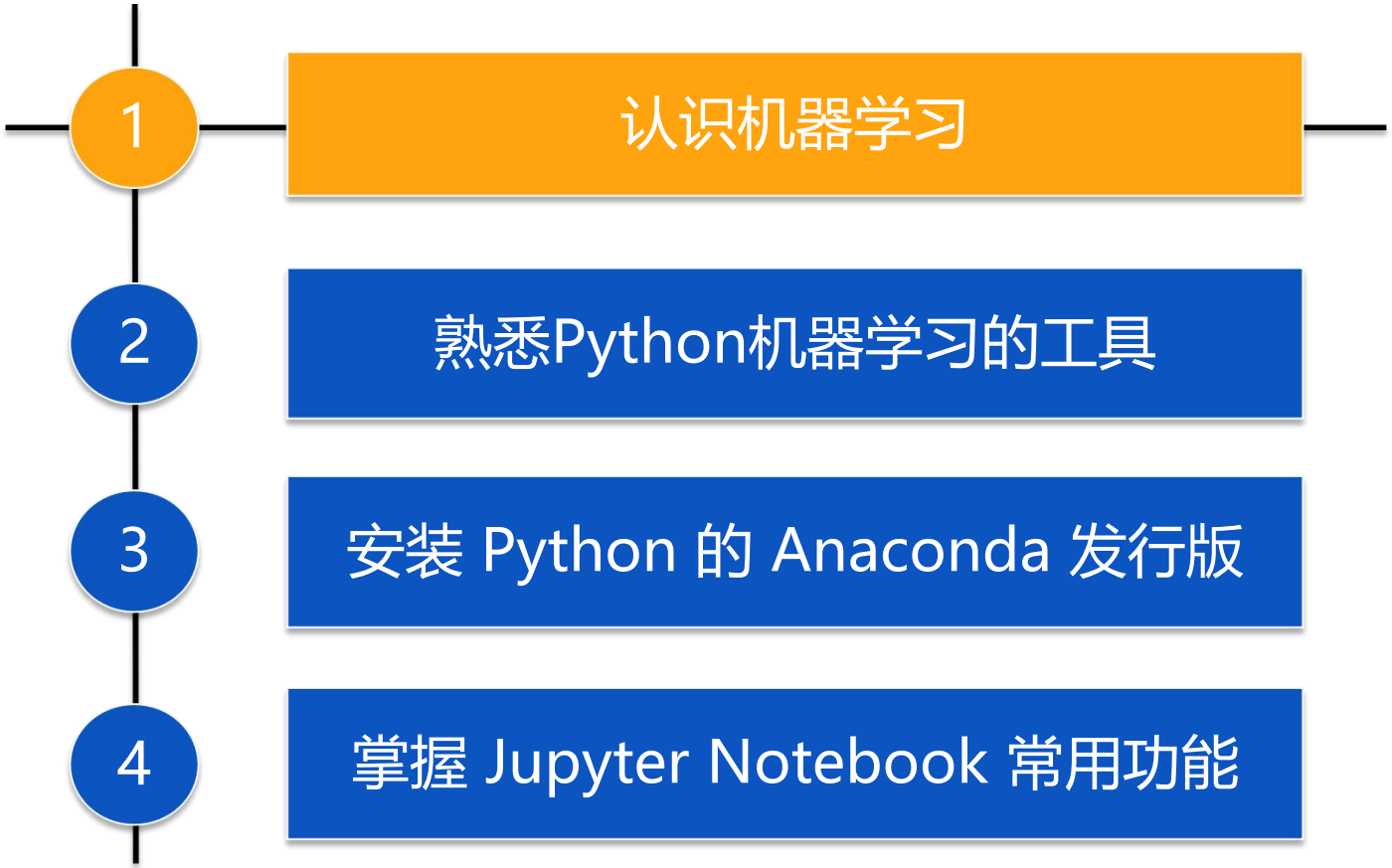
Python数据分析与机器学习

- 李海涛 13933519566
- 网络教学平台：腾讯课堂+学习通
- 成绩：考勤+平时小测+作业(40%)
  结课实操项目(60%)

Python数据分析与机器学习概述

2018/1/8

目录


1
认识机器学习
2
熟悉Python机器学习的工具
3
安装 Python 的 Anaconda 发行版
4
掌握 Jupyter Notebook 常用功能

什么是机器学习

PaddlePaddle

机器学习是从人工智能中产生的一个重要学科分支，是实现智能化的关键。

机器学习（Machine Learning）是一门多领域**交叉学科**，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科 。专门研究计算机怎样模拟或实现人类的学习行为，以获取新知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

百度百科

Machine learning is the study of **algorithms** and mathematical **models** that computer systems use to progressively improve their performance on a specific task. Machine learning algorithms build a mathematical model of **sample data**, known as “training data”, in order to make predictions or decisions without being explicitly programmed to perform the task.

Wikipedia

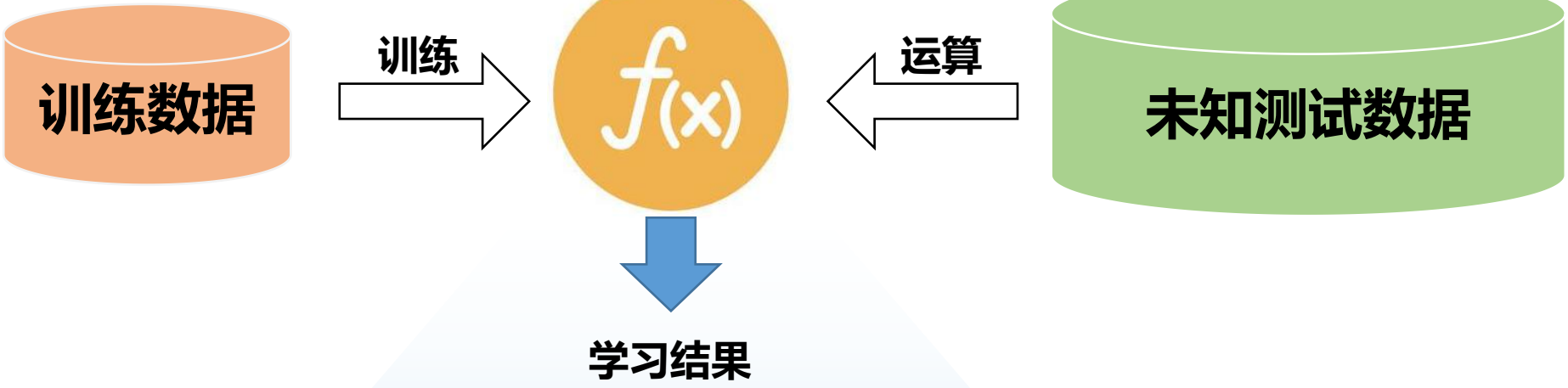
概念(Data Analysis and Machine Leaning)

机器学习方法是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。

人类学习就是对经验进行“归纳”，获得“规律”，使用这些“规律”，对未知问题与未来进行“推测”，从而指导自己的生活和工作。

机器学习中对数据的“训练”与“预测”过程可以对应到人类的“归纳”和“推测”过程。

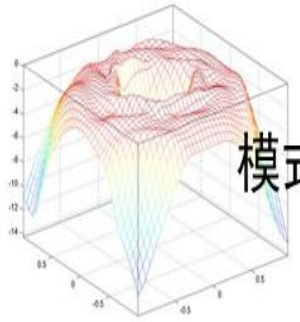
机器学习的一般过程



模型　　策略　　算法

方法：贝叶斯统计、回归、SVM、聚类、决策树

工具：Python+扩展(Scikit-learn tensoflow)



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

机器学习方法

有监督学习（supervised learning）：从给定的**有标注的训练数据集**中学习出一个函数（模型参数），当新的数据到来时可以根据这个函数预测结果。 常见任务包括**分类**与**回归**。

Classification: Y is discrete

Y：年轻人(1)，老年人(-1)

X： x_1 黑头发的比例， 值域（0，1）；

x_2 行走速度， 值域（0，100）米/每分钟.

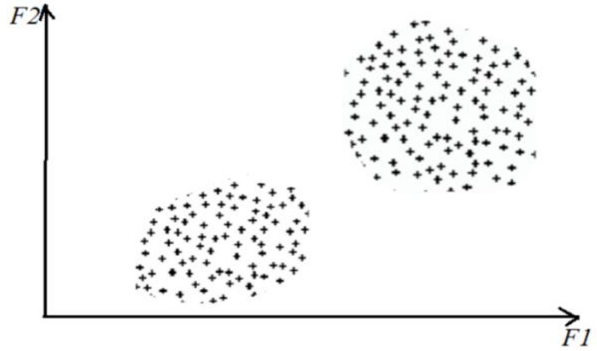
Training Data：

Y=1：(1，99)、(0.9，80)、(0.80，100) …

Y=-1：(0.2，30)、(0.5，50)、(0.4，30) …

Test:

X=(0.85, 98), Y=?



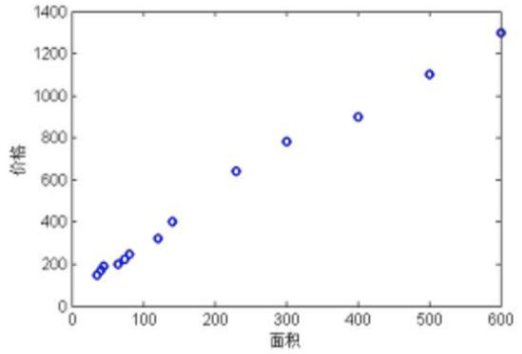
Regression: Y is continue

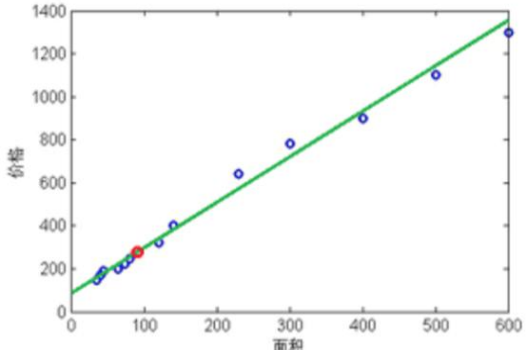
Y: 房屋价钱（万元），值域 Y>=0.

X: x1=房屋面积 m^2.

Training Data：

35	150
40	170
45	190
65	200
74	224
80	245
120	320
140	400
230	640
300	780
400	900
500	1100
600	1300



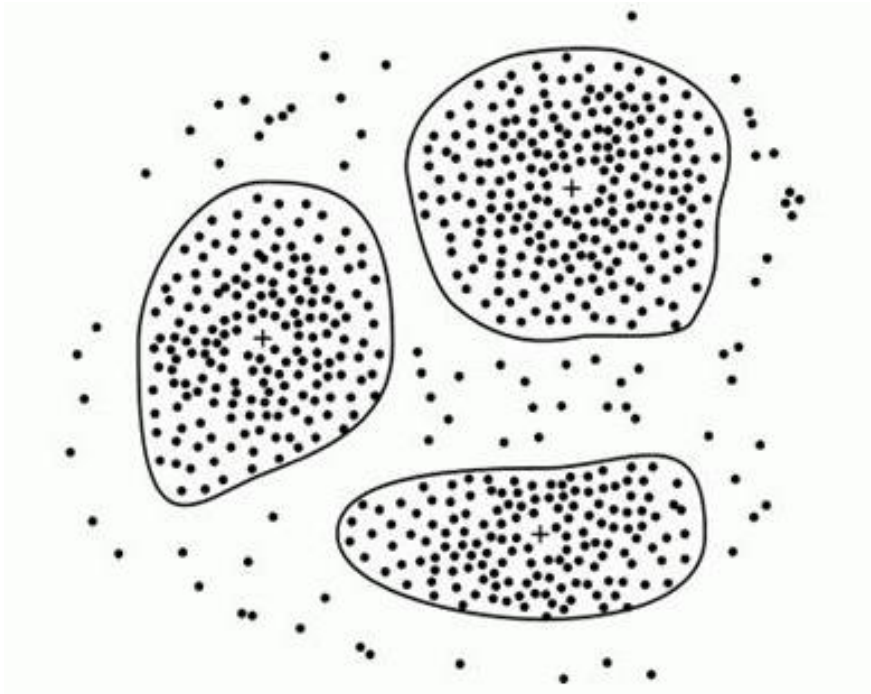


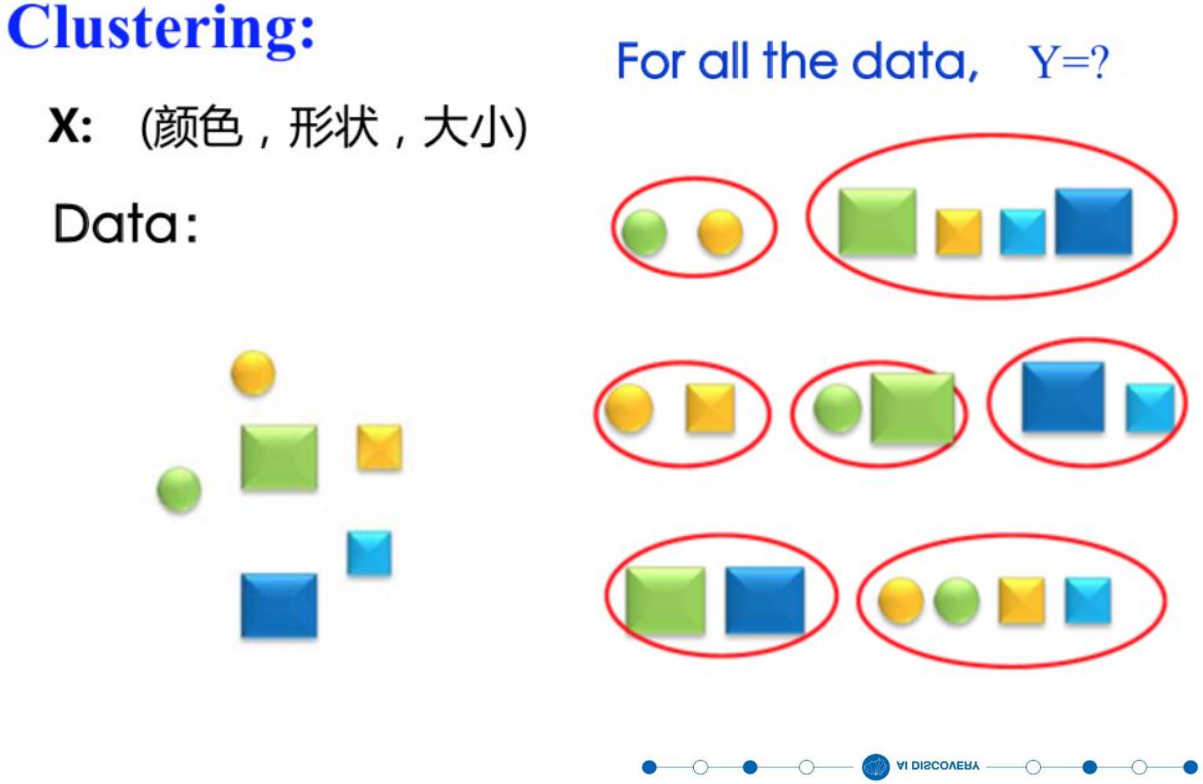
$y=ax+b$

Test: X=90 Y= ?

机器学习方法

无监督学习（unsupervised learning）：没有标注的训练数据集，需要根据样本间的统计规律对样本集进行分析，常见任务如**聚类**等。





机器学习已无处不在

PaddlePaddle

搜索引擎：网页、图片、视频、新闻、学术、地图

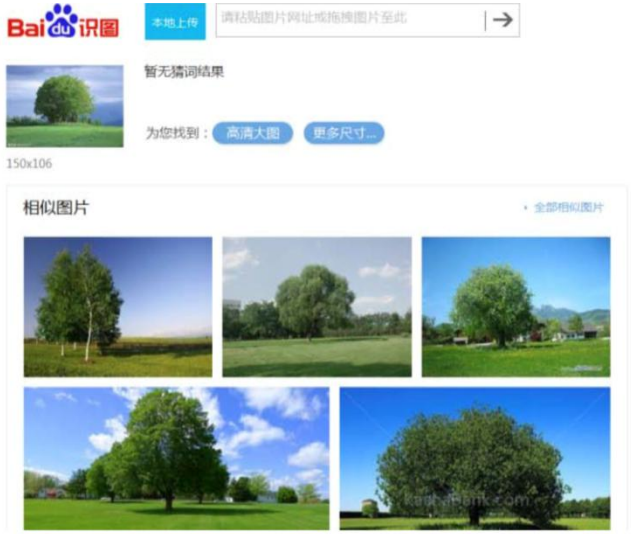
信息推荐：新闻、商品、游戏、书籍

图片识别：人像、用品、动物、交通工具

用户分析：社交网络、影评、商品评论

机器翻译、**摘要生成**……

生物信息学……

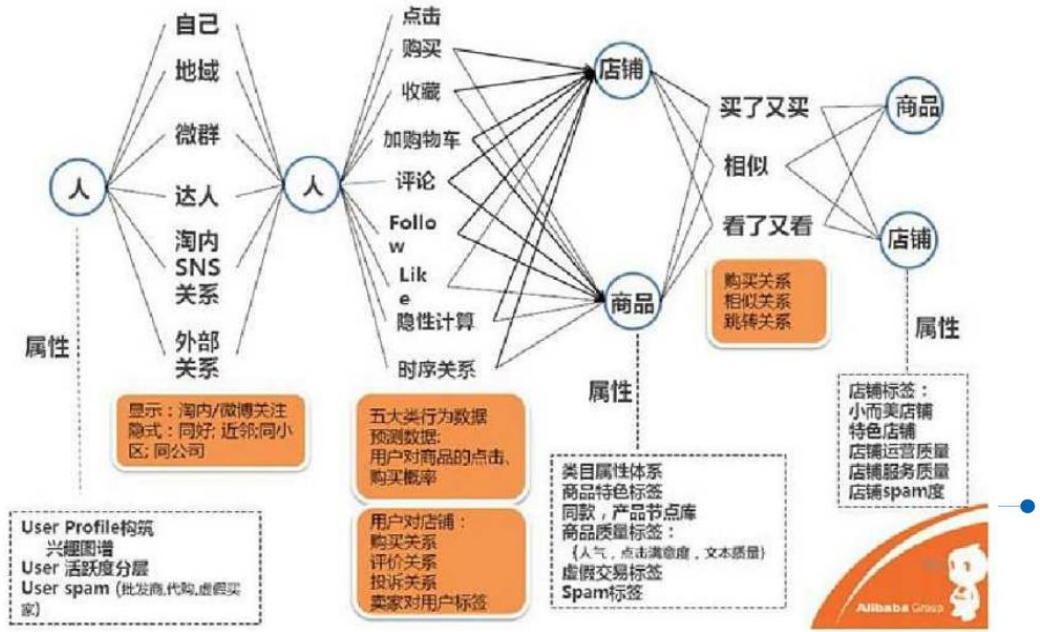


Google的成功，使得Internet搜索引擎成为一个新兴的产业

不仅有众多专营搜索引擎的公司出现（例如专门针对中文搜索的就有慧聪、百度等），而且Microsoft等巨头也开始投入巨资进行研发

Google掘到的第一桶金，来源于其创始人Larry Page和Sergey Brin提出的PageRank算法

机器学习技术正在支撑着各类搜索引擎（尤其是贝叶斯学习技术）




查询
综合画像
学习
作息
三餐
社交关系
身体素质
综合画像分析
2162100025
王清
男
汉族
山西
机械工程（…
装控61
个人信息
学习水平
成绩不稳定
学困
上课纪律很好
挂科1门
社交关系
很少运动

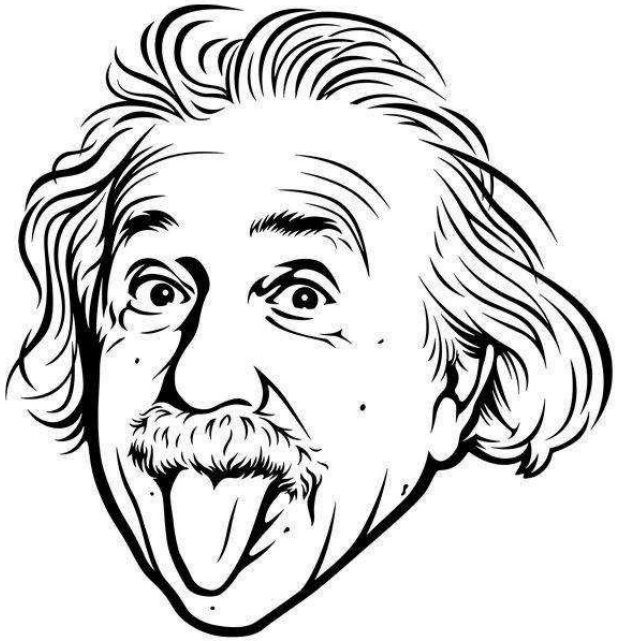

学生工作大数据分析平台
10.49.9.35/#/gropPortrait
群体毕业去向预测
北京大学读研
工作
本校读研
预测挂科情况
挂科2门
挂科1门

机器学习无所不能?

PaddlePaddle

◆**问题思考**：机器学习是否无所不能?

V.S.

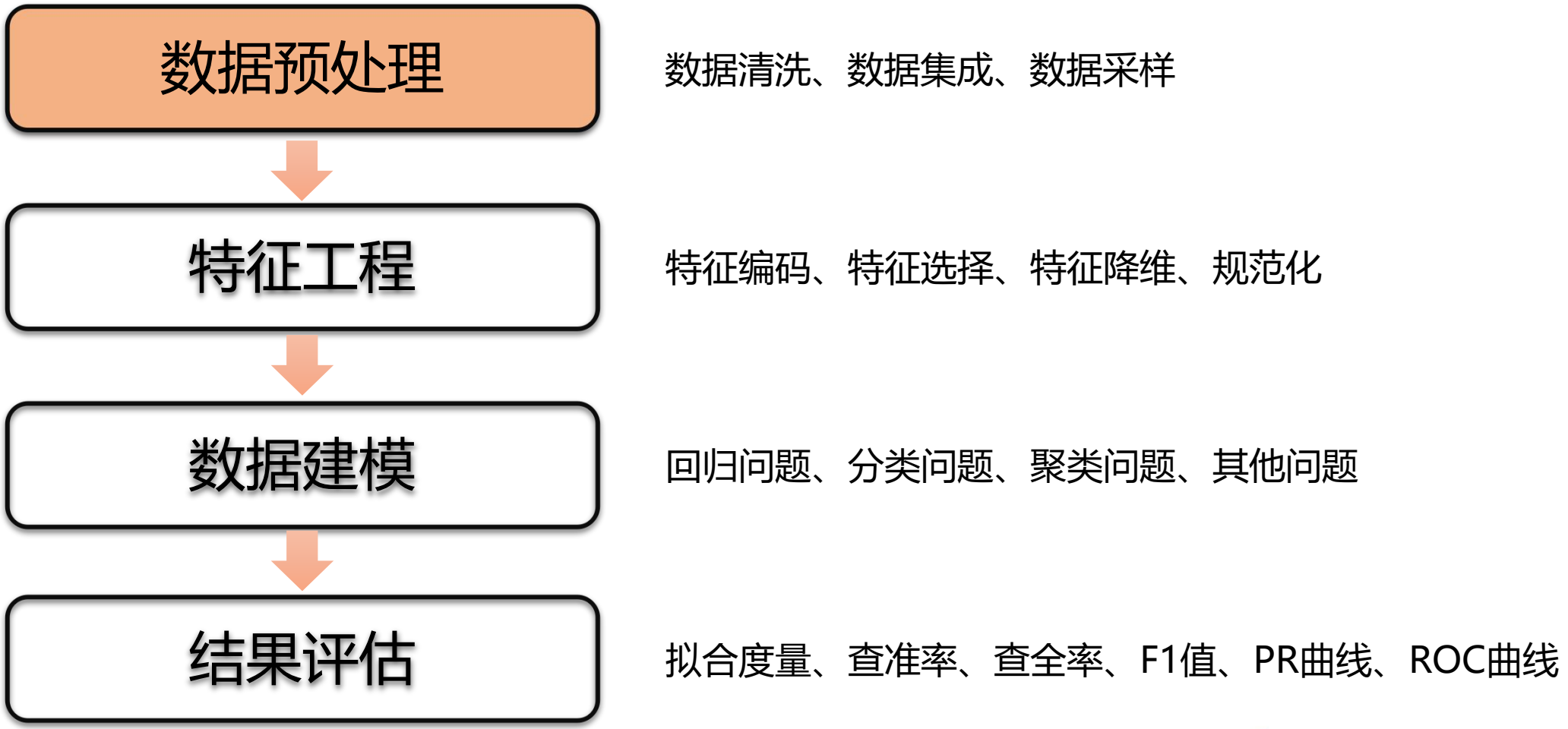


规则、计算、模式

思想、创意、情感

机器如何学习

PaddlePaddle



数据清洗、数据集成、数据采样

特征编码、特征选择、特征降维、规范化

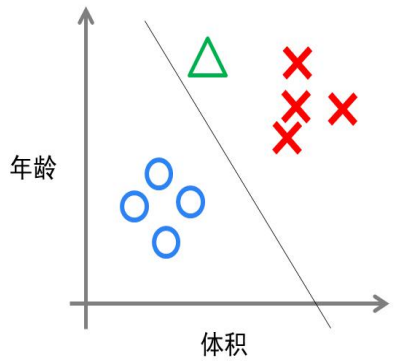
回归问题、分类问题、聚类问题、其他问题

拟合度量、查准率、查全率、F1值、PR曲线、ROC曲线

课程内容与规划

Numpy(同类型数组) Pandas(多类带标签)	Matplotlib	Scikit-learn
• 数组计算与统计 • 数据索引与合并	• 数据可视化基础 • Seabon数据分析图	• 线性回归 决策树 • 聚类分析

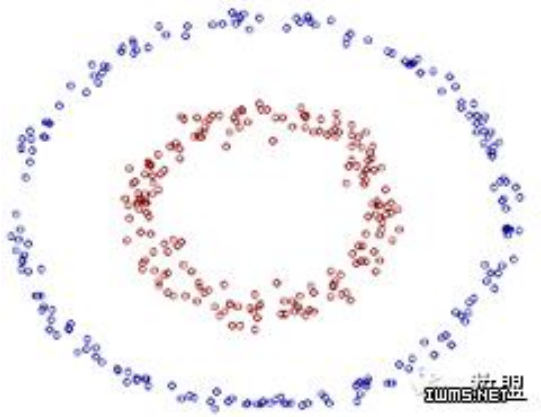
```
In [8]: import numpy as np
   ...: import pandas as pd
   ...: from pandas import Series,DataFrame
   ...: frame_1=DataFrame(np.arange(16).reshape(4,4),index=['c','f','a','d'],
   ...:                   columns=['Tue','Wed','Fri','Mon'])
   ...: frame_1
Out[8]:
   Tue  Wed  Fri  Mon
c    0    1    2    3
f    4    5    6    7
a    8    9   10   11
d   12   13   14   15
```



目标：预测肿瘤的性质

输入：肿瘤的体积，
患者的年龄

输出：良性或恶性



Email:lht@ysu.edu.cn

了解Python数据分析的优势

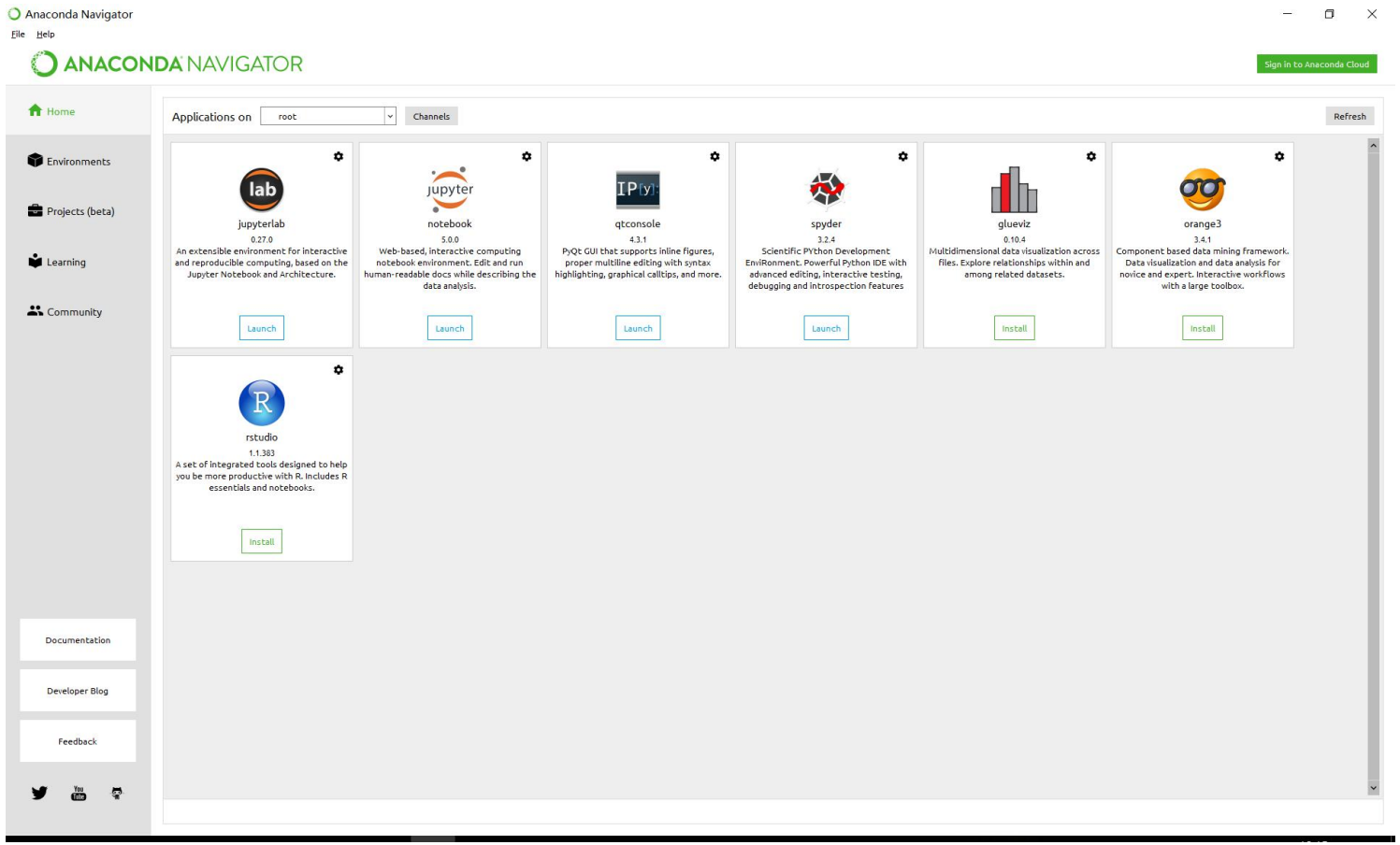
Python 数据分析主要包含以下 5 个方面优势

- 语法简单精练。对于初学者来说，比起其他编程语言，Python更容易上手。
- 有很强大的库。可以只使用Python这一种语言去构建以数据为中心的应用程序。
- 功能强大。Python是一个混合体，丰富的工具集使它介于传统的脚本语言和系统语言之间。Python不仅具备所有脚本语言简单和易用的特点，还提供了编译语言所具有的高级软件工程工具。
- 不仅适用于研究和原型构建，同时也适用于构建生产系统。研究人员和工程技术人员使用同一种编程工具，会给企业带来非常显著的组织效益，并降低企业的运营成本。
- Python是一门胶水语言。Python程序能够以多种方式轻易地与其他语言的组件“粘接”在一起。

了解 Python 的 Anaconda 发行版

Anaconda

- 预装了大量常用 Packages。
- 完全开源和免费。
- 额外的加速和优化是收费的，但对于学术用途，可以申请免费的 License。
- 对全平台和几乎所有Python版本支持。



在 Windows 系统上安装 Anaconda

安装流程

安装包——“next”——“I agree”——“All Users(requires admin privileges)”——选择安装路径——“Install”——“finish”。

IPython常见技巧

IPython-Python的交互式接口，用于交互式科学计算和数据密集型计算。Jupyter Notebook /Spyder

命令	作用
?/??　　len?	显示帮助/获取源码
<TAB>	自动补全
%run c:/test.py	执行外部文件
%timeit a=[i**2 for i in range(100)]	计算代码执行时间，%%timeit 可以处理多行输入
Jupter Notebook工作目录 打开 cmd 输入命令 jupyter notebook --generate-config 设置配置文件jupyter_notebook_config.py c.NotebookApp.notebook_dir = 'E:\\百度云同步盘\\百度云同步盘\\2018秋季\\教学文档\\tensorflow'	
Spyder 工作目录 %cd e:\	

Thank you!