# 案例二 公共自行车租借预测

## 【1】案例背景

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Now the problem is to predict the number of the bike-rental under a certain environment setting.The csv file contains the bike sharing background data of the feeling temperature,the weather sit,the humidity,the wind speed and the number of rented bike in a certain day or a certain hour in history.The feeling temperature,the humidity,and the wind speed data has already been normalized, which is suitable for fitting a linear regression.

自行车共享系统是传统自行车租赁的新一代，整个过程从会员注册、租赁到归还都变得自动化。通过这些系统，用户可以轻松地从一个特定位置租用自行车，并在另一个位置归还。现在的问题是要预测在特定环境设置下的自行车租赁数量。CSV文件包含了历史上某一天或某一小时的自行车共享背景数据，包括气温感觉、天气状况、湿度、风速以及租用的自行车数量。气温感觉、湿度和风速数据已经被标准化，适合用于拟合线性回归模型。

## 【2】方法陈述

I use Multiple Linear Regression to describe the relationship of the feeling temperature,the weather sit,the humidity,the wind speed with the bike-rental number.It can be easily fitted thanks to the help of the scikit-learn library function LinearRegression().

我使用多元线性回归来描述气温感觉、天气情况、湿度和风速与自行车租赁数量之间的关系。感谢scikit-learn库中的LinearRegression()函数，模型的拟合变得非常容易。

## 【3】实验代码

### 步骤一 import python library pandas,sklearn,numpy

import numpy as np

import pandas as pd

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error

#引用numpy库、pandas库和sklearn库中的线性回归计算相关的函数

### 引入csv数据（注意！day.csv与hour.csv是在两次不同的运行实例当中引入的）

df = pd.read\_csv('day.csv')#使用“day.csv”

df = pd.read\_csv('hour.csv')#使用“hour.csv”

### Extracting data from csv file as datasets

x = df[['atemp','hum','weathersit','windspeed']]

y = df['cnt']

#划分题干要求的4个自变量为DataFrame\_X,题干所要预测的自行车租赁量为应变量DataFrame\_Y，方便后续的拟合

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.02, random\_state=20)

#将两个DataFrame按照2%的比例划出训练集和测试集，方便模型的训练和检验

### Use function LinearRegression() to fit a linear regression model

model = LinearRegression()

model.fit(x\_train, y\_train)

#利用训练集，使用sklearn库中的LinerRegression函数进行回归线性拟合

### Performing model evaluation

y\_pred = model.predict(x\_test)

mse = mean\_squared\_error(y\_test,y\_pred)

print(f"Mean Square Error {mse}")

#计算均方误差（MSE）来进行模型评估

### Create a prediction dataset(数据是手动编的，仅作模拟预测)

new\_weathersit = [4.0]

new\_atemp = [0.189405]

new\_hum = [0.590435]

new\_windspeed = [0.1869]

#创建预测用的数据集

### Conduct a prediction and output a integer result

new\_data = pd.DataFrame({ 'atemp': new\_atemp, 'hum': new\_hum,'weathersit': new\_weathersit, 'windspeed': new\_windspeed})

predicted\_rentals = model.predict(new\_data)

print(f'Predicted Bike Rentals: {predicted\_rentals[0]:.0f}')

#进行预测并输出整型结果

## 【4】结果分析

Using the data set splitting by day as the smallest time unit has a pretty high MSE number of 1083958,and the predicted result is 1381,which is relatively high comparing with the number of the data set.The result is absolutely unreasonable as the prediction weather sit is 4,which means Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog as no one will rent a bike under such environment.But using the data set splitting by hour,the MSE number reduces to 20982.We can learn that the fitting effectiveness using the data set of “hour.csv” is better than using data set splitting by day.Using the current model,the final prediction result is 100.

使用以天为最小时间单位拆分的数据集具有相对较高的均方误差（MSE）值，为1083958，并且预测结果为1381，与数据集中的实际值相比相对较高。这个结果与数据集中的实际情况明显不符，因为预测的天气条件是4，意味着大雨+冰粒+雷暴+薄雾，雪+雾，这种天气条件下几乎没有人会租自行车。但是，如果使用以小时为单位拆分的数据集，MSE值减少到20982。我们可以看出，使用“hour.csv”数据集的拟合效果要好于使用以天为单位拆分的数据集。使用当前模型，最终的预测结果是100。