1.5 学习环境简介

CSDN学院

学习环境

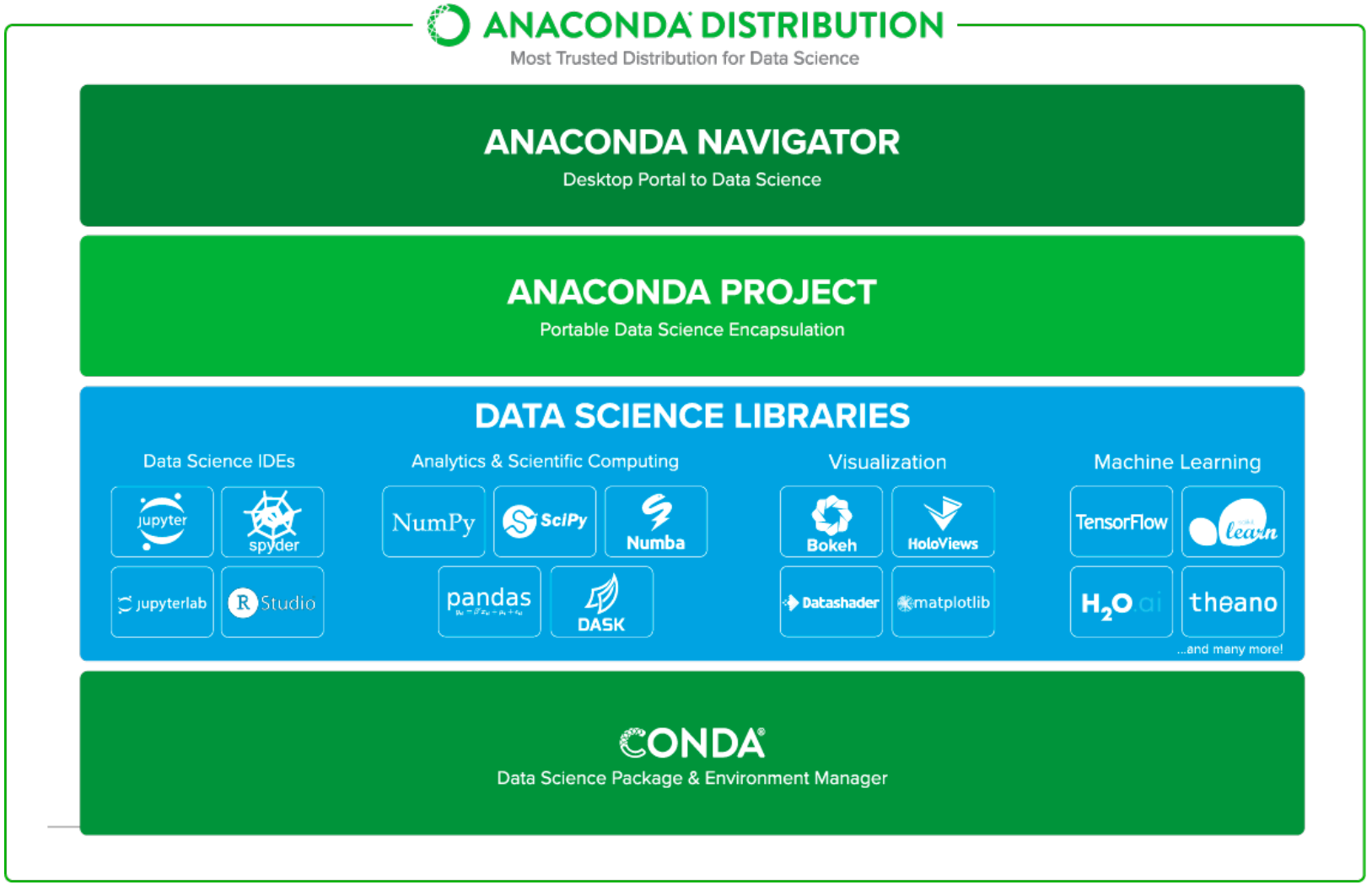
- 编程语言：Python
- 数据处理工具包
  - NumPy
  - SciPy
  - pandas
- 数据可视化工具包
  - Matplotlib
  - Seaborn
- 机器学习工具包
  - scikit learn
- 示例代码：INotebook

软件安装

- 推荐直接安装Anaconda, 包括
  - 很多有用的工具包（包括 pandas 和scikit learn）
  - IPython 和 The Jupyter Notebook (IPython Notebook)
  - conda package manager
  - Spyder IDE

https://www.anaconda.com/download/

Anaconda


ANACONDA DISTRIBUTION
Most Trusted Distribution for Data Science
ANACONDA NAVIGATOR
Desktop Portal to Data Science
ANACONDA PROJECT
Portable Data Science Encapsulation
DATA SCIENCE LIBRARIES
Data Science IDEs
Analytics & Scientific Computing
Visualization
Machine Learning
jupyter
spyder
NumPy
SciPy
Numba
Bokeh
HoloViews
TensorFlow
learn
jupyterlab
RStudio
pandas
DASK
Datashader
matplotlib
H2O.ai
theano
...and many more!
CONDA
Data Science Package & Environment Manager

Jupyter Notebook

- 组成部件
  - **IPython interpreter**
  - **Browser-based notebook interface:** 代码、格式化文本（解释）和图形可以放在一起

- 启动**Notebook**
  - 命令行键入：**ipython notebook**
  - 在Notebook 运行时不要关闭命令行

课程示例代码的格式为ipynb

Notebook快捷键

- 命令模式 (蓝色框)
  - 在当前cell的上面 (**a**) 或下面 (**b**)创建新的cell
  - 用上箭头**up arrow**和下箭头 **down arrow**可上下浏览
  - 将cell类型切换到Markdown (**m**)或代码 (**y**)模式
  - **h**：键盘快捷键帮助
  - **Enter**：切换到编辑模式 **Enter**
- 编辑模式 (绿色框)
  - **Ctrl+Enter** ：运行一个cell
  - **Esc**：切换到命令模式

NumPy

- NumPy(Numeric Python)是Python的开源数值计算扩展，可用来存储和处理大型矩阵
  - 官网：http://www.numpy.org/
- NumPy 包括：
  - N维数组（ ndarray ）
  - 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

- SciPy 是建立在NumPy基础上、是科学和工程设计的 Python 工具包，提供统计、优化和数值微积分计算等功能
- 课程中我们主要用到其稀疏矩阵表示及运算
  - NumPy 处理10^6 级别的数据通常没有大问题，但当数据量达到10^7 级别时速度开始发慢， 内存受到限制 (具体情况取决于实际内存大小)
  - 当处理超大模数据集，比如10^{10}级别， 且数据中包含大量的 0 时，可采用稀疏矩阵可显著的提高速度和效率
  - import scipy.sparse

Pandas（Panel data structures）

- Pandas 是Python语言的“关系数据库”数据结构和数据分析工具，非常高效且易于使用
  - 基于 NumPy补充了大量数据操作功能，能实现统计、分组、排序、透视表（SQL语句的大部分功能）
- 官网： http://pandas.pydata.org/
- Pandas 主要有 2 种重要数据类型:
  - Series：一维序列
  - DataFrame：二维表（机器学习数据的常用数据结构）

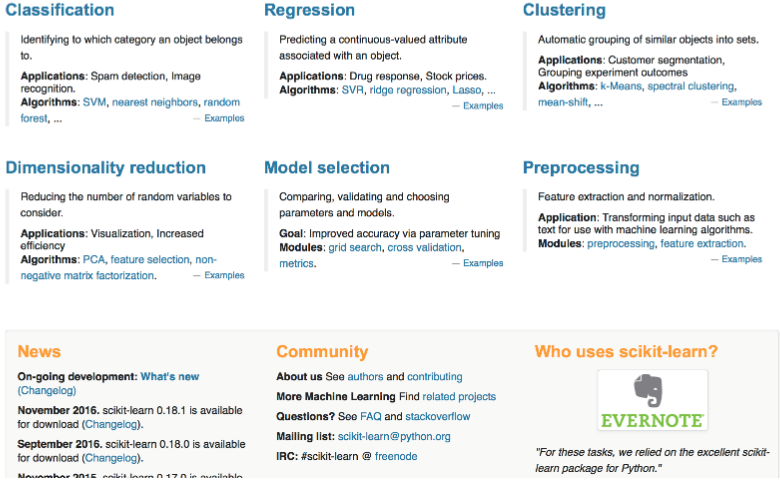
Matplotlib

- Matplotlib是Python语言的2D图形绘制工具
- 官网：http://matplotlib.org/

Seaborn

- Seaborn是一个基于Matplotlib的Python可视化工具包，提供更高层次的用户接口，可以给出漂亮的数据统计图
- 官网： https://seaborn.pydata.org/

Scikit-Learn
—*Machine Learning in Python*

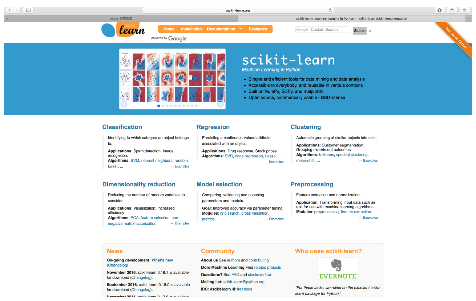


Scikit-Learn 是基于Python 的开源机器学习模块，最早于2007年由 David Cournapeau 发起

官网： http://scikit-learn.org/stable/

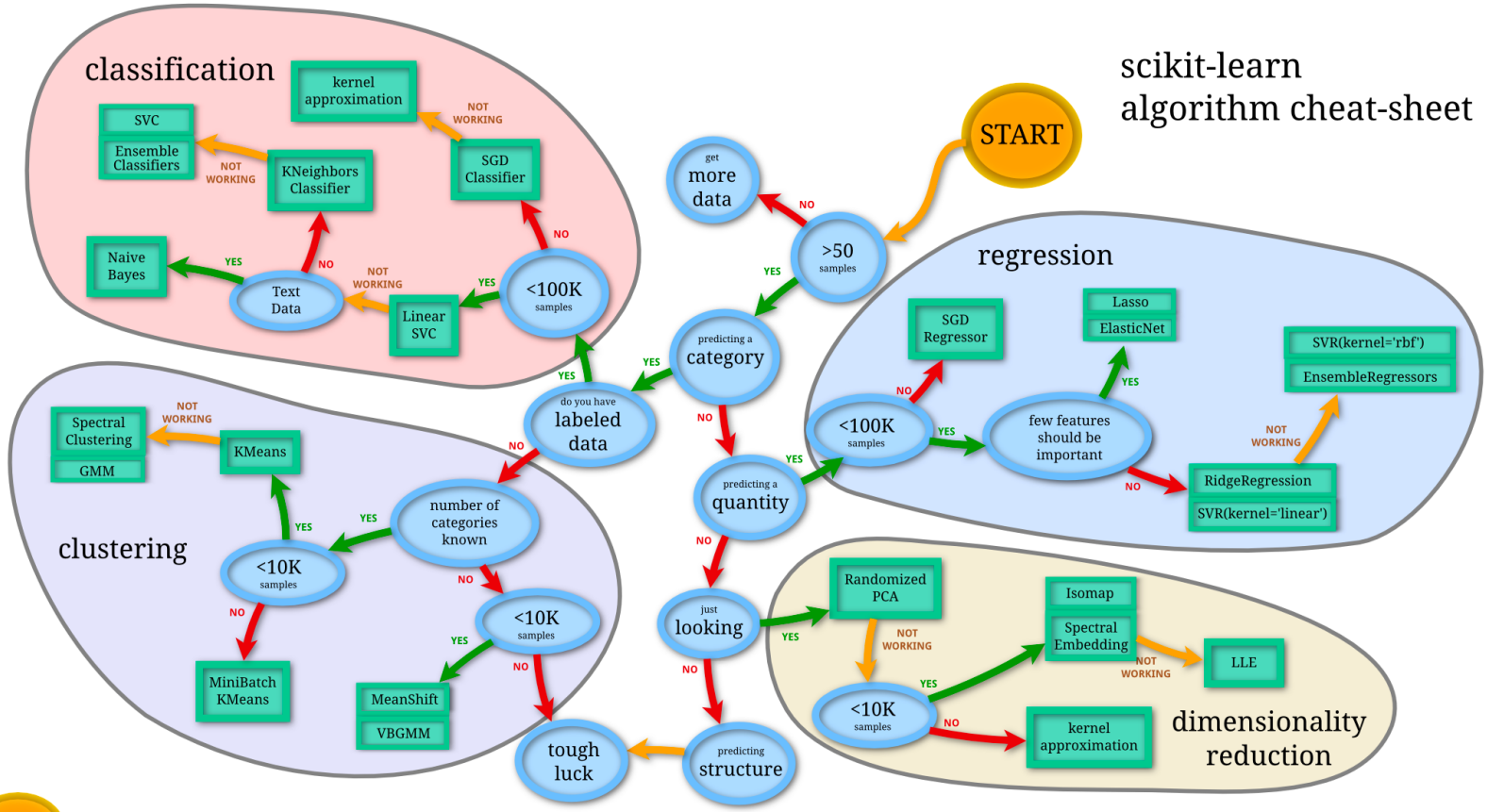
中文版用户手册：
http://sklearn.apachecn.org/cn/0.19.0/

Scikit-Learn



- 基本功能有六个部分：分类（Classification）、回归（Regression）、聚类（Clustering）、数据降维（Dimensionality reduction）、模型选择（Model Selection）、数据预处理（Preprocessing）。
- 对于具体的机器学习问题，通常可以分为三个步骤
  - 数据准备与预处理（Preprocessing、Dimensionality reduction）
  - 模型选择与训练（Classification、Regression、Clustering）
  - 模型验证与参数调优（Model Selection）

机器学习模型选择


scikit-learn
algorithm cheat-sheet
START
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
<100K samples
Text Data
Naive Bayes
Linear SVC
KNeighbors Classifier
SVC
Ensemble Classifiers
kernel approximation
SGD Classifier
classification
Spectral Clustering
GMM
KMeans
number of categories known
<10K samples
MiniBatch KMeans
MeanShift
VBGMM
clustering
<100K samples
SGD Regressor
Lasso
ElasticNet
few features should be important
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
regression
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
dimensionality reduction
YES
NO
NOT WORKING
Back

采用scikit-learn的优点

- 各种机器学习模型有统一的接口
- 模型既有默认参数，也提供多种参数调优方法
- 卓越的文档
- 丰富的随附任务功能集合
- 活跃的社区提供开发和支持

► 小节

- Python开源工具多、语法简单
- 机器学习工具包scikit learn 集成机器学习大多数模型的实现，且为各种模型提供统一接口
- 实现AI变得很轻松

- 助教已经给我们准好了详细的安装文档☺