值
- uniform产生[0,1)均匀分布的样本值

5.pandas入门

pandas数据结构介绍

- Series (data,index=v)
  - 一组数据和一组索引组成的一维数组
  - 看成是一个定长的有序字典
  - 可以直接通过字典创建
  - 在数值运算中会自动对齐
  - 索引可以通过直接赋值的方式修改
- DataFrame
  - 表格型的数据结构
  - 直接传入由等长列表或np数组组成的字典
  - 可以指定列序列columns
  - 通过类似字典标记da.date或属性da['date']的方式获取一个列为Series(拥有原来df相同的索引)
  - 为不存在的列赋值时会创建一个新列
  - 嵌套字典的外层键作为列，内层作为行索引
  - 可以穿给DF的数据[二维ndarray、数组元组列表组成的字典、np的结构化数组、Series组成的字典、字典组成的字典、字典或Series组成的列表、列表或元组组成的列表、DF、np的MaskedArray]
  - index对象不可修改
- index的方法和属性：
  - append链接另外一个index产生新的index
  - diff计算差集得到一个index
  - intersection计算交集
  - union\isin计算是否包含在参数集合中的布尔型数组

- delete删除索引i出的元素并得到新的index
- drop、insert、is_monotonic、is_unique、unique

基本功能

- 重新索引reindex：创建一个适应新索引的新对象
- reindex的method选项
  - ffill pad前向填充
  - bfill backfill后向填充
- reindex函数的参数
  - index索引的新序列
  - method插值方式
  - fill_value替代缺失值的
  - limit向前向后的最大填充量
  - level、copy
- drop删除指定行或列data.drop ('two')
- 索引、选取、过滤
  - 利用标签的索引和普通的索引不同
  - obj[val]选取单列或者多列(布尔型、切片、布尔型df有奇效)
  - obj.ix[val]选取单个行或者一组行
  - obj.ix[:,val]选取单个列或列子集
  - obj.ix[val1,val2] 同时选取行和列
  - xs方法根据标签选取单行或者单列，并返回一个Series
  - icol、irow根据整数位置选取单行或者单列，并返回一个Series
  - get_value、set _value根据标签选取设置单个值
- 算术运算和数据对齐
  - 不重叠标签NA
  - 算术方法中可以填充值
  - DF和Series可以运算，沿行进行广播
  - apply方法可以将函数运用到列或行形成的一维数组上
  - applymap，Series.map
- 排序
  - sort_index默认是行索引升序(axis=1)列索引升序ascending=False降序
  - 按值对Series进行排序order，缺失值都在末尾
  - 给sort_index的by传名称即可按照相应的名字排
- 排名.rank()

    - 与排序对比会增设一个排名值
    - 相同名次以method解决
    - average默认平均化
    - min、max、first
- 索引也可以是重复的

汇总和计算描述统计

- 规约方法
    - axis df行用0，列用1
    - skipna跳过na值，默认是True
    - level层次化索引就根据level分组规约
    - describe返回多个列汇总信息

> count、describe、min、max、argmin、argmax、idxmin、idxmax、quantile、sum、mean、median、mad、var、std、skew、kurt、cumsum、cummin、cummax、cumprod、diff、pct_change

- 相关系数和协方差
    - 3.x只保留了一个 items() 方法
    - Series中corr用于计算相关系数[重叠、非NA、索引对齐]
    - cov计算协方差
    - df的cov、corr会返回完整的矩阵
    - df的corrwith计算其列或行和另一个Series或df
    - unique.sort()返回一组唯一值有序数组
    - value_counts()返回一个Series各值出现的频率，pd.

处理缺失数据

- NA处理方法
    - dropna、fillna
    - isnull、notnull
- 过滤缺失数据

    - dropna返回一仅含非空数据和索引值的Series=data.notnull()；对于df会丢弃任何含有na的行，传入how='all'只丢弃全为NA的行；丢弃列则传入axis=1
    - thresh参数
- 填充缺失数据
    - fillna方法参数
        - value用于填充的值或字典对象
        - method填充方法，默认ffill
        - axis默认0即行，axis=1为列
        - inplace是否产生副本
        - limit填充最大连续数量

层次化索引

- 能以低维度形式处理高维度数据
- 可以通过unstack方法重新排到一个df中[stack逆运算]
- 还可以为轴标签指定名称
- 重排分层排序
    - swaplevel可以交换两个层级并返回新的
    - sortlevel
- df将一个列或多个当做行索引
    - set_index(['c','d'),drop=False)
    - reset_index()

pandas的其他话题

- 整数索引
    - Series的iget_value
    - df的irow和icol
- Panel数据
    - Panel中的每一项都是一个df
    - df有to_ panel方法[逆运算是to_frame]

6.数据加载存储和文件格式

读取文本格式的数据

- pandas解析函数

    - read_csv从文件、url、文件型对象加载带分隔符的对象，默认分隔符是逗号，
    - read_table同上，默认分隔符是制表符\t；指定分隔符sep=','=read_csv
    - read_fwf读取定宽列格式数据
    - read_clipboard读取剪切板数据
- read_csv/read_table
    - 可以指定索引和列名，也可传入列名列表做成多层索引
    - 当处理不固定分隔符时使用正则表达式来作为分隔符
    - skiprows跳行
    - na_values接收用于表示缺失值的字符串

> 函数参数：path、sep|delimiter、header、index_col、names、skiprows、na_values、comment、parse_dates、keep_date_col、converters、dayfirst、date_parser、nrows、iterator、chunksize、skip_footer、verbose、encoding、squeeze、thousands

- 逐块读取文本文件
    - nrows指定读取几行
    - chunksize指定逐块读取的大小
- 将数据写出到文本文件
    - to_csv可以指定分隔符[from_csv]
    - 缺失值默认是空字符串，可以通过na_rep指定标记值
    - 默认会输出行列索引，可以通过index=False，header=False禁用
    - 也可以只输出部分列
- 手动处理分隔符格式
    - csv
- JSON 数据
    - json.load加载json数据
    - json.dump转换为json对象
    - pandas.to_json[from_json]
- XML、HTML
    - findall和XPath
    - py2.x中的urllib2 =py3.x 中的urllib.request

- The StringIO and cStringIO modules are gone. Instead, import the io module and use io.StringIO or io.BytesIO for text and data respectively.
- ixml.objectify解析xml

二进制数据格式

- pandas.save和pandas.load 读写pickle形式数据
- HDF5格式(hierarchical data format层次数据格式)
  - python中有两个接口PyTables&h5py
  - 处理海量数据要好好研究这两个接口
  - pd.ExcelFile读取Excel文件

使用HTML和Web API

- json、request
- df便于分析

使用数据库

> In python 2, zip returned a list. In python 3, it returns an iterable object. But you can make it into a list just by calling list on it.
> list(zip(*ngram))[0]=zip(*nagram)[0]

存取MongoDB的数据

7.数据规整化：清洗、转换、合并、重塑

合并数据集

- pandas内置方法合并
  - pandas.merge根据一个或多个键连接不同的df，实现数据库的连接操作
  - pandas.concat沿一条轴合并多个对象
  - combine_first将重复数据接在一起
- pd.merge(df1，df2，on='key')

    - 不指定哪个列进行连接，默认是重叠的列名进行连接
    - 两个对象的列名不同可以分别指定
    - 默认情况merge是how='inner'结果中的键是交集，outer是并集，还有left、right
    - 多对多连接产生的是行的笛卡尔积
    - 要对多个键进行合并传入一个键的列表即可
    - merge函数参数
        - left、right、how、on、left_on、right_on
        - left_index、right_index、sort、suffixes、copy
- 索引上的合并
    - 层次索引必须以列表的形式指明用作合并键的多个列
    - df.join按索引实现合并并且合并多个带有相同或相似的df对象；还可以传入一组df
- 轴向连接
    - concat默认在axis=0工作，将值和索引连接到一起
    - 如果传入axis=1则结果会变成一个df
    - concat函数参数
        - objs参与连接的pd对象的列表或字典，唯一必须参数
        - axis、join、join_axes、keys、levels
        - names、verify_integrity、ignore_index
- 合并重叠数据

    - np.where&pd.combine_first

重塑reshape和轴向旋转pivot

- 重塑层次化索引
    - stack 列-->行 df-->Series 默认滤除缺失值
    - unstack 行-->列 Series-->df
- 将长格式转换成宽格式

    - pivot

数据转换

- 移除重复数据
    - df.duplicated()返回一个布尔型Series表示是否重复行
    - drop_duplicates返回一个移除了重复行的df；默认是判断全部列，也可以指定列；默认保留第一个值，也可以保留最后一个
- 利用匿名函数或映射进行数据转换

    - map&lambda
- 替换值
    - replace
- 重命名轴索引
    - map直接修改原始数据，rename创建数据集的转换版[可以结合字典实现对部分轴索引的修改]，也可inplace=True修改原数据
- 离散化和面元划分
    - 离散化函数pd.cut&pd.qcut
- 检测和过滤离群值
    - np.random.permutation
    - df.take
- 计算指标/哑变量
    - 将分类变量转换为虚拟矩阵或指标矩阵
    - pd.get_dummies(prefix加前缀)结合cut

字符串操作

- 字符串对象方法
    - split()结合strip(修剪空白符，换行符)
    - '::'.join()
    - find[找不到返回-1]和index[找不到会引发异常]
    - count返回子字符串出现的次数
    - replace将指定字符替换成指定字符，删除就替换空字符
    - 内置字符串方法
        - count、endswith、startswith、join、index、find、rfind、replace、strip、rstrip、lstrip、split、lower、upper、ljust、rjust
- 正则表达式
    - 通过re.compile创建regex对象可以节省大量时间如果对许多字符串应用同一个正则表达式
    - findall返回所有匹配项的列表，finditer逐个迭代返回
    - search返回第一个匹配项
    - match从字符串起始位置开始匹配，返回第一个，否则None
    - sub将匹配到的替换成指定字符串，并返回新的字符串subn前n个
    - re.IGNORECASE忽略大小写
    - split将匹配到的拆分成数段
- pandas中矢量化的字符串函数
    - 获取矢量化的元操作:str.get;str[]

- 矢量化的字符串方法
  - cat、contains、count、endswith、startswith、findall、get、join、len、lower、upper、match、pad、center、repeat、replace、slice、split、strip、rstrip、lstrip

8.绘图和可视化

matplotlib入门

- matplotlib的实例库和文档是成为绘图高手的最佳资源
- Figure & Subplot
  - matplotlib的图像都位于Figure对象中
  - pyplot.subplots的参数
    - nrows、ncols、sharex、sharey、subplot_kw、 **fig_wk
    - subplots_adjust调整图像间距
- 颜色、标记和线型
  - plot(linestyle=、color=)常用颜色有缩写，任意RGB
  - 转折点的标记marker=o；drawstyle插值绘图方式
- 刻度、标签和图例
  - 设置刻度和刻度标签
    - set_xticks选择要设置刻度的位置
    - ser_xticklables就是设置刻度的标签
    - set_xlable设置轴标签
    - set_title设置标题
  - 添加图例
    - 在添加subplot的时候传入lable
    - ax.legend|plt.legend(loc='best')自动选一个最好的地方
- 注解或在Subplot上绘图
  - 注解可以通过text、arrow、annotate添加
  - text可以文本绘制在指定坐标
  - 在图表上添加一个图形，需要先创建一个块对象shp然后通过ax.add_oatch(shp)将其添加到subplot中
- 图片保持Figure.savefig
  - fname、dpi、facecolor、edgecolor、format、bbox_inches
- matplotlib配置
  - plt.rc函数配置，第一个参数是要配置的对象

pandas中的绘图函数

- 线形图
  - Series.plot方法默认就是线形图
    - label、ax、style、alpha、kind、logy、use_index、rot、xticks、yticks、xlim、ylim、grid
  - df.plot会在一个subplot中为各列绘制一条线并自动创建图例
    - subplots、sharex、sharey、figsize、title、legend、sort_columns
  - 要更深入需要多学matplotlib API
- 柱状图
  - kind='bar'垂直|kind='barh'水平
  - Series索引会被用作刻度=df.行索引，列索引会作分组
  - stacked=True堆积柱状图
- 直方图和密度图
  - hist生成直方图
  - plot kind='kde'生成密度图
  - 二者通常一起使用
- 散布图
  - plt.scatter观察两个一维数据序列之间的关系
  - pd.scatter_matrix散布图矩阵
  - basemap地图插件
  - 图形库mayavi

9.数据聚合与分组运算

GroupBy分组

- split-apply-combine
- 分组键中的缺失值可以使结果包含在NA组了吧
- 对分组进行迭代
- 选取一个或一组列
- 通过字典或Series分组，索引和分组轴要对齐
- 通过函数进行分组
- 将函数、数组、列表、字典、Series混合使用进行分组
- 根据索引级别分组[层次化索引]

数据聚合：从数组产生标量值的数据转换过程

- 如果要使用自己的聚合函数，传入aggregate和agg方法
- 非聚合运算的describe方法也可用

- 优化过的GroupBy方法
  - count、sum、mean、median、std、var、min、max、prod、first、last
- 面向列的多函数应用
  - 不同的列使用不同的函数或一次应用多个函数
  - 如果传入的是函数或者函数名，相应的列就会以函数名命名
  - 如果传入的是元组(name,function)就会以第一个参数名命名
  - 如果要对不同的列使用不同的函数，那么就向agg传入一个从列名映射到函数的字典
- as_index=False结果返回是无索引的

数组运算和转换

- groupby的transform方法，会将一个函数运用到各个分组
- apply：一般性的'拆分-应用-合并'
- group_keys=False禁止分组键
- 分位数quantile和桶bucket分析

透视表和交叉表

- 透视表pivot table根据一个或多个键并根据行、列键将数据分配到各个举行区域里
  - pd.pivot_table|df.pivot_table参数
    - margins=True aggfunc= 、values、index、columns、fill_value
- 交叉表crosstab：用于计算分组频率的特殊透视表
  - crosstab前两参数可以是数组、Series、数组列表

关于basemap的种种问题

- geos始终无法安装好
- basemap无法直接安装whl也安装不了
- win10 64 py3.5

时间序列

日期和时间数据类型及工具

- datetime、time、calendar
- date.timedelta表示两个datetime对象之间的时间差

- 字符串和datetime之间的相互转换
  - str和strftime可以将datetime转换成字符串
  - datetime.strptime可以将格式化字符串装换成datetime对象；解析已知格式
  - dateutil包的parser.parse方法解析所有格式；实用但不完美
  - datetime格式定义
    - %Y 四位数年 %y %m %d 两位数
    - %H 24H制 %I 12H制 %M %S
    - %w 星期几[0,6] %U 每年的第几周，星期天为界，%W，星期一为界
    - %z UTC时区偏移量 %F Y-m-d %D m/d/y
  - 特定于当前环境的
    - %a 星期几简称、%A星期全称
    - %b月份简称 %B
    - %c完整日期和时间 %p am，pm
    - %x适合于当前地区的日期格式，%X时间格式

时间序列基础

- pd最基本的时间序列就是以时间戳为索引的Series
- 索引、选取、子集构造
  - 传入可以被解析成日期的字符串
  - 传入年、月即可选取数据切片
  - 可以使用字符串日、datetime、Timestamp
  - 对非唯一时间戳进行聚合使用groupby

日期的范围、频率、移动

- resample将时间序列转换成一个具有固定频率的时间序列
- pd.date_range会生成指定长度的daretimeindex
- 时间序列的基础频率
  - D日历日 B工作日 H T分 S WOM每月的星期几
- 移动数据 pd.shift

时区处理

- pytz包
- 本地化和转换

    - tz_localize、tz_convert转换到别的时区
    - Timestamp对象的转换
    - 不同时区之间的运算
- 时期及其算术运算
    - 时期的频率转换
    - 按季度计算的时期频率
        - to_timestamp
    - 将Timestamp转换成Period[or相反]
        - to_period方法
    - 通过数组创建PeriodIndex

重采样和频率转换

- 重采样resample是指将时间序列从一个频率转换到另外一个频率
    - 降采样、升采样，非降非升采样
    - 方法参数
        - freq、how、axis、fill_method、closed、label、loffset、limit、kind、convention
        - OHLC重采样，open、high、low、close
        - 通过groupby进行重采样
    - 升采样和插值
    - 通过时期进行重采样

时间序列绘图

- pd时间序列的绘图功能比mt好
- 移动窗口函数rolling_mean
- 用户自定义移动窗口函数rolling_apply，能从片段中产生单个值

性能和内存使用的注意事项

- pandas对数据对齐和重采样进行了高度优化
- 规则频率之间的转换优化

金融和经济数据应用

数据规整方面

- 时间序列以及截面对齐
  - pandas可以在算术运算中自动对齐数据
  - 通过一组不同索引的Series构建df
- 频率不同的时间序列的运算
  - 频率转换resample和重对齐reindex
  - 使用Period索引的两个不同频率的时间序列之间的运算必须进行显示转换
- 时间和当前数据选取
  - at_time、between_time
  - 将Timestamp传入asof可以得到时间点最近的值(若是na的话)
- 拼接多个数据源
  - pd.concat
  - df.combine_first

分组变换和分析

- 分组因子暴露
  - 因子分析是投资组合定量管理的一种技术
- 十分位和十分位分析
  - 基于样本分位数的分析

NumPy高级应用

ndarray对象的内部机制

- numpy数据类型体系

高级数组操作

- 向数组的reshape传入一个表示新形状的元组即可
  - 扁平化|散开
- C[行优先&内存相邻]和Fortran[列优先&内存相邻]顺序
- order='C' || order='F'
- 数组的拆分和合并
  - np.concatenate可以指定轴将一个数组序列(元组或列表等)连接到一起
  - 比较方便的是np.vstack|np.hstack
  - np.split将数组沿指定轴拆分为多个数组
  - concatenate、vstack、row_stack、hstack、column_stack、dstack、split、hsplit、vsplit、dsplit

    - r_、c_
- 元素的重复操作tile&repeat
- 花式索引的等价函数take&put
    - take可以使用axis put只能按C顺序

广播

- 后缘维度的轴长相符，其中一方长度为1
- 沿其他轴向广播
- 利用广播机制设置数组的值

ufunc高级应用

- ufunc实例方法
    - reduce、accumulate、reduceat、outer
- 自定义ufunc
    - np.frompyfunc接收一数组个函数及参数

结构化和记录式数组

- 定义结构化dtype，使用元组列表，字典式访问
- 嵌套dtype和多维字段

和排序有关的话题

- ndarray的sort实例方法是直接排序，原始数据会消失
- 而np.sort会创建一个已排序的副本，还可以指定排序轴
    - 二者都无法设置为降序
- 间接排序：argsort、lexsort
- python对象数组只能用快排排序
- np.searchsorted在有序数组中查找元素，返回位置坐标
- Numpy的Matrix类

高级输入输出

- 内存映像文件可以处理内存放不下的大文件

- np.memmap会将大文件分成小段来读写

性能建议

- 将python的循环和逻辑转换成数组运算和布尔数组运算
- 多用广播
- 多使用数组切片避免复制数据
- 使用ufunc
- 考虑Cython
- 连续内存

分类: Python&Machine Learning

cutd
关注 - 3
粉丝 - 10
+加关注

0 0

« 上一篇： 电脑莫名其妙的被装上了流氓软件DNSunlocker的解决办法
» 下一篇： 《BuildingMachineLearningSystemsWithPython》学习笔记

posted @ 2016-04-06 23:48 cutd 阅读(3271) 评论(0) 编辑 收藏

最新IT新闻:
· Adobe推出VR编辑器Project CloverVR，方便VR内容制作
· 李开复：AlphaGo证明AI进化速度远比人类想象的快
· Adobe在MAX大会上发布的这些黑科技，使其股价暴涨近10%
· Keep你是不是膨胀了！为什么不让我安安静静的假装运动？
· Cortana智能音箱Harman Kardon Invoke将于10月22日上市
» 更多新闻...

最新知识库文章:
· 实用VPC虚拟私有云设计原则
· 如何阅读计算机科学类的书
· Google 及其云智慧
· 做到这一点，你也可以成为优秀的程序员
· 写给立志做码农的大学生
» 更多知识库文章...