广告投入与产品销量预测——特征工程

该数据集来自 Advertising.csv 是来
自 http://faculty.marshall.usc.edu/gareth-james/ISL/Advertising.csv

数据集包含 200 个样本，每个样本有 3 个输入属性：

1. 电视广告投入
2. 收音机广告投入
3. 报纸广告 以及一个输出/响应：
4. 产品销量

1. 导入必要的工具包

```
#数据处理
import numpy as np
import pandas as pd

#数据可视化
import matplotlib.pyplot as plt
%matplotlib inline

#显示中文
plt.rcParams['font.sans-serif'] = ['Arial Unicode MS']
```

2. 读取数据

```
#读取数据
dpath = "./data/"
df = pd.read_csv(dpath + "Advertising.csv")

#通过观察前5行，了解数据每列（特征）的概况
df.head()
#去掉索引列
df.drop(['Unnamed: 0'], axis = 1, inplace = True)
# 数据总体信息
df.info()
```

3. 数据标准化

```
# 从原始数据中分离输入特征x和输出y
y = df['sales']
X = df.drop('sales', axis = 1)

#特征名称，用于后续显示权重系数对应的特征
```

```
feat_names = X.columns
# 数据标准化
# 本数据集中3个特征的单位相同，可以不做特征缩放，不影响正则
# 但3个特征的取值范围不同，如果采用梯度下降/随机梯度下降法求解，
# 还是将所有特征的取值范围缩放到相同区间
from sklearn.preprocessing import StandardScaler

# 分别初始化对特征和目标值的标准化器
ss_X = StandardScaler()

# 对训练数据，先调用fit方法训练模型，得到模型参数；然后对训练数据和测
试数据进行transform
X = ss_X.fit_transform(X)
```

4. 保存特征工程的结果到文件，供机器学习模型使用

```
fe_df = pd.DataFrame(data = X, columns = feat_names, index = df.index)

#加上标签y
fe_df["sales"] = y

#保存结果到文件
fe_df.to_csv(dpath + 'FE_Advertising.csv', index=False)
fe_df.head()
```