AI\_Project2 实验报告

金英豪 10172100162

本次实验主要针对三个不同种类的数据集，调用不同的机器学习算法对其进行初步的分析与模型建构。其中，Air Quality Dataset 对应回归算法，BLE&RSSL Dataset Labeled 对应分类算法，BLE&RSSL Dataset Unlabeled 对应聚类算法。以下内容将对数据集、算法及实验过程进行大致介绍。

1. 数据集简介
2. Air Quality Dataset

该数据集包含9358个小时平均响应实例，这些实例来自嵌入空气质量化学多传感器装置中的5个金属氧化物化学传感器阵列。该装置位于意大利城市内一个污染严重的区域，在道路一级。数据记录于2004年3月至2005年2月（一年），代表现场部署的空气质量化学传感器设备响应的最长免费记录。地面真相每小时平均浓度为CO，非有机碳氢化合物，苯，总氮氧化物（NOx）和二氧化氮（NO2）。

1. BLE&RSSL Dataset

该数据集是使用西密歇根大学Waldo图书馆一楼13个ibeacons阵列的RSSI读数创建的。使用iPhone 6S收集数据。数据集包含两个子数据集：标记数据集（1420个实例）和未标记数据集（5191个实例）。录音是在图书馆的营业时间内进行的。对于带标签的数据集，输入数据包含位置（标签列），时间戳，后跟13个iBeacons的RSSI读数。RSSI测量值为负值。较大的RSSI值表示更接近给定的iBeacon（例如，与-85的RSSI相比，-65的RSSI表示与给定iBeacon的距离更近）。对于超出范围的iBeacons，RSSI由-200表示。与RSSI读数相关的位置组合在一列中，该列包括列的字母和位置行的数字。Labeled与Unlabeled的区别在于有无提供location信息。

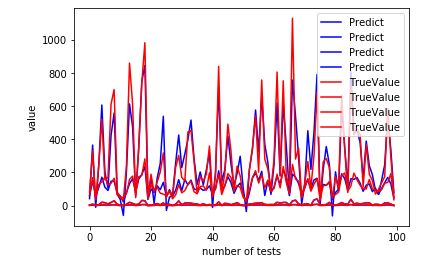
1. 实验过程及使用算法
2. 回归

在Air Quality Dataset上运用线性回归算法进行分析。主要步骤如下：

* 数据预处理：导入数据集，删除无关信息（如：Date），异常值替换（-200替换为Nan，然后删除该项），数据集分割（将原数据集分割为训练集与测试集，同时分别再进行输入与输出的分割）
* 训练与测试：调用SKlearn-LinearRegression模型对训练集进行回归学习，再使用score函数导入test对训练模型的准确率进行评估
* 绘图：使用matplotlib模块进行预测值与真实值的比较绘图。

实验结果大致如下：

微信截图_20190515111547



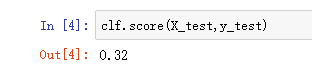
可以看出准确率较高，拟合度较好。

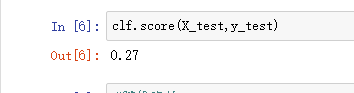
1. 分类

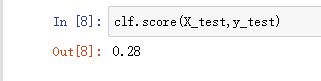
在BLE&RSSL Dataset Labeled上运用分类算法进行分析。主要步骤如下：

* 数据预处理（同回归）
* 调用不同学习库并对训练模型测试

测试结果如下：

SVM：

决策树：

随机森林：

可以看出，在本数据集中SVM的的效果略好于另两种方法，但三者的情况都欠佳，原因是该数据集中含有大量缺失值（-200）导致无法训练符合要求的模型。

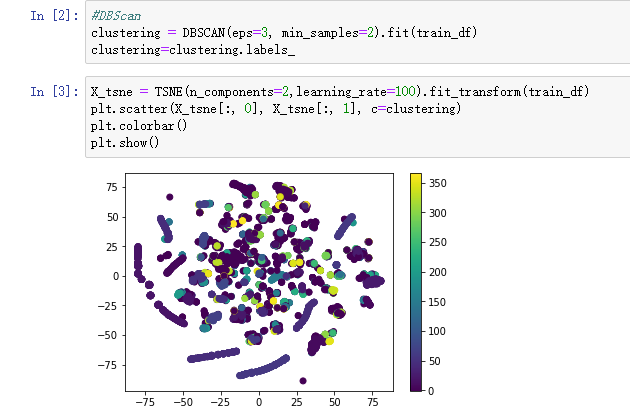
1. 聚类

在BLE&RSSL Dataset Unlabeled上运用聚类算法进行分析。主要步骤如下：

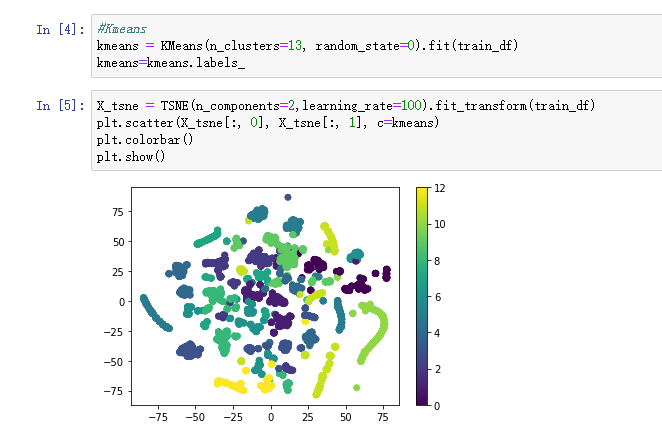
* 数据预处理
* 对不同学习算法设置不同的参数，这里基本上都运用了默认值，少数根据真实情况做了修改（如：聚类数量等）
* 以t-SNE实现聚类结果可视化

结果如下：

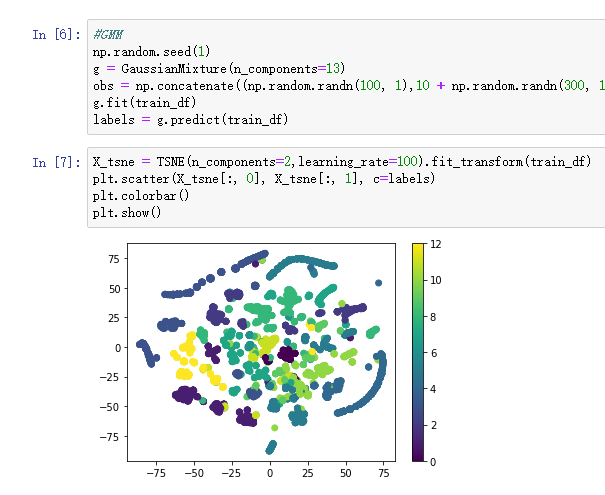
DBScan：



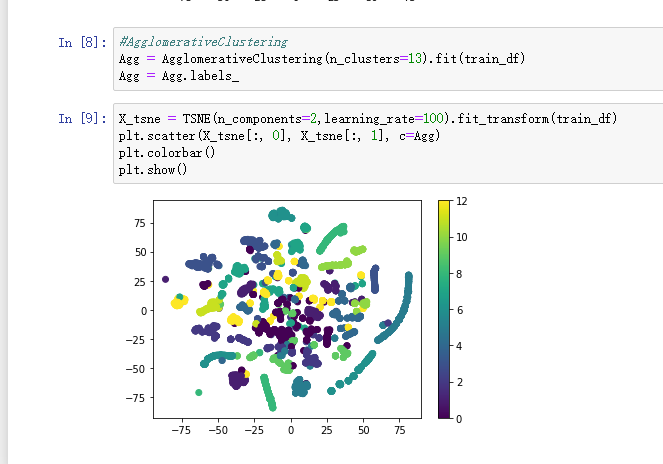
Kmeans：



GMM：



层次聚类（AgglomerativeClustering）：



聚类算法与回归、分类不同之处在于，由于标签的缺失无法直接用score来判断聚类结果好坏。但是如果要进一步分析，可以计算同一类点间密度、与centroid中心点的距离等来评估。

根据初步评估，四种聚类算法各展现出不同的优势与劣势，这与他们自身的特点有关：

Kmeans：

优点：简单直接（体现在逻辑思路以及实现难度上），易于理解，在低维数据集上有不错的效果（简单的算法不见得就不能得到优秀的效果）。

缺点：对于高维数据，其计算速度十分慢，主要是慢在计算距离上，它的另外一个缺点就是它需要我们设定希望得到的聚类数k，若我们对于数据没有很好的理解，那么设置k值就成了一种估计性的工作。

层次聚类算法：

优点：1，距离和规则的相似度容易定义，限制少；2，不需要预先制定聚类数；3，可以发现类的层次关系。

缺点：1，计算复杂度太高；2，奇异值也能产生很大影响；3，算法很可能聚类成链状。

GMM：

优点：投影后样本点不是得到一个确定的分类标记，而是得到每个类的概率，这是一个重要信息。

缺点：GMM每一步迭代的计算量比较大，大于k-means，并且GMM的求解办法基于EM算法，因此有可能陷入局部极值，因此和初始值的选取十分相关。

DBScan：

优点：基于密度定义，相对抗噪音，能处理任意形状和大小的簇。

缺点：当簇的密度变化太大时，会有麻烦，对于高维问题，密度定义是个比较麻烦的问题。