第二次作业：回归、分类和聚类

任课老师：罗轶凤

2019年5月12日

回归、分类和聚类

本次作业在Jupiter Notebook上通过scikit-learn对两个数据集分别实现了：1、回归（Air quality dataset）：逻辑回归；2、分类（BLE&RSSI dataset）：SVM、决策树、随机森林；3、聚类（BLE&RSSI dataset）：DBScan、kmeans、GMM、层次聚类算法（其中聚类算法以t-SNE实现了结果可视化）。

具体实现过程如下：

1. 回归（Air quality dataset）

代码路径：./AirQualityUCI/reg\_test.ipynb

1.1 数据集地址：https://archive.ics.uci.edu/ml/datasets/Air+quality

1.2 数据集介绍：The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level,within an Italian city. Data were recorded from March 2004 to February 2005 (one year)representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value. This dataset can be used exclusively for research purposes. Commercial purposes are fully excluded.

1.3 数据预处理：

1.3.1 首先读取数据。表内数据存在一些值为-200的缺失值，为了回归的准确性，采用去除含缺失值数据项的方法。

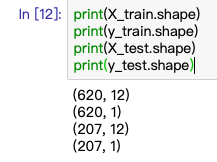
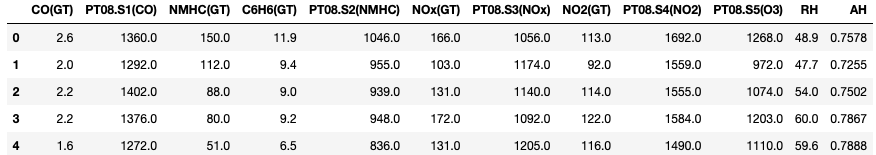
data = pd.read\_csv(‘AirQualityUCI.csv',na\_values=-200)

data=data.dropna(axis=0,how=‘any')

处理后得到(827, 15)大小的data数据集

1.3.2 然后将data数据分为X，y两部分

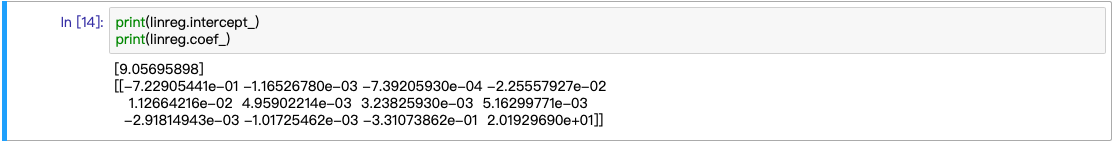
X=data[['CO(GT)','PT08.S1(CO)','NMHC(GT)','C6H6(GT)','PT08.S2(NMHC)','NOx(GT)','PT08.S3(NOx)','NO2(GT)','PT08.S4(NO2)','PT08.S5(O3)','RH','AH']]

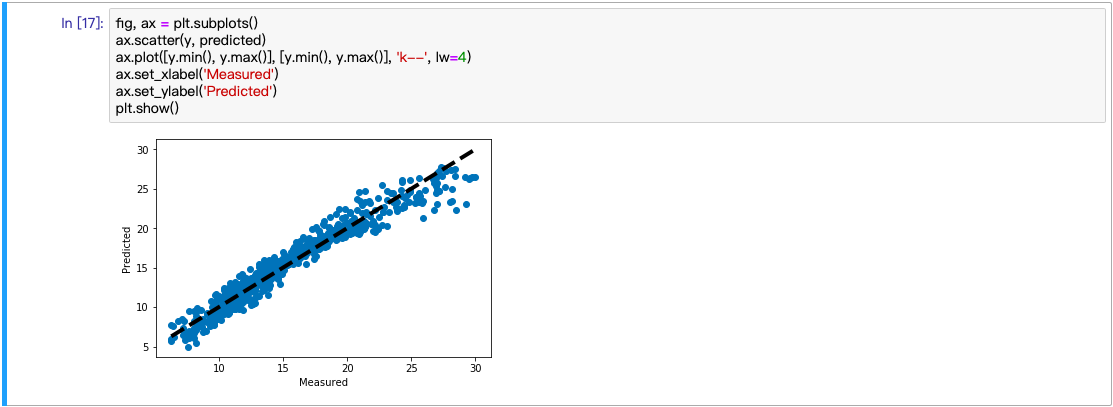
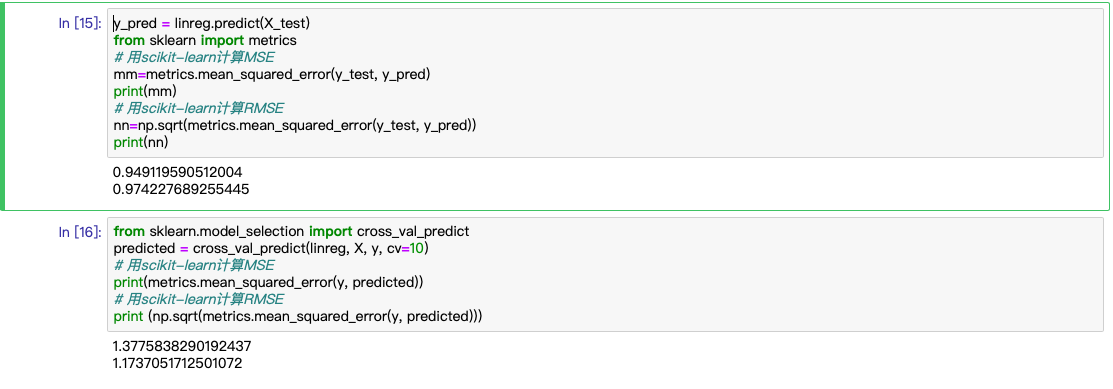
y=data[[’T']]

1.3.3 划分训练集和测试集

from sklearn.cross\_validation import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, random\_state=1)

1.4 进行回归计算参数

1.5 查看结果

2. 分类（BLE&RSSI dataset）：SVM、决策树、随机森林

代码路径：./BLE\_RSSI\_dataset/BR\_2\_classification.ipynb

2.1 数据集

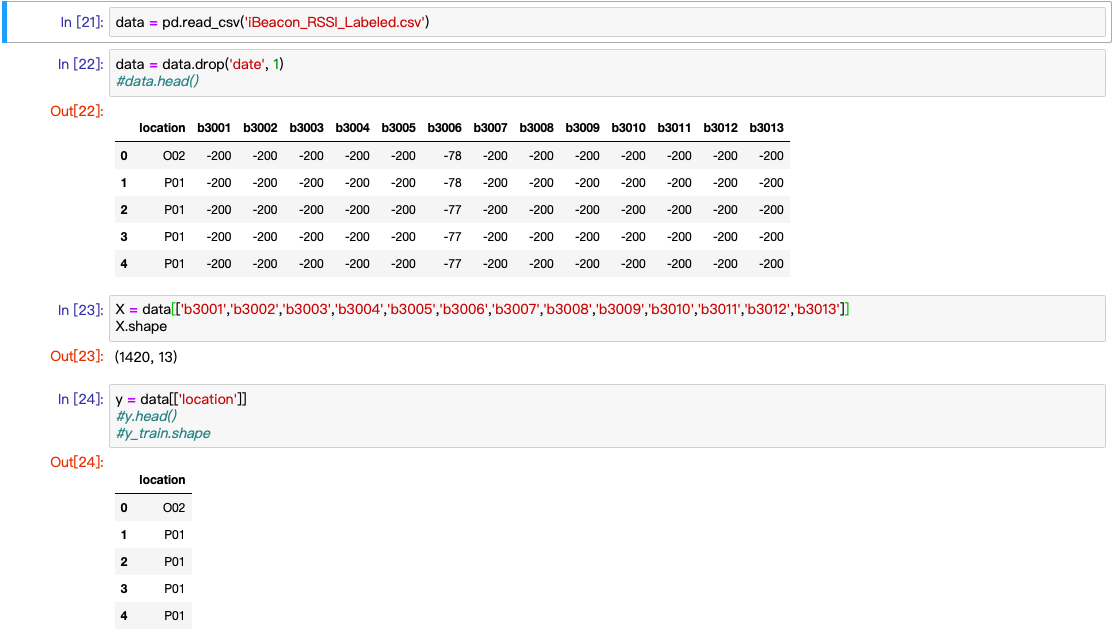
2.1.1 数据集地址及介绍：

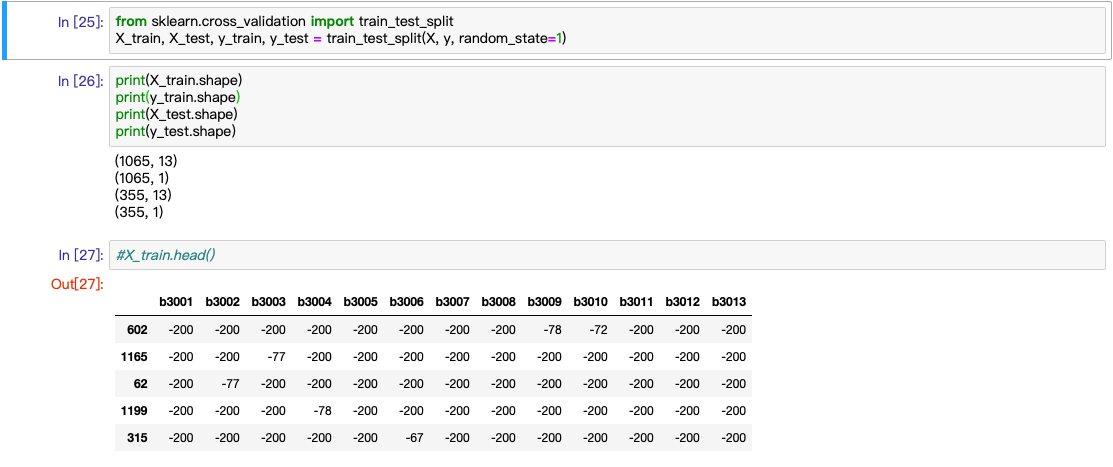
<https://archive.ics.uci.edu/ml/datasets/BLE+RSSI+Dataset+for+Indoor+localization+and+Navigation>

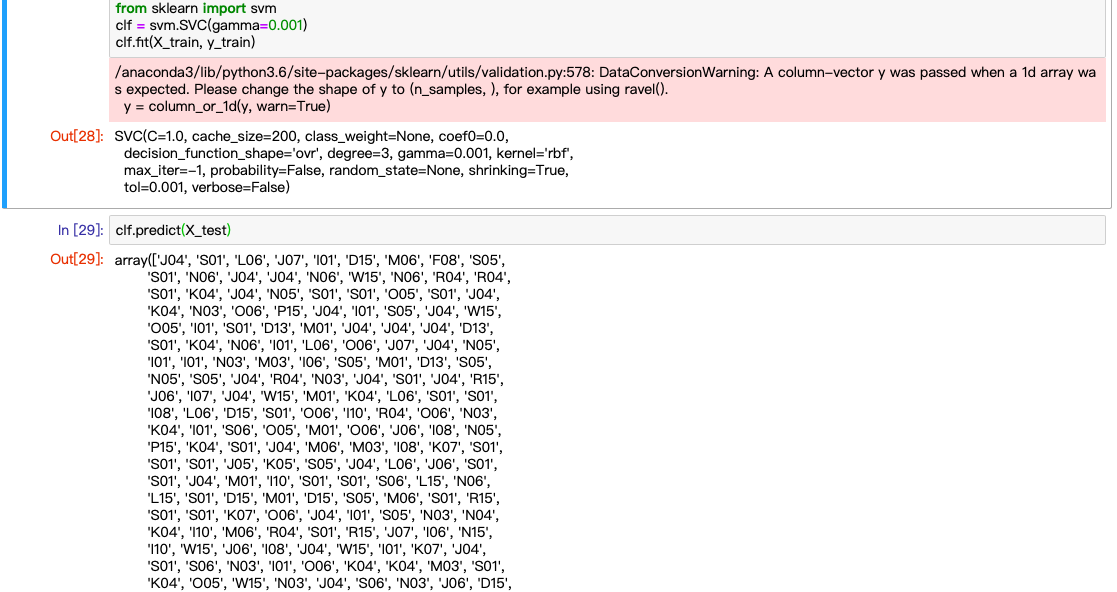
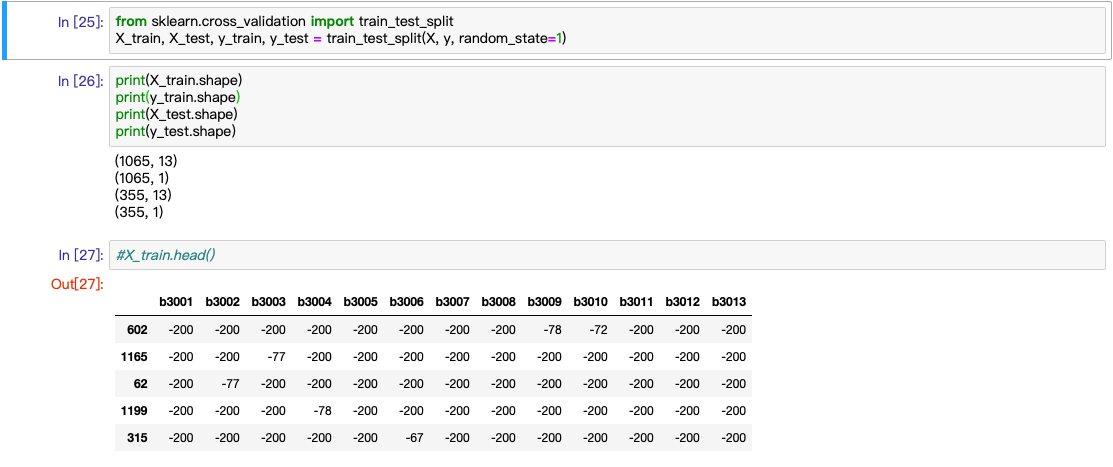
The dataset was created using the RSSI readings of an array of 13 ibeacons in the first floor of Waldo Library, Western Michigan University. Data was collected using iPhone 6S. The dataset contains two sub-datasets: a labeled dataset (1420 instances) and an unlabeled dataset (5191 instances). The recording was performed during the operational hours of the library. For the labeled dataset, the input data contains the location (label column), a timestamp, followed by RSSI readings of 13 iBeacons. RSSI measurements are negative values. Bigger RSSI values indicate closer proximity to a given iBeacon (e.g., RSSI of -65 represent a closer distance to a given iBeacon compared to RSSI of -85). For out-of-range iBeacons, the RSSI is indicated by -200. The locations related to RSSI readings are combined in one column consisting a letter for the column and a number for the row of the position. The attached figure depicts the layout of the iBeacons as well as the arrange of locations.

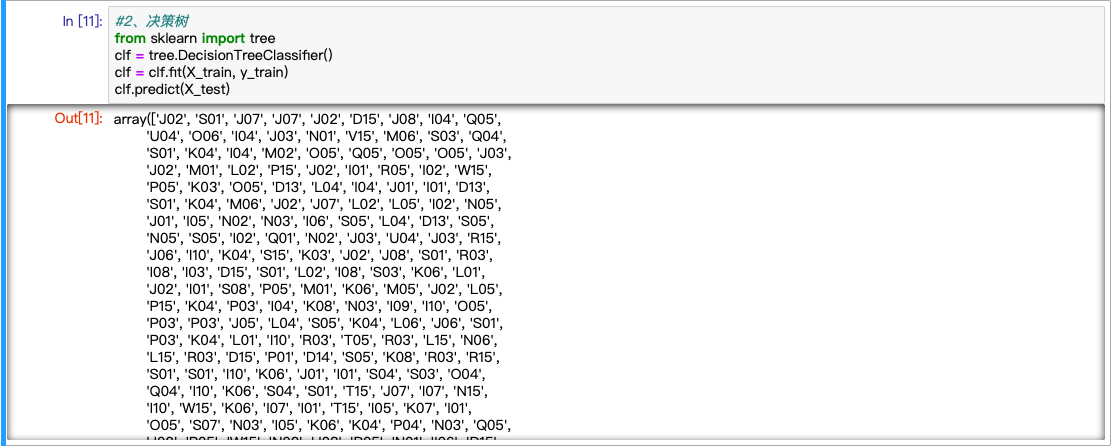
Consectetuer arcu ipsum ornare pellentesque vehicula, in vehicula diam, ornare magna erat felis wisi a risus. Justo fermentum id. Malesuada eleifend, tortor molestie, a fusce a vel et. Mauris at suspendisse, neque aliquam faucibus adipiscing, vivamus in. Wisi mattis leo suscipit nec amet, nisl fermentum tempor ac a, augue in eleifend.

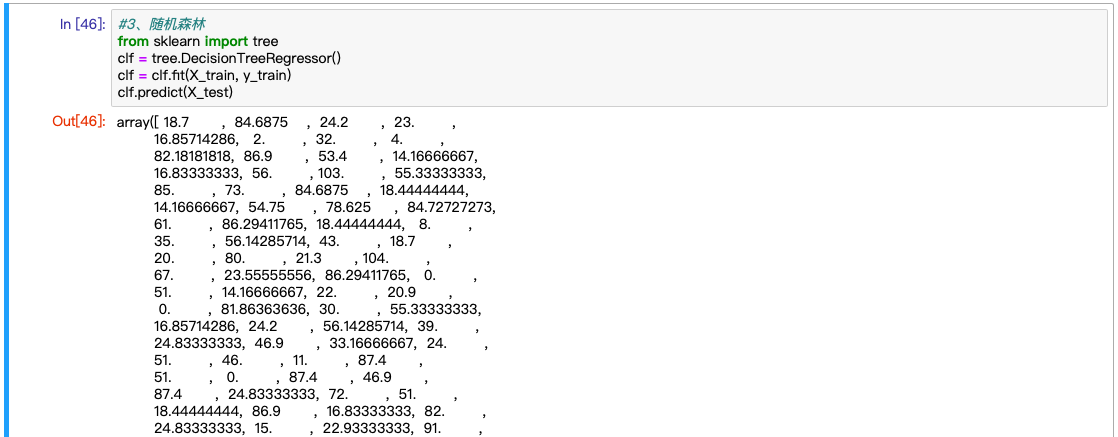
2.1.2 数据预处理

原数据集共15列，删去与分类无关的date列，取后13列作X，location信息作y：

划分训练集与测试集：

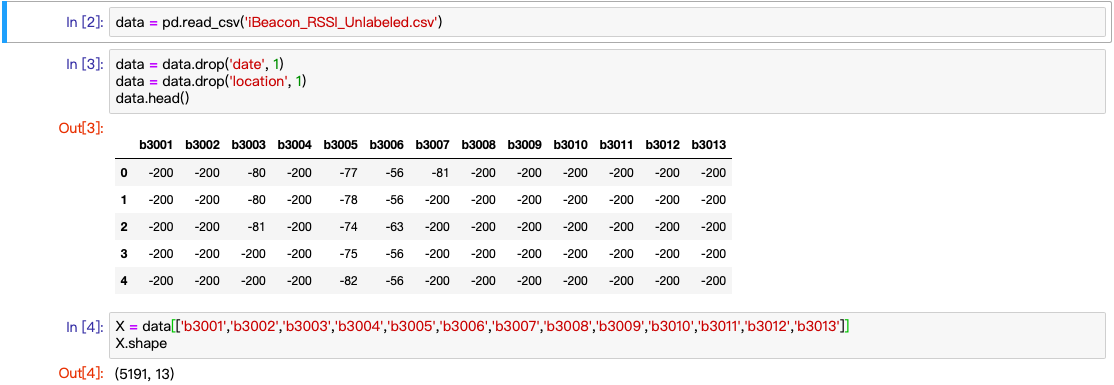
2.2 SVM

2.3 决策树

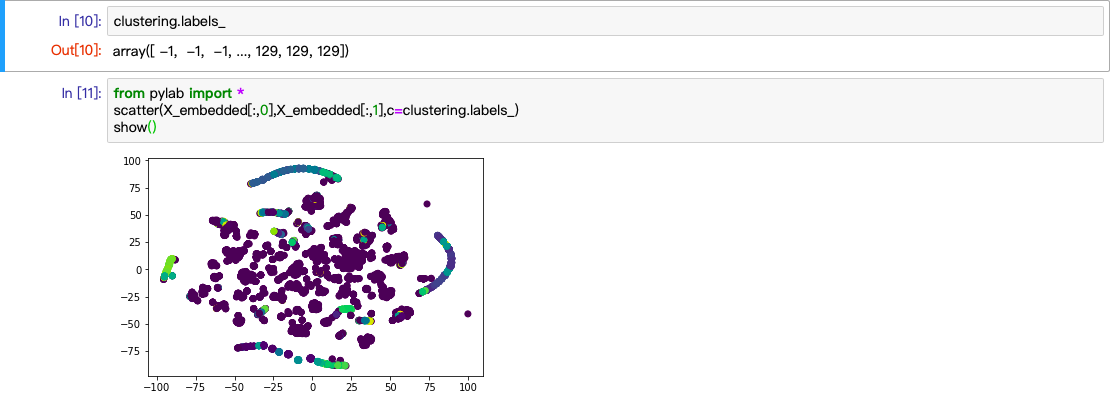
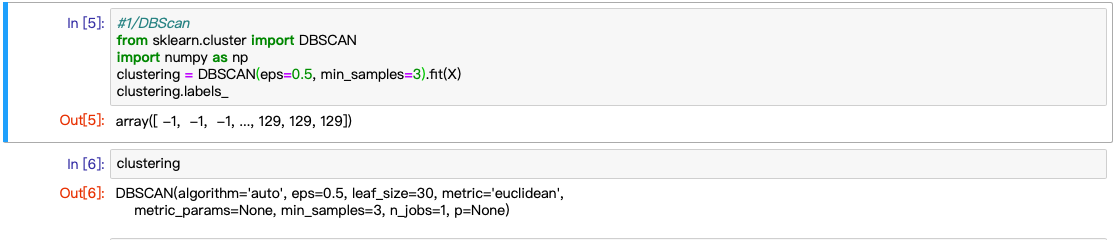
2.4 随机森林

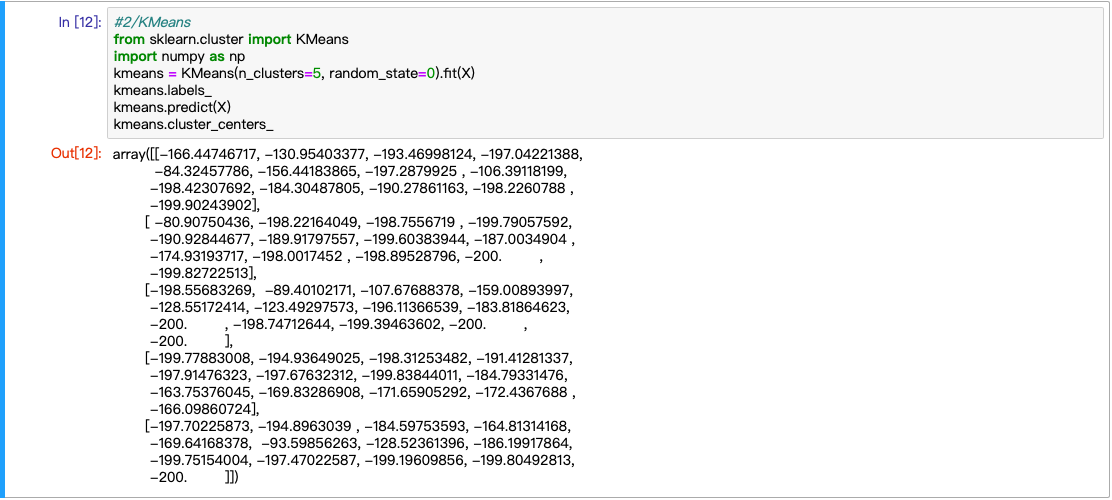
3、聚类（BLE&RSSI dataset）：DBScan、kmeans、GMM、层次聚类算法

数据集信息同2

3.1 数据处理

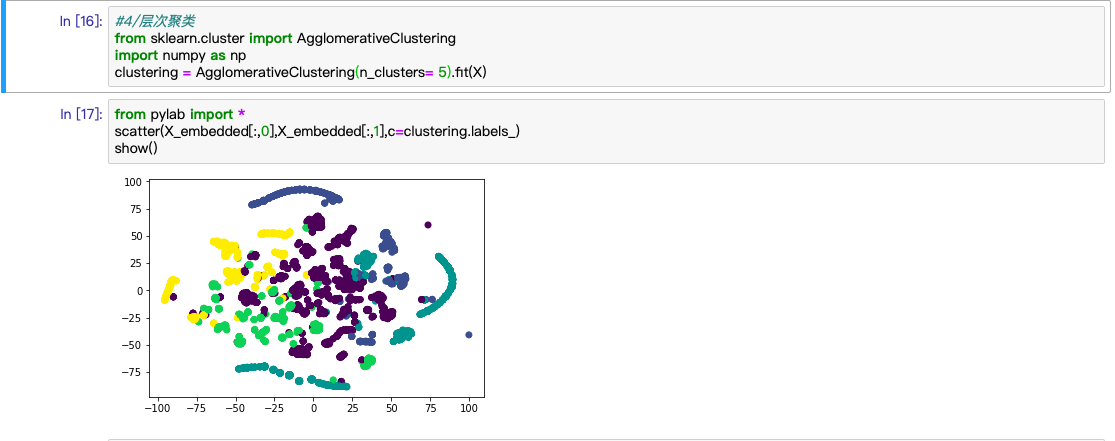
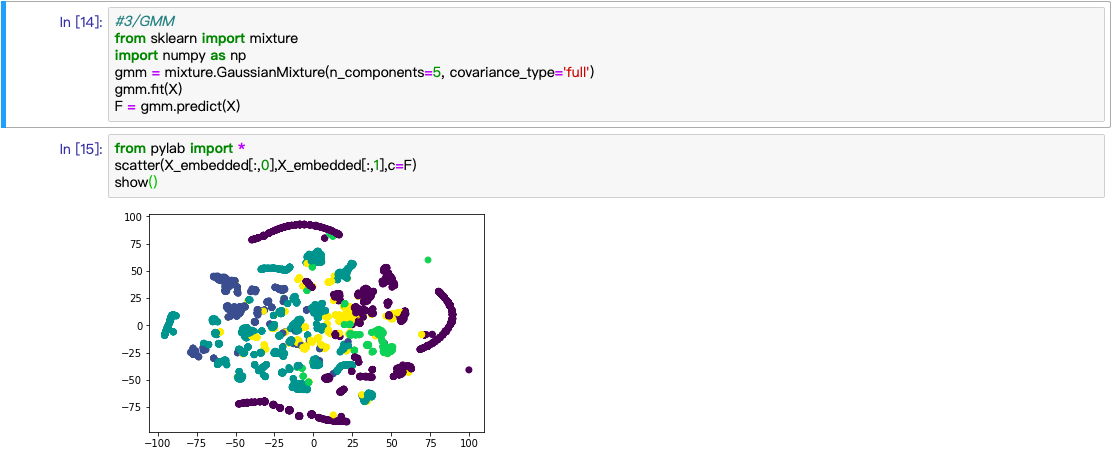
3.2 t-SNE 降维：将X投影至二维数组，已被可视化画图使用

3.3 DBScan

3.4 k-means

3.5 GMM

3.6 层次聚类

通过本次作业，亲自动手，对人工智能算法有了进一步的认识与理解，还锻炼了查找官网英文材料的能力，加强了sklearn和Jupiter Notebook的熟练应用。