机器学习

第1章 绪论

朱桂祥 (9120201070@nufe.edu.cn)
南京财经大学信息工程学院
江苏省电子商务重点实验室
电子商务信息处理国家级国际联合研究中心
电子商务交易技术国家地方联合工程实验室

https://github.com/zgx881205/Machine-Learning/tree/main

目录

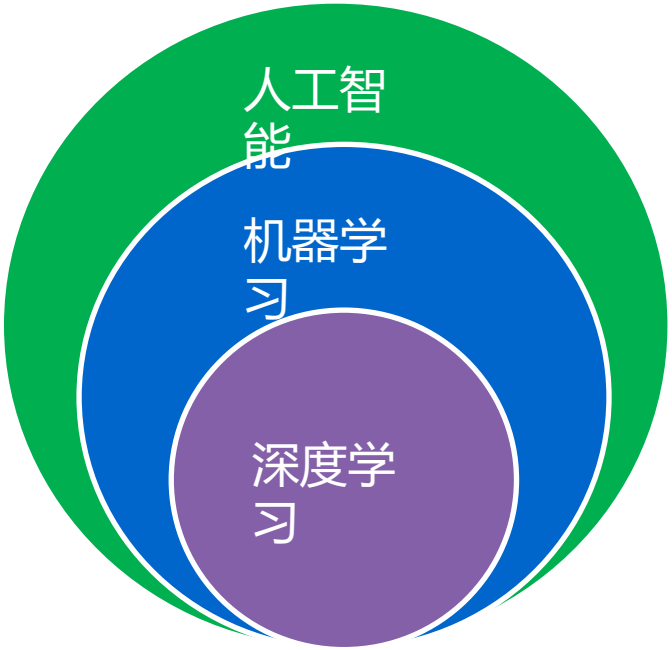
1. 机器学习概述

机器学习与人工智能、深度学习的关系

人工智能： 机器展现的人类智能

机器学习： 计算机利用已有的数据(经验)，得出了某种模型，并利用此模型预测未来的一种方法。

深度学习： 实现机器学习的一种技术

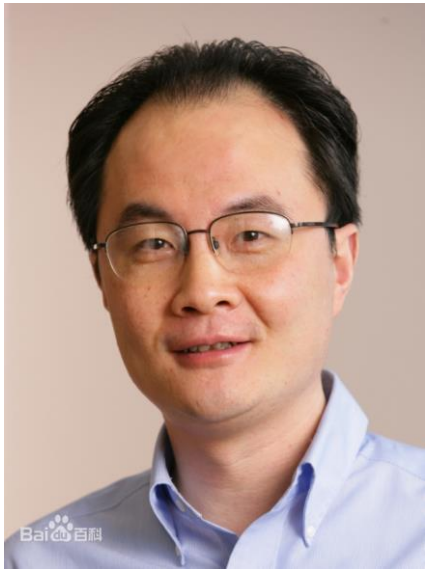


机器学习界的执牛耳者

杨立昆（Yann LeCun）
杰弗里·欣顿（Geoffrey Hinton）
本吉奥（ Bengio ）
共同获得了2018年计算机科学的最高奖项——**ACM图灵奖**。

Andrew Ng
中文名**吴恩达**，斯坦福大学副教授，前“百度大脑”的负责人与百度首席科学家。

机器学习界的国内泰斗



李航, 现任字节跳动科技有限公司人工智能实验室总监，北京大学、南京大学客座教授，IEEE 会士，ACM 杰出科学家，CCF 高级会员。
代表作：《统计学习方法》

周志华，南京大学计算机科学与技术系主任 、人工智能学院院长。
代表作：《机器学习》（西瓜书）

机器学习界的青年才俊

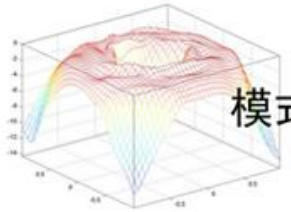
陈天奇,陈天奇是机器学习领域著名的青年华人学者之一，本科毕业于上海交通大学ACM班，博士毕业于华盛顿大学计算机系。
主要贡献：设计了XGBoost算法。

何恺明，本科就读于清华大学，博士毕业于香港中文大学多媒体实验室。2016年，加入Facebook AI Research（FAIR）担任研究科学家。
主要贡献：设计了ResNets

国内外知名人工智能企业榜单

编码	企业名称	人工智能技术	应用领域	所属国家	成立时间	资本市场状态	市值/估值/融资额
1	Microsoft（微软）	计算机视觉技术、自然语言处理技术等	办公	美国	1975年	上市	市值1.21万亿美元
2	Google（谷歌）	计算机视觉技术、自然语言处理技术等	综合	美国	1998年	上市	市值9324亿美元
3	Facebook（脸书）	人脸识别、深度学习等	社交	美国	2004年	上市	市值5934亿美元
4	百度	计算机视觉技术、自然语言处理技术、知识图谱等	综合	中国	2001年	上市	市值438亿美元
5	大疆创新	图像识别技术、智能引擎技术等	无人机	中国	2006年	战略融资	估值210亿美元
6	商汤科技	计算机视觉技术、深度学习	安防	中国	2014年	D轮融资	估值70亿美元
7	旷视科技	计算机视觉技术等	安防	中国	2011年	D轮融资	估值40亿美元
8	科大讯飞	智能语音技术	综合	中国	1999年	上市	市值108亿美元
9	Automation Anywhere	自然语言处理技术、非结构化数据认知	企业管理	美国	2003年	B轮融资	估值68亿美元
10	IBM Watson（IBM沃森）	深度学习、智适应学习技术	计算机	美国	1911年	上市	市值1198亿美元
11	松鼠AI 1对1	智适应学习技术、机器学习	教育	中国	2015年	A轮融资	估值11亿美元
12	字节跳动	跨媒体分析推理技术、深度学习、自然语言处理、图像识别	资讯	中国	2012年	Pre-IPO轮融资	估值750亿美元
13	Netflix（网飞）	视频图像优化、剧集封面图片个性化、视频个性化推荐	媒体及内容	美国	1997年	上市	市值1418亿美元
14	Graphcore	智能芯片技术、机器学习	芯片	英国	2016年	D轮融资	估值17亿美元
15	NVIDIA（英伟达）	智能芯片技术	芯片	美国	1993年	上市	市值1450亿美元
16	Brainco	脑机接口	教育、医疗、智能硬件	美国	2015年	天使轮融资	融资额600万美元
17	Waymo	自动驾驶	交通	美国	2016年	C轮融资	估值1050亿美元
18	ABB Robotics	机器人及自动化技术	机器人	瑞士	1988年	上市	市值514亿美元
19	Fanuc（发那科）	机器人技术	制造	日本	1956年	上市	市值362亿美元
20	Preferred Networks	深度学习、机器学习技术	物联网	日本	2016年	C轮融资	估值20亿美元

机器学习的范围



模式识别

计算机视觉

数据挖掘

机器学习

语音识别

统计学习

自然语言处理

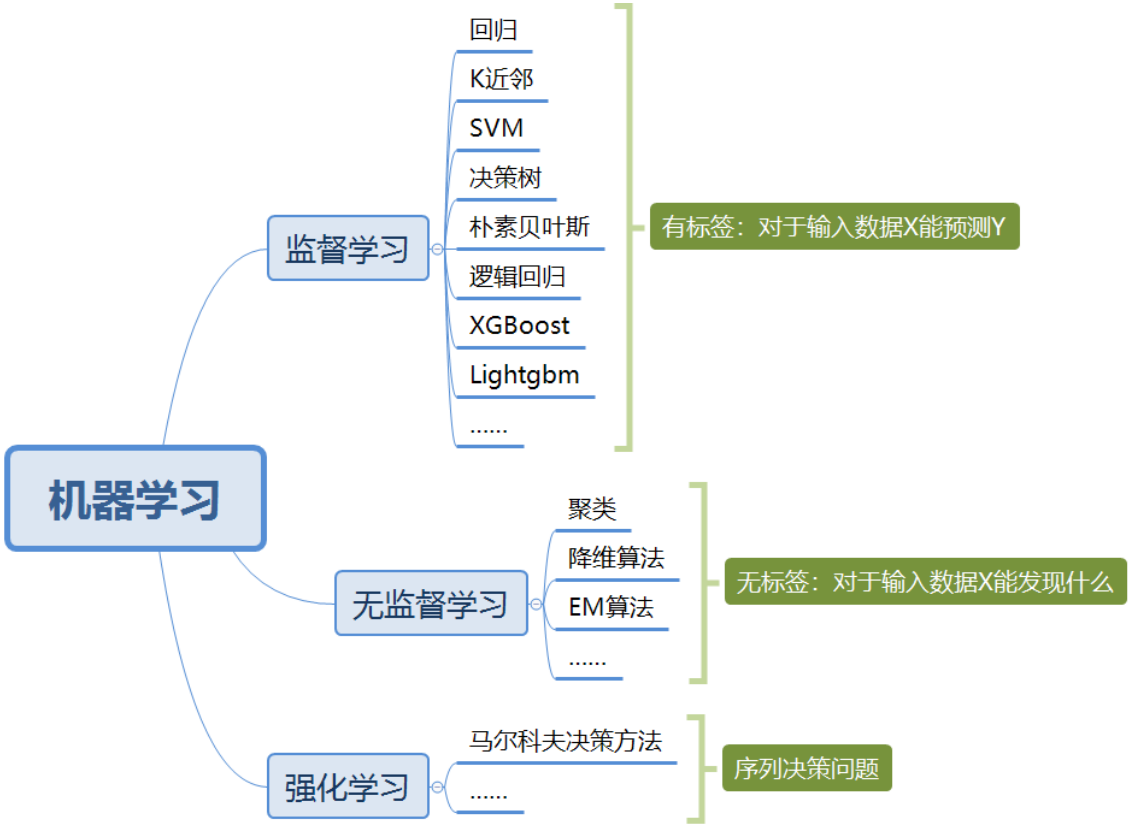
机器学习可以解决什么问题

- 给定数据的预测问题
  - ✓ 数据清洗/特征选择
  - ✓ 确定算法模型/参数优化
  - ✓ 结果预测

- 不能解决什么
  - ✓ 大数据存储/并行计算
  - ✓ 做一个机器人

2. 机器学习的类型

2. 机器学习的类型


机器学习
监督学习
回归
K近邻
SVM
决策树
朴素贝叶斯
逻辑回归
XGBoost
Lightgbm
.....
有标签：对于输入数据X能预测Y
无监督学习
聚类
降维算法
EM算法
.....
无标签：对于输入数据X能发现什么
强化学习
马尔科夫决策方法
.....
序列决策问题

2. 机器学习的类型-监督学习

- ✓分类（Classification）
  - ✓身高1.65m，体重100kg的男人肥胖吗?
  - ✓根据肿瘤的体积、患者的年龄来判断良性或恶性?
- ✓回归（Regression、Prediction）
  - ✓如何预测上海浦东的房价?
  - ✓未来的股票市场走向?

2. 机器学习的类型-无监督学习

- ✓聚类（Clustering）
  - ✓如何将教室里的学生按爱好、身高划分为5类?
- ✓降维（ Dimensionality Reduction ）
  - ✓如何将将原高维空间中的数据点映射到低维度的空间中?

2. 机器学习的类型-强化学习

- ✓强化学习（Reinforcement Learning）
  - ✓用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题 。

3. 机器学习的背景知识

3. 机器学习的背景知识-希腊字母

大写	小写	英文注音	国际音标注音	中文注音
Α	α	alpha	alfa	阿耳法
Β	β	beta	beta	贝塔
Γ	γ	gamma	gamma	伽马
Δ	δ	deta	delta	德耳塔
Ε	ε	epsilon	epsilon	艾普西隆
Ζ	ζ	zeta	zeta	截塔
Η	η	eta	eta	艾塔
Θ	θ	theta	θita	西塔
Ι	ι	iota	iota	约塔
Κ	κ	kappa	kappa	卡帕
Λ	λ	lambda	lambda	兰姆达
Μ	μ	mu	miu	缪
Ν	ν	nu	niu	纽
Ξ	ξ	xi	ksi	可塞
Ο	ο	omicron	omikron	奥密可戎
∏	π	pi	pai	派
Ρ	ρ	rho	rou	柔
∑	σ	sigma	sigma	西格马
Τ	τ	tau	tau	套
Υ	υ	upsilon	jupsilon	衣普西隆
Φ	φ	phi	fai	斐
Χ	χ	chi	khai	喜
Ψ	ψ	psi	psai	普西
Ω	**ω**	**omega**	**omiga**	欧米

3. 机器学习的背景知识-数学基础

高等数学

导数、微分、泰勒公式……

线性代数

向量、矩阵、行列式、秩、线性方程组、特征值和特征向量……

概率论与数理统计

随机事件和概率、概率的基本性质和公式、常见分布、期望、协方差……

3. 机器学习的背景知识-Python基础

Python 的环境的安装

- Anaconda
- Jupyter notebook
- Pycharm

详细教程：https://zhuanlan.zhihu.com/p/59027692

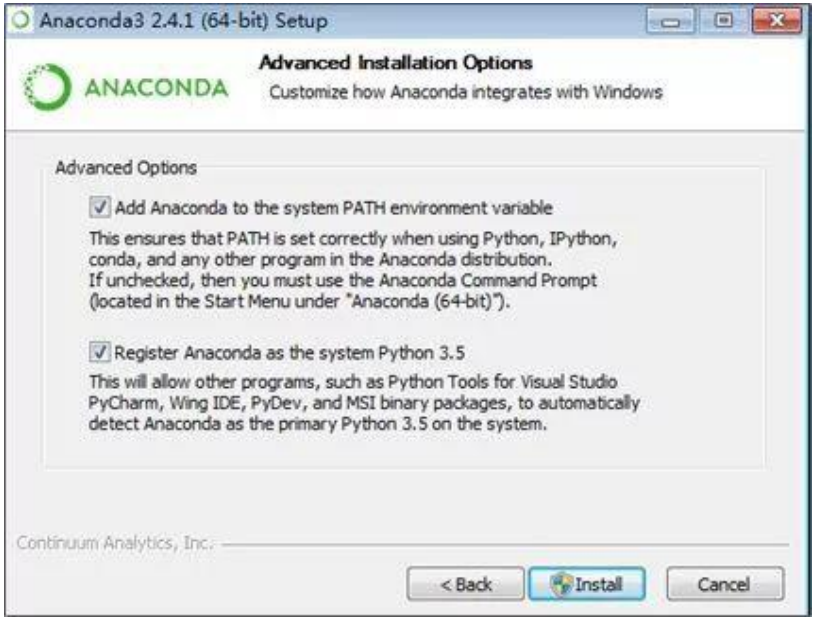
Python 的环境的安装

- Anaconda

https://www.anaconda.com/distribution/

通常选3.7版本，64位

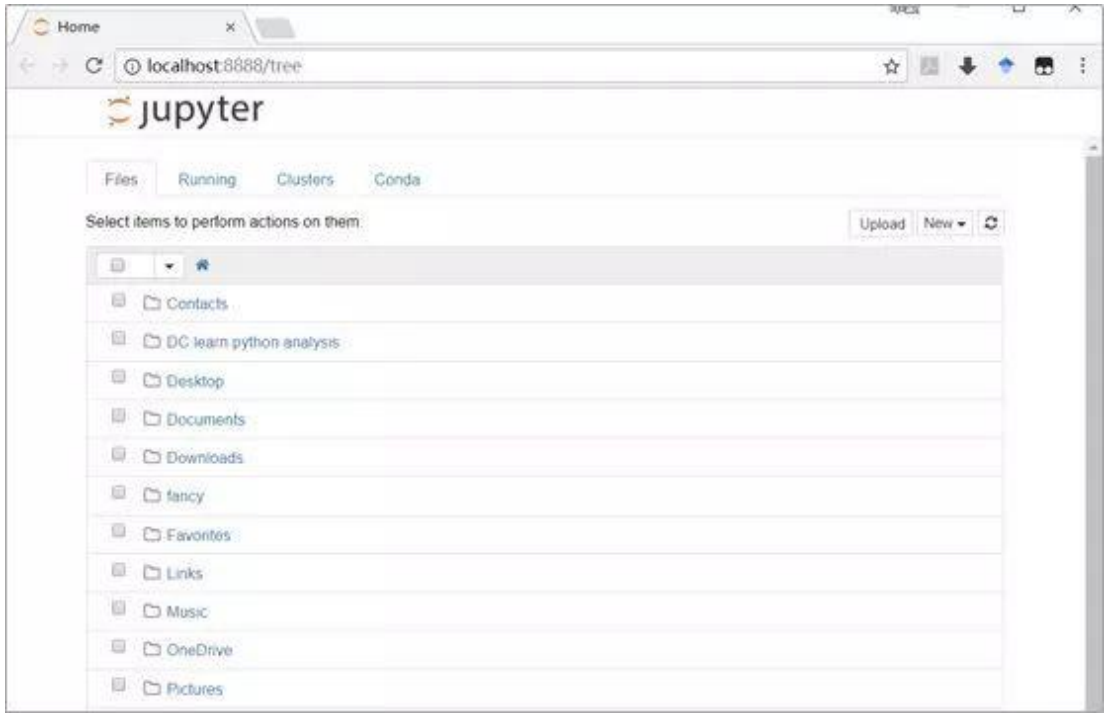
可以用默认安装，右图两个选择框都勾上



Python 的环境的安装

- Jupyter notebook

在cmd环境下，切换到代码的目录，输入命令：
jupyter notebook之后就可以启动jupyter botebook编辑器，启动之后会自动打开浏览器，并访问http://localhost:8088，默认跳转到http://localhost:8088/tree



Python 的环境的安装

- Pycharm

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

Pycharm 提供 免费的社区版 与 付费的专业版。专业版额外增加了一些功能，如项目模板、远程开发、数据库支持等。个人学习 Python 使用免费的社区版已足够。
如果有edu邮箱，那么推荐使用专业版，edu邮箱是可以免费使用专业版的。

安装过程照着提示一步步操作就可以了。
注意：安装路径尽量不使用带有 中文或空格 的目录，这样在之后的使用过程中减少一些莫名的错误。

Python 的主要数据类型

- 字符串
- 整数与浮点数
- 布尔值
- 日期时间
- 其它

Python 的数据结构

●列表(list)

用来存储一连串元素的容器，列表用[ ]来表示，其中元素的类型可不相同。

●元组(tuple)

元组类似列表，元组里面的元素也是进行索引计算。列表里面的元素的值可以修改，而元组里面的元素的值不能修改，只能读取。元组的符号是( )

●集合(set)

集合主要有两个功能，一个功能是进行集合操作，另一个功能是消除重复元素。 集合的格式是：set( )，其中()内可以是列表、字典或字符串，因为字符串是以列表的形式存储的

●字典(dict)

字典dict也叫做关联数组，用大括号{ }括起来，在其他语言中也称为map，使用键-值（key-value）存储，具有极快的查找速度，其中key不能重复。

Python控制流

- 顺序结构
- 分支结构
- 循环结构
- break、continue和pass
- 列表生成式

Python函数

- **调用函数**

  调用内置函数

- **定义函数**

  def 函数名():
  函数内容
  <return 返回值>

- **高阶函数**

  匿名函数：高阶函数传入函数时，不需要显式地定义函数，直接传入匿名函数更方便（lambda函数）

Python模块

- numpy
- pandas
- scipy
- matplotlib
- scikit-learn

Python模块-NumPy

●numpy

Numpy是一个用python实现的科学计算的扩展程序库，包括：

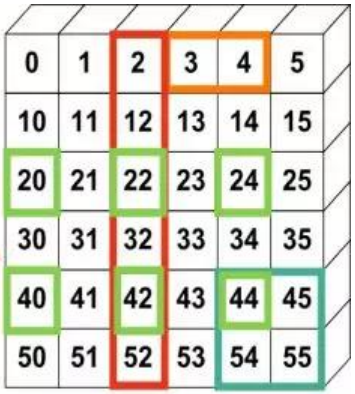
1、一个强大的N维数组对象Array；
2、比较成熟的（广播）函数库；
3、用于整合C/C++和Fortran代码的工具包；
4、实用的线性代数、傅里叶变换和随机数生成函数。numpy和稀疏矩阵运算包scipy配合使用更加方便。

NumPy（Numeric Python）提供了许多高级的数值编程工具，如：矩阵数据类型、矢量处理，以及精密的运算库。专为进行严格的数字处理而产生。多为很多大型金融公司使用，以及核心的科学计算组织如：Lawrence Livermore，NASA用其处理一些本来使用C++，Fortran或Matlab等所做的任务。

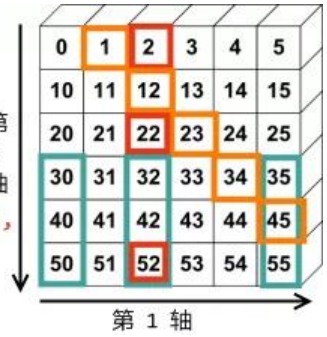
Python模块-NumPy

切片

```
>>> a[0,3:5]
array([3,4])
>>> a[4:,4:]
array([[44,45],[54,55]])
>>> a[:,2]
array([2,12,22,32,42,52])
>>> a[2::2,::2]
array([[20,22,24],
       [40,42,44]])
```

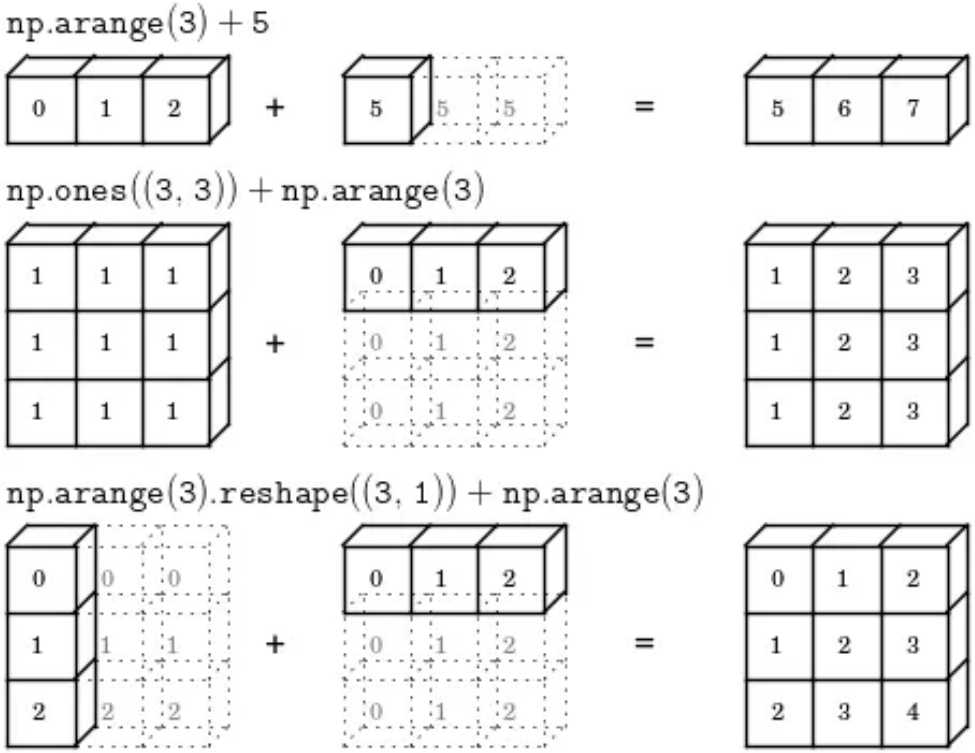



```
>>> a[(0,1,2,3,4),(1,2,3,4,5)]
array([1,12,23,34,45])
>>> a[3:,[0,2,5]]
array([[30,32,35],
       [40,42,45],
       [50,52,55]])
>>> mask=np.array([1,0,1,0,0,1],
          dtype=np.bool)
>>> a[mask,2]
array([2,22,52])
```



Python模块-NumPy

广播



Python模块-pandas

- **pandas**

Pandas 是基于NumPy 的一种工具，该工具是为了解决数据分析任务而创建的。

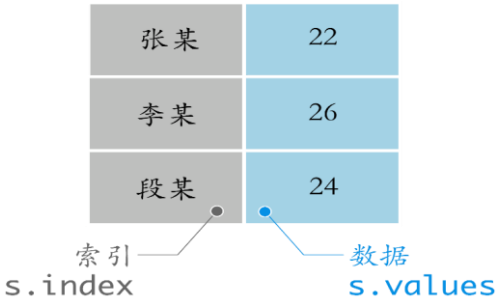
Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。Pandas提供了大量能使我们快速便捷地处理数据的函数和方法。你很快就会发现，它是使Python成为强大而高效的数据分析环境的重要因素之一。

Python模块-pandas

- **基本数据结构**

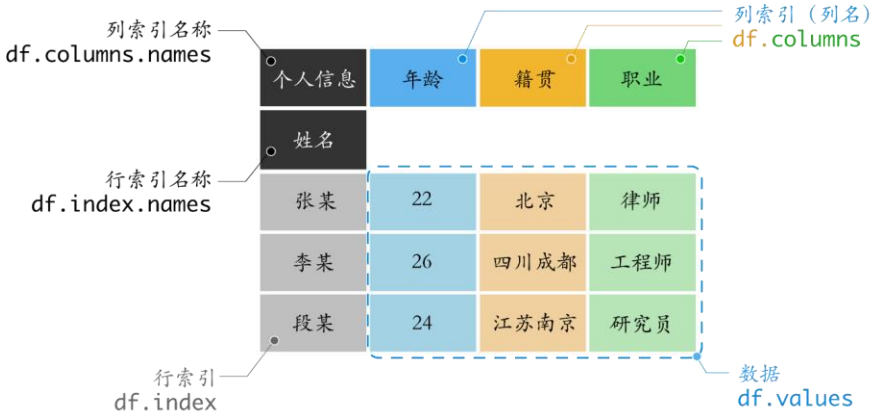
Series

一维数据结构，包含行索引和数据两个部分



DataFrame

二维数据结构，包含带索引的多列数据，各列的数据类型可能不同



Python模块-pandas

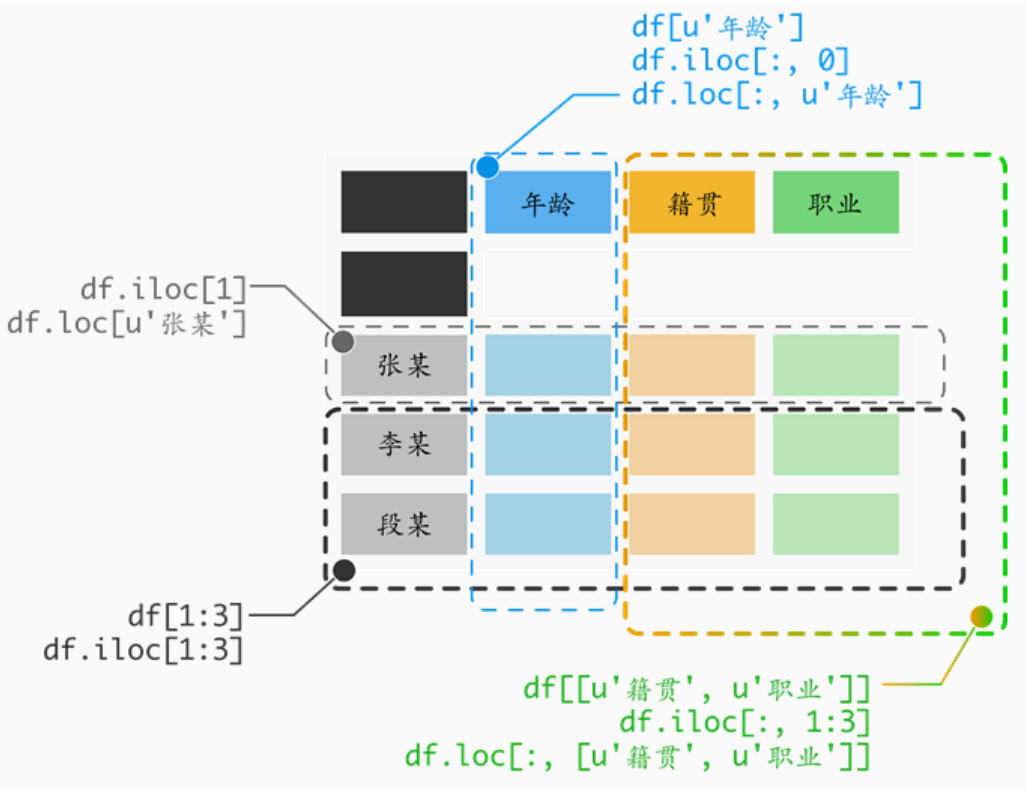
- 数据索引

df[5:10]
通过切片方式选取多行.

df[col_label] or df.col_label
选取列.

df.loc[row_label, col_label]
通过标签选取行/列.

df.iloc[row_loc, col_loc]
通过位置（自然数）选取行/列
.



Python模块-Pandas

● 数据合并

`pd.merge(left, right)` 类数据库的数据融合操作.

参数：`how`，融合方式，包括左连接、右连接、内连接（默认）和外连接；`on`，连接键；`left_on`，左键；`right_on`，右键；`left_index`，是否将left行索引作 为左键；`right_index`，是否将right行索引作为右键.

	姓名	年龄
0	张某	22
1	李某	26
2	段某	24

left

	姓名	籍贯
7	张某	北京
8	李某	四川成都
9	钱某	江苏南京

right

inner

	姓名	年龄	籍贯
0	张某	22	北京
1	李某	26	四川成都

```
pd.merge(left, right,
         how='inner', on=u'姓名')
```

outer

	姓名	年龄	籍贯
0	张某	22.0	北京
1	李某	26.0	四川成都
2	段某	24.0	NaN
3	钱某	NaN	江苏南京

```
pd.merge(left, right,
         how='outer', on=u'姓名')
```

left

	姓名	年龄	籍贯
0	张某	22	北京
1	李某	26	四川成都
2	段某	24	NaN

```
pd.merge(left, right,
         how='left', on=u'姓名')
```

right

	姓名	年龄	籍贯
0	张某	22.0	北京
1	李某	26.0	四川成都
2	钱某	NaN	江苏南京

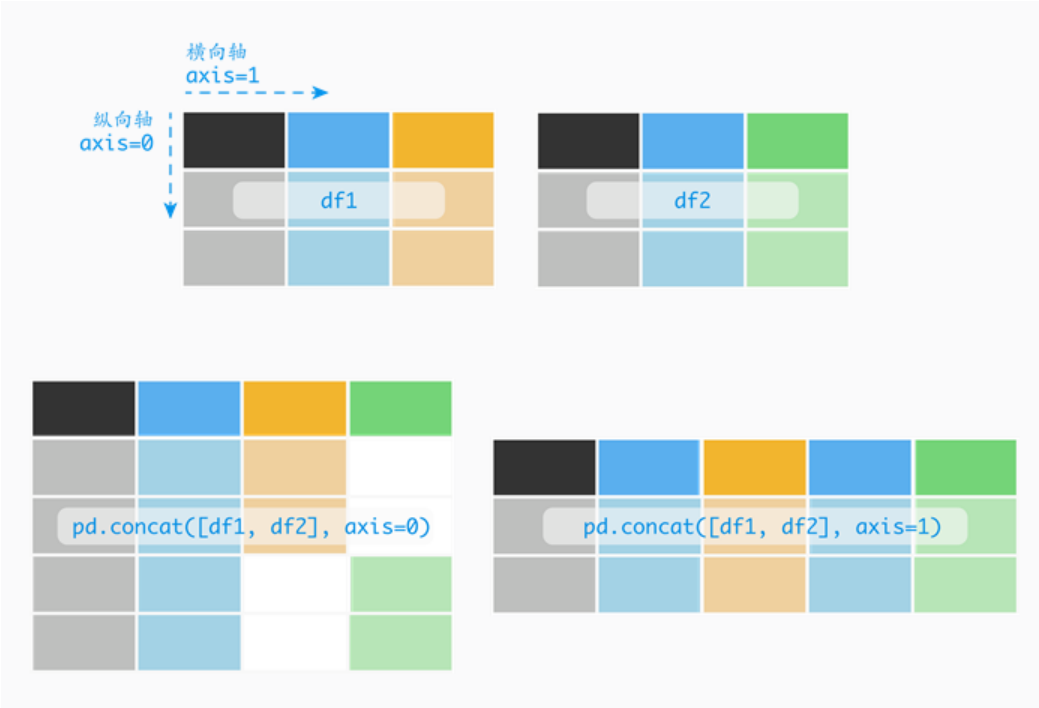
```
pd.merge(left, right,
         how='right', on=u'姓名')
```

Python模块-pandas

- 数据融合

pd.concat([df1, df2])

轴向连接多个DataFrame.



Python模块-pandas

文件读写

从文件中读取数据（DataFrame）

pd.read_csv() | 从CSV文件读取

pd.read_table() | 从制表符分隔文件读取，如TSV

pd.read_excel() | 从 Excel 文 件 读 取

pd.read_sql() | 从 SQL 表 或 数 据 库 读 取

pd.read_json() | 从JSON格式的URL或文件读取

pd.read_clipboard() | 从剪切板读取

将DataFrame写入文件

df.to_csv() | 写入CSV文件

df.to_excel() | 写入Excel文件

df.to_sql() | 写入SQL表或数据库

df.to_json() | 写入JSON格式的文件

df.to_clipboard() | 写入剪切板

Python模块-Scipy

- **Scipy**

scipy是构建在numpy的基础之上的，它提供了许多的操作numpy的数组的函数。

SciPy是一款方便、易于使用、专为科学和工程设计的python工具包，它包括了统计、优化、整合以及线性代数模块、傅里叶变换、信号和图像图例，常微分方差的求解等

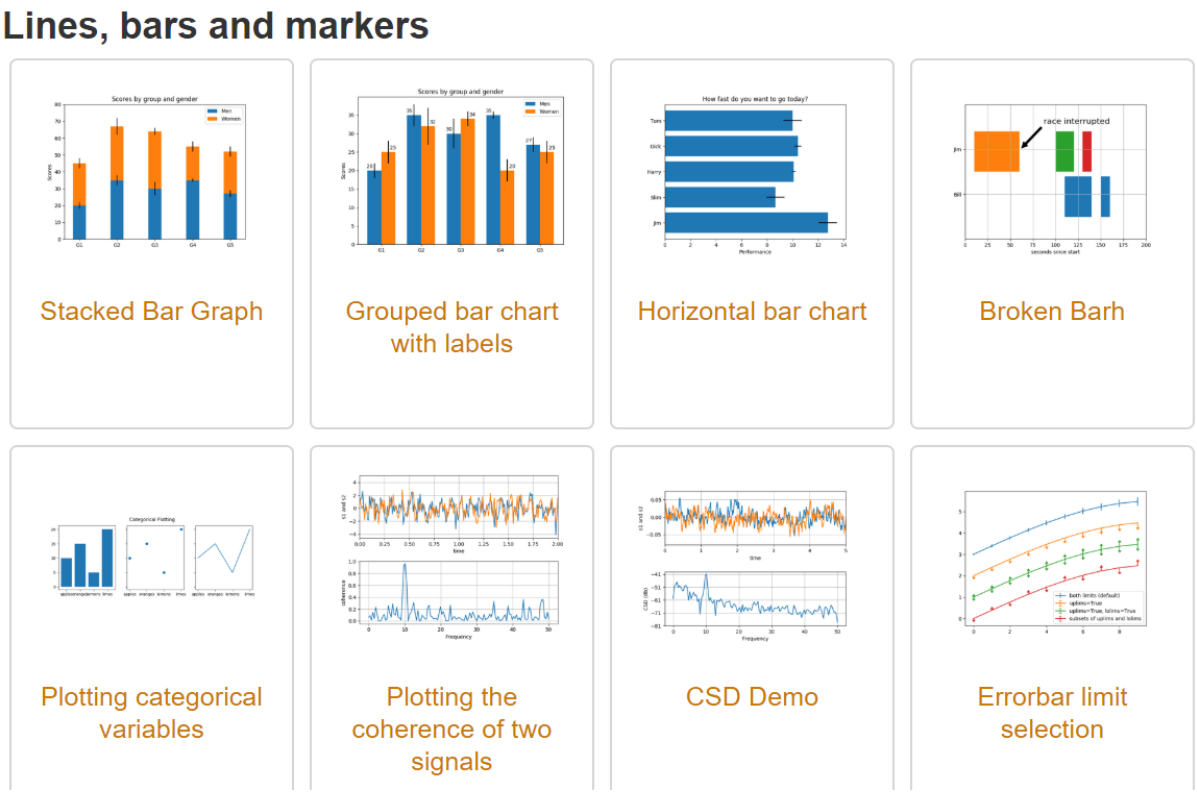
scipy.cluster	向量量化
scipy.constants	数学常量
scipy.fftpack	快速傅里叶变换
scipy.integrate	积分
scipy.interpolate	插值
scipy.io	数据输入输出
scipy.linalg	线性代数
scipy.ndimage	N维图像
scipy.odr	正交距离回归
scipy.optimize	优化算法
scipy.signal	信号处理
scipy.sparse	稀疏矩阵
scipy.spatial	空间数据结构和算法
scipy.special	特殊数学函数
scipy.stats	统计函数

Python模块-matplotlib

- Matplotlib

Matplotlib 是一个 Python 的 2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形 。

通过 Matplotlib，开发者可以仅需要几行代码，便可以生成绘图，直方图，功率谱，条形图，错误图，散点图等。



https://matplotlib.org/gallery/index.html

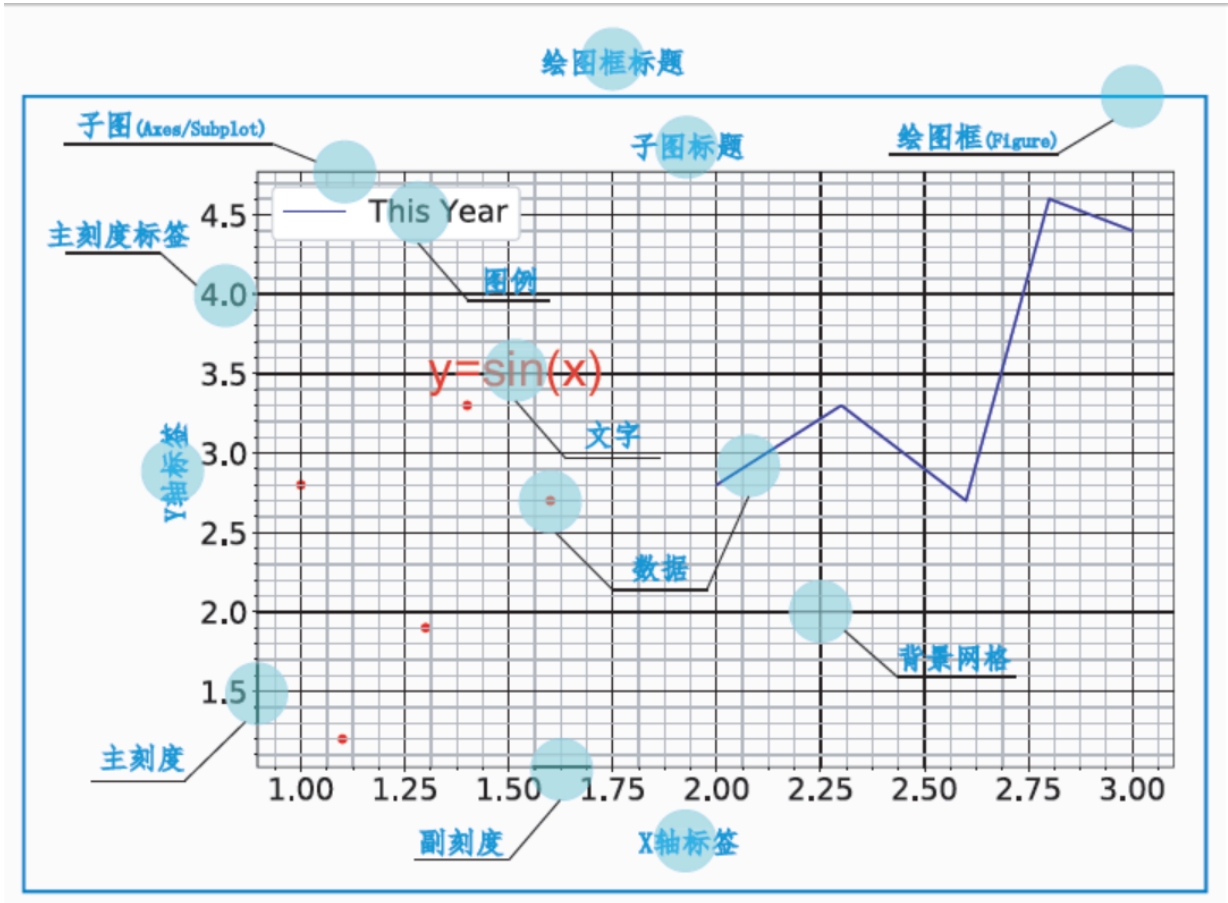
Python模块-matplotlib

图形的各元素名称如下：

绘图框 是图形的最高容器，所有图形必须放置在绘图框中.

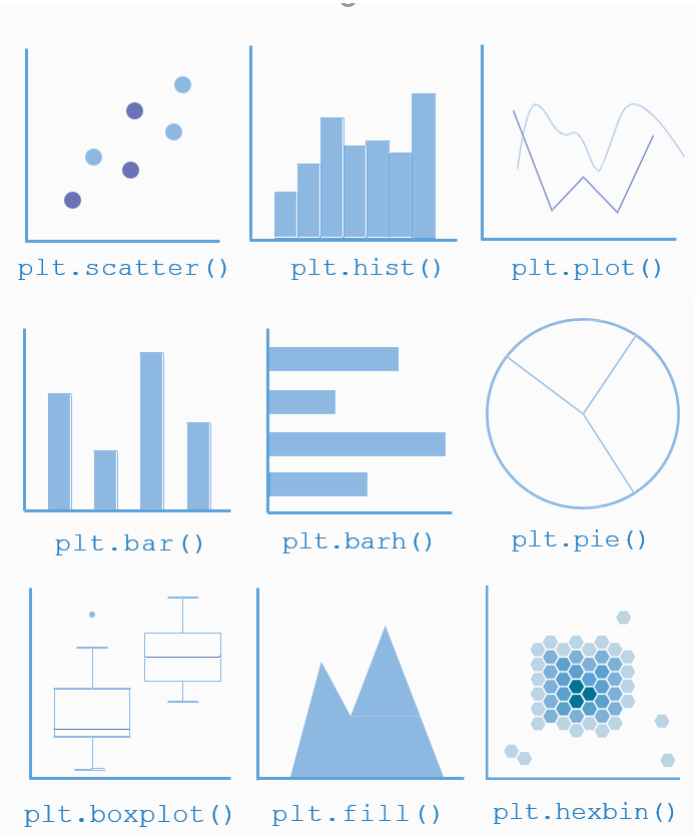
子图 是绘图框中所包含的图形，即便绘图框只包含一幅图，也称之为子图.

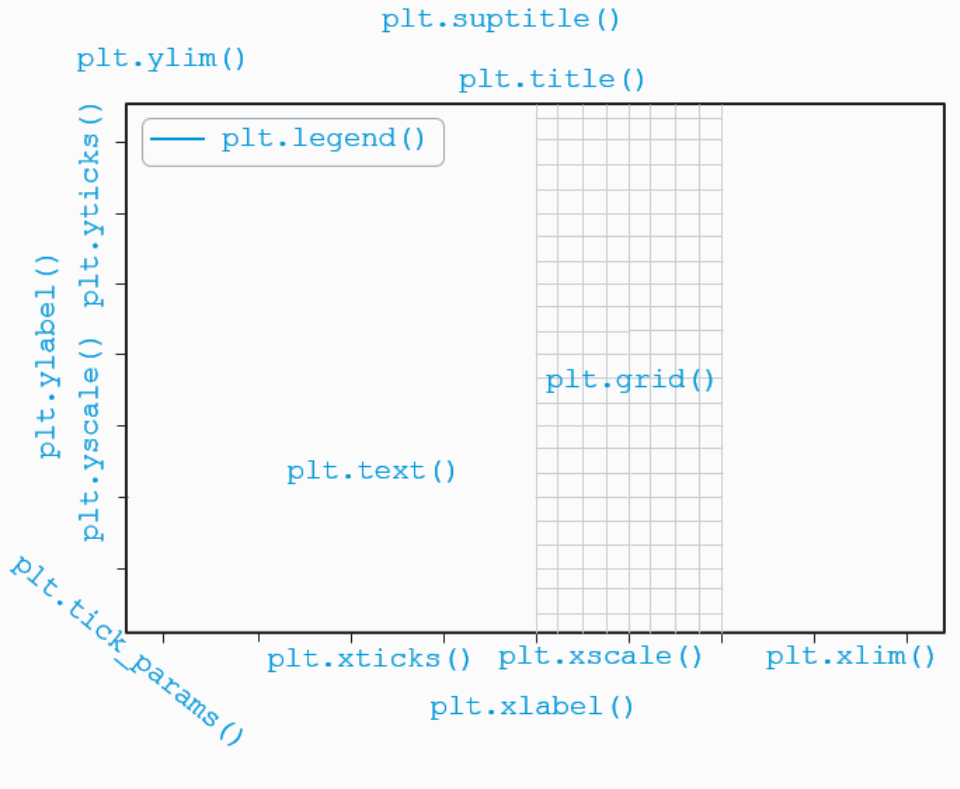
元素 是组成子图的部件，从子图最内部的数据线条到外围的坐标轴标签等都属于元素.



Python模块-matplotlib

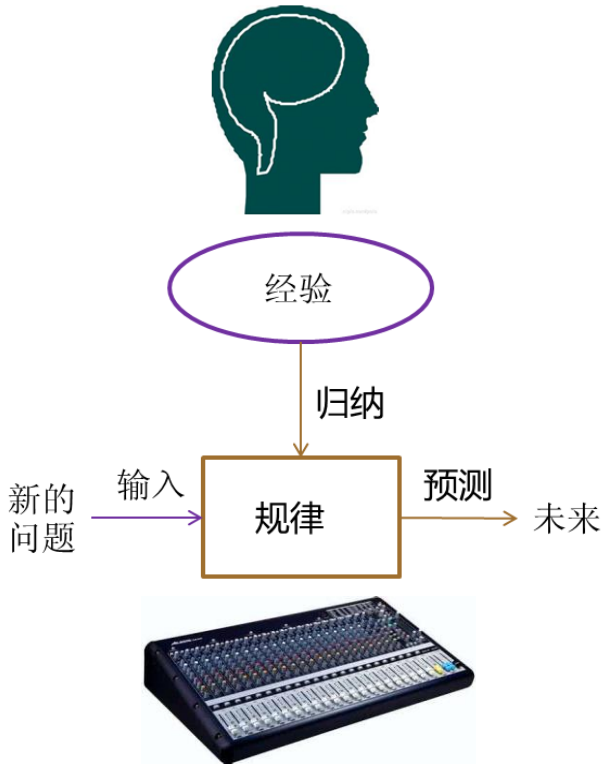
图形样式





4. 机器学习的开发流程

机器学习的一般步骤



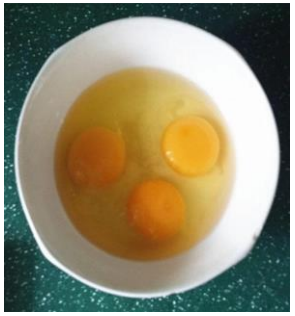
机器学习的一般步骤

数据搜集 → 数据清洗 → 特征工程 → 数据建模



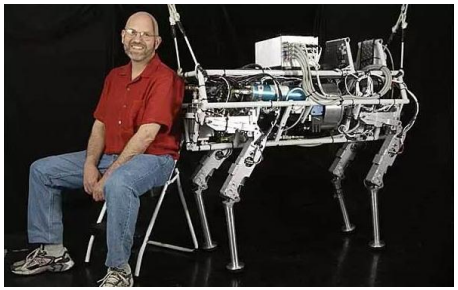




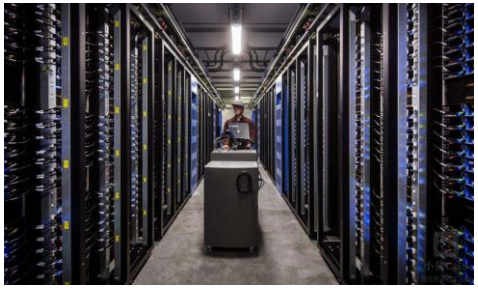


不同视角的机器学习

不同行业的人以为我做的事情



父母以为我做的事情



朋友以为我做的事情

$$\frac{\partial}{\partial w}L(w,b,\alpha)=w-\sum\alpha_i y_i x_i=0,\ w=\sum\alpha_i y_i x_i$$
$$\frac{\partial}{\partial b}L(w,b,\alpha)=\sum\alpha_i y_i=0$$

代入$L(w,b,\alpha)$

$$\begin{aligned}\min\ L(w,b,\alpha)&=\tfrac{1}{2}||w||^2+\textstyle\sum_{i=1}^{m}\alpha_i(-y_i(w^Tx_i+b)+1)\\&=\tfrac{1}{2}w^Tw-\textstyle\sum_{i=1}^{m}\alpha_i y_i w^Tx_i-b\sum_{i=1}^{m}\alpha_i y_i+\sum_{i=1}^{m}\alpha_i\\&=\tfrac{1}{2}w^T\textstyle\sum\alpha_i y_i x_i-\sum_{i=1}^{m}\alpha_i y_i w^Tx_i+\sum_{i=1}^{m}\alpha_i\\&=\textstyle\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i w^Tx_i\\&=\textstyle\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i x_j)\end{aligned}$$

再把 max 问题转成 min 问题：

$$\max\ \textstyle\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i x_j)=\min\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i x_j)-\sum_{i=1}^{m}\alpha_i$$
$$s.t.\ \textstyle\sum_{i=1}^{m}\alpha_i y_i=0,$$

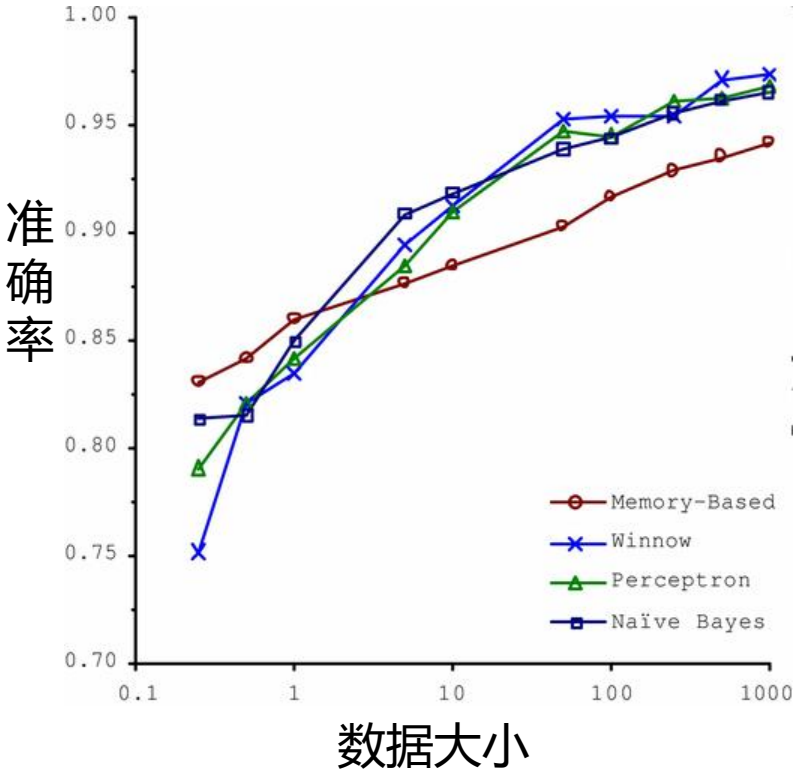
程序员以为我做的事情

我自己以为我做的事情

```
import xgboost as xgb
import numpy as np
```

实际上我做的事情

数据决定一切



通过这张图可以看出，各种不同算法在输入的数据量达到一定级数后，都有相近的高准确度。于是诞生了机器学习界的名言：

成功的机器学习应用不是拥有最好的算法，而是拥有最多的数据！

谢谢观赏 下节课见