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

第四讲

数据可视化

--梁斌

目录

- Matplotlib
- Seaborn
- 交互式数据可视化—Bokeh
- Logistic Regression
- 项目案例：世界高峰数据可视化

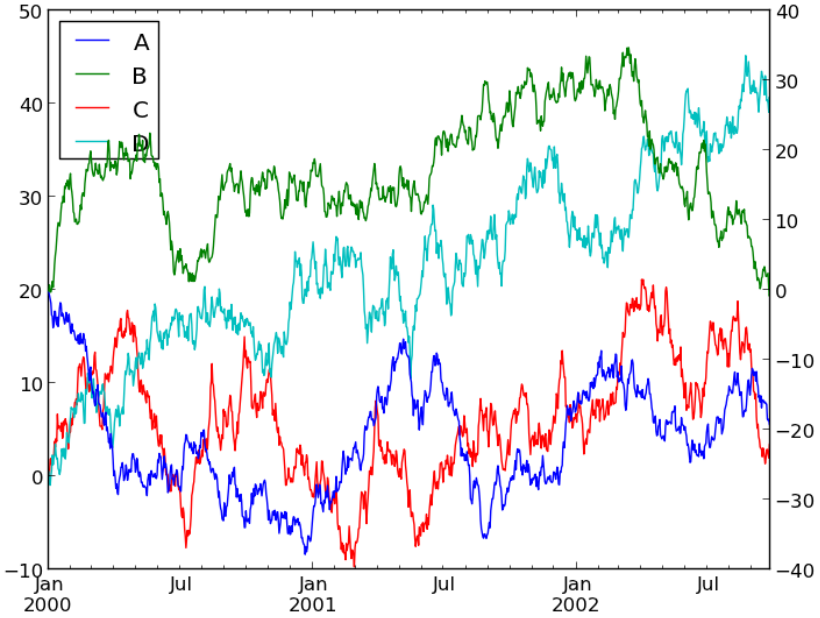
目录

- Matplotlib
- Seaborn
- 交互式数据可视化—Bokeh
- Logistic Regression
- 项目案例：世界高峰数据可视化

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()

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

示例代码： 01_matplotlib.ipynb

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：

01_matplotlib.ipynb,

01_matplotlib.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：

01_matplotlib.ipynb,

01_matplotlib.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：

01_matplotlib.ipynb,

01_matplotlib.py

Matplotlib

颜色、标记、线型

- ax.plot(x, y, ‘r--’)
  - 等价于ax.plot(x, y, linestyle= ‘--’, color= ‘r’)

颜色

- b: blue
- g: green
- r: red
- c: cyan
- m: magenta
- y: yellow
- k: black
- w: white

标记

marker	description
"."	point
","	pixel
"o"	circle
"v"	triangle_down
"^"	triangle_up
"<"	triangle_left

线型

linestyle	description
'-' or 'solid'	solid line
'--' or 'dashed'	dashed line
'-.' or 'dashdot'	dash-dotted line
':' or 'dotted'	dotted line
'None'	draw nothing
' '	draw nothing
''	draw nothing

示例代码： 01_matplotlib.ipynb

Matplotlib

刻度、标签、图例

- 设置刻度范围
  - plt.xlim(), plt.ylim()
  - ax.set_xlim(), ax.set_ylim()
- 设置显示的刻度
  - plt.xticks(), plt.yticks()
  - ax.set_xticks(), ax.set_yticks()
- 设置刻度标签
  - ax.set_xticklabels(), ax.set_yticklabels()
- 设置坐标轴标签
  - ax.set_xlabel(), ax.set_ylabel()

示例代码： 01_matplotlib.ipynb

Matplotlib

刻度、标签、图例 (续)

- 设置标题
  - ax.set_title()
- 图例
  - ax.plot(label= ‘legend’)
  - ax.legend(), plt.legend()
    - loc= ‘best’ 自动选择放置图例最佳位置

matplotlib设置

- plt.rc()
- http://matplotlib.org/users/customizing.html

示例代码： 01_matplotlib.ipynb

目录

- Matplotlib
- Seaborn
- 交互式数据可视化—Bokeh
- Logistic Regression
- 项目案例：世界高峰数据可视化

Seaborn

什么是Seaborn

- Python中的一个制图工具库，可以制作出吸引人的、信息量大的统计图
- 在Matplotlib上构建，支持numpy和pandas的数据结构可视化，甚至是scipy和statsmodels的统计模型可视化

特点

- 多个内置主题及颜色主题
- 可视化单一变量、二维变量用于比较数据集中各变量的分布情况
- 可视化线性回归模型中的独立变量及不独立变量

Seaborn

特点 (续)

- 可视化矩阵数据，通过聚类算法探究矩阵间的结构
- 可视化时间序列数据及不确定性的展示
- 可在分割区域制图，用于复杂的可视化

安装

- conda install seaborn
- pip install seaborn

Seaborn

数据集分布可视化

- 单变量分布 sns.distplot()
  - 直方图 sns.distplot(kde=False)
  - 核密度估计 sns.distplot(hist=False) 或 sns.kdeplot()
  - 拟合参数分布 sns.distplot(kde=False, fit=)
- 双变量分布
  - 散布图 sns.jointplot()
  - 二维直方图 Hexbin sns.jointplot(kind= ‘hex’)
  - 核密度估计 sns.jointplot(kind= ‘kde’)
- 数据集中变量间关系可视化 sns.pairplot()

示例代码： 02_seaborn.ipynb

Seaborn

类别数据可视化

- 类别散布图
  - sns.stripplot() 数据点会重叠
  - sns.swarmplot() 数据点避免重叠
  - hue指定子类别
- 类别内数据分布
  - 盒子图 sns.boxplot(), hue指定子类别
  - 小提琴图 sns.violinplot(), hue指定子类别
- 类别内统计图
  - 柱状图 sns.barplot()
  - 点图 sns.pointplot()

示例代码： 02_seaborn.ipynb

目录

Bokeh

什么是Bokeh

- 专门针对Web浏览器的交互式、可视化Python绘图库
- 可以做出像D3.js简洁漂亮的交互可视化效果

特点

- 独立的HTML文档或服务端程序
- 可以处理大量、动态或数据流
- 支持Python (或Scala, R, Julia...)
- 不需要使用Javascript

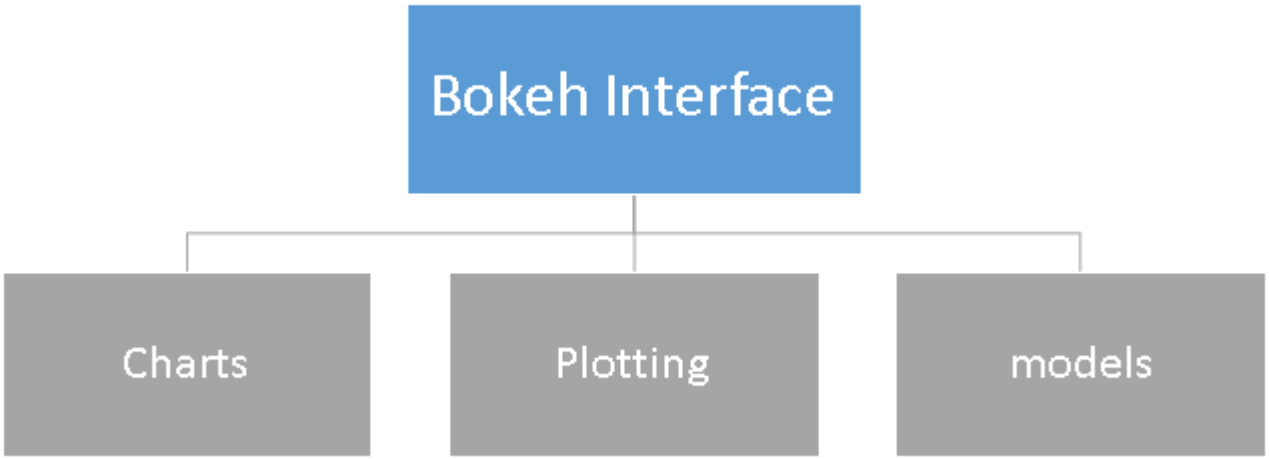
安装

- conda install bokeh
- pip install bokeh

Bokeh

Bokeh接口

- Charts: 高层接口，以简单的方式绘制复杂的统计图
- Plotting: 中层接口，用于组装图形元素
- Models: 底层接口，为开发者提供了最大的灵活性



Bokeh

包引用

- from bokeh.io import output_file 生成.html文档
- from boken.io import output_notebook 在jupyter中使用

bokeh.charts

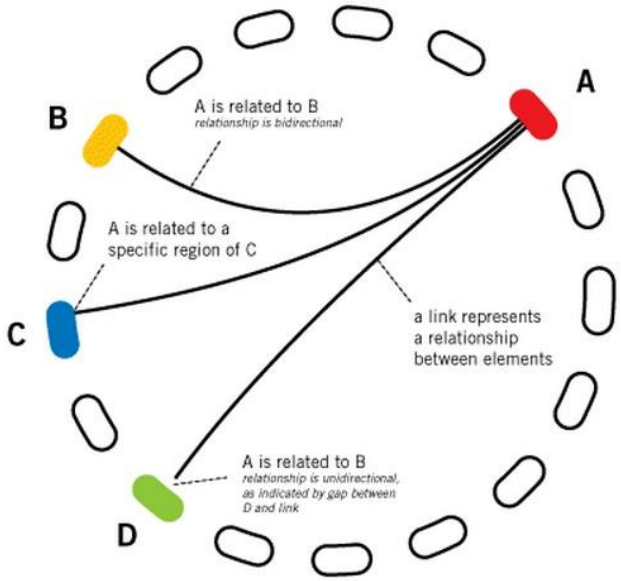
- http://bokeh.pydata.org/en/latest/docs/reference/charts.html
- 散点图 Scatter
- 柱状图 Bar
- 盒子图 BoxPlot
- ...

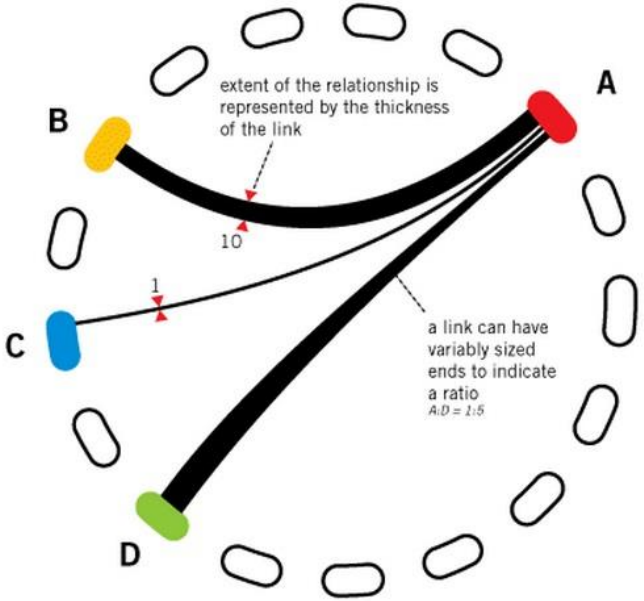
示例代码： 03_bokeh.ipynb

Bokeh

bokeh.charts (续)

- 弦图 Chord
  - 展示多个节点之间的联系
  - 连线的粗细代表权重





示例代码： 03_bokeh.ipynb

Bokeh

bokeh.plotting

- 方框 square
- 圆形 circle
- …
- 更多图形元素参考
  http://bokeh.pydata.org/en/latest/docs/reference/plotting.html

示例代码： 03_bokeh.ipynb

目录

Logistic 回归

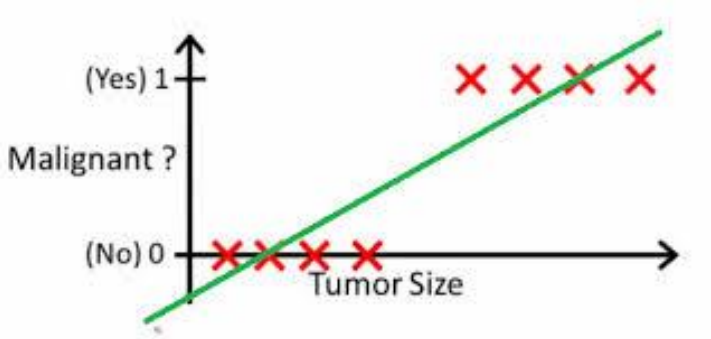
概率(probability)

- 定义：对一件事情发生可能性的衡量
- 范围：0<=p<=1
- 条件概率

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Logistic Regression（逻辑回归）

- 例子

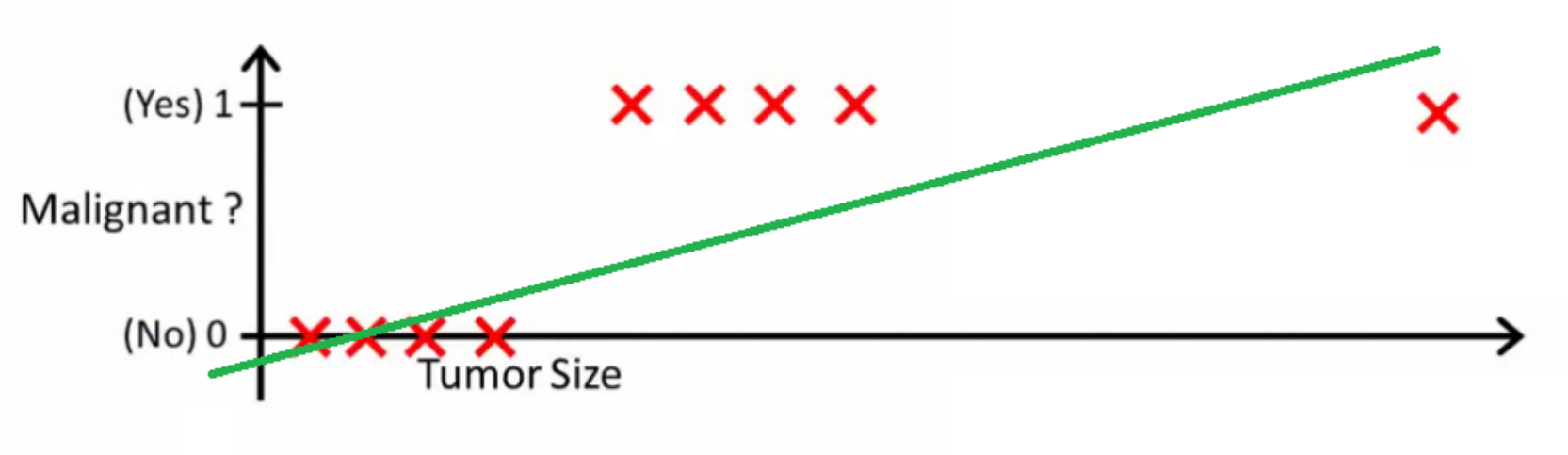


h(x) > 0.5 or h(x) < 0.5

Logistic 回归

Logistic Regression（逻辑回归）（续）

- 例子



$h(x) > 0.2$ or $h(x) < 0.2$

Logistic 回归

基本模型

- 训练样本为： X（x_0，x_1，x_2，…，x_n）
- 学习的参数为：θ（θ_0，θ_1，θ_2，…，θ_n）

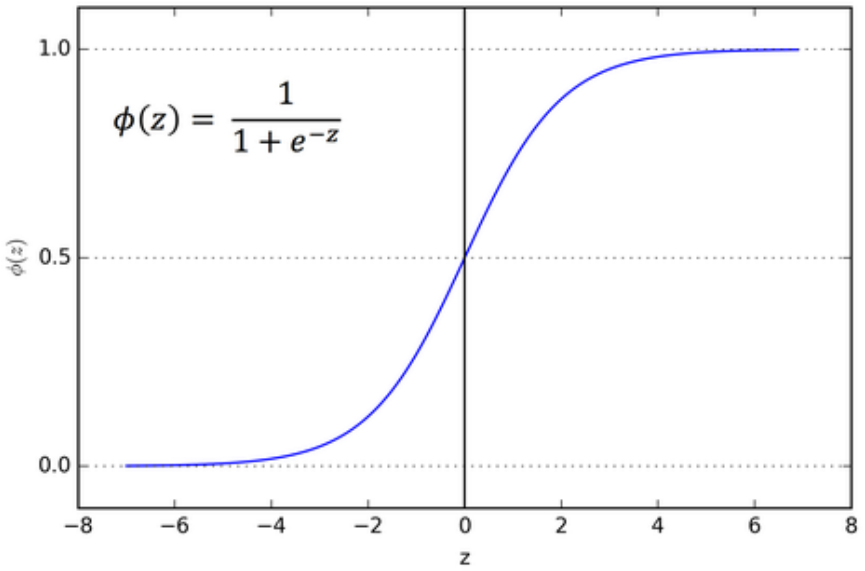
$$Z = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

- 向量表示

$$Z = \Theta^T X$$

- Sigmoid函数将线型转换成非线性

$$g(Z) = \frac{1}{1+e^{-Z}}$$



Logistic 回归

基本模型（续）

- 预测函数 $$h_\theta(X) = g(\Theta^T X) = \frac{1}{1+e^{-\Theta^T X}}$$
- 用概率的形式表示
  - 正样本 $$h_\theta(X) = P(y=1 \mid X;\Theta)$$
  - 负样本 $$1-h_\theta(X) = P(y=0 \mid X;\Theta)$$
- 损失函数

$$Cost(h_\Theta(X),y) = \begin{cases} -\log(h_\Theta(X)) \_ \text{when\_y=1} \\ -\log(1-h_\Theta(X)) \_ when \_ y=0 \end{cases}$$

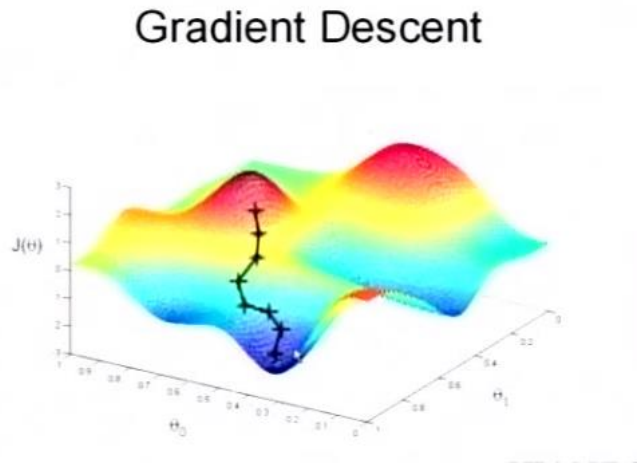
$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(h_\Theta(x^{(i)}), y^{(i)}) = -\frac{1}{m}[\sum_{i=1}^{m}(y^{(i)}\log(h_\Theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\Theta(x^{(i)})))]$$

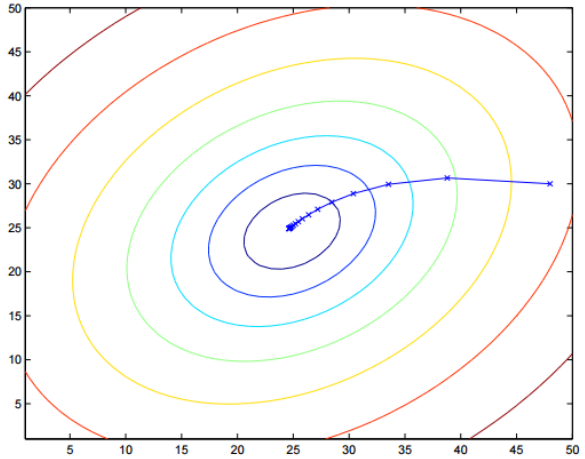
- 目标：通过训练样本求出参数theta使损失函数最小化

Logistic 回归

基本模型（续）

- 解法：梯度下降（gradient descent）





$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), (j = 0 \ldots n)$$

更新方式：

$$\theta_j = \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}, (j = 0 \ldots n)$$

- alpha：学习率
- 同时更新所有theta
- 迭代更新直到收敛

示例代码： lect04_lr

小象学院 ChinaHadoop.cn

目录

项目案例

项目介绍

- 项目地址：https://www.kaggle.com/abcsds/highest-mountains

项目任务

- 可视化高峰数据

涉及知识点

- Matplotlib 数据可视化

示例代码： lect04_proj

参考

- Matplotlib示例库
  http://matplotlib.org/gallery.html
- Matplot线型
  http://matplotlib.org/api/lines_api.html#matplotlib.lines.Line2D.set_linestyle
- Matplotlib标记
  http://matplotlib.org/api/markers_api.html
- Seaborn教程
  http://seaborn.pydata.org/tutorial.html
- 利用Seaborn可视化数据分布
  http://seaborn.pydata.org/tutorial/distributions.html

参考

- Bokeh教程

  http://nbviewer.jupyter.org/github/bokeh/bokeh-notebooks/blob/master/tutorial/00%20-%20intro.ipynb

- Logistic Regression 模型简介

  http://tech.meituan.com/intro_to_logistic_regression.html

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