第一周作业 利用线性回归技术实现共享单车数量预测

一、数据说明： Capital Bikeshare （美国Washington, D.C.的一个共享单车公司）提供的共享单车数据。数据包含每天的日期、天气等信息，需要预测每天的共享单车骑行量。

原始数据集地址：http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

1) 文件说明

day.csv: 按天计的单车共享次数（作业只需使用该文件）

hour.csv: 按小时计的单车共享次数（无需理会）

readme：数据说明文件

2) 字段说明

Instant记录号

Dteday：日期

Season：季节（1=春天、2=夏天、3=秋天、4=冬天）

yr：年份，(0: 2011, 1:2012)

mnth：月份( 1 to 12)

hr：小时 (0 to 23) （只在hour.csv有，作业忽略此字段）

holiday：是否是节假日（0/1）

weekday：星期中的哪天，取值为0～6

workingday：是否工作日（0/1）

1=工作日 （是否为工作日，1为工作日，0为非周末或节假日）

weathersit：天气（1：晴天，多云 2：雾天，阴天 3：小雪，小雨 4：大雨，大雪，大雾）

temp：气温摄氏度

atemp：体感温度

hum：湿度

windspeed：风速

casual：非注册用户贡献的骑行量（作业无需理会该字段）

registered：注册用户贡献的骑行量（作业无需理会该字段）

cnt：给定日期（天, day.csv）时间（每小时,hour.csv）总租车人数，响应变量y

casual、registered和cnt三个特征均为要预测的y（cnt =casual+registered ），作业里只需对cnt进行预测。

二、作业要求：

1. 对数据做数据探索分析（可参考EDA\_BikeSharing.ipynb，不计分）
2. 适当的特征工程（可参考FE\_BikeSharing.ipynb，不计分）
3. 对全体数据，随机选择其中80%做训练数据，剩下20%为测试数据，评价指标为RMSE。（10分）
4. 用训练数据训练最小二乘线性回归模型（20分）、岭回归模型、Lasso模型，其中岭回归模型（30分）和Lasso模型（30分）需采用5折交叉验证得到最佳正则超参数。
5. 比较用上述三种模型得到的各特征的系数，以及各模型在测试集上的性能。并简单说明原因。（10分）