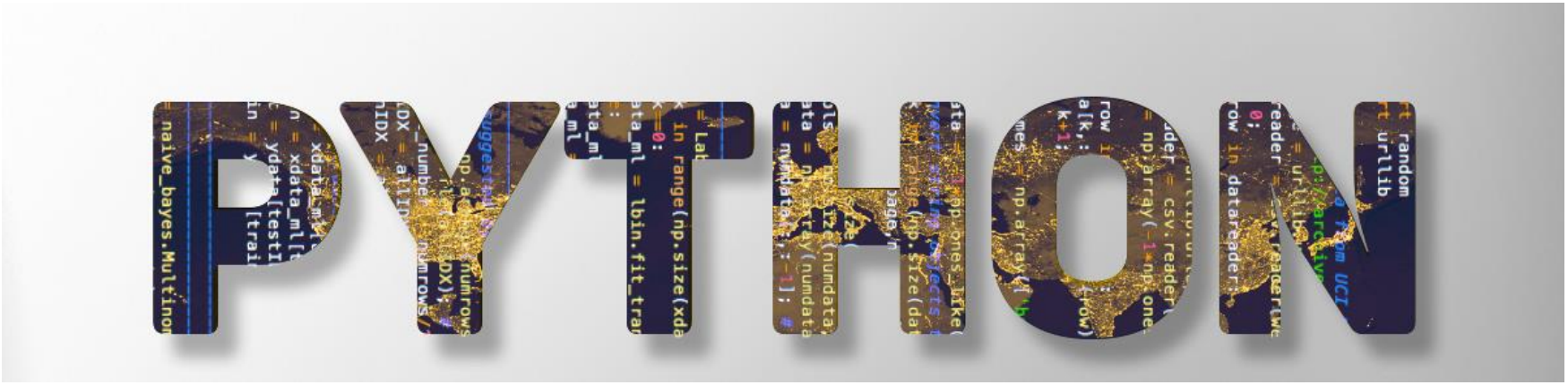
法律声明

- 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

第一讲



工作环境准备及
数据分析建模理论基础

--梁斌

目录

- 课程介绍
- Python语言基础及Python3.x新特性
- 使用NumPy和SciPy进行科学计算
- 数据分析建模理论基础
- 实战案例：科技工作者心理健康数据分析

目录

课程介绍

面向人群：

1. 想了解和学习典型的数据分析流程和实践方法的学习者
2. 想接触和学习非结构化数据(比如：文本、图像等)分析的学习者
3. 想学习数据分析中常用建模知识的相关从业人员
4. 尚不会使用Python的数据分析师从业者
5. 想转行从事数据分析师行业的学习者
6. 想使用Python实现机器学习的工程师

课程介绍

课程目标：

1. 熟悉数据分析的流程，包括数据采集、处理、可视化等
2. 掌握Python语言作为数据分析工具，从而有能力驾驭不同领域的数据分析实践
3. 掌握非结构化数据的处理与分析
4. 快速积累多个业务领域的数据分析项目经验
5. 掌握使用Python实现基于机器学习的数据分析和预测
6. 掌握数据分析中常用的建模知识

课程介绍

升级内容：

1. 使用最新版本的Python 3.x作为分析工具
2. 新增数据分析常用的建模知识
3. 新增使用Python处理和分析时间序列数据
4. 新增使用Python进行文本数据分析
5. 新增使用Python进行图像数据处理及分析
6. 升级全部随课项目，并提供更详细的分析步骤

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

课程介绍 什么是数据分析?

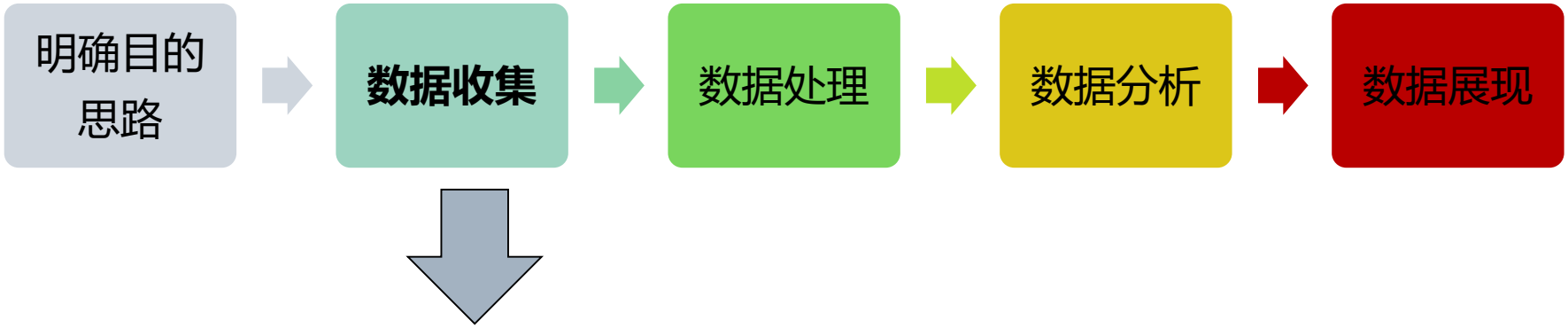
什么是数据分析?

Analysis of data is a process of **inspecting**, **cleansing**, **transforming**, and **modeling** data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making.

-- WIKEPIDIA

课程介绍

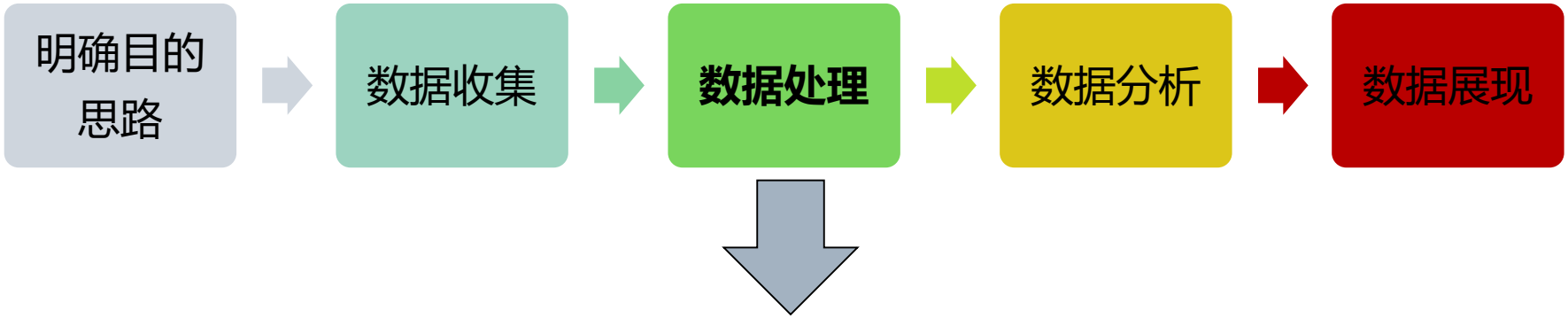
基本步骤：



第二课：数据采集与操作

课程介绍

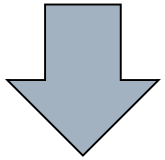
基本步骤：



第三课：数据分析工具Pandas

课程介绍

基本步骤：



第三课：数据分析工具Pandas

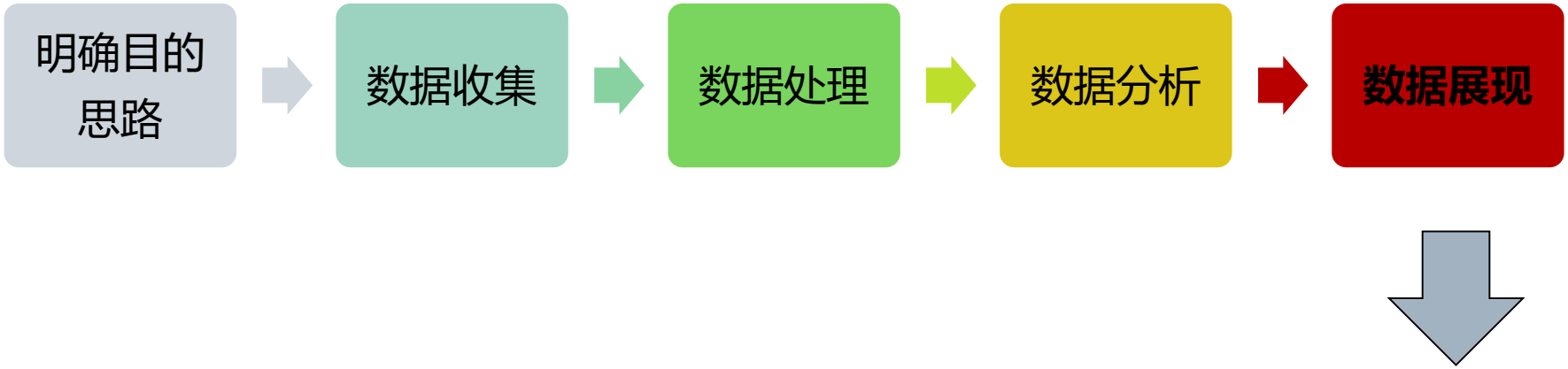
第五课：时间序列数据分析

第六课：文本数据分析

第七课：图像数据处理及分析

课程介绍

基本步骤：



第四课：数据可视化

第五、六、七课

第八课：机器学习基础及机器学习库scikit-learn

第九课: 通过移动设备行为数据预测使用者的性别和年龄

目录

- 课程介绍
- Python语言基础及Python3.x新特性
- 使用NumPy和SciPy进行科学计算
- 数据分析建模理论基础
- 实战案例：科技工作者心理健康数据分析

Python基础

- **Python环境**

  Anaconda是一个集成了大量常用扩展包的环境，避免单独安装时需要配置或兼容等各种问题

  Anaconda https://www.continuum.io/downloads

  64位、Python3.5 版本

- **Python包管理**

  安装：pip install xxx, conda install xxx

  卸载：pip uninstall xxx, conda uninstall xxx

  升级：pip install –upgrade xxx, conda update xxx

  详细用法： https://pip.pypa.io/en/stable/reference/

Python基础

- **Python虚拟环境**

  Virtualenv: https://virtualenv.pypa.io/en/stable/userguide/

  conda 虚拟环境: https://conda.io/docs/using/envs.html

- **多版本Python管理**

  conda管理: https://conda.io/docs/py2or3.html

- **IDE**

  Jupyter notebook

  1. Anaconda自带，无需单独安装
  2. 记录思考过程，实时查看运行过程
  3. 基于web的在线编辑器（本地）

Python基础

- **IDE（续）**

  Jupyter notebook

  4. .ipynb文件分享
  5. 可交互式
  6. 记录历史运行结果
  7. 支持Markdown，Latex

  IPython

  1. Anaconda自带，无需单独安装
  2. Python的交互式命令行 Shell
  3. 可查看历史操作
  4. 及时验证想法

Python基础

- IDE -- 没有最好的，只有最适合自己的（以下选一个就可以）

PyCharm社区版，部分免费，可满足不涉及web的开发，适合大多数开发者

https://www.jetbrains.com/pycharm/download/

Eclipse + PyDev，完全免费，适合熟悉Eclipse或Java的开发者

1. Eclipse，https://eclipse.org/downloads/
2. PyDev插件，https://marketplace.eclipse.org/content/pydev-python-ide-eclipse

Spyder，完全免费，适合熟悉Matlab的开发者

https://github.com/spyder-ide/spyder

Python3.x新特性

- print()，是函数，不是一个语句
- Python3对文本和二进制数据做了更为清晰的区分。

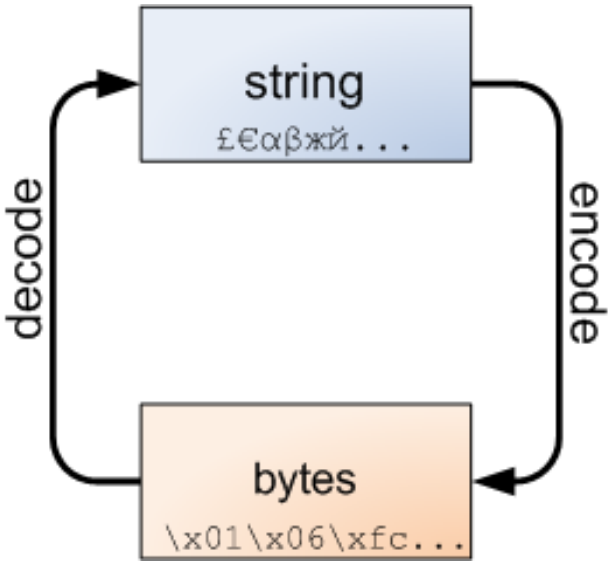
  文本由unicode表示，为str类型;

  二进制数据由bytes（字节包)表示，为bytes类型
- 新增数据类型 bytes（字节包），代表二进制数据以及被编码的文本

  字符串前有个前缀“b”
- Python3中 bytes与str转换

  str可以编码成bytes

  bytes可以解码成str



示例代码： lect01_py3.ipynb

Python3.x新特性

- 字符串格式化输出
  - 新增format()方式
- dict类型变化
  - 删除之前的iterkeys(), itervalues(), iteritems()
  - 改为keys(), values(), items()

示例代码： lect01_py3.ipynb

目录

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`

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`

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： lec01_np.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： lec01_np.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： lec01_np.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： `lec01_np.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

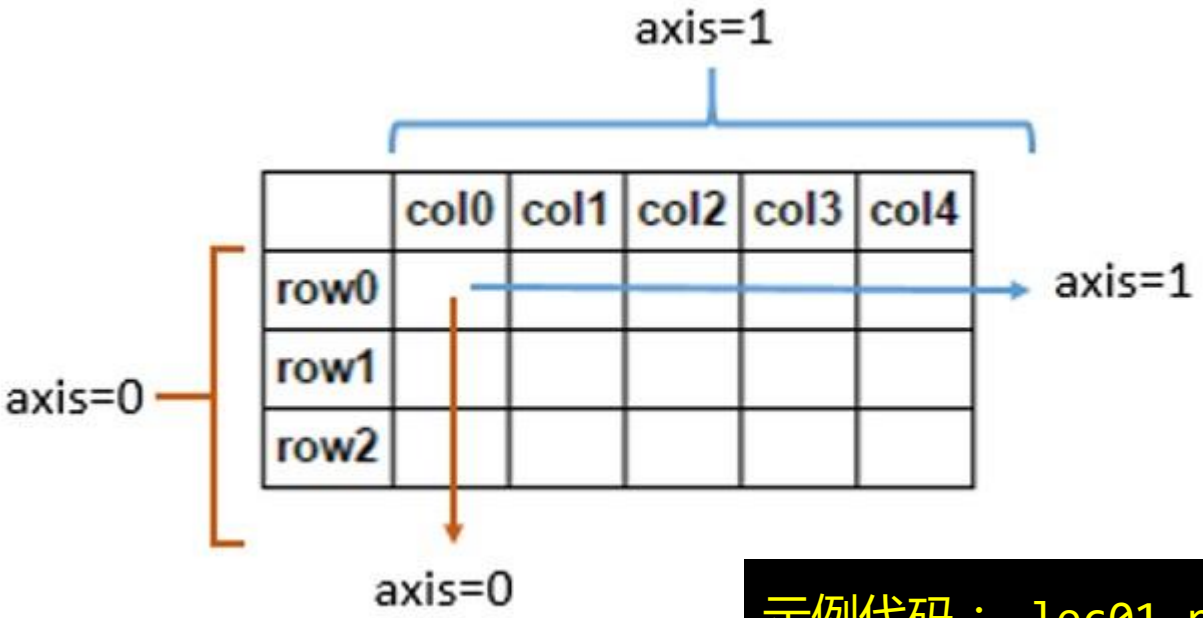
8

示例代码： lec01_np.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： lec01_np.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： lec01_np.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： `lec01_np.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

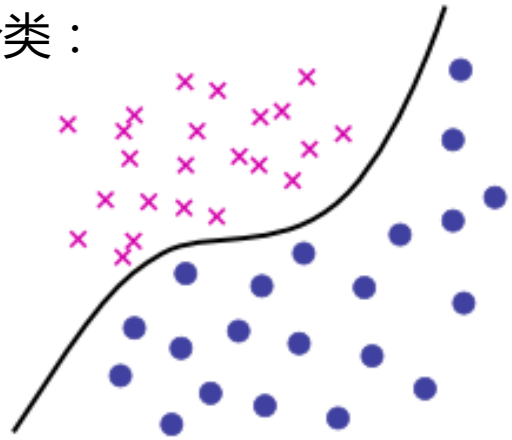
- 找到唯一值并返回排序结果

示例代码： lec01_np.ipynb

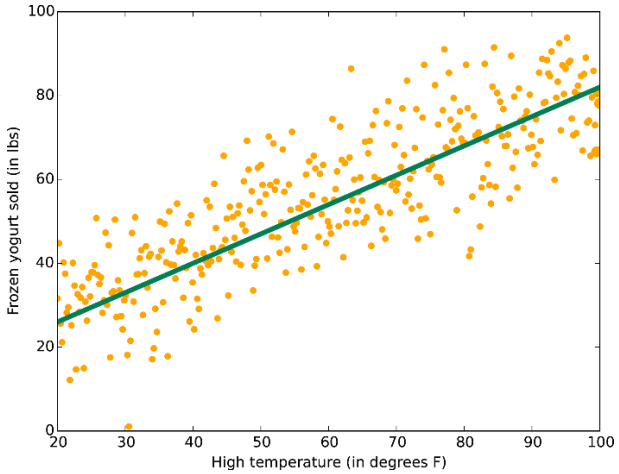
目录

建模基础

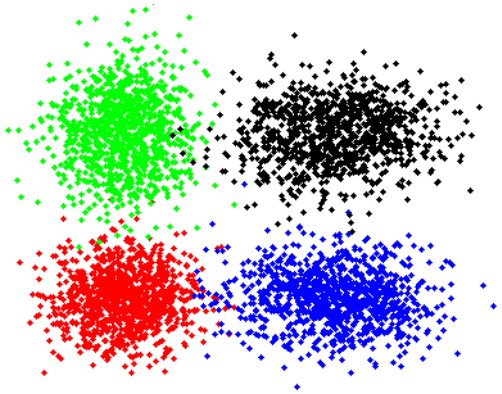
任务分类：



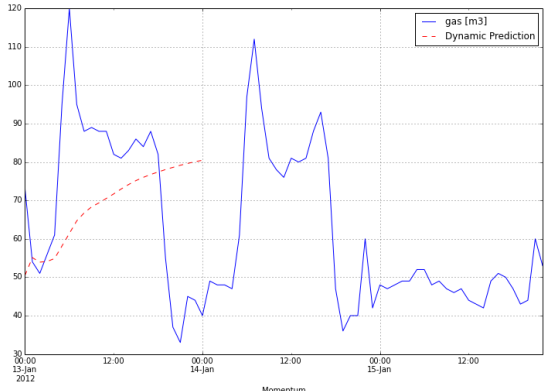
分类



回归



聚类



时序分析

建模基础

分类与回归

- 应用：信用卡申请人风险评估、预测公司业务增长量、预测房价等
- 原理：

  分类，将数据映射到预先定义的群组或类。算法要求基于数据属性值来定义类别，把具有某些特征的数据项映射到给定的某个类别上。

  回归，用属性的历史数据预测未来趋势。算法首先假设一些已知类型的函数可以拟合目标数据，然后利用某种误差分析确定一个与目标数据拟合程度最好的函数。
- 区别：分类模型采用离散预测值，回归模型采用连续的预测值。

建模基础

聚类

- 应用：根据症状归纳特定疾病、发现信用卡高级用户、根据上网行为对客户分群从而进行精确营销等
- 原理：

  在没有给定划分类的情况下，根据信息相似度进行信息聚类。

  聚类的输入是一组未被标记的数据，根据样本特征的距离或相似度进行划分。划分原则是保持最大的组内相似性和最小的组间相似性。

建模基础

时序模型

- 应用：下个季度的商品销量或库存量是多少？明天用电量是多少？
- 原理：

  描述基于时间或其他序列的经常发生的规律或趋势，并对其建模。

  与回归一样，用已知的数据预测未来的值，但这些数据的区别是变量所处时间的不同。重点考察数据之间在时间维度上的关联性。

目录

实战案例

项目名称：科技工作者心理健康数据分析

项目地址：https://www.kaggle.com/osmi/mental-health-in-tech-survey

项目任务：

统计各国家男性、女性心理健康数据分布

课后任务：

统计各国家存在心理健康问题的平均年龄

参考

- pip用法

  https://pip.pypa.io/en/stable/reference/
- Virtualenv用法

  https://virtualenv.pypa.io/en/stable/userguide/
- conda虚拟环境

  https://conda.io/docs/using/envs.html
- 多版本Python管理

  https://conda.io/docs/py2or3.html
- 字符串和编码

  http://www.liaoxuefeng.com/wiki/001431608955726 4a6b348958f449949df42a6d3a2e542c000/001431664106267f12e9bef7ee14cf6a8776a479bdec9b9000

参考

- 格式化字符串format函数

  http://blog.csdn.net/handsomekang/article/details/9183303
- 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》

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