

机器学习：Day1实训内容

李猛 2017-08-10
V1.0.0.0

The World's Local Training Provider

目录

MENU

1

序言

2

机器学习概论

3

分类算法介绍

4

总结Q&A

5

结语

1.1 序言：个人背景介绍

- 十年工作经验，后端为主
- 互联网行业
- 资深架构师
- 机器学习
- 精通四门主流语言Java/C#/Scala/Python
- 系统架构(Java/C#)
- 大数据(Scala)
- 机器学习/深度学习(Python)
- 业余兼职：技术顾问/技术培训

- Q&A：学员的跨语言项目经验？

1.2 序言：个人与机器学习

- .Net平台机器学习：电商推荐+关联挖掘
- Java平台机器学习：大数据平台Hadoop mahout/Spark Mlib
- Python平台机器学习：画图matplotlib
- 混合机器学习平台

1.3 序言：学员分组

- 单个小组4~5人/或者2人一组/结对编程
- 分工协作
- Q&A：互动性
- 书活动 《数学之美》 《算法帝国》

1.4序言： 培训方式

- 基本理论概念介绍，示意图为主，文字描述为辅，点到为止
- Jupyter案例演示
- 学员练习
- 互动Q&A

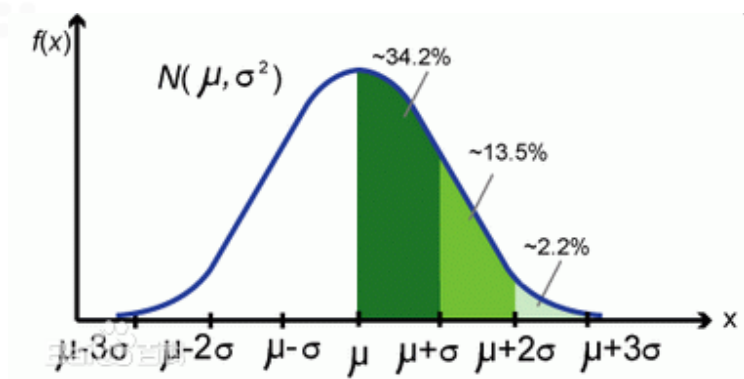
1.5 序言： 培训大纲介绍

- 考虑对于Python与机器学习的了解
- 培训大纲设计原则： Day1广度优先， Day2深度优先
- 如果编码实战多则注重理论学习 或者 如果是学院派则注重实战编码
- Day1： 机器学习基本概念介绍， 简要原理介绍
- Day2： 案例实战， 机器学习构建整个过程
- [培训大纲图](#)

2.1.1 统计学习理论：介绍

- 统计学习理论主要是指统计学+概率学
 - 通常所说的机器学习指的就是统计学习理论
 - 统计学：大量数据；如：高斯分布；
 - 概率学：概率分布；如：条件概率；
 - 机器学习是指通过现有条件对统计学习理论的一种工程实践，通过计算机来模拟统计学与概率学求解，找出定义数据分布的规律
-
- Q&A：学员的数学基础

2.1.2 统计学习理论：高斯分布/概率



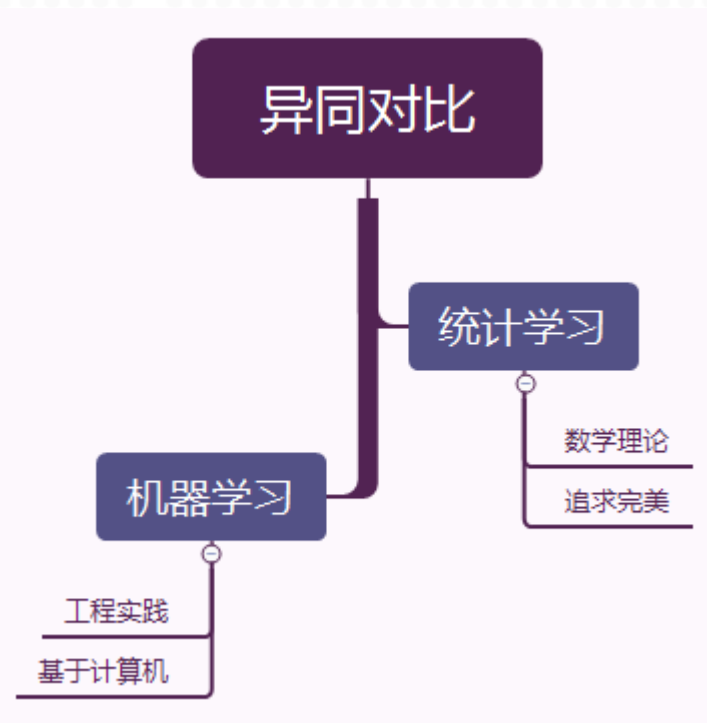
图示:高斯分布/正太分布



图示：抛硬币

2.1.3 统计学习理论：对比

- 统计学习：数理理论
- 机器学习：工程实践



2.1.4统计学习理论：数学基础

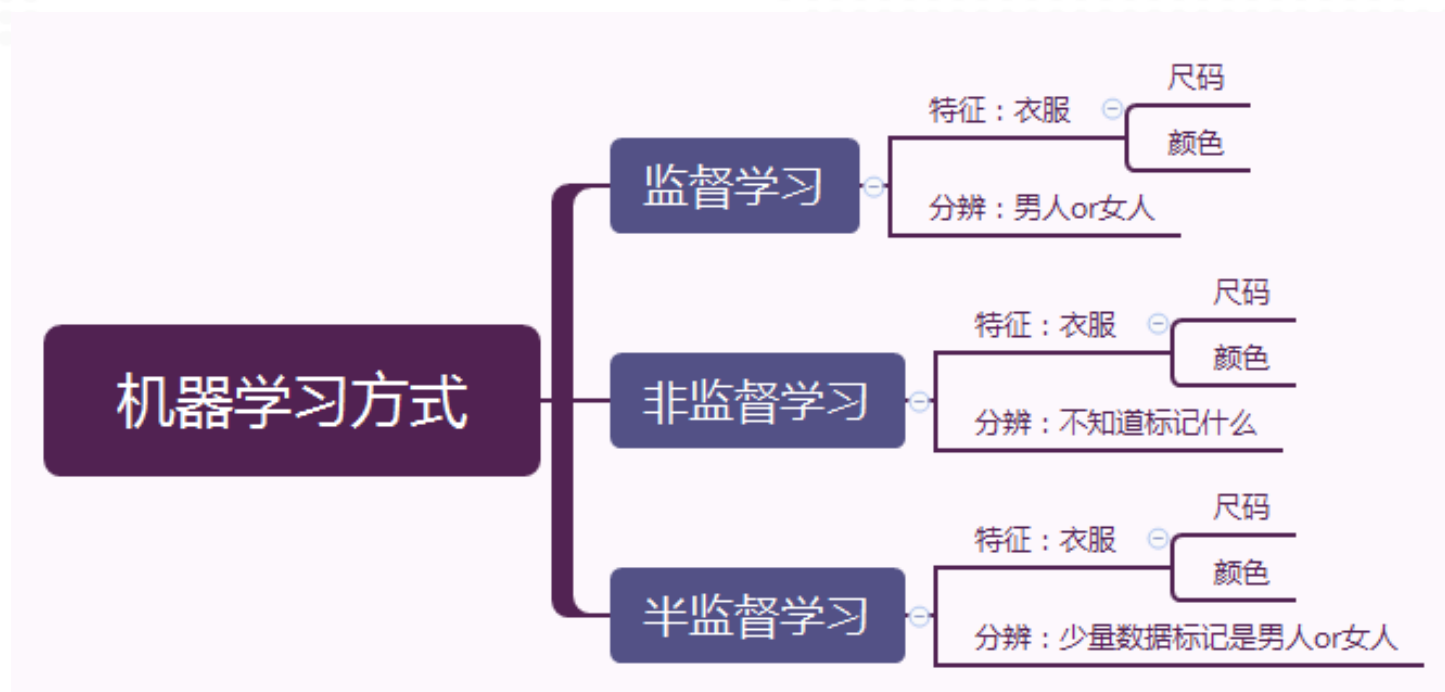
- 线性代数
- 概率
- Q&A：学员举例？

$$I_1 = [1], I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \dots, I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

图示：矩阵

2.2.1 机器学习概论：学习方式

- 监督学习 : **supervised learning**
- 非监督学习 : **unsupervised learning**
- 半监督学习 : **semi-supervised learning**
- **Q&A: 学员举例**



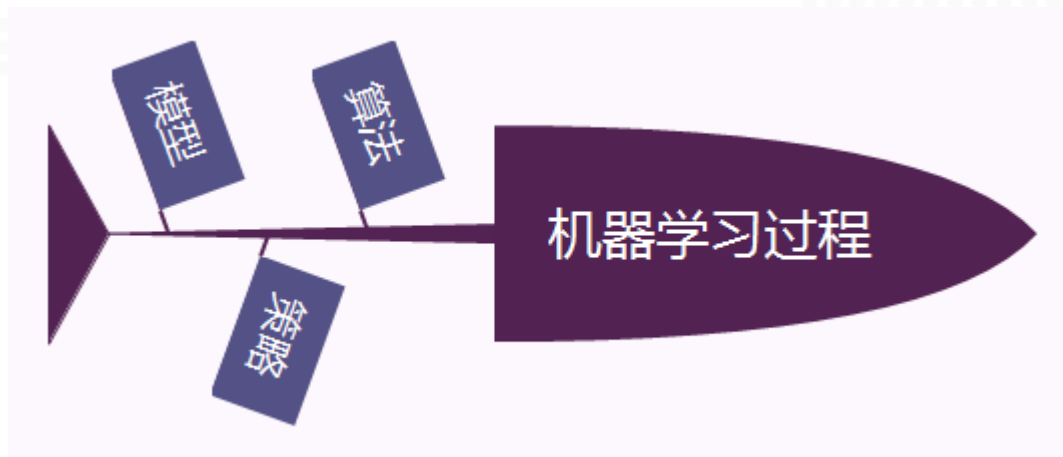
2.2.2 机器学习概论：类型划分

- 分类：离散值 **classification**
- 回归：连续值 **regression**
- Q&A 学员举例



2.2.3 机器学习概论：三要素

- 模型：数学建模，如：设计一个数学模型识别男人与女人/人与动物
- 策略：如何设计模型，通过哪些特征Feature区分
- 算法：具体的数学理论上什么，线性代数，逻辑回归
- Q&A：学员举例？



2.3.1 Python机器学习：介绍

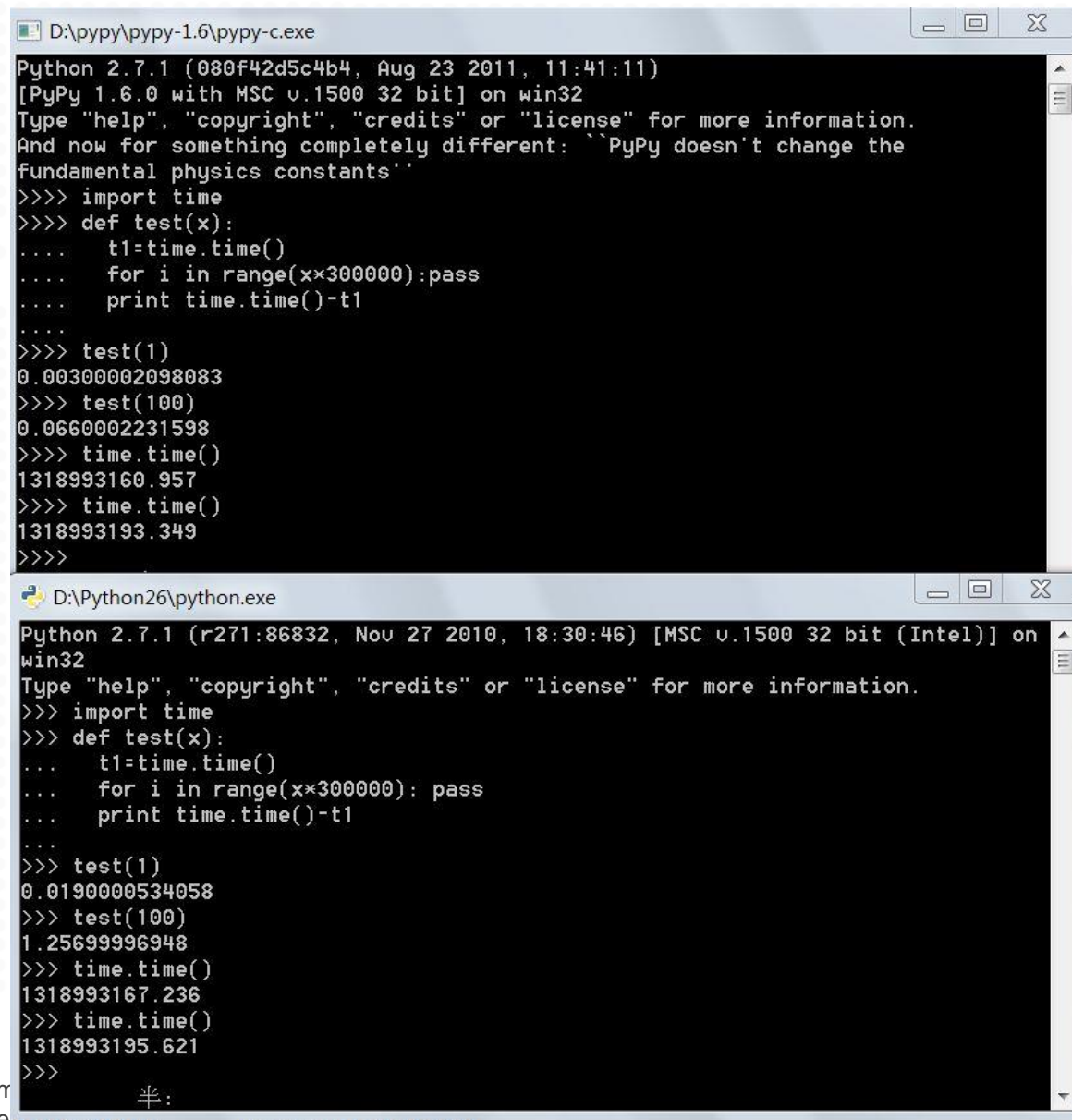
- 为什么选择Python?
- 社区：资源
- 平台：跨平台性
- 语言：解析式；函数式；多范式；
- 应用：数据分析；数据挖掘；机器学习；
- Q&A：学员举例？

2.3.2 Python机器学习： toolkit工具包

- 机器学习基础库toolkit?
- numpy 矩阵加 基本数学函数
- matplotlib 画图
- pandas 矩阵+数据
- scipy 高级数学函数
- Sympy 数学表达式
- scikit-learning机器学习库
- 更多.....
- Q&A: 学员举例?

2.3.3 Python机器学习：跨平台性

- 使用C/CPP高性能计算库
- 基于pypy平台，JIT编译机制
- Q&A：学员举例？



The image shows two terminal windows side-by-side, comparing the performance of PyPy and CPython. Both windows run the same Python script that tests a loop with 300,000 iterations. The PyPy window (top) shows a much faster execution time (approx. 0.003s) compared to the CPython window (bottom, approx. 1.25s).

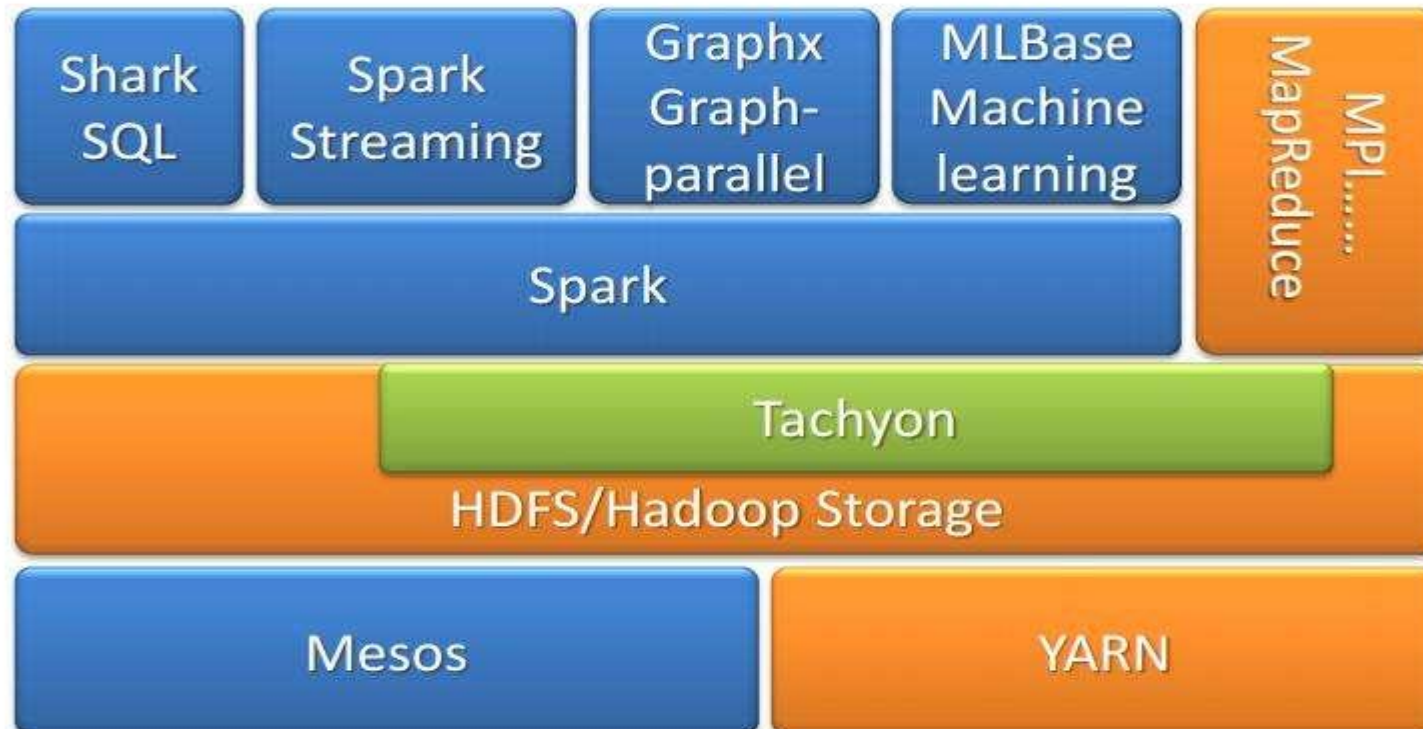
```
D:\pypy\pypy-1.6\pypy-c.exe
Python 2.7.1 (080f42d5c4b4, Aug 23 2011, 11:41:11)
[PyPy 1.6.0 with MSC v.1500 32 bit] on win32
Type "help", "copyright", "credits" or "license" for more information.
And now for something completely different: ``PyPy doesn't change the
fundamental physics constants``
>>> import time
>>> def test(x):
...     t1=time.time()
...     for i in range(x*300000):pass
...     print time.time()-t1
...
>>> test(1)
0.00300002098083
>>> test(100)
0.0660002231598
>>> time.time()
1318993160.957
>>> time.time()
1318993193.349
>>>

D:\Python26\python.exe
Python 2.7.1 (r271:86832, Nov 27 2010, 18:30:46) [MSC v.1500 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import time
>>> def test(x):
...     t1=time.time()
...     for i in range(x*300000): pass
...     print time.time()-t1
...
>>> test(1)
0.0190000534058
>>> test(100)
1.25699996948
>>> time.time()
1318993167.236
>>> time.time()
1318993195.621
>>>
```

2.3.4 Python机器学习：跨平台性

- 大数据Spark平台
- 利用Map/Reduce机制提高计算性能
- pyspark

Spark Ecosystem



2.4.1练习：Python开发环境

- Pycharm IDE
- Visual Studio 2015
- Jupyter notebook

2.4.1练习：Python基本功

- 重点练习
- Tuple（元组）
- List（列表）
- Dictionary（字典）
- Def（函数）
- Loop（循环）
- Range（范围生成）
- File（IO）
- 案例：jupyter
- Q&A：与Java对比下

2.4.2练习：矩阵：numpy

- 多维度矩阵： ndarray
- 生成矩阵： arrage
- 矩阵广播： 矩阵计算(加减乘除)
- 矩阵数组转换
- 线性函数:sin,cos
- 文件处理： loadtxt,save
- 案例： jupyter
- Q&A： 与其它平台对比

2.4.3练习：可视化：matplotlib

- 画点
- 画线
- 画面
- 与numpy函数
- 矩阵
- 布局
- 画圆
- 2D、3D（此处省略）
- 案例：jupyter
- Q&A：与其它平台对比

2.4.3练习： 矩阵+数据:pandas

- 矩阵结构信息
- 时间矩阵
- 矩阵查找
- IO支持多种数据读取 csv、json、sql等
- 案例： jupyter
- Q&A： 与其它平台对比

2.4.5练习：数学函数： scipy

- 线性函数库linalg
 - 高级函数库
 - 聚类
 - [傅立叶变换示意图](#)
-
- 案例： jupyter
 - Q&A： 与其它平台对比

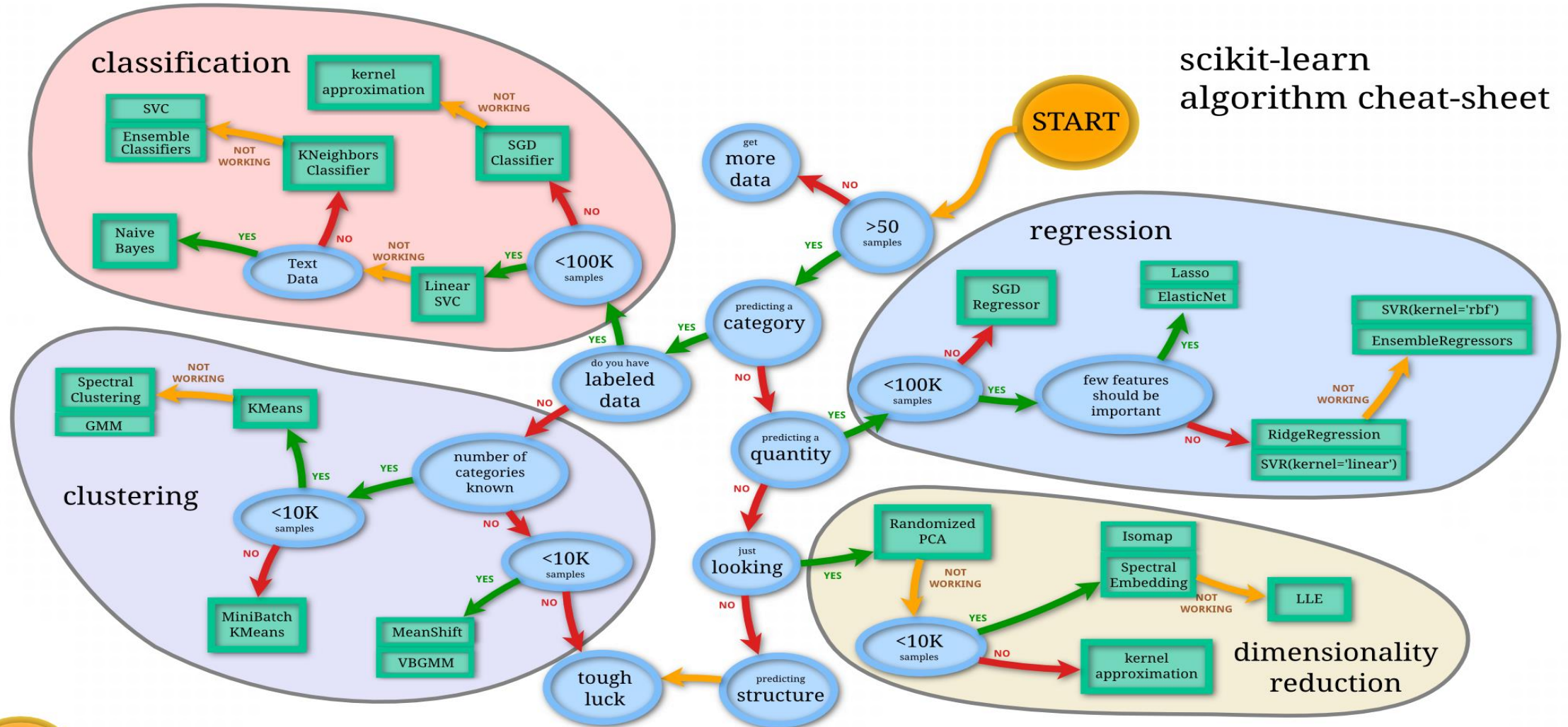
2.4.6练习：数学表达式：sympy

- 数学表达式
- 使用表达式计算
- 表达式求解
- 案例：jupyter
- Q&A：与其它平台对比

2.4.7练习：机器学习库：scikit-learning

- 基本结构
- 模型
- 算法
- 练习数据：原有数据，创建数据
- 示意图：下一页

2.4.7练习：机器学习库：scikit-learning





“给我一个支点，我就能撬起整个地球。”

▲ 阿基米德

联系方式

- Mail:1789909854@qq.com
- QQ:1789909854
- [Tel:17621063575](tel:17621063575)
- Wechat:ynuosoft

