**《机器学习》上机实验报告**

**学号：10930130448 姓名：吕思勤**

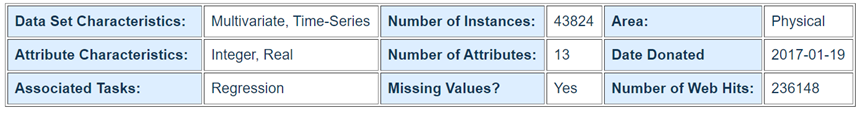
**第三次实验：北京PM2.5评估**

**一、 数据探索**

1.数据来源

https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data

2.数据描述



构建北京地区的PM2.5的预测模型，通过该地区的温度、湿度等气候特征对该地区的PM2.5进行准确预测。

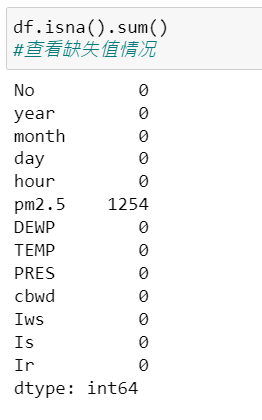
**二、数据预处理**

1. 数据读取

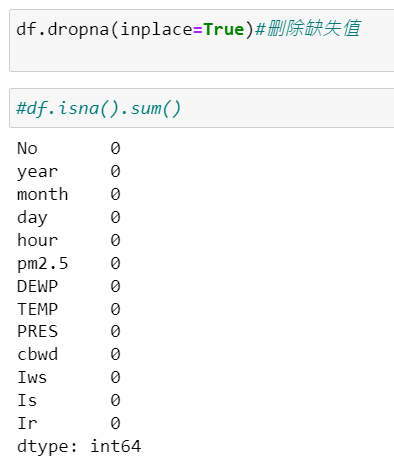
一張含有 文字 的圖片

自動產生的描述

1. 缺失值处理



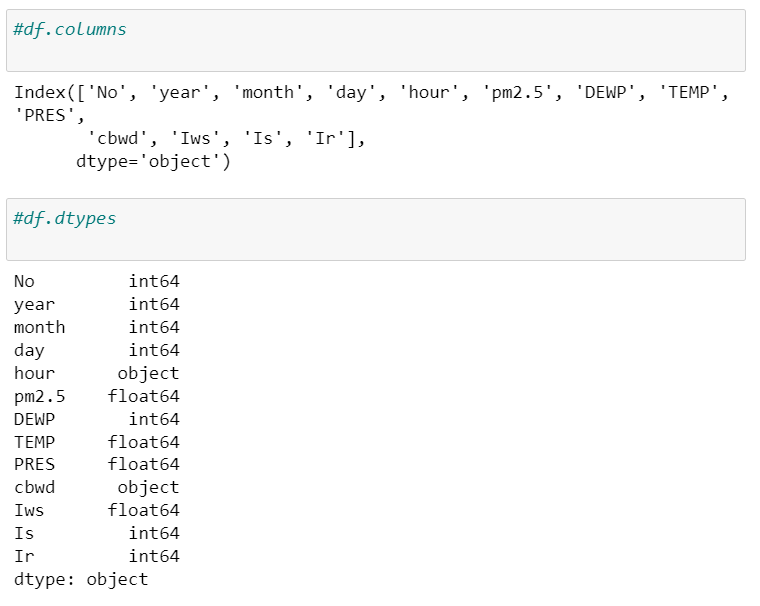
只有pm25这一列数据有空值，一共1254行。而总的数据集的数据一共43800行，包含空值的数据所占总数据集的比例较少，可以直接删除。



1. 更改数据类型



1. 数据了解



1. 属性选择
2. 特征提取One-Hot 编码

一張含有 桌 的圖片

自動產生的描述

one-hot的基本思想：将离散型特征的每一种取值都看成一种状态，若你的这一特征中有N个不相同的取值，那么我们就可以将该特征抽象成N种不同的状态，one-hot编码保证了每一个取值只会使得一种状态处于“激活态”，也就是说这N种状态中只有一个状态位值为1，其他状态位都是0。

**三、****模型训练（回归）**

1. 普通的线性回归

模型参数

fit\_intercept:是否有截据，如果没有则直线过原点;  
normalize:是否将数据归一化（标准化）;  
copy\_X:默认为True，当为True时，X会被copied,否则X将会被覆写;  
n\_jobs:默认值为1。计算时使用的核数

**一張含有 文字 的圖片

自動產生的描述**

2. XGB Regressor

模型参数

learning\_rate 迭代决策树时的步长(shrinkage)，又叫学习率，控制迭代速率，防止过拟合

xgb.XGBRegressor() learning\_rate，default=0.1 取值范围[0,1]

max\_depth 树的最大深度

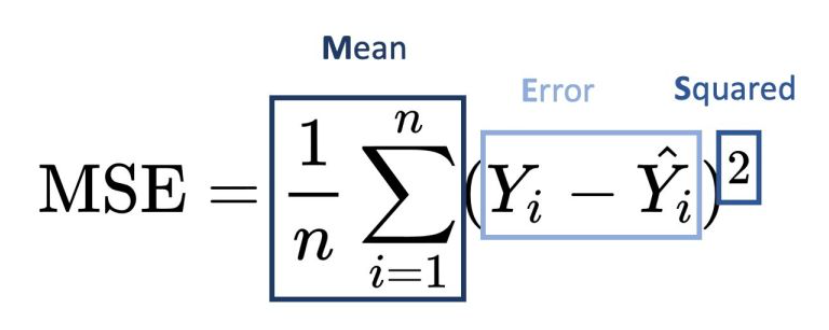
n\_estimators表示集成的弱评估器的个数，n\_estimators越大，模型的学习能力就会越强，模型也越容易过拟合。 在随机森林中，调整的第一个参数就是n\_estimators，这个参数非常强大，常常能够一次性将模型调整到极限， 在XGBoost中它也是如此。

xgb.XGBRegressor()：n\_estimators，default=100**一張含有 文字 的圖片

自動產生的描述**

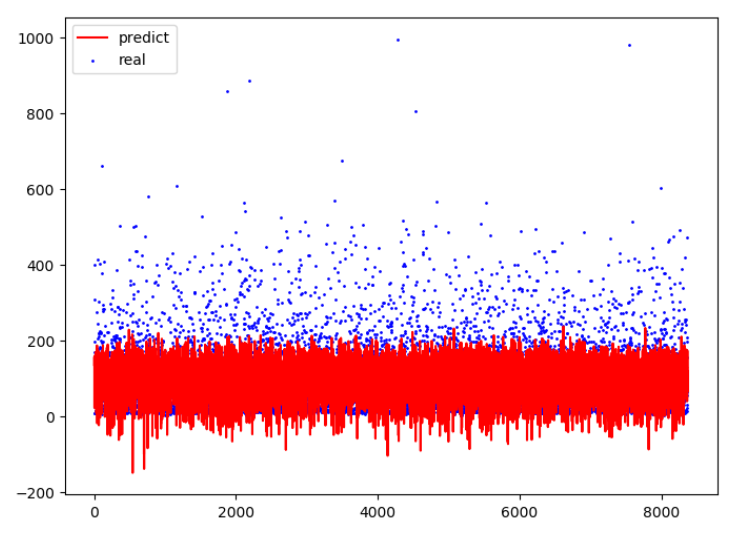
1. **分析与总结**
2. 评价指标 :均方误差 (MSE)

测量预测值Ŷ与某些真实值匹配程度。MSE 通常用作回归问题的损失函数。



1. 结果分析

普通的线性回归:



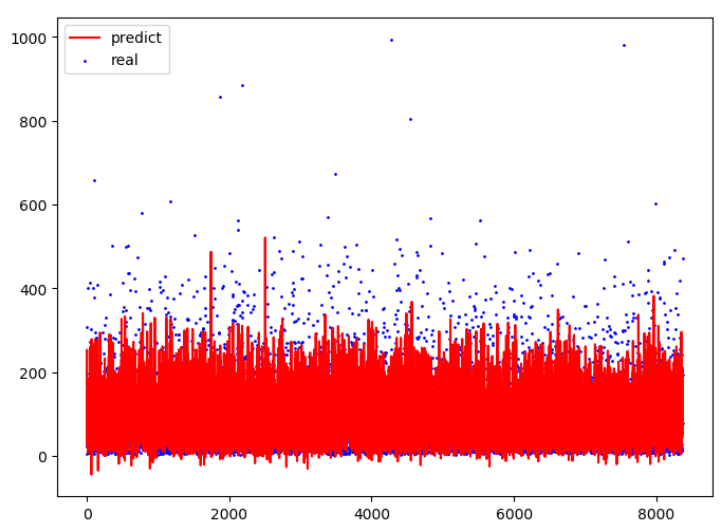
将数据进行标准化处理以后，测试精度和训练精度几乎没有改进

一張含有 文字 的圖片

自動產生的描述

MSE=5889

XGB Regressor:



一張含有 文字 的圖片

自動產生的描述

MSE=3607

由实验结果可知，使用XGB Regressor训练模型得出的效果较佳。