课程总结

ML30

礼欣

www.python123.org

主要内容

- 无监督学习 （Unsupervised Learning）
- 监督学习 （Supervised Learning）
- 强化学习（Reinforcement Learning，增强学习）

无监督学习

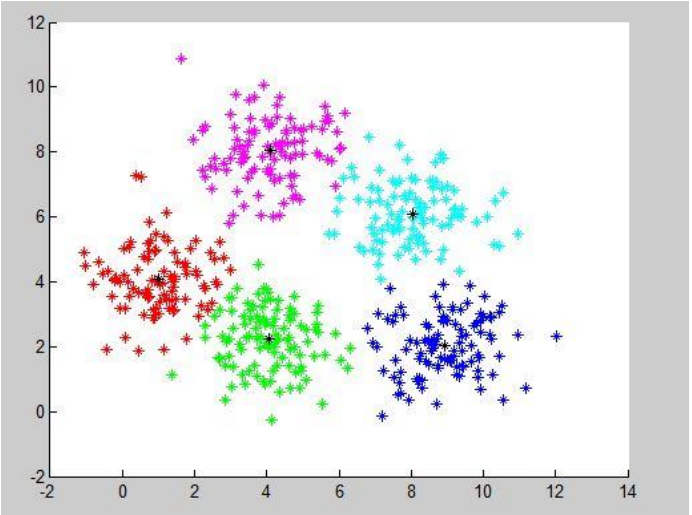
无监督学习的两大任务

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。

- 无监督学习最常应用的场景是聚类(clustering)
- 降维(Dimension Reduction)

聚类(clustering)

聚类(clustering)，就是根据数据的“相似性”将数据分为多类的过程。估算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”，最常用的就是欧式距离，此外还有马氏距离，曼哈顿距离，余弦距离等。右图是利用欧式距离的一种数据展示结果。



Sklearn vs. 聚类

scikit-learn库提供的常用聚类算法函数包含在sklearn.cluster模块中，如：K-Means，DBSCAN，等，我们在前面的讲解中通过实例具体讲解了K-Means，DBSCAN这些经典的聚类函数的在sklearn中的使用方法和也简单介绍了他们的算法思想。对大多数聚类算法来说，需要指定聚类的数目， DBSCAN是少数不需要指定聚类数目的算法之一。

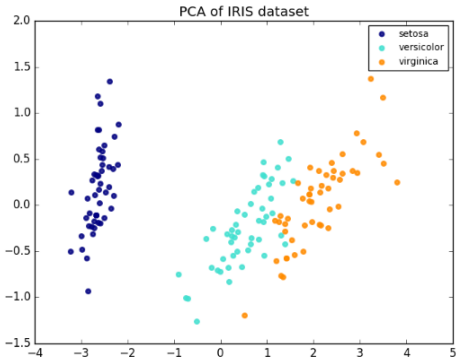
- K-means：对簇中心的初始化比较敏感
- DBSCAN：它可以发现使用K均值不能发现的许多簇，单不适合密度变化太大的数据，而且对于高维数据，该方法也有问题，因为密度定义比较困难。

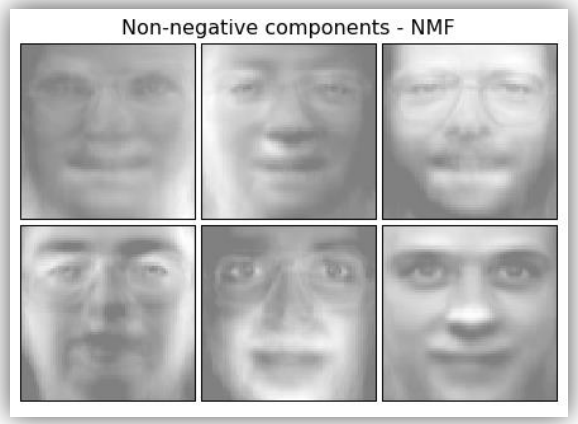
降维

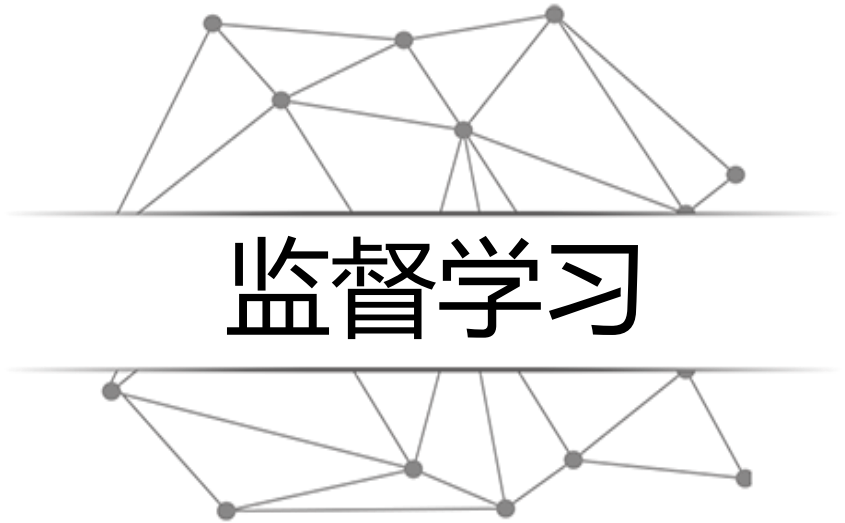
降维，就是在保证数据所具有的代表性特性或者分布的情况下，将高维数据转化为低维数据的过程。

sklearn vs.降维

sklearn库提供多种降维算法，被封装在sklearn.decomposition模块中，在前面的学习中我们也展示了PCA和NMF在鸢尾花数据集和人脸数据集上的特征提取过程的相关操作。注意：降维方法通常用于高维数据集的探索与可视化，降维过的数据可以为其他任务作数据准备。







监督学习

监督学习的两大任务

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。

分类学习

输入：一组有标签的训练数据(也称观察和评估)，标签表明了这些数据（观察）的所署类别。分类模型根据这些训练数据，训练自己的模型参数，学习出一个适合这组数据的分类器，当有新数据（非训练数据）需要进行类别判断，就可以将这组新数据作为输入送给学好的分类器进行判断。

Sklearn vs. 分类

在分类学习个单元我们重点介绍了kNN,朴素贝叶斯，决策树模型的基本思想，通过对分类问题的实例编写，实现了对不同分类算法的调用，并进行了实验对比。

初学者经常会将分类问题和聚类问题混淆，而训练数据有无标签是区别这两个问题的关键，在将实际问题转换为学习问题的过程中，需要大家判断准确，选择合适的机器学习算法。

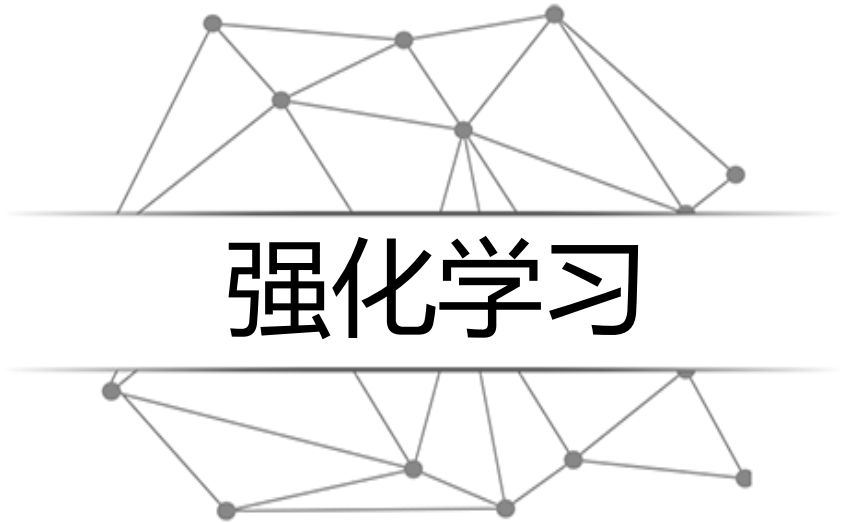
回归分析

回归：是一种统计学上分析数据的方法，目的在于了解两个或多个变数间是否相关、研究其相关方向与强度，并建立数学模型以便观察特定变数来预测研究者感兴趣的变数。回归方法常被用来进行如股票趋势预测，交通流量预测这种带有时序信息的数据分析上。

Sklearn vs. 回归

回归函数主要分为两类，线性回归和非线性回归。线性回归函数主要封装在sklearn.linear_model模块中，如普通线性回归，岭回归，Lasso。

注意：sklearn中的非线性回归是通过利用sklearn.preprocessing模块生成原始数据的非线性特征，在调用线性回归模型将这些非线性特征按线性方式进行拟合实现的。



强化学习

强化学习

- 传统强化学习模型（model-based, model-free）
- 深度强化学习的基本思路
- TensorFlow的简单使用技巧
- 以Tensorflow实现Flappy bird游戏的自主学习的实例，希望同学们通过实例的学习，可以掌握利用Tenforlow构造深度神经网络，并进行训练以解决实际问题的能力。