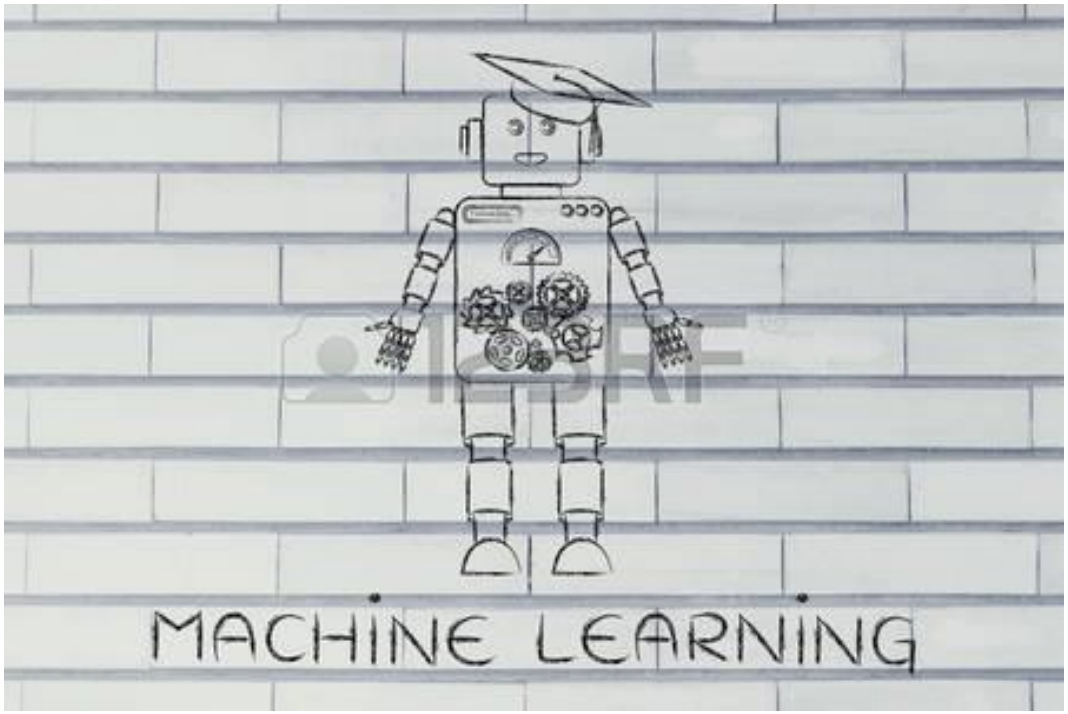
法律声明

- 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

第八讲



机器学习基础及机器学习库scikit-learn入门

--梁斌

目录

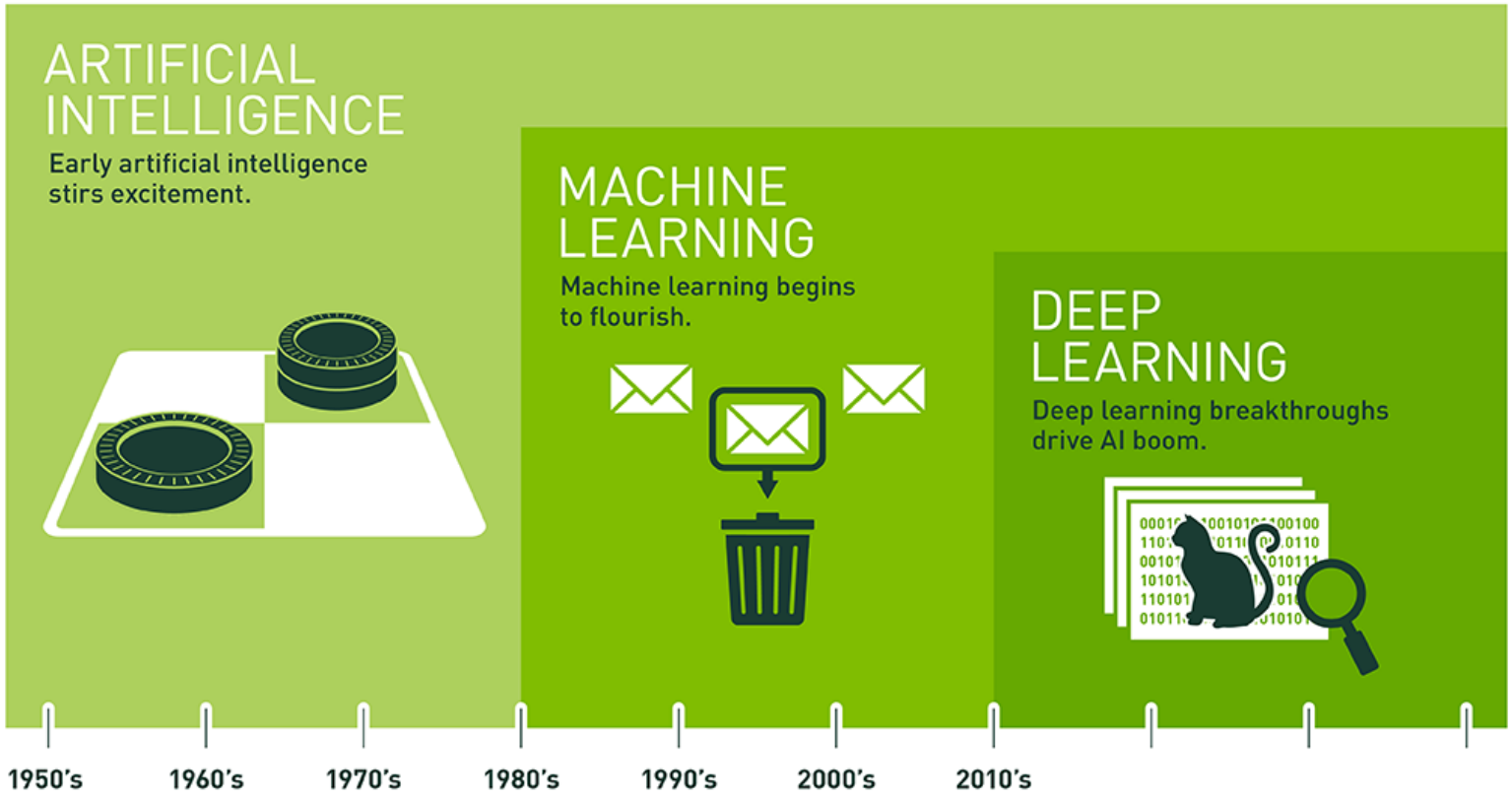
- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 实战案例：利用声音数据进行性别识别

目录

- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 实战案例：利用声音数据进行性别识别

什么是机器学习？

人工智能 vs 机器学习 vs 深度学习

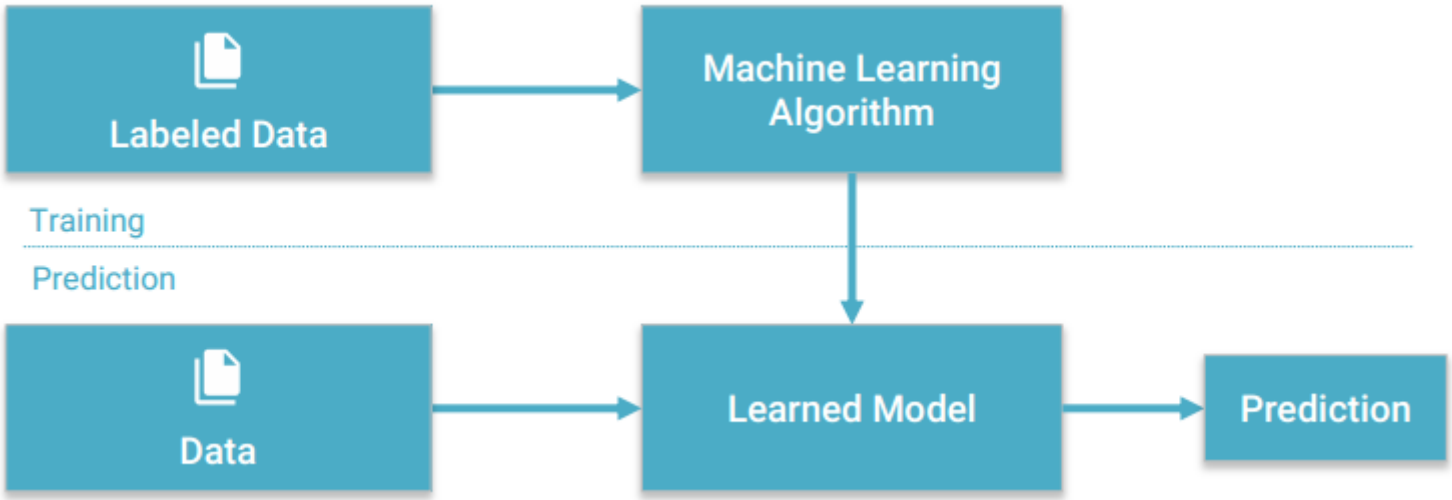


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

什么是机器学习？

定义

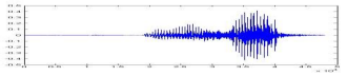
- Machine Learning is a type of Artificial Intelligence that provides computers with the ability to learn without being explicitly programmed.
- Provides various techniques that can learn from and make predictions on DATA.



什么是机器学习？

≈ 寻找一个函数

- 语音识别

  f() = “你好吗？”

- 图像识别

  f() = “猫”

- 围棋对战

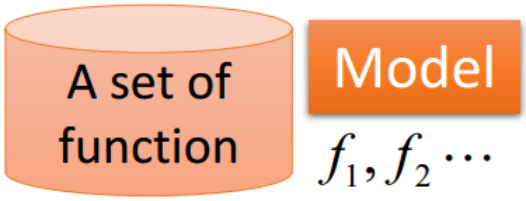
  f() = “5-5”(下一步)

- 对话系统（如Siri）

  f(“你好！”(用户发问)) = “您好！”(系统回应)

什么是机器学习？

如何选择？



图像识别

$f($  $) =$ “猫”

$f_1($  $) =$ “猫”

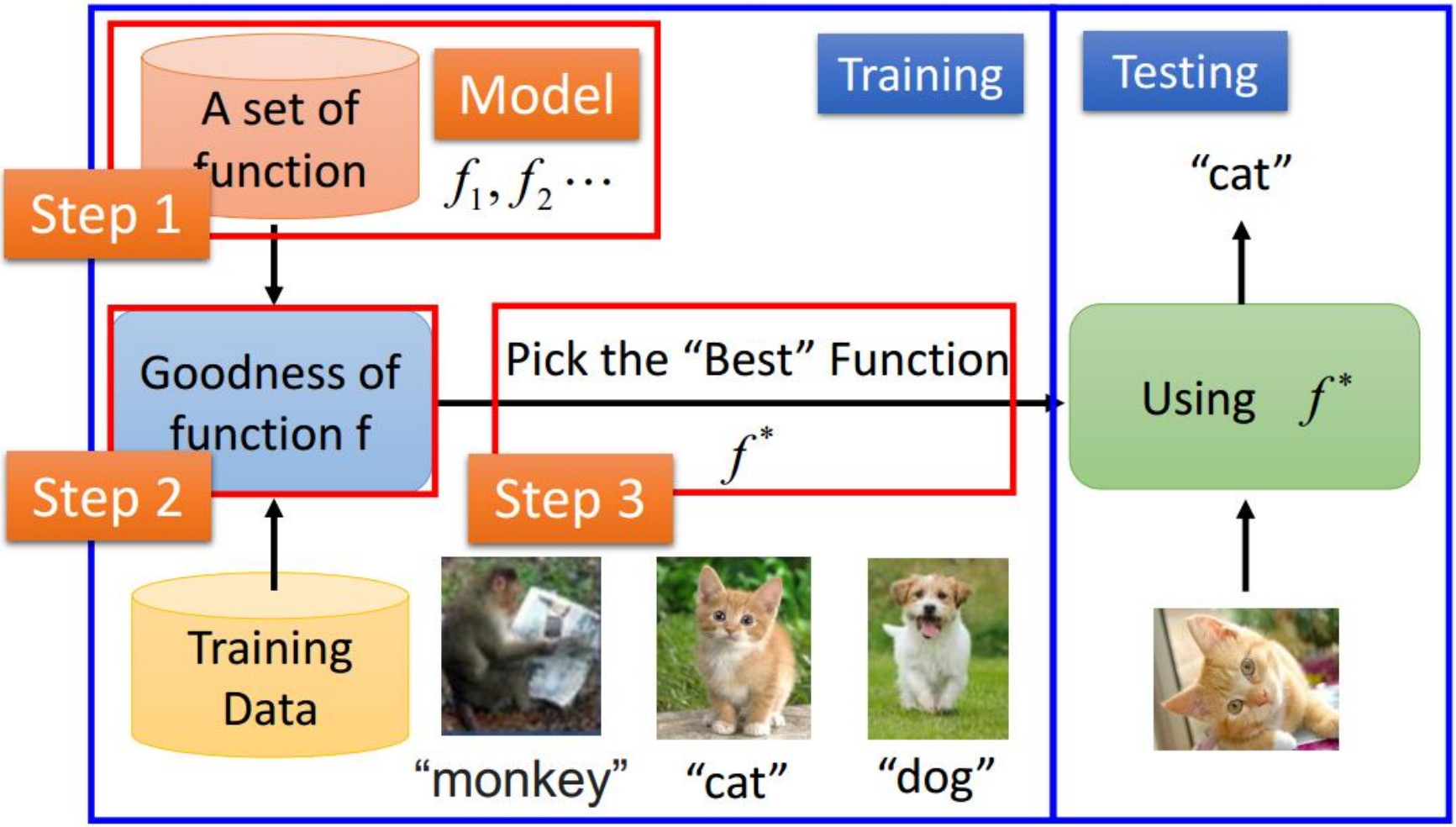
$f_2($  $) =$ “猴子”

$f_1($  $) =$ “狗”

$f_2($  $) =$ “蛇”

什么是机器学习？

基本框架



什么是机器学习？

基本步骤

Step 1.

定义一系列函数

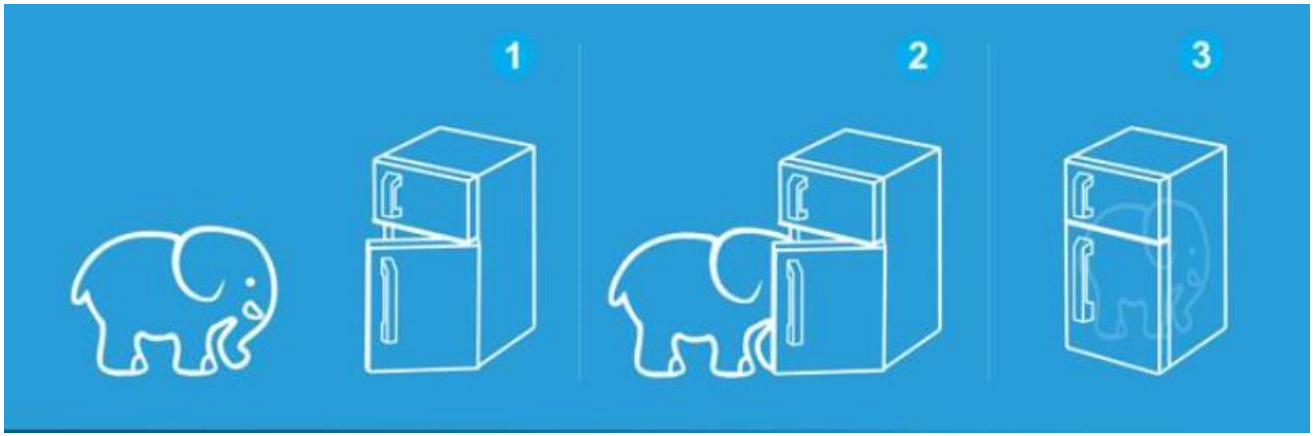
Step 2:

定义函数的优劣

Step3:

选择最优的函数

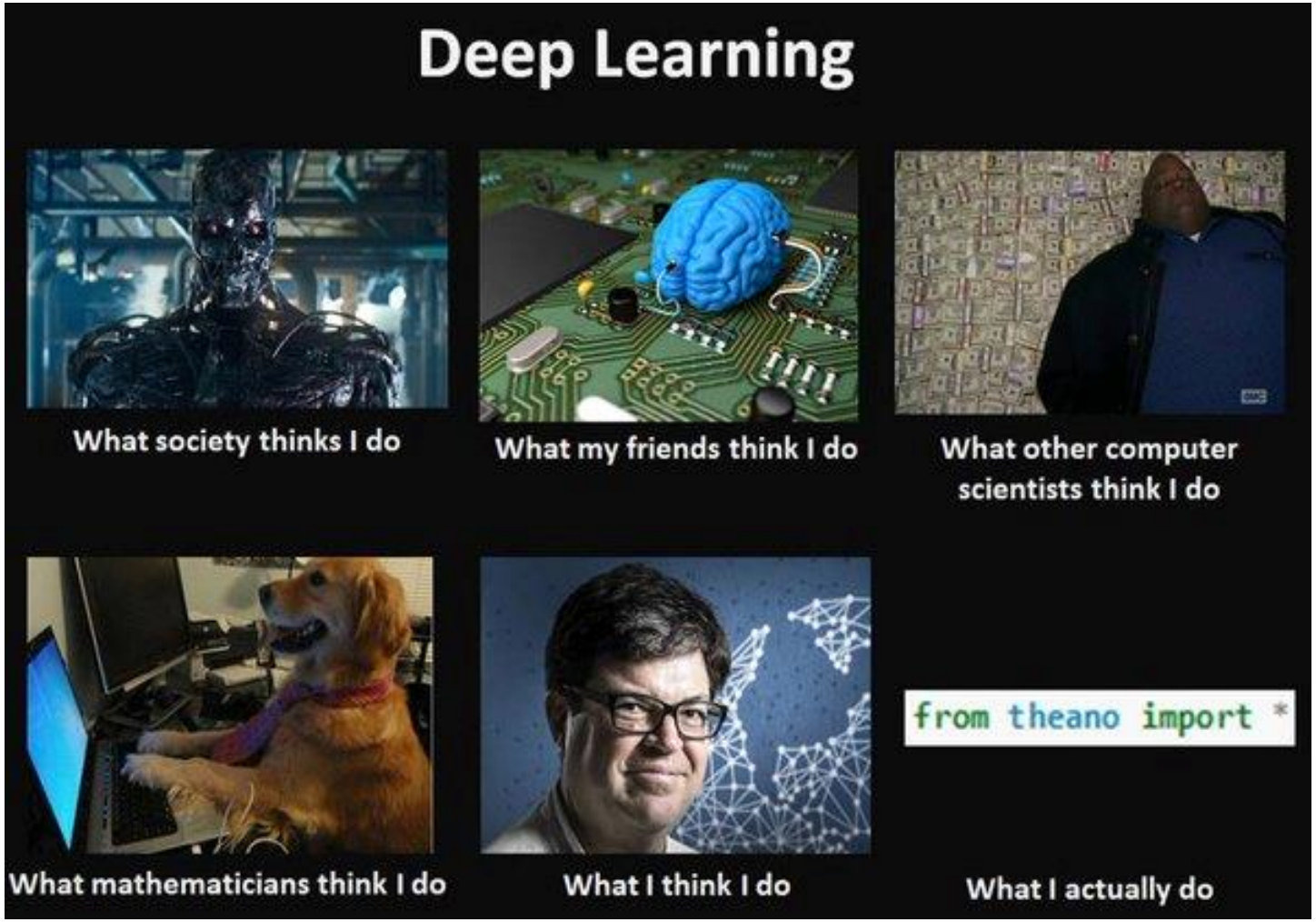
机器学习就是这么简单...



目录

- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 实战案例：利用声音数据进行性别识别

什么是scikit-learn?



什么是scikit-learn?

- 面向Python的免费机器学习库
- 包含分类、回归、聚类算法，比如：SVM、随机森林、k-means等
- 包含降维、模型筛选、预处理等算法
- 支持NumPy和SciPy数据结构
- 用户

  http://scikit-learn.org/stable/testimonials/testimonials.html
- 安装
  - pip install scikit-learn
  - conda install scikit-learn

通过scikit-learn认识机器学习

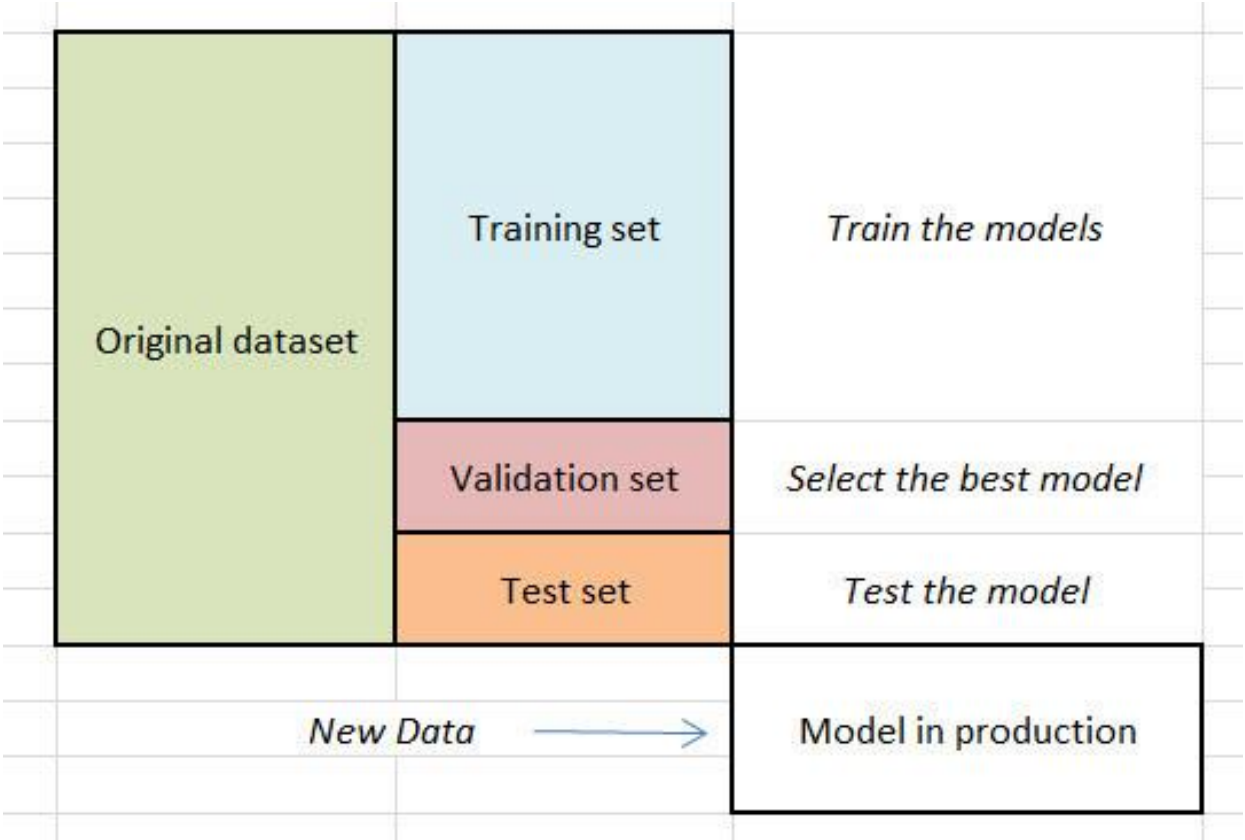
机器学习：问题描述

- “学习” 问题通常包括n个样本数据（训练样本），然后预测未知数据（测试样本）的属性
- 每个样本包含的多个属性（多维数据）被称作“特征”
- 分类：
  - 监督学习，训练样本包含对应的“标签”，如识别问题
    - 分类问题，样本标签属于两类或多类（离散）
    - 回归问题，样本标签包括一个或多个连续变量（连续）
  - 无监督学习，训练样本的属性不包含对应的“标签”，如聚类问题

通过scikit-learn认识机器学习

机器学习：问题描述（续）

- 训练集 vs 验证集 vs 测试集



通过scikit-learn认识机器学习

scikit-learn 上手

- 加载示例数据集
  - iris
  - digits
- 在训练集上训练模型
  - svm模型
  - .fit() 训练模型
- 在测试集上测试模型
  - .predict() 进行预测
- 保存模型
  - pickle.dumps()

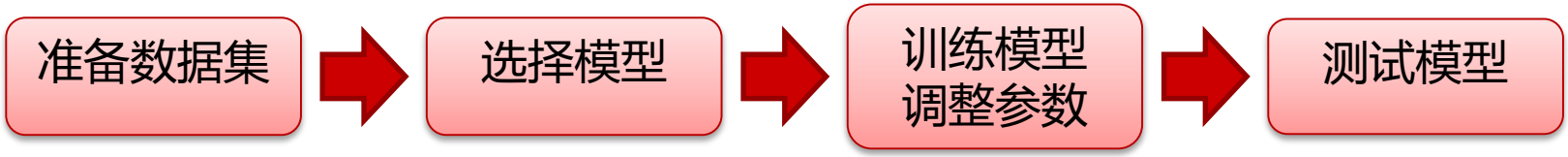
示例代码： 01_scikit_ml.ipynb

目录

- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 实战案例：利用声音数据进行性别识别

scikit-learn入门

使用scikit-learn的流程



- 数据处理
- 特征工程
- 训练集、测试集分割

- 根据任务选择模型
- 分类模型
- 回归模型
- 聚类模型
- ……

- 根据经验设定参数
- 交叉验证确定最优参数

- 预测
- 识别
- ……

scikit-learn入门

准备数据集

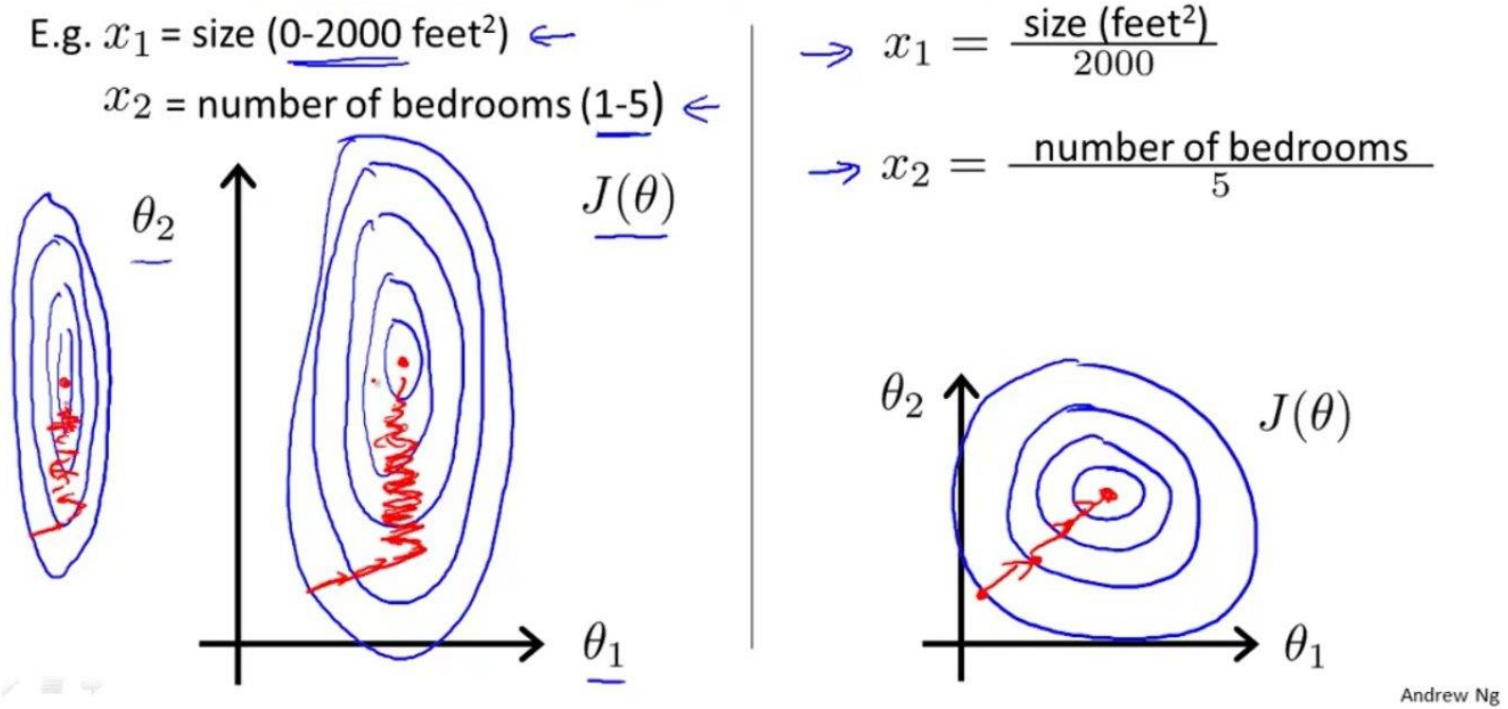
- 数据处理
  - 数据集格式
  - 二维数组，形状 (n_samples, n_features)
  - 使用np.reshape()转换数据集形状
- 特征工程
  - 特征提取
  - 特征归一化 (normalization)
  - ……
- train_test_split() 分割训练集、测试集

示例代码： 02_scikit_tutorial.ipynb

scikit-learn入门

准备数据集（续）

- 特征归一化（normalization）
  - preprocessing.scale()



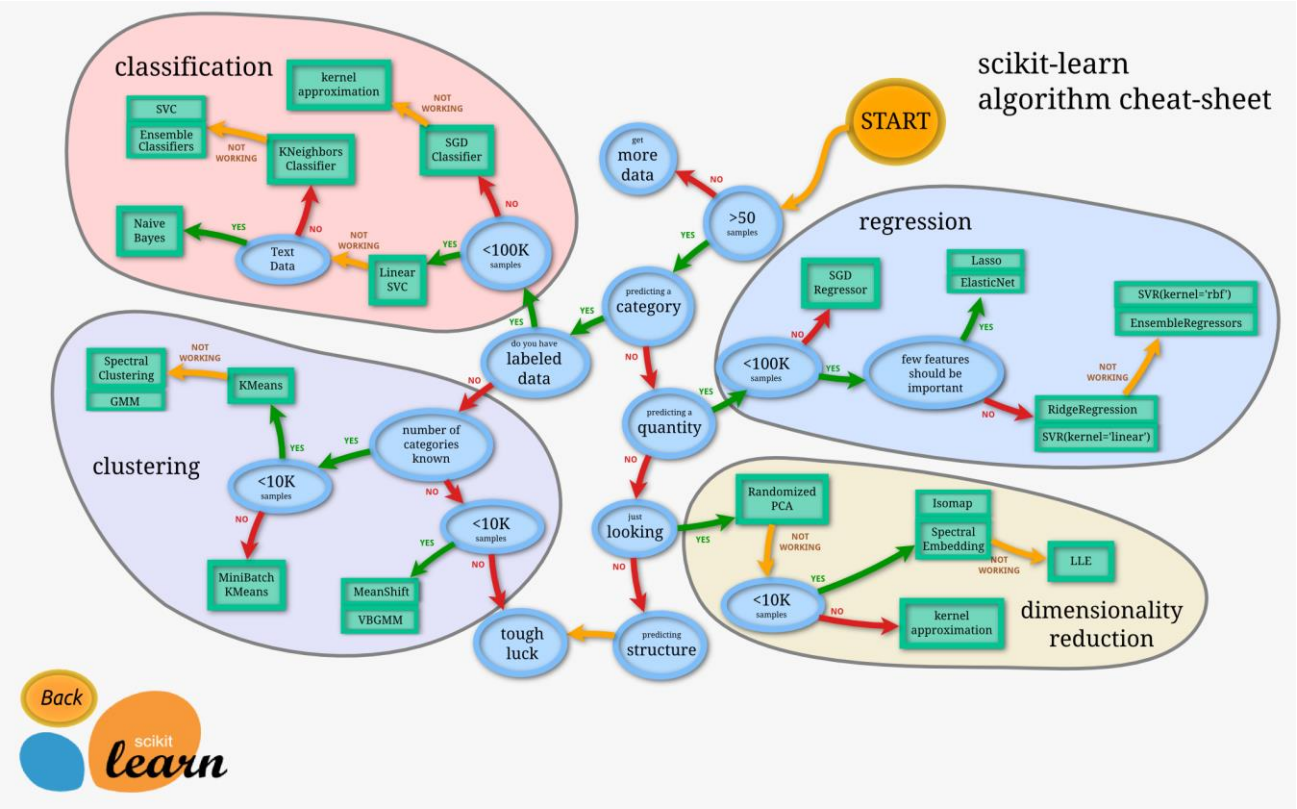
示例代码： 02_scikit_tutorial.ipynb

scikit-learn入门

选择模型

- 模型选择路线图

  http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



scikit-learn入门

训练模型

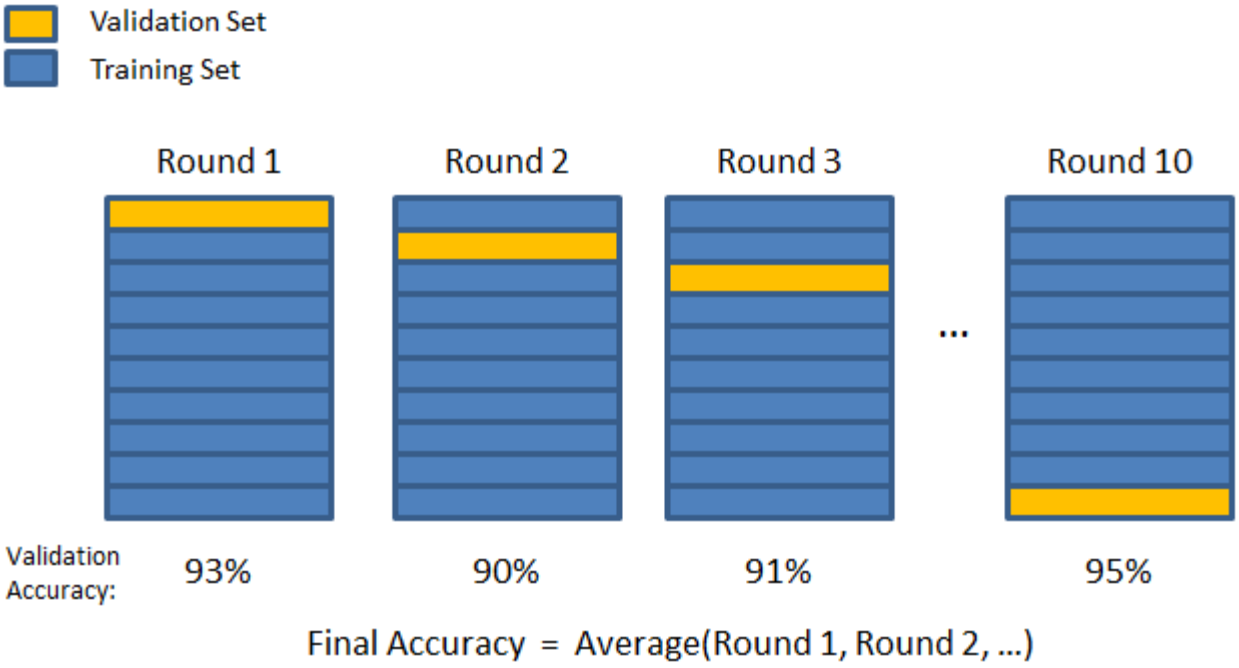
- Estimator对象
- 从训练数据学习得到的
- 可以是分类算法、回归算法或者是特征提取算法
- fit方法用于训练Estimator
- Estimator的参数可以训练前初始化，或者之后更新
- get_params()返回之前定义的参数
- score()对Estimator进行评分
  - 回归模型：使用“决定系数”评分（Coefficient of Determination）
  - 分类模型：使用“准确率”评分（accuracy）

示例代码： 02_scikit_tutorial.ipynb

scikit-learn入门

调整参数

- 依靠经验
- 依靠实验，交叉验证（cross validation）
  - cross_val_score()



示例代码： 02_scikit_tutorial.ipynb

scikit-learn入门

测试模型

- `model.predict(X_test)`
  - 返回测试样本的预测标签
- `model.score(X_test, y_test)`
  - 根据预测值和真实值计算评分

示例代码： `02_scikit_tutorial.ipynb`

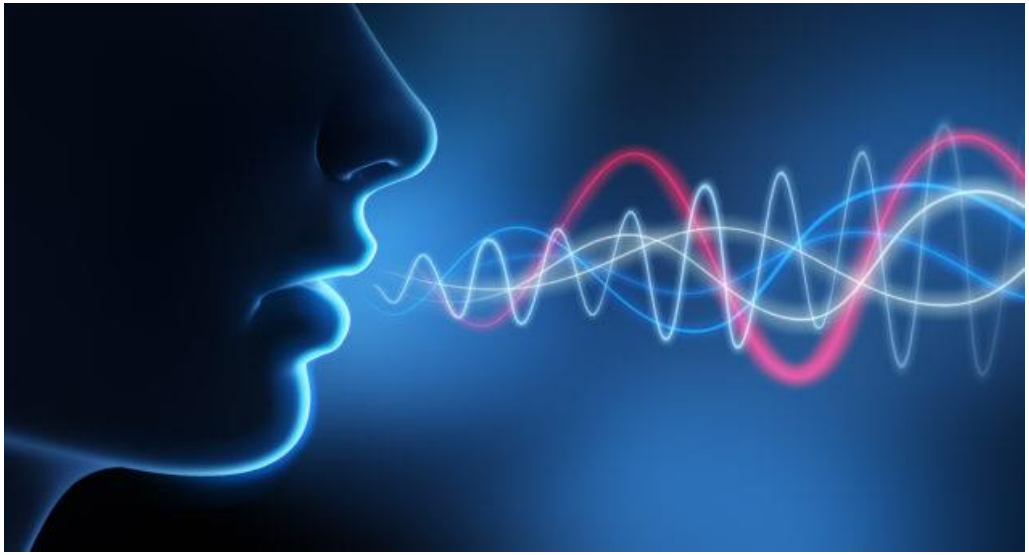
目录

- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 实战案例：利用声音数据进行性别识别

实战案例

项目介绍

- https://www.kaggle.com/primaryobjects/voicegender
- 样本包括已经提取的特征及标签



项目任务

- 根据声音属性识别说话者的性别

涉及知识点

- Pandas数据处理
- Seaborn绘图
- 使用scikit-learn完成机器学习

示例代码：lecture08_proj.zip

实战案例

分析步骤

1．查看数据

2．明确分析目标

3．处理缺失数据（可选）

4．数据统计分析

- 特征分布可视化

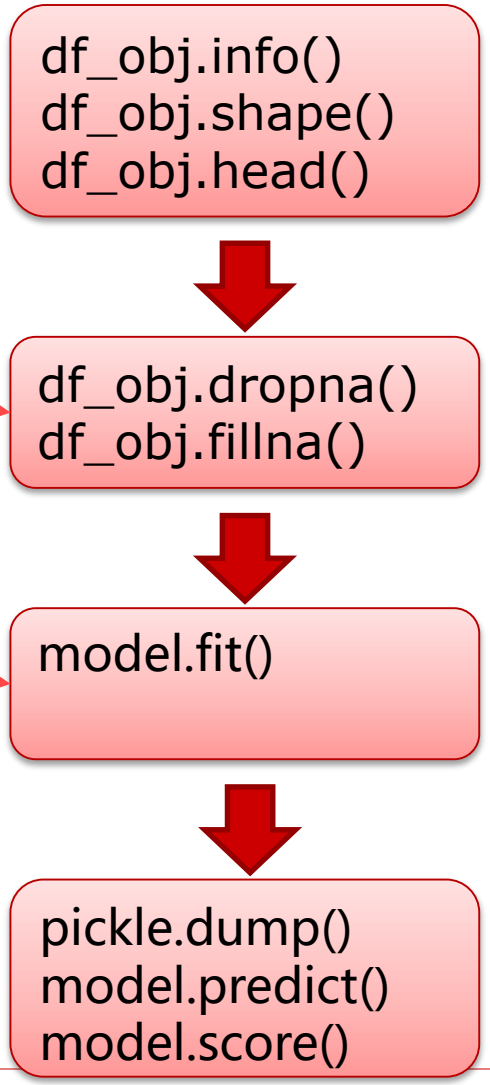
5．选择模型

- 训练模型
- 交叉验证（可选）

6．保存分析结果

1．保存模型

2．测试模型



示例代码：lecture08_proj.zip

参考

- 一天搞懂深度学习
  http://www.slideshare.net/tw_dsconf/ss-62245351
- scikit-learn 教程
  http://scikit-learn.org/stable/tutorial/
- 使用sklearn做单机特征工程
  http://www.cnblogs.com/jasonfreak/p/5448385.html
- 机器学习模型选择路线图
  http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- 项目参考自
  https://www.kaggle.com/lewuathe/d/primaryobjects/voicegender/evaluation-of-gender-classification-model

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