資料探勘HW2 科管所一 0753524 邱薇如

1. **資料前處理**
2. **取出10.11.12月資料**

|  |
| --- |
| data = pd.read\_excel("D://106年新竹站\_20180309.xls")  data = pd.DataFrame(data)  data = data.iloc[4914:6570,]  data = data.drop(['日期','測站', '測項'],axis=1) |
|  |

1. **缺失值以及無效值以前後一小時平均值取代 / c. NR表示無降雨，以0取代**

|  |
| --- |
| def data\_preprocess(data):  data = data.replace('NR',0)  data.iloc[:,0:] = data.iloc[:,0:].apply(pd.to\_numeric,errors='coerce')  temp\_data = pd.DataFrame(data.values.reshape(1,-1))  temp\_f = temp\_data.ffill(axis = 1)  temp\_b = temp\_data.bfill(axis = 1)  data = (temp\_f + temp\_b) / 2  data = pd.DataFrame(data.values.reshape(1656,24))  return data  data = data\_preprocess(data) |
|  |

1. **將資料切割成訓練集(10.11月)以及測試集(12月)**

|  |
| --- |
| def train\_test\_split(data):  train = data.iloc[:18\*61,:]  test = data.iloc[18\*61:,:]  return train, test  train, test = train\_test\_split(data) |
|  |

1. **製作時序資料:**

**將資料形式轉換為行(row)代表18種屬性，欄(column)代表逐時數據資料**

|  |
| --- |
| def reshape\_data(train, test):  train = pd.DataFrame(train.values.reshape(18,-1))  test = pd.DataFrame(test.values.reshape(18,-1))  Index = ['AMB\_TEMP','CH4','CO','NMHC','NO','NO2','NOx','O3','PM10',  'PM2.5','RAINFALL','RH','SO2','THC','WD\_HR','WIND\_DIREC',  'WIND\_SPEED','WS\_HR']  train.index = Index  test.index = Index  return train, test  train, test = reshape\_data(train, test) |
|  |

**2. 時間序列**

**a. 取6小時為一單位切割 / b. X請分別取PM2.5和所有18種屬性**

|  |
| --- |
| def generate\_data(data, label):  for i in range(data.shape[1]-6):  if label == 'pm':  if i==0:  X = data.iloc[:,0:6].values  else:  temp = data.iloc[:,i:i+6].values  X = np.concatenate((X,temp),axis=0)  if label == 'all':  if i==0:  X = data.iloc[:,0:6].values  else:  temp = data.iloc[:,i:i+6].values  X = np.concatenate((X,temp),axis=0)  Y = data.iloc[:,6:].T.values  return X, Y  X\_train, Y\_train = generate\_data(X\_train, ’pm’)  X\_test, Y\_test = generate\_data(X\_test, ‘all’) |
|  |

**c. 使用兩種模型 Linear Regression 和 Random Forest Regression 建模 /**

**d. 用測試集資料計算MAE (會有4個結果，2種模型\*2種X資料)**

|  |
| --- |
| def fit(X\_train, X\_test, label):  X\_train, Y\_train = generate\_data(X\_train, label)  X\_test, Y\_test = generate\_data(X\_test, label)  lr = LinearRegression().fit(X\_train, Y\_train.ravel())  lr\_pred = lr.predict(X\_test)  lr\_MAE = mean\_absolute\_error(Y\_test.ravel(), lr\_pred)  print(label," LR: ",lr\_MAE)  rf = RandomForestRegressor(n\_estimators=30).fit(X\_train, Y\_train.ravel())  rf\_pred = rf.predict(X\_test)  rf\_MAE = mean\_absolute\_error(Y\_test.ravel(), rf\_pred)  print(label," RF: ",rf\_MAE)  return lr\_MAE, rf\_MAE  fit(Pm\_X\_train, Pm\_X\_test, 'pm')  fit(train, test, 'all') |
|  |

|  |  |  |
| --- | --- | --- |
|  | PM2.5 | 18種屬性 |
| Linear Regression | 3. 51327 | 3. 26724 |
| Random Forest Regression | 4. 87097 | 4. 61701 |

**結論：**

不論是只有PM2.5或者是18種屬性的資料，使用Linear Regression的模型得到的誤差都比Random Forest Regression 的模型低，Linear Regression表現優於Random Forest Regression