Python程序设计

第十讲 机器学习应用

Python机器学习概述

张 华

WHU

机器学习的目标

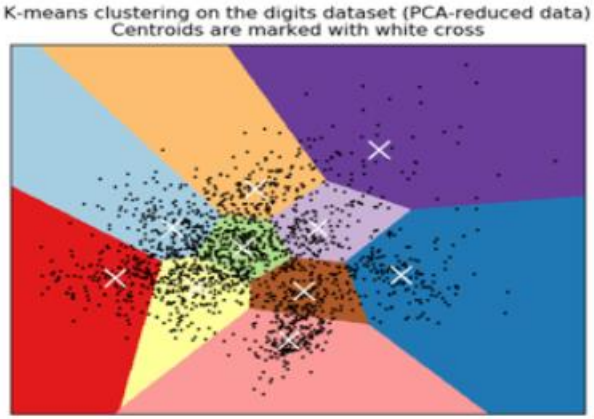
- **人工智能（Artificial Intelligence，AI）**
  - 是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学。
  - 是计算机科学的一个分支，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。
- **机器学习是实现人工智能的手段。**
  - **Machine Learning，ML**的主要研究内容是如何利用数据或经验进行学习，改善具体算法的性能。
  - 多领域交叉，涉及概率论、统计学，算法复杂度理论等多门学科。
  - 广泛应用于网络搜索、垃圾邮件过滤、推荐系统、广告投放、信用评价、欺诈检测、股票交易和医疗诊断等应用。

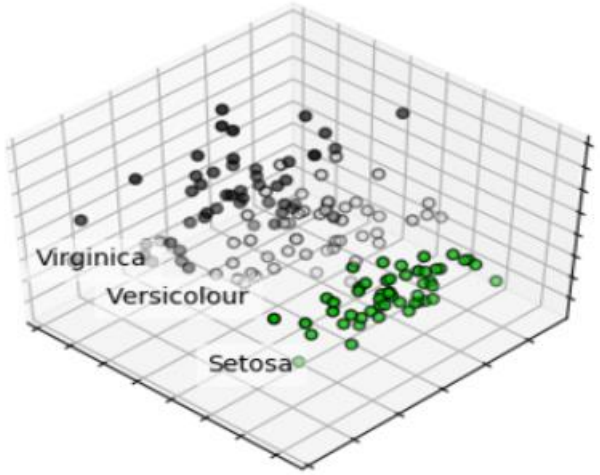
机器学习分类

- 机器学习一般分为下面几种类别
  - 无监督学习（Unsupervised Learning）
  - 监督学习（Supervised Learning）
  - 半监督学习（Semi-supervised Learning）
  - 深度学习（Deep Learning）
  - 强化学习（Reinforcement Learning，增强学习）

无监督学习

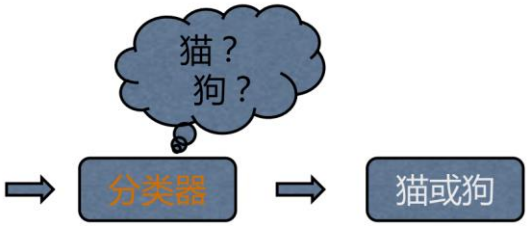
- 利用无标签的数据学习数据的分布或数据与数据之间的关系被称作无监督学习。
  - 有监督学习和无监督学习的最大区别在于数据是否有标签。
- 无监督学习最常应用的场景是
  - 聚类（Clustering）
  - 降维（Dimension Reduction）

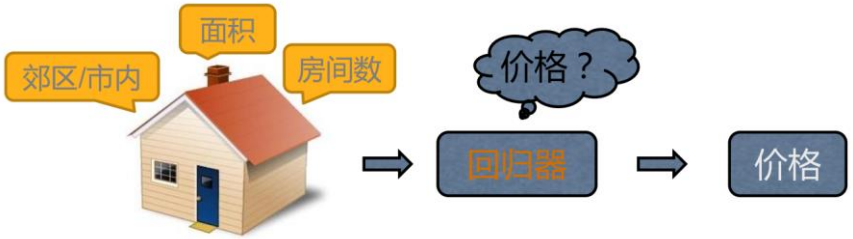




监督学习

- 利用一组带有标签的数据，学习从输入到输出的映射，然后将这种映射关系应用到未知数据上，达到分类或回归的目的。
  - 分类：当输出是离散的，学习任务为分类任务。
  - 回归：当输出是连续的，学习任务为回归任务。





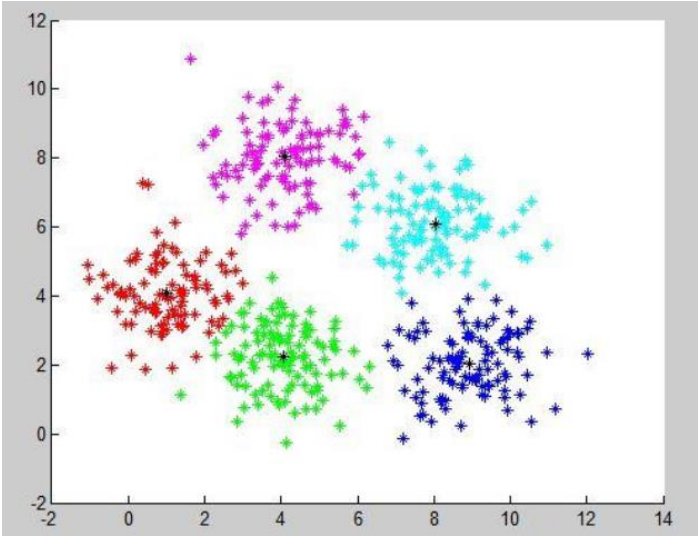
强化学习

- 强化学习就是程序或智能体（agent）通过与环境不断地进行交互学习一个从环境到动作的映射，学习的目标就是使累计回报最大化。
  - 强化学习是一种试错学习，因其在各种状态（环境）下需要尽量尝试所有可以选择的动作，通过环境给出的反馈（即奖励）来判断动作的优劣，最终获得环境和最优动作的映射关系（即策略）。

无监督学习的聚类算法

聚类

- 就是根据数据的“相似性”将数据分为多类的过程。
- 评估两个不同样本之间的“相似性”，通常使用的方法就是计算两个样本之间的“距离”。
- 使用不同的方法计算样本间的距离会关系到聚类结果的好坏。

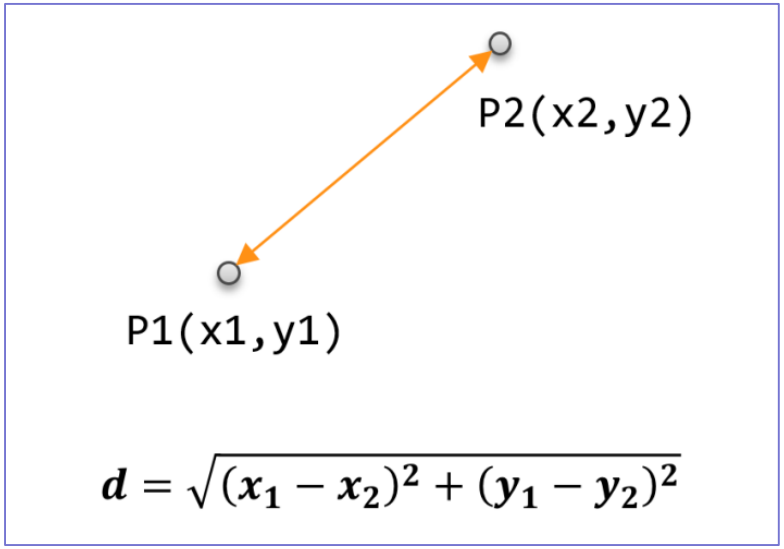


距离计算方法：欧式距离

欧式距离

- 欧氏距离是最常用的一种距离度量方法，源于欧式空间中两点的距离。
- 其计算方法如下：

$$d = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2}$$



二维空间中的欧式距离

距离计算方法

- 其他常用的距离计算方法
  - 曼哈顿距离
  - 马氏距离
  - 夹角余弦
  - ……

机器学习的学习资料

参考书

- 《机器学习》

  周志华 编著　　清华大学出版社　2016

在线课程

- 《机器学习》

  Andrew Ng（吴恩达） 编著　　Stanford

  https://www.coursera.org/learn/machine-learning

Python机器学习库

Scikit-learn

- http://scikit-learn.org/stable/
- sklearn是scikit-learn的简称，是一个基于Python的第三方模块。
- sklearn库集成了一些常用的机器学习方法，在进行机器学习任务时，并不需要实现算法，只需要简单的调用sklearn库中提供的模块就能完成大多数的机器学习任务。
- sklearn库是在Numpy、 Scipy和matplotlib的基础上开发而成的，因此在介绍sklearn的安装前，需要先安装这些依赖库。
  - Anaconda3已集成sklearn库。

Scikit-learn

sklearn自带的数据集

	数据集名称	调用方式	适用算法	数据规模
小数据集	波士顿房价数据集	load_boston()	回归	506*13
	鸢尾花数据集	load_iris()	分类	150*4
	糖尿病数据集	load_diabetes()	回归	442*10
	手写数字数据集	load_digits()	分类	5620*64
大数据集	Olivetti 脸部图像数据集	fetch_olivetti_faces()	降维	400*64*64
	新闻分类数据集	fetch_20newsgroups()	分类	-
	带标签的人脸数据集	fetch_lfw_people()	分类；降维	-
	路透社新闻语料数据集	fetch_rcv1()	分类	804414*47236

注：小数据集可以直接使用，大数据集要在调用时程序自动下载（一次即可）。

Scikit-learn

sklearn的基本功能

- sklearn库的共分为6大部分，分别用于完成分类任务、回归任务、聚类任务、降维任务、 模型选择以及数据的预处理。

分类模型	加载模块
最近邻算法	neighbors.NearestNeighbors
支持向量机	svm.SVC
朴素贝叶斯	naive_bayes.GaussianNB
决策树	tree.DecisionTreeClassifier
集成方法	ensemble.BaggingClassifier
神经网络	neural_network.MLPClassifier

回归模型	加载模块
岭回归	linear_model.Ridge
Lasso回归	linear_model.Lasso
弹性网络	linear_model.ElasticNet
最小角回归	linear_model.Lars
贝叶斯回归	linear_model.BayesianRidge
逻辑回归	linear_model.LogisticRegression
多项式回归	preprocessing. PolynomialFeatures

聚类方法	加载模块
K-means	cluster.KMeans
AP聚类	cluster.AffinityPropagation
均值漂移	cluster.MeanShift
层次聚类	cluster.AgglomerativeClustering
DBSCAN	cluster.DBSCAN
BIRCH	cluster.Birch
谱聚类	cluster.SpectralClustering

降维方法	加载模块
主成分分析	decomposition.PCA
截断SVD和LSA	decomposition.TruncatedSVD
字典学习	decomposition.SparseCoder
因子分析	decomposition.FactorAnalysis
独立成分分析	decomposition.FastICA
非负矩阵分解	decomposition.NMF
LDA	decomposition.LatentDirichletAllocation

Scikit-learn

Scikit-learn和聚类

- scikit-learn库（以后简称sklearn库）提供的常用聚类算法函数包含在sklearn.cluster这个模块中。
  - 如：K-Means，近邻传播算法，DBSCAN等。
- 以同样的数据集应用于不同的算法，可能会得到不同的结果，算法所耗费的时间也不尽相同，这是由算法的特性决定的。

Scikit-learn

sklearn.cluster

- sklearn.cluster模块提供的各聚类算法函数可以使用不同的数据形式作为输入:
  - 标准数据输入格式：[样本个数，特征个数]定义的矩阵形式。
  - 相似性矩阵输入格式：即由[样本数目，样本数目]定义的矩阵形式。
    - 矩阵中的每一个元素为两个样本的相似度，如**DBSCAN，AffinityPropagation(近邻传播算法)**接受这种输入。
    - 如果以余弦相似度为例，则对角线元素全为**1**，矩阵中每个元素的取值范围为**[0,1]**。

Scikit-learn

sklearn.cluster

算法名称	参数	可扩展性	相似性度量
K-means	聚类个数	大规模数据	点间距离
DBSCAN	邻域大小	大规模数据	点间距离
Gaussian Mixtures	聚类个数及其他超参	复杂度高，不适合处理大规模数据	马氏距离
Birch	分支因子，阈值等其他超参	大规模数据	两点间的欧式距离

Python机器学习应用

- 无监督学习的K-means聚类方法及其应用案例