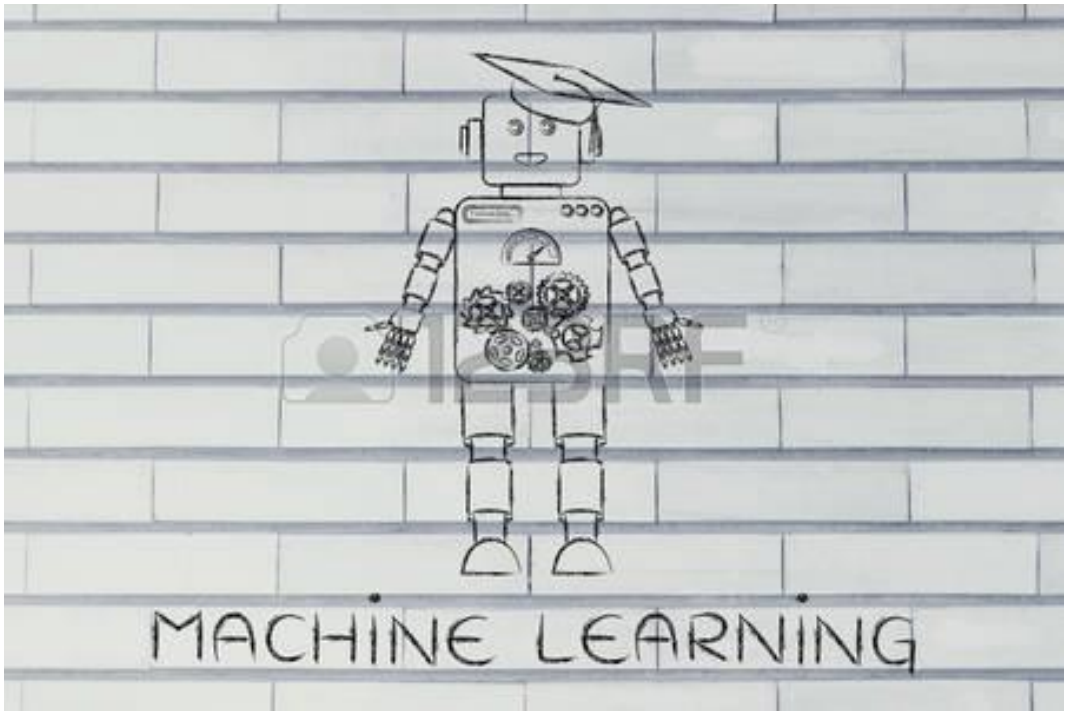
法律声明

- 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

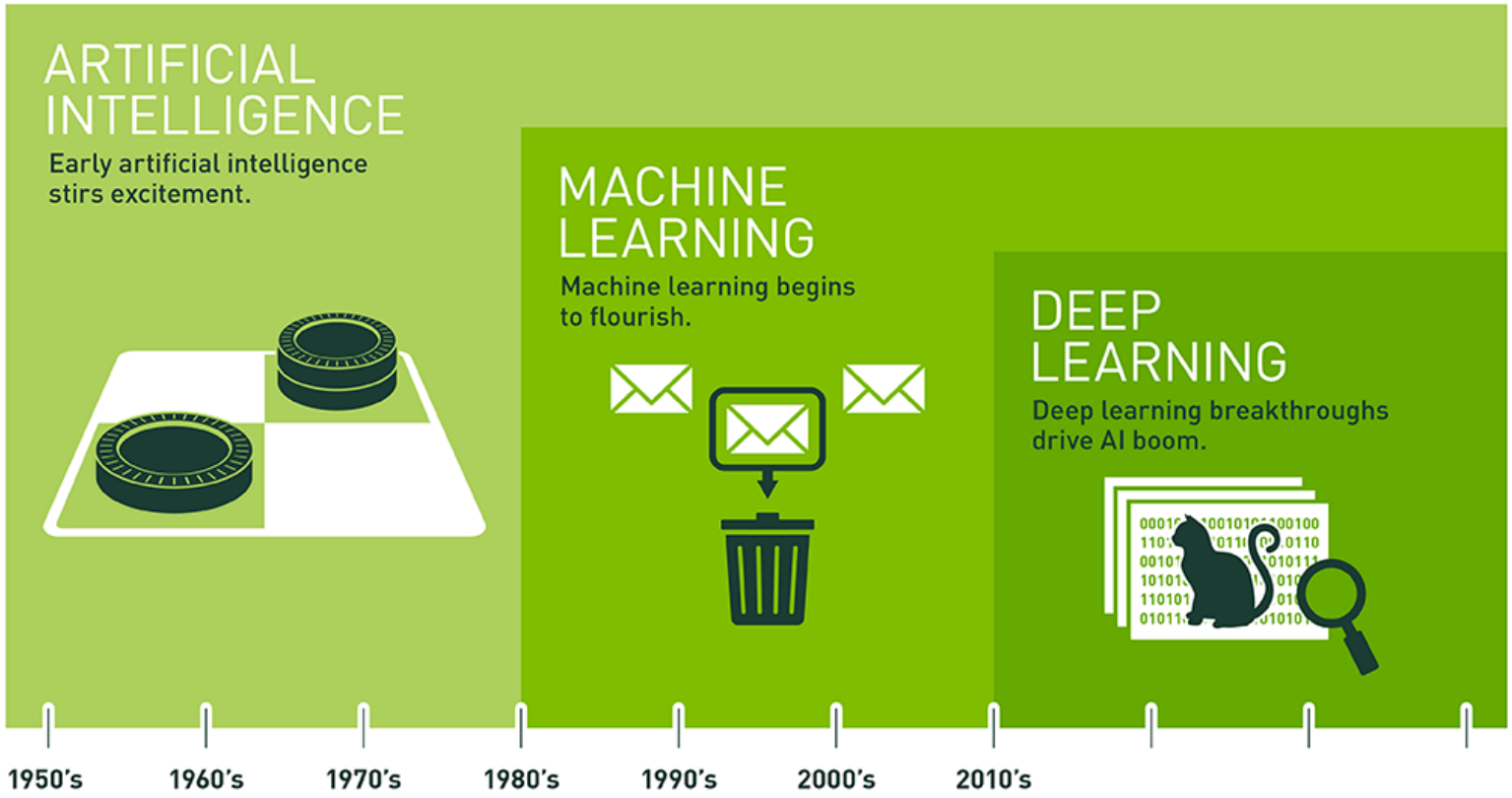
目录

- 什么是机器学习？
- 通过scikit-learn认识机器学习
- scikit-learn入门
- 特征降维—主成分分析
- 实战案例：识别Twitter用户性别

目录

什么是机器学习？

人工智能 vs 机器学习 vs 深度学习

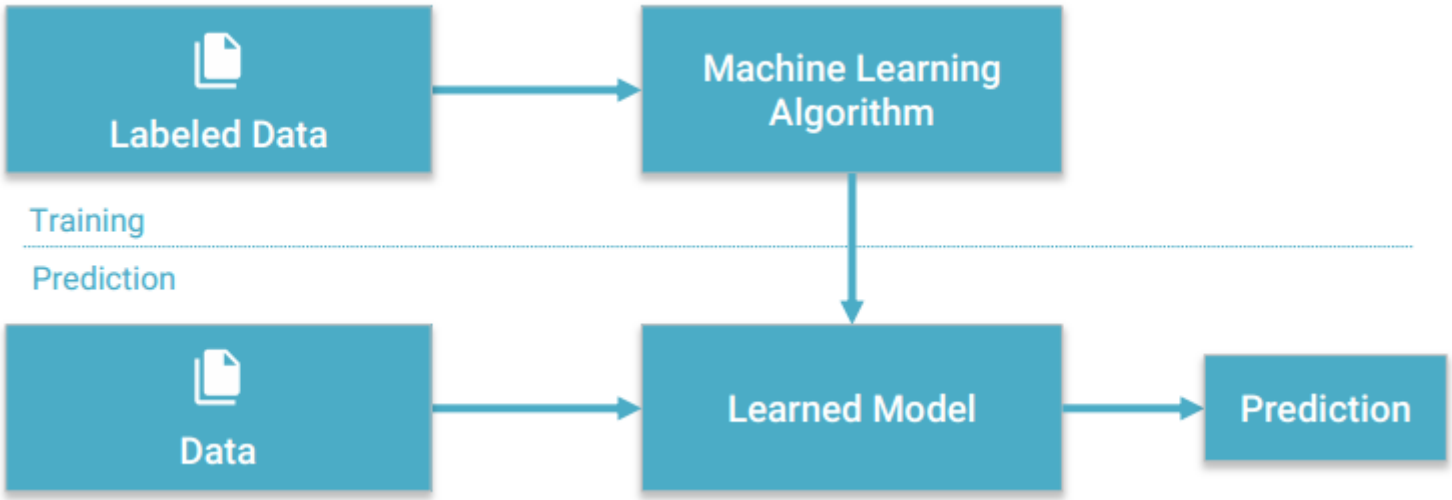


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

什么是机器学习？

定义

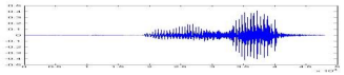
- Machine Learning is a type of Artificial Intelligence that provides computers with the ability to learn without being explicitly programmed.
- Provides various techniques that can learn from and make predictions on DATA.



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

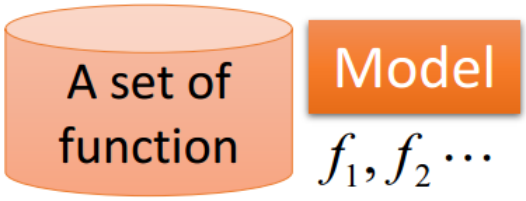
  f() = “5-5”(下一步)

- 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”

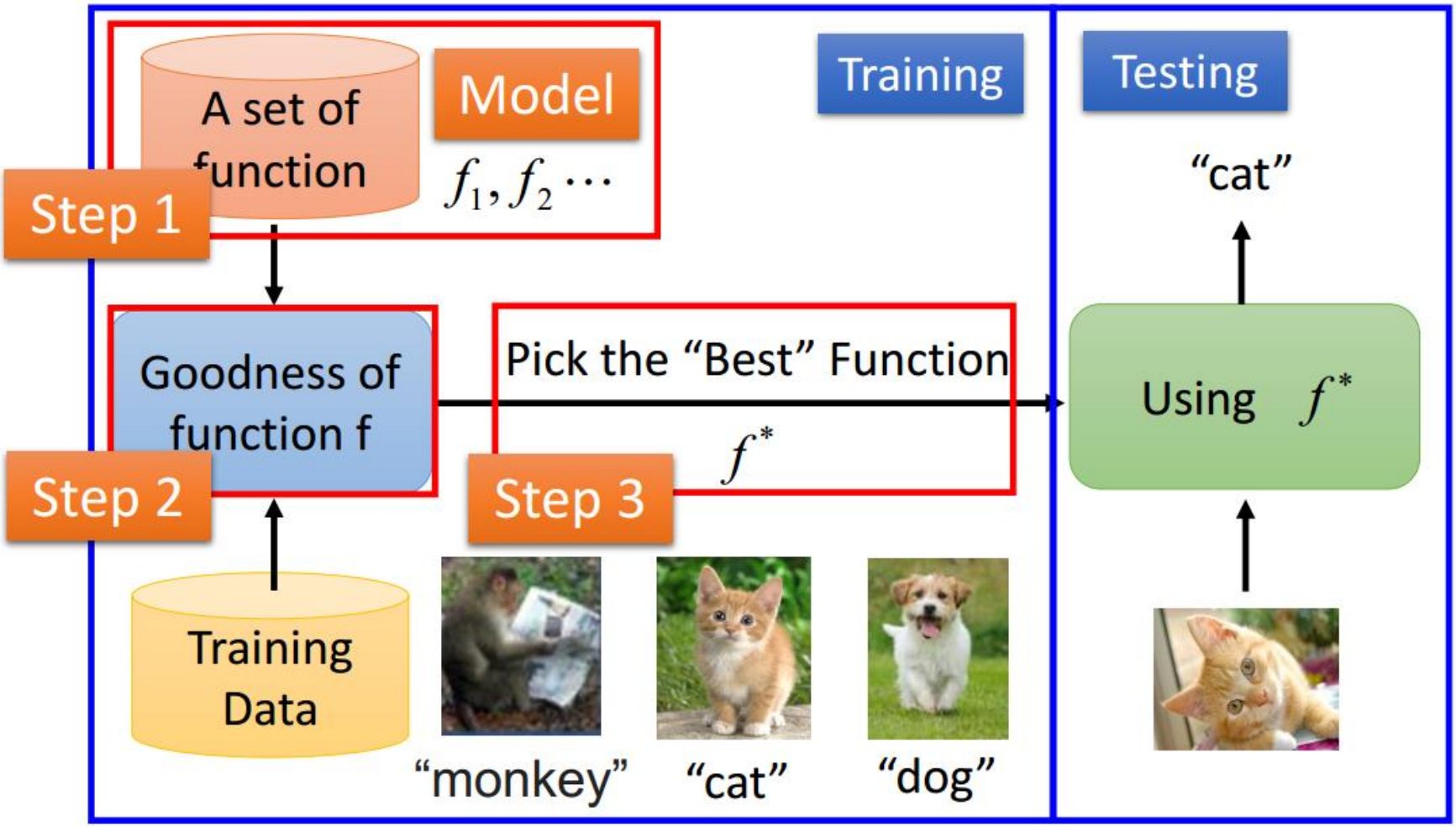
$f_2($  $) =$ “猴子”

$f_1($  $) =$ “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

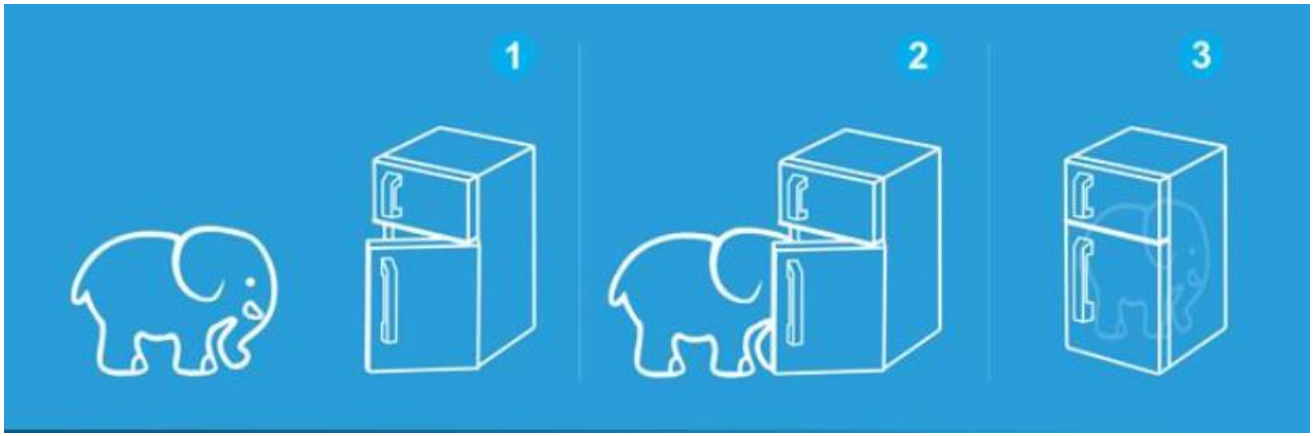
Step 2:

定义函数的优劣

Step3:

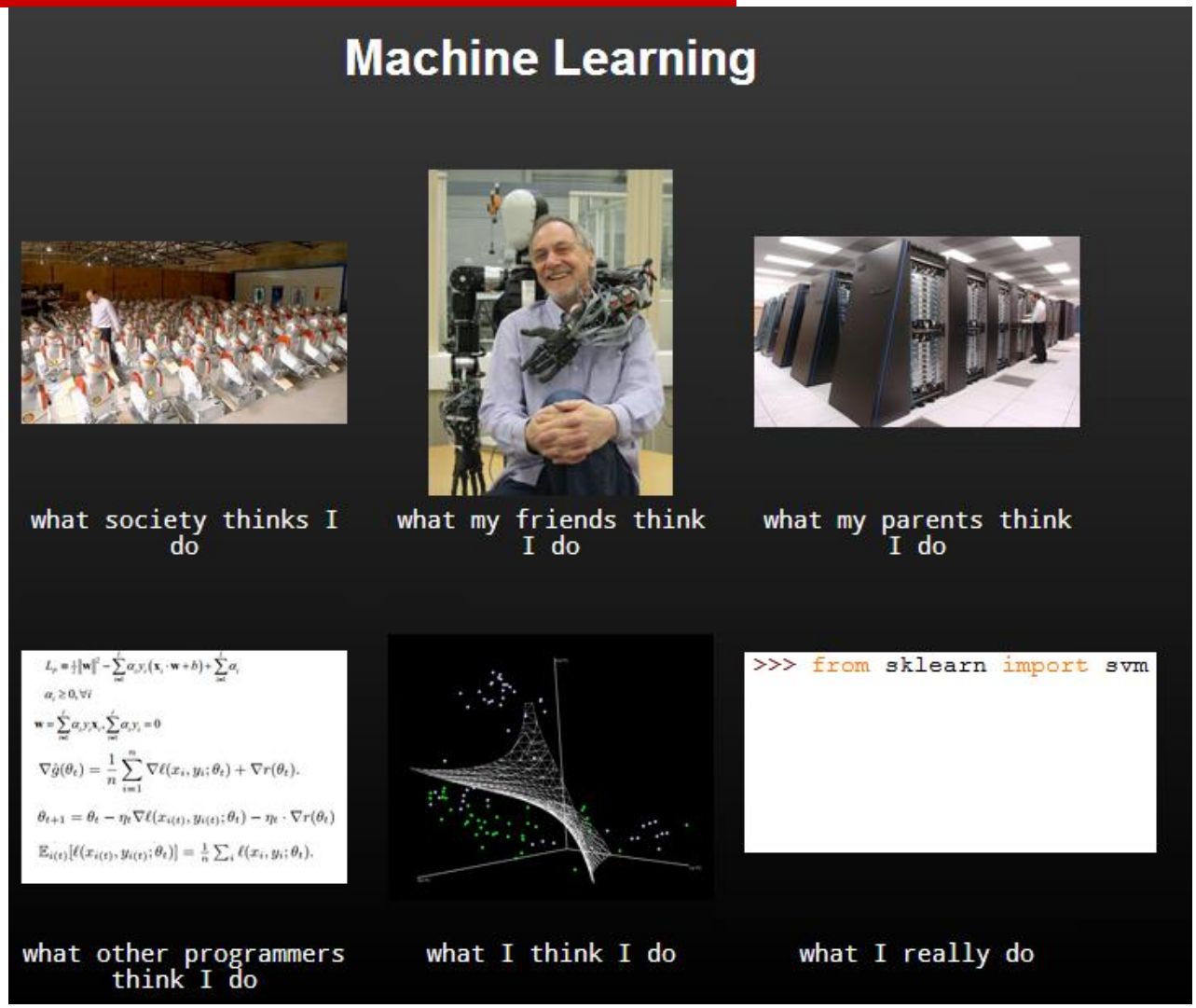
选择最优的函数

机器学习就是这么简单...

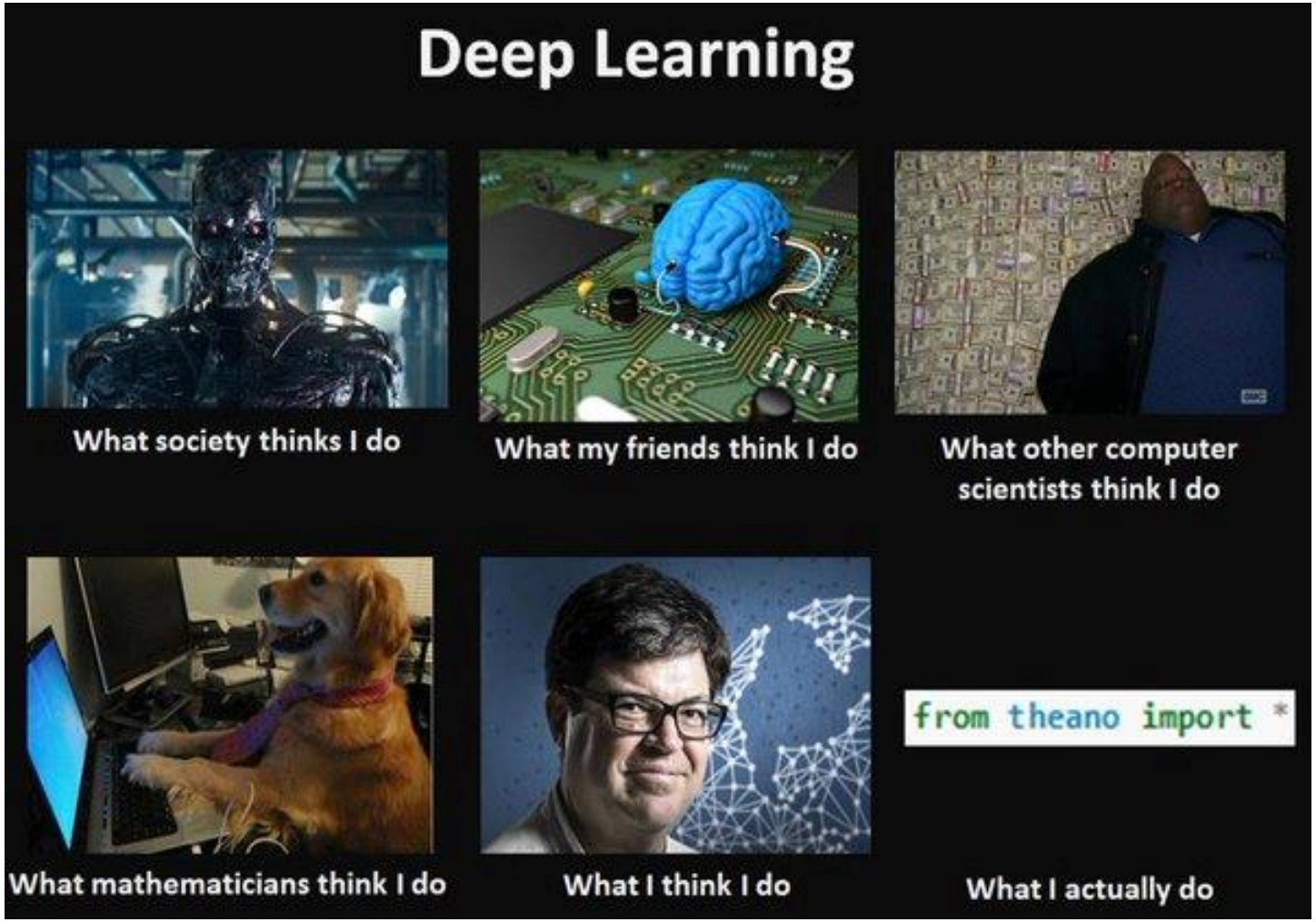


目录

什么是scikit-learn?



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

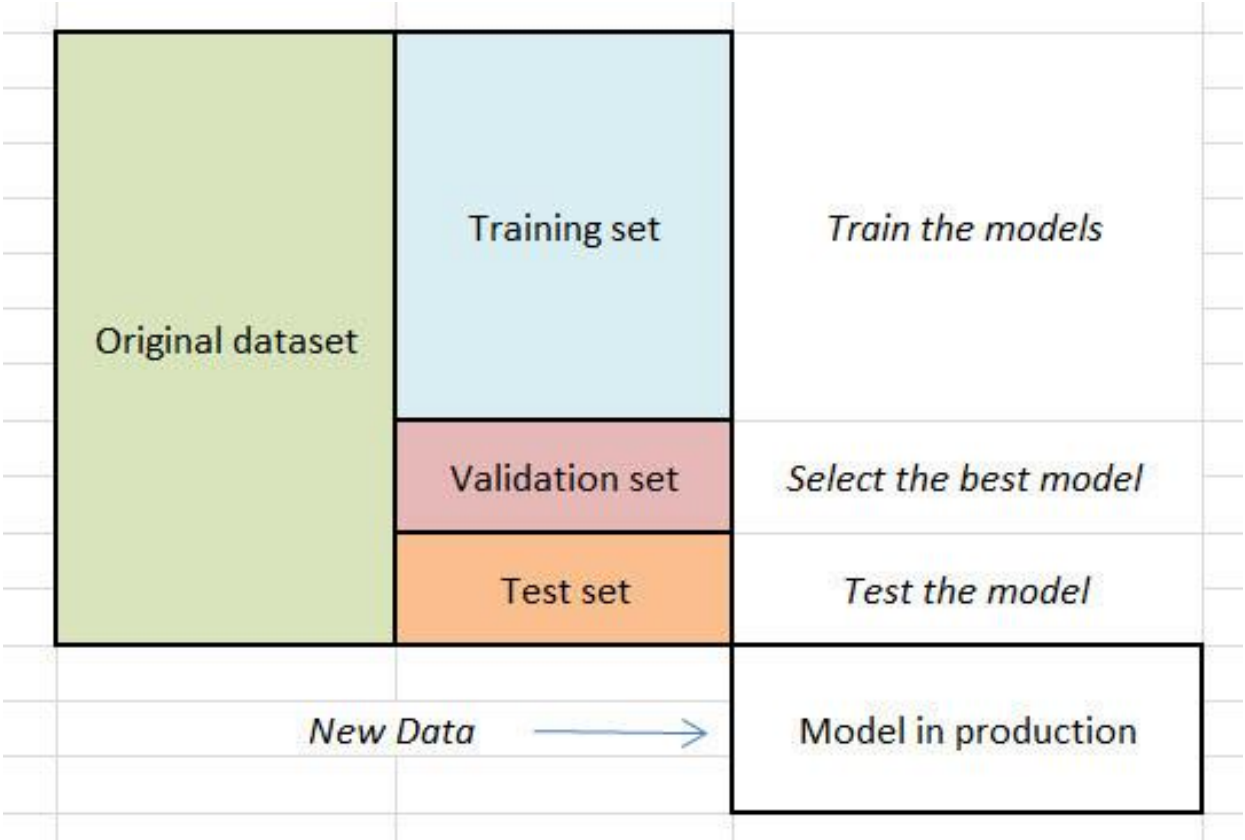
机器学习：问题描述

- “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）

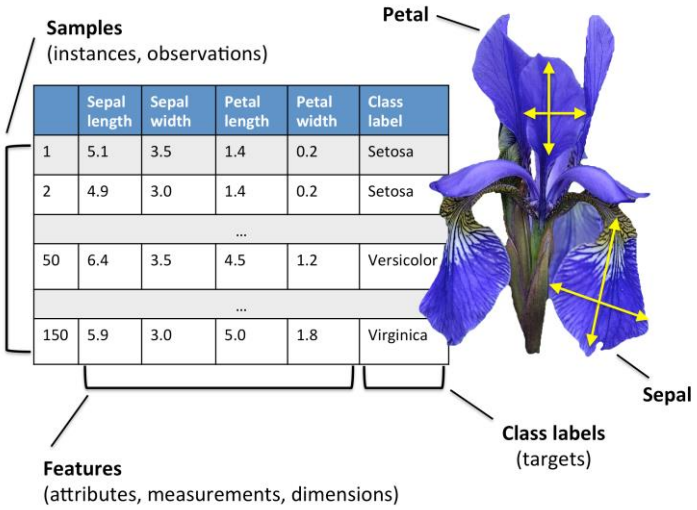
- 训练集 vs 验证集 vs 测试集



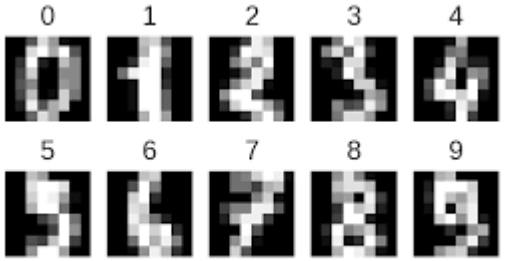
通过scikit-learn认识机器学习

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()



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

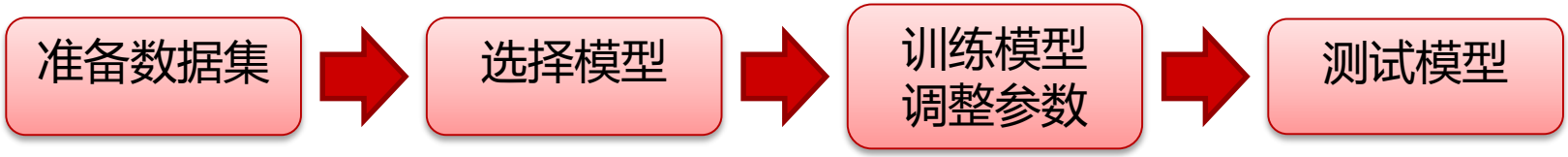


示例代码： 01_scikit_ml.ipynb

目录

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

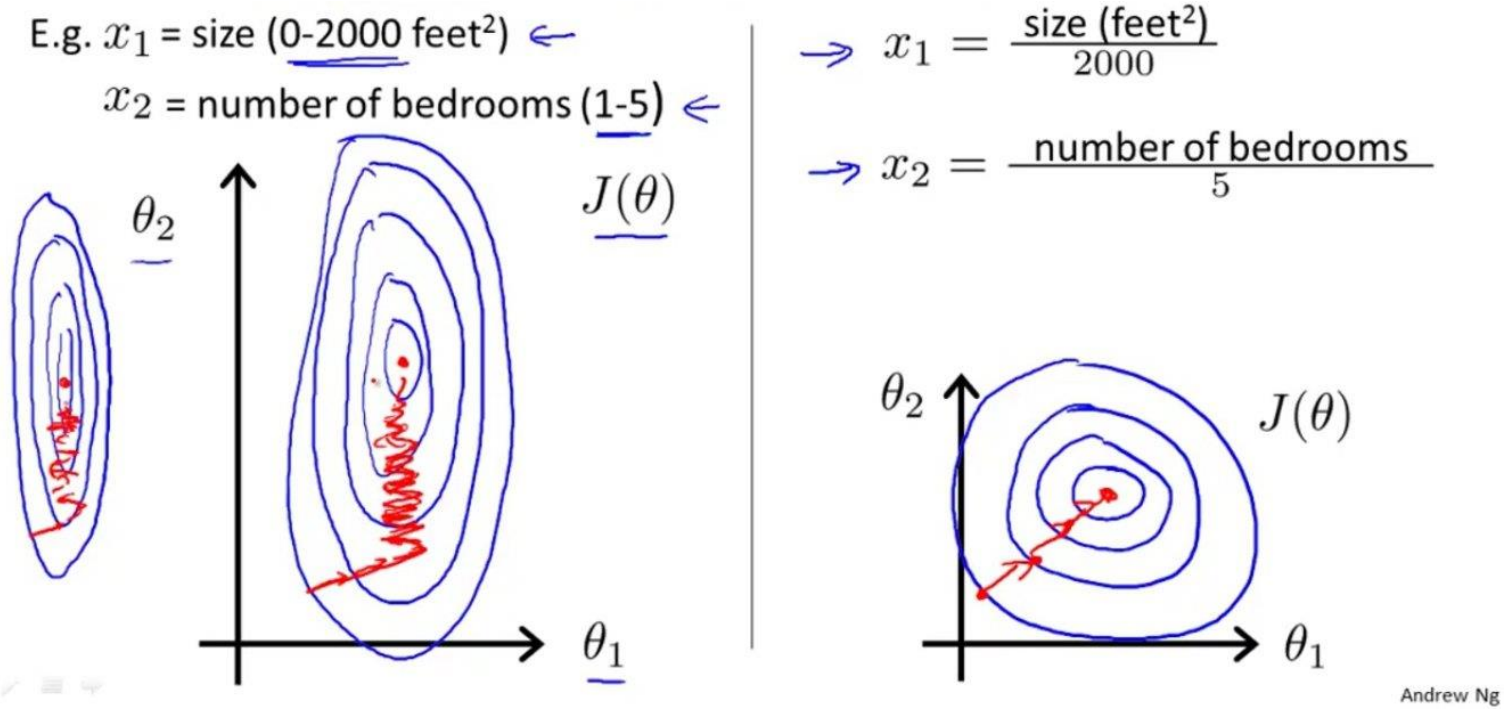
- 数据处理
  - 数据集格式
  - 二维数组，形状 (n_samples, n_features)
  - 使用np.reshape()转换数据集形状
- 特征工程
  - 特征提取
  - 特征归一化 (normalization)
  - ……
- train_test_split() 分割训练集、测试集

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()



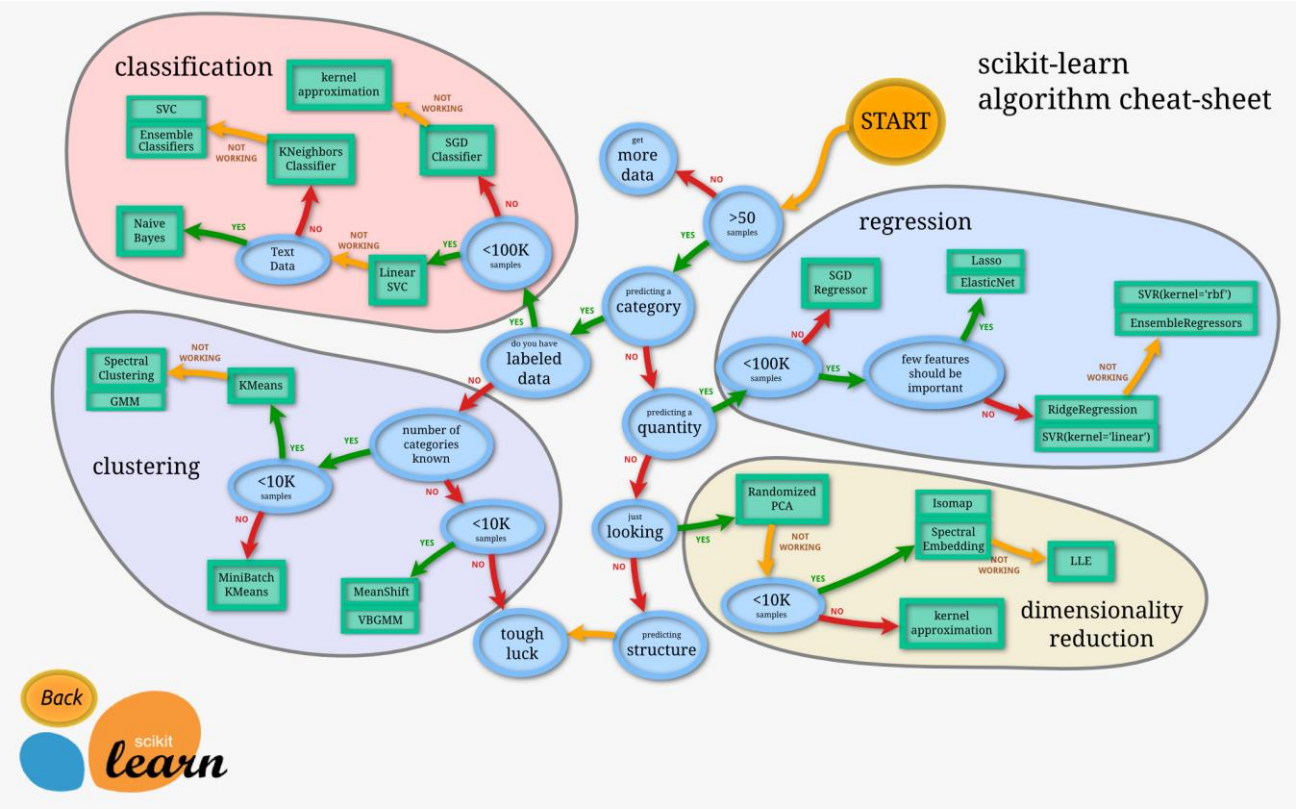
示例代码： 02_scikit_tutorial.ipynb

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

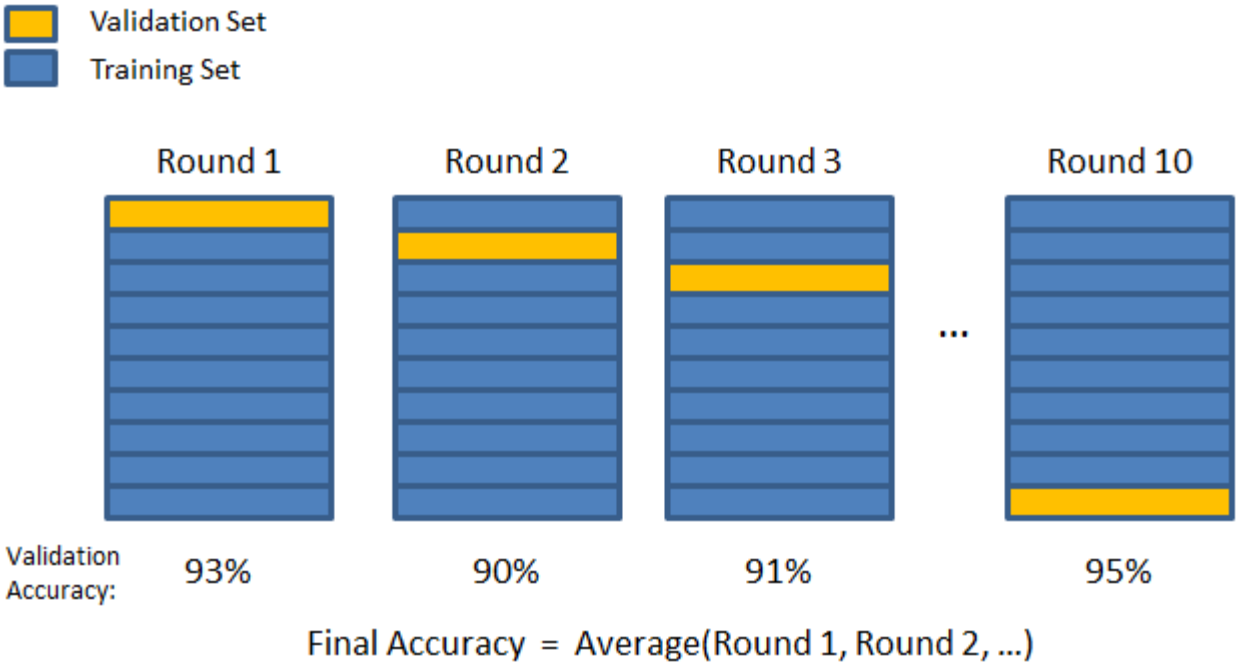
- Estimator对象
- 从训练数据学习得到的
- 可以是分类算法、回归算法或者是特征提取算法
- fit方法用于训练Estimator
- Estimator的参数可以训练前初始化，或者之后更新
- get_params()返回之前定义的参数
- score()对Estimator进行评分
  - 回归模型：使用“决定系数”评分（Coefficient of Determination）
  - 分类模型：使用“准确率”评分（accuracy）

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

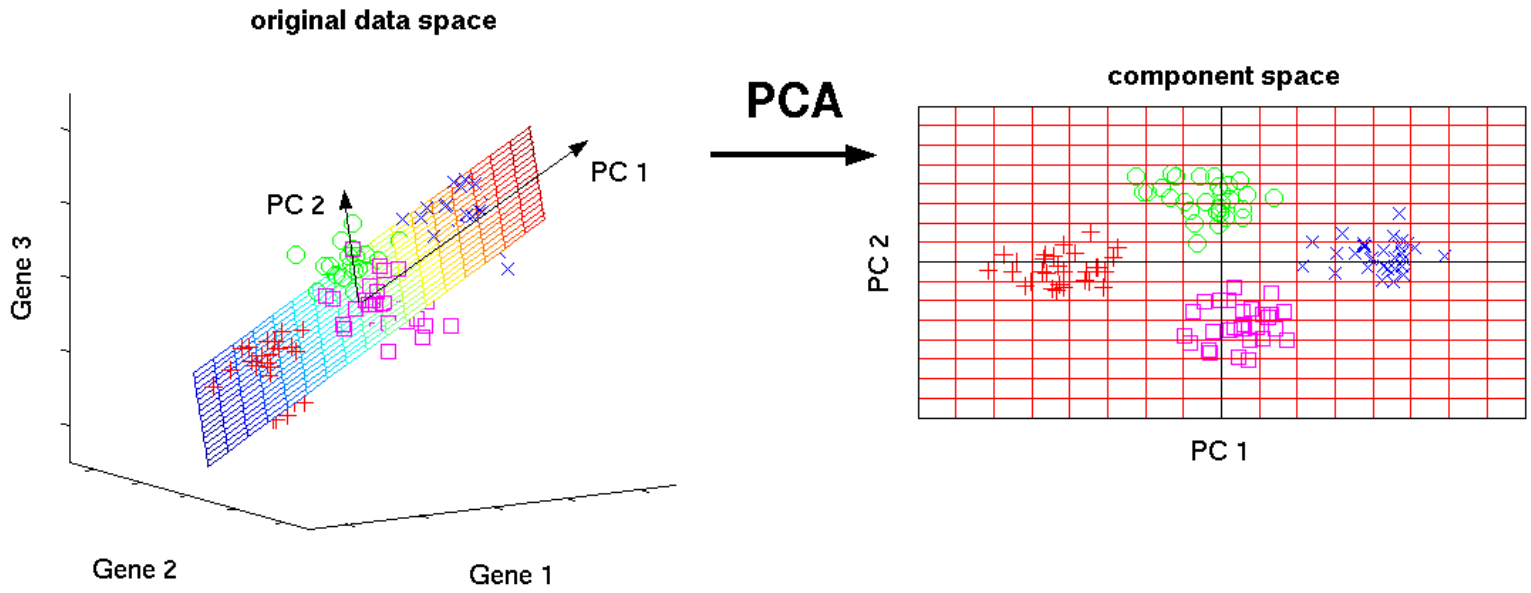
示例代码： `02_scikit_tutorial.ipynb`

目录

特征降维—主成分分析

Principal components analysis (PCA)

- 用于减少数据集的维度，同时保持数据集中的对方差贡献最大的特征
- 保留低阶主成分，忽略高阶成分，这样的低阶成分往往能够保留住数据的最重要方面



特征降维—主成分分析

方差与协方差

- 用于衡量一系列点在它们的重心或均值附近的分散程度
- 方差：衡量这些点在一个维度的偏差
- 协方差：衡量一个维度是否会对另一个维度有所影响，从而查看这两个维度之间是否有关系
  - 某个维度和自身之间的协方差就是其方差

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

协方差矩阵

- 如果数据集是d维的，(x1，x2，…，xd)，则可计算出(x1，x2)，(x1，x3)，…，(x1，xd)，(x2，x3)，…(x2，xd)，…(xd-1，xd)之间的协方差。由于协方差的对称性，再加上各维度自身的协方差，可以构成协方差矩阵

特征降维—主成分分析

$$\begin{bmatrix} \mathrm{cov}(x1,x1) & \mathrm{cov}(x1,x2) & \mathrm{cov}(x1,x3) \\ \mathrm{cov}(x2,x1) & \mathrm{cov}(x2,x2) & \mathrm{cov}(x2,x3) \\ \mathrm{cov}(x3,x1) & \mathrm{cov}(x3,x2) & \mathrm{cov}(x3,x3) \end{bmatrix}$$

协方差矩阵（续）

- 其中对角线上的是方差
- 协方差为正,代表两个变量变化趋势相同；反之亦然

PCA

- 通过线型变换将原数据映射到新的坐标系统中，使映射后的第一个坐标上的方差最大（即第一个主成分）,第二个坐标上的方差第二大（即第二个主成分）...

特征降维—主成分分析

PCA步骤：

1．数据集 $\mathbf{X} \in R^{m \times n}$，其中每个样本 $\mathbf{x}^{(\mathrm{i})} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}]$

计算每个维度的均值

$$\overline{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}] \in R^n$$

每个维度减去这个均值，得到一个矩阵

相当于将坐标系进行了平移

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \overline{\mathbf{x}} \\ \mathbf{x}^{(2)} - \overline{\mathbf{x}} \\ \cdots \\ \mathbf{x}^{(\mathrm{m})} - \overline{\mathbf{x}} \end{bmatrix}$$

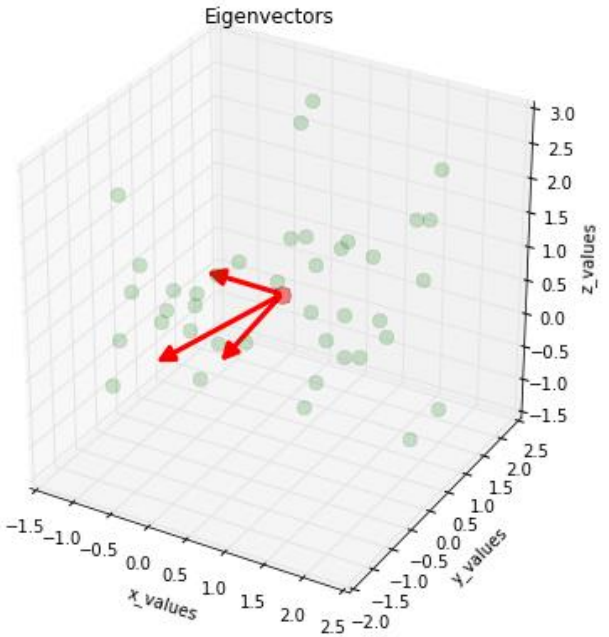
特征降维—主成分分析

PCA步骤：

2．构建协方差矩阵 $\mathbf{Q} = \mathbf{Y}^T\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \bar{\mathbf{x}} & \mathbf{x}^{(2)} - \bar{\mathbf{x}} & \dots & \mathbf{x}^{(m)} - \bar{\mathbf{x}} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} - \bar{\mathbf{x}} \\ \mathbf{x}^{(2)} - \bar{\mathbf{x}} \\ \dots \\ \mathbf{x}^{(m)} - \bar{\mathbf{x}} \end{bmatrix}$

3．矩阵分解（如SVD），得到特征值(eigenvalues)及特征向量（eigenvectors）

4．将特征值从大到小排序，对应的特征向量就是第一个主成分，第二个主成分…



如何选择主成分个数？

- 交叉验证
- 根据主成分的累计贡献率

应用：特征提取、数据降维

示例代码： 03_scikit_pca.ipynb

目录

实战案例

项目介绍

- https://www.kaggle.com/crowdflower/twitter-user-gender-classification

项目任务

- 根据twitter的数据识别用户的性别

涉及知识点

- 网络爬虫
- 文本特征提取
- 图像特征提取
- 使用scikit-learn完成机器学习

示例代码：lecture08_proj

实战案例

分析步骤

1．查看数据

2．明确分析目标

3．数据清洗(可选)

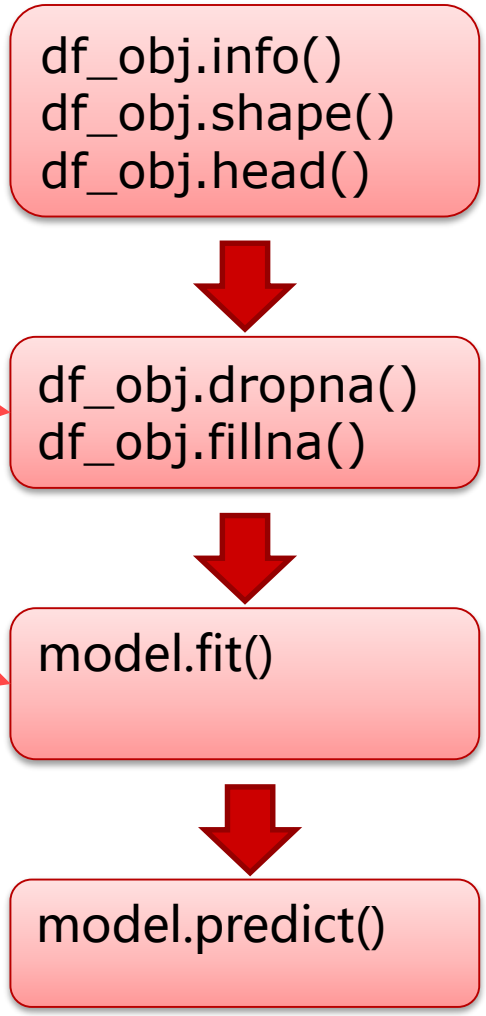
4．特征工程

- 特征提取
- 归一化
- 降维处理(可选)

5．选择模型

- 训练模型
- 交叉验证（可选）

6．模型测试



示例代码：lecture08_proj

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
- 主成分分析可视化
  http://setosa.io/ev/principal-component-analysis/

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