一:

1.空气质量数据集

Data Set Information:

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.

Attribute Information:

0 Date (DD/MM/YYYY)

1 Time (HH.MM.SS)

2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)

3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)

4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)

5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)

6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)

7 True hourly averaged NOx concentration in ppb (reference analyzer)

8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)

9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)

10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)

11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)

12 Temperature in Â°C

13 Relative Humidity (%)

14 AH Absolute Humidity

1. 实验目的:

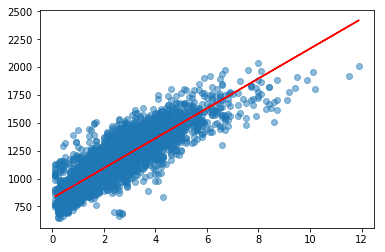
PT08.S1(CO)作为自变量,预测CO(GT)的值

1. 预处理

去除自变量和因变量中值为-200的数据

1. 实验结果:

训练集准确率：0.7724 测试集准确率：0.7755



二:

1. BLE RSSI Dataset数据集

Data Set Information:

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.

Attribute Information:

location: The location of receiving RSSIs from ibeacons b3001 to b3013; symbolic values showing the column and row of the location on the map (e.g., A01 stands for column A, row 1).

Date: Datetime in the format of â€˜d-m-yyyy hh:mm:ssâ€™

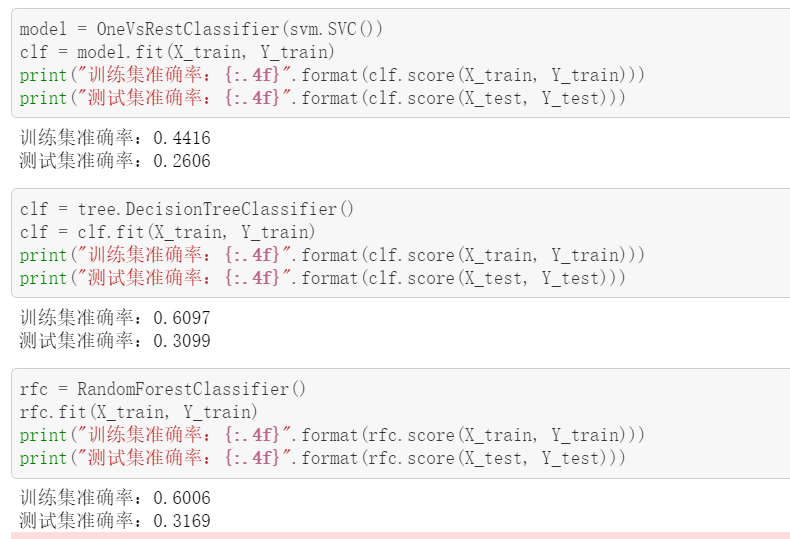
b3001 - b3013: RSSI readings corresponding to the iBeacons; numeric, integers only.

1. 实验目的:在有标签的数据集上分别采用SVM,决策树,随机森林算法对数据分类的预处;

在无标签数据集上,分别通过DBScan、kmeans、GMM、层次聚类算法实现对数据的聚类.

1. 实验结果:

如下图



在聚类算法中,最重要参数是聚类的个数,经过测试,聚类个数为三时,效果较好.

为了可视化实验结果,采用tSNE将13维的数据降到2维.

可视化结果如下:

