本周汇报

3 Days

◎ 视频

李宏毅：

Regression作业的复现

P4-P7

花书：

第一周视频（前四章）

In [2]:
```
import sys
import pandas as pd
import numpy as np
# 繁体字以big5编码 又称大五码，是繁体中文字符集编码标准
data = pd.read_csv('train.csv', encoding='big5')
# 读取前十行
data.head(10)
```

In [3]:
```
# 显示数据的尺寸
data.shape
```

In [4]:
```
# 只要数值部分
data = data.iloc[:, 3:]
data[data == 'NR'] = 0
# 转换成numpy的数组
raw_data = data.to_numpy()
raw_data.shape
```

In [5]:
```
# 此处使用了字典 month为键 键的值为每个月的数据
month_data = {}
for month in range(12):
    sample = np.empty([18, 480])
    for day in range(20):
        # 所有行（一共18行） 每天（24小时）的数据 = 原数据所有列（24列） 每天18行数据
        sample[:, day * 24 : (day + 1) * 24] = raw_data[18 * (20 * month + day) : 18 * (20 * month + day + 1), :]
        # 每个月数据拼接起来
    month_data[month] = sample
# 分成了12个月，每个月有18行×480列的数据。
# 对于每个月，每10个小时分成一组，由前9个小时的数据来预测第10个小时的PM2.5，把前9小时的数据放入x，把第10个小时的数据放入y。因
# 把一组18×9的数据平铺成一行向量，然后放入x的一行中，每个月有471组，共有12×471组向量，因此x有12×471行，18×9列。
# 将预测值放入y中，y有12（月）×471（组）行，1列。
x = np.empty([12 * 471, 18 * 9], dtype = float)
y = np.empty([12 * 471, 1], dtype = float)
for month in range(12):
    for day in range(20):
        for hour in range(24):
            # 不懂
            if day == 19 and hour > 14:
                continue
            # 最大值还是12*471 从0 开始  往x， y里放数据
            # reshape(1, -1)： 转换为一行 18*9 第二个18*9.....
            x[month * 471 + day * 24 + hour, :] = month_data[month][:,day * 24 + hour : day * 24 + hour + 9].reshape(1, -1)
            y[month * 471 + day * 24 + hour, 0] = month_data[month][9, day * 24 + hour + 9] #value
print(x)
```