HW1：线性回归

1、 任务描述

请在 Capital Bikeshare （美国 Washington, D.C.的一个共享单车公司）提供的自行车数据上进行回归分析。训练数据为 2011 年的数据，要求预测 2012 年每天的单车共享数量。

需要提交代码文件，并给出必要的结果解释。

1) 训练数据和测试数据分割（请将 2012 年的数据作为测试数据）;（20 分）
2) 适当的特征工程（及数据探索）;（20 分）

   提示：

   a) 有些特征看起来是数据值特征，其实是类别型特征，如月份、季节
   b) 数值型特征归一化
   c) 可以丢弃一些不必要的特征
3) 岭回归，并选择最佳的正则参数；（30 分）

   a) 参数调优
   b) 结果可视化
4) Lasso，并选择最佳的正则参数；（30 分）

2、 数据说明：

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

weekday：星期中的哪天，取值为 0～6

workingday：是否工作日

1=工作日（非周末和节假日）

0=周末

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：非注册用户个数

registered：注册用户个数

cnt：给定日期（天）时间（每小时）总租车人数，响应变量 y

注意：蓝色标记的后三个特征均为要预测的 **y**，作业里只需对 cnt 进行预测

黑色标记的特征为输入特征 **x**