无监督学习-课程导学

ML04

礼欣
www.python123.org

无监督学习简介

无监督学习的目标

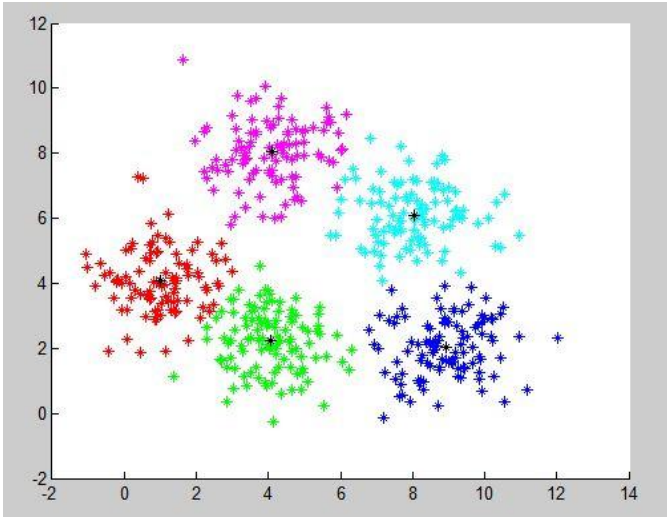
利用无标签的数据学习数据的分布或数据与数据之间的关系被称作无监督学习。

- 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)和降维(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。

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。



欧氏距离

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：

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

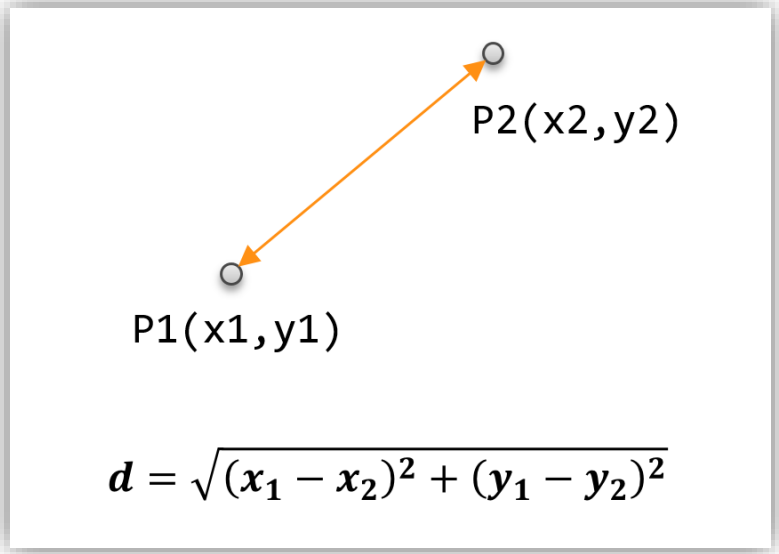


图. 二维空间中欧式距离的计算

曼哈顿距离

曼哈顿距离也称作“城市街区距离”，类似于在城市之中驾车行驶，从一个十字路口到另外一个十字楼口的距离。其计算方法如下：

$$d = \sum_{k=1}^{n} |x_{1k} - x_{2k}|$$

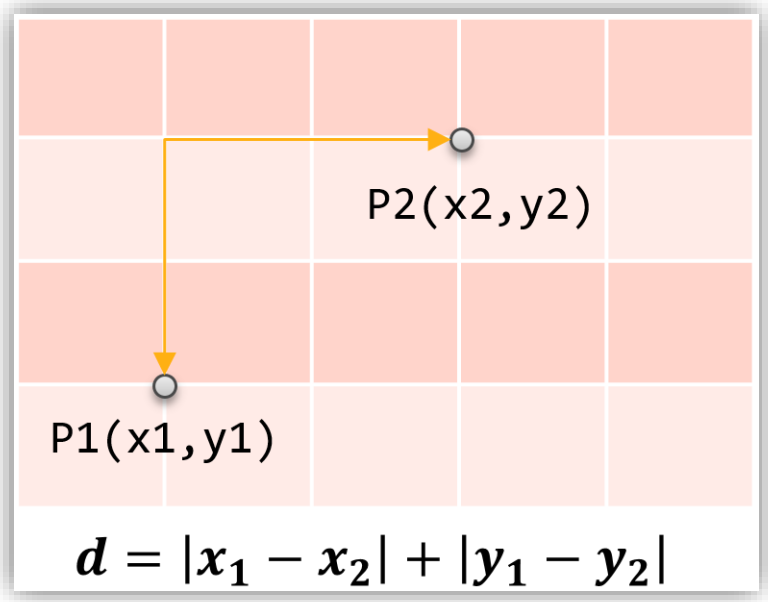


图. 二维空间中曼哈顿距离的计算

马氏距离

马氏距离表示数据的协方差距离，是一种尺度无关的度量方式。也就是说马氏距离会先将样本点的各个属性标准化，再计算样本间的距离。其计算方式如下：（s是协方差矩阵，如图）

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T s^{-1} (x_i - x_j)}$$

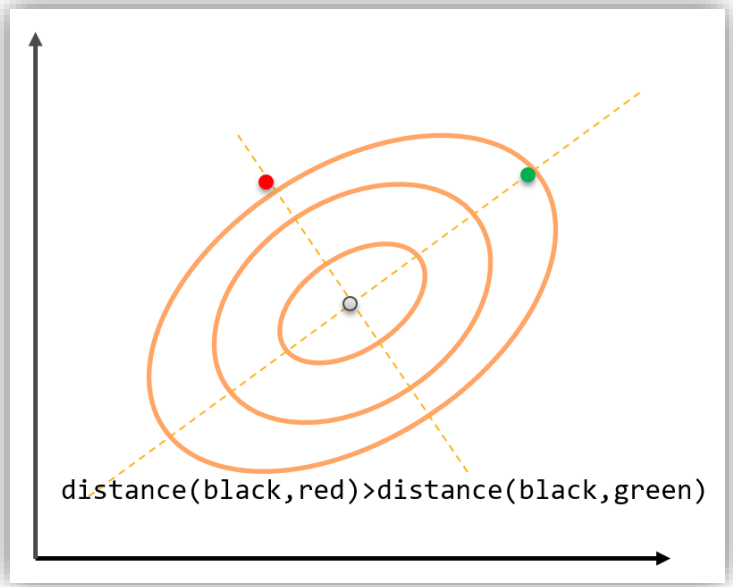


图. 二维空间中的马氏距离

夹角余弦

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个样本差异的大小。余弦值越接近1，说明两个向量夹角越接近0度，表明两个向量越相似。其计算方法如下：

$$cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2} \sqrt{\sum_{k=1}^{n} x_{2k}^2}}$$

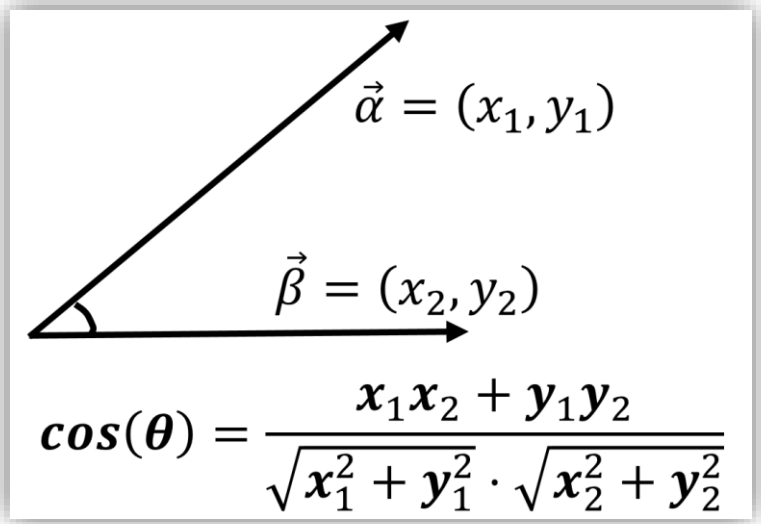
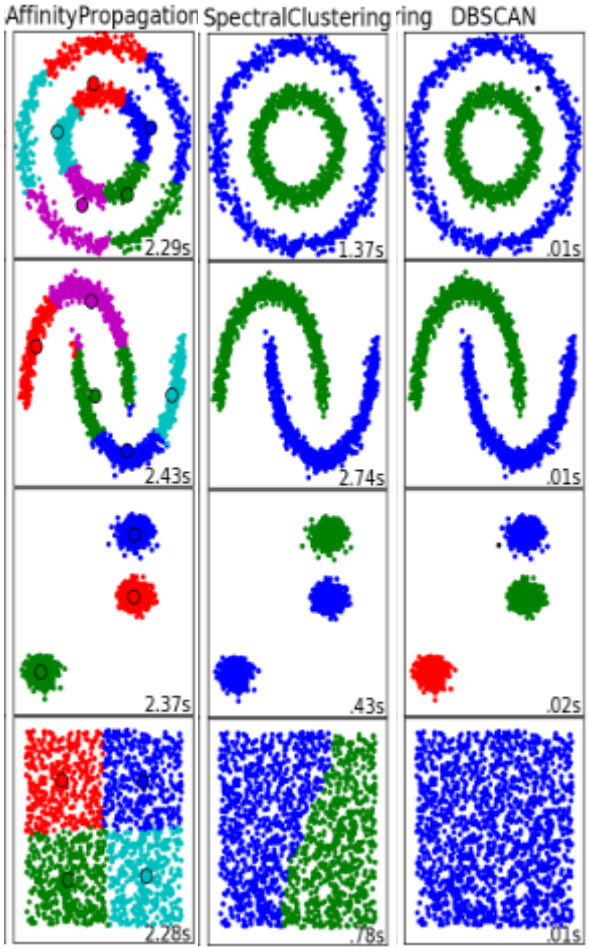


图. 二维空间中的夹角余弦

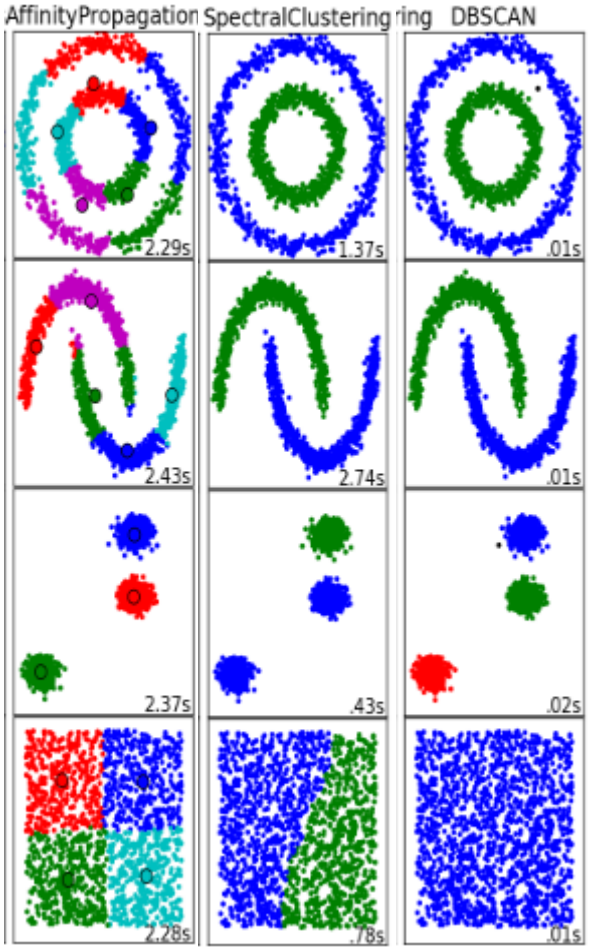
Sklearn vs. 聚类

- 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。



Sklearn vs. 聚类

右图是我们调用sklearn库的标准函数对不同数据集执行的聚类结果。



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]。

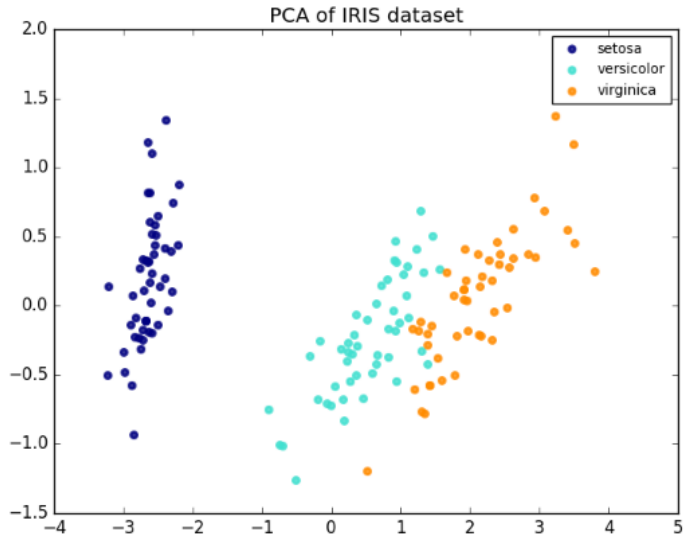
sklearn.cluster

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

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：

- 数据的可视化
- 精简数据



聚类 vs.降维

聚类和降维都是无监督学习的典型任务，任务之间存在关联，比如某些高维数据的聚类可以通过降维处理更好的获得，另外学界研究也表明代表性的聚类算法如k-means与降维算法如NMF之间存在等价性，在此我们就不展开讨论了，有兴趣的同学可以参考我们推荐的阅读内容。

sklearn vs.降维

- 降维是机器学习领域的一个重要研究内容，有很多被工业界和学术界接受的典型算法，截止到目前sklearn库提供7种降维算法。
- 降维过程也可以被理解为对数据集的组成成份进行分解（decomposition）的过程，因此sklearn为降维模块命名为decomposition，在对降维算法调用需要使用sklearn.decomposition模块

sklearn.decomposition

算法名称	参数	可扩展性	适用任务
PCA	所降维度及其他超参	大规模数据	信号处理等
FastICA	所降维度及其他超参	超大规模数据	图形图像特征提取
NMF	所降维度及其他超参	大规模数据	图形图像特征提取
LDA	所降维度及其他超参	大规模数据	文本数据，主题挖掘

未来任务

在后续的讲解中我们将通过实例展示如何利用sklearn库提供的分类和降维算法解决具体问题（大家可以通过本次的讲授先行思考下面的问题哪些是聚类问题，哪些是降维任务？）：

- 31省市居民家庭消费调查
- 学生月上网时间分布调查
- 人脸图像特征抽取
- 图像分割