实验报告

施若男

AirQualityUCI 数据集

- **数据集介绍**

数据集包含在意大利城市现场部署的气体多传感器装置的响应。记录每小时响应平均值以及来自认证分析仪的气体浓度参考。缺少值标记为-200 值。

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. True hourly averaged concentration CO in mg/m^3 (reference analyzer)
4. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)
6. True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
7. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. True hourly averaged NOx concentration in ppb (reference analyzer)
9. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)
11. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
12. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13. Temperature in Â°C
14. Relative Humidity (%)
15. AH Absolute Humidity

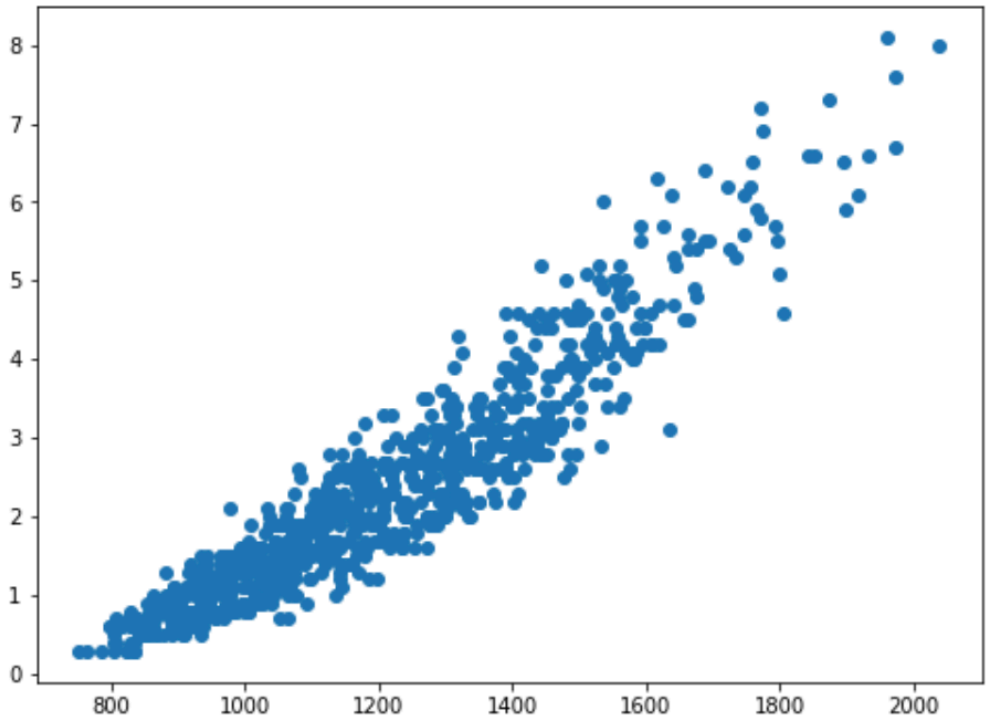
- **数据预处理**

删除-200 数据和 NA 数据

- **数据可视化**

散点图

x 轴为 CO 气体传感器数据，y 轴为 CO 真实数据，基本为线性关系



- **建模**

建立线性回归模型

LinearRegression 对 CO(GT)预测：训练集的准确率为: 0.9368 测试集的准确率为: 0.9517

LinearRegression 对 NOx(GT)预测：训练集的准确率为: 0.9056 测试集的准确率为: 0.9069

- **对比**

一种气体观测值预测：
LinearRegression 对CO(GT)预测：
训练集的准确率为: 0.8747
测试集的准确率为: 0.8810

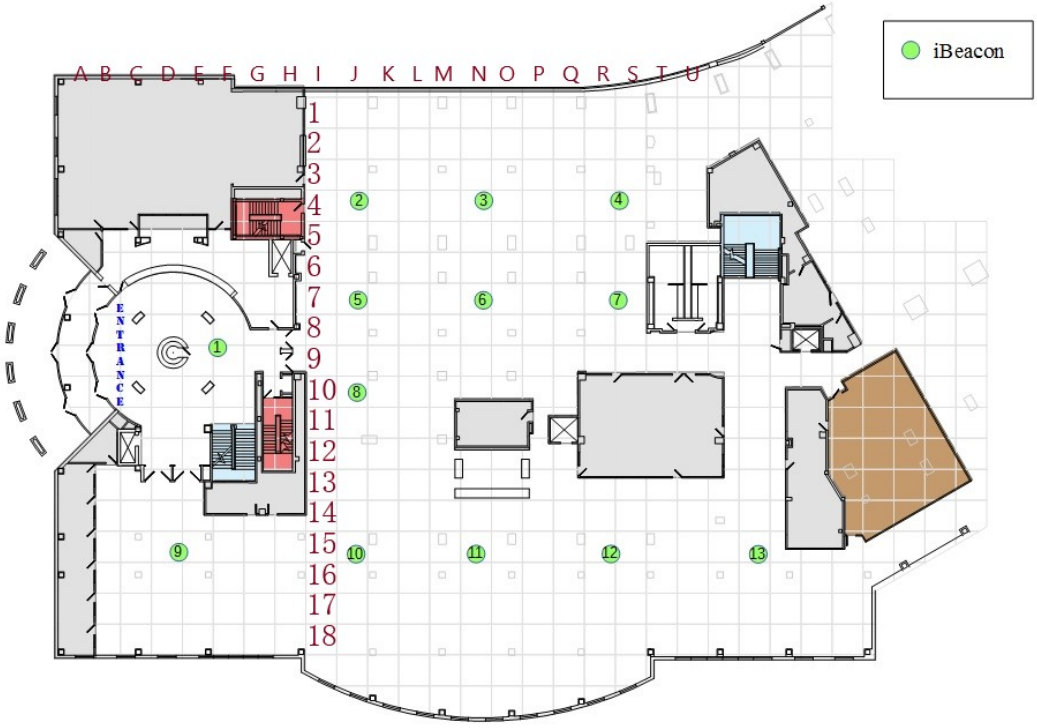
多种气体观测值预测：
LinearRegression 对CO(GT)预测：
训练集的准确率为: 0.9368
测试集的准确率为: 0.9517

对比发现多种气体观测值预测效果要好一些，说明各种气体的观测值之间有一些联系。

BLE_RSSI 数据集

- **数据集介绍**

数据集为蓝牙信号强度数据



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

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

- **数据预处理**

删除数据量<5 的 location 数据

- **分类**

结果如下：

```
rfc = RandomForestClassifier(n_estimators=12).fit(x_train,y_train)
score(rfc,x_train, y_train,x_test, y_test)
```

```
RandomForestClassifier:
训练集的准确率为: 0.5945
测试集的准确率为: 0.3187
```

```
dt = DecisionTreeClassifier(max_depth=20).fit(x_train,y_train)
score(dt,x_train, y_train,x_test, y_test)
```

```
DecisionTreeClassifier:
训练集的准确率为: 0.5838
测试集的准确率为: 0.3216
```

```
svc = SVC().fit(x_train,y_train)
score(svc,x_train, y_train,x_test, y_test)
```

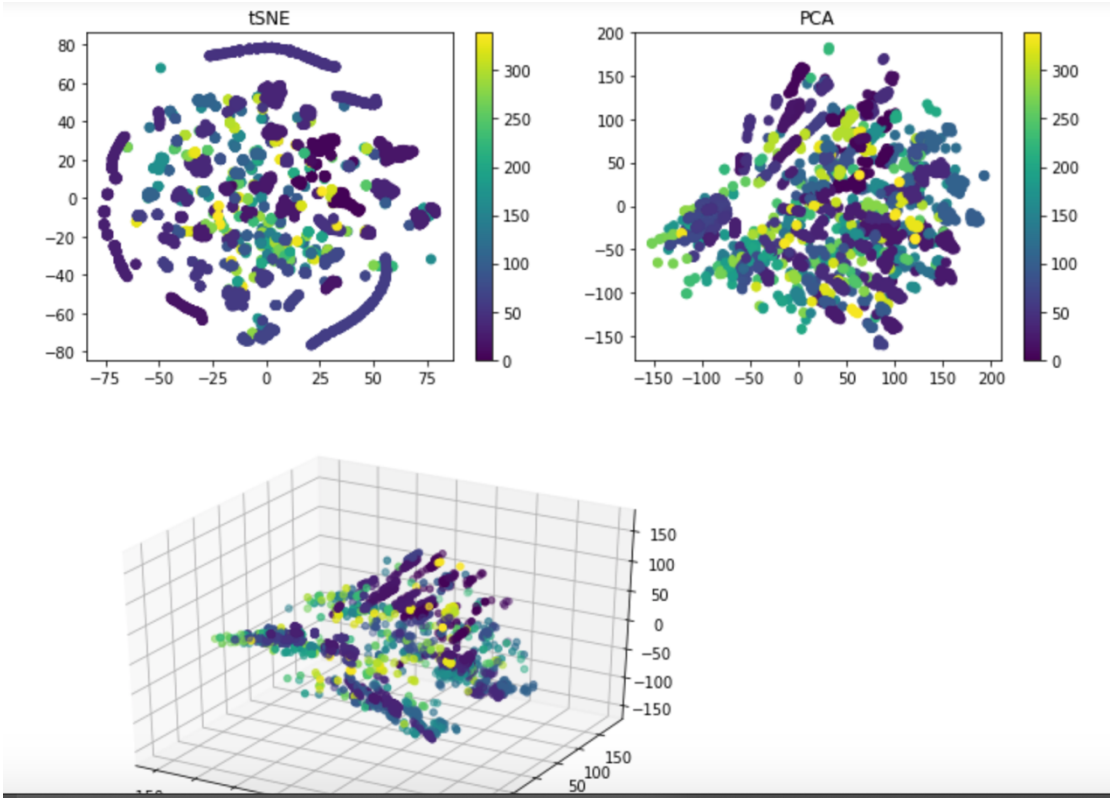
```
SVC:
训练集的准确率为: 0.4971
测试集的准确率为: 0.3158
```

- **聚类**

聚类：采用 sklearn 库中的 kmeans、密度聚类、层次聚类和高斯混合模型

降维与可视化：tSNE，PCA

以密度聚类为例，可视化结果如下图，发现 tSNE 降维效果比 PCA 的降维效果要差


tSNE
PCA