ChallengeB

Guohui Jiang; Jiayin Zhai

11/25/2017

Github[**link**](https://github.com/jiangguohui/rprogramming_challengeB). In Task 3B, we need about 6 mins to import the SIREN dataset into out computer.

# Task 1B - Predicting house price in Ames, Lowa (continued)

## Step 1

The intuition behind Random Forests: we randomly split the training data into some subsamples, then we will use models to predict the observations of each subsample. The final prediction should be a function of each prediction. For example, we have 10,000 observations in the training data, then we randomly draw 100 subsamples from it. We will run a model on one subsample. It will repeat the process 100 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction, or we can pick final prediction based on majority rule.

## Step 2

Before the following operations, we first deal with the dataset according to the solutions to Challenge A. **Please change the path when you import dataset.** We train the data through randomForest, and the features we include are the same as thoses in linear model.

## Step 3

We compare the two predictions through descriptive statistics and density distribution. We think they are very similar. See the details in the script.

# Task 2B - Overfitting in Machine Learning (continued)

First we simulate the data and split it into training and testing as we did in Challenge A.

## Step 1-3

See the details in the script. See the figure in the following pages.

## Step 4

The high-flexibility model(blue one) is more variable, and the low-flexibility model(red one) is less biased.

## Step 5

Here the high-flexibility model(blue one) is more variable, but the low-flexibility model is less biased. See the figure in the following pages.

## Step 6-9

See the details in the script.

## Step 10:

Conclusion: As the bandwidth increases, the mse on training data always increases, which means as the training model becomes less flexible, it will have a larger MSE, i.e. the model is more biased compared with the true model when it’s not flexible.  
However, the mse on testing data first decreases and then increases as the bandwidth increases, which means when the model is very flexible or very not flexible, the predictions are both more biased compared with the true model. Therefore we should not set the bandwidth too low or too high. Especially we should not set the bandwidth too low, in which case the training model is less biased but the predictions are far from the true values.

# Privacy regulation compliance in France

## Step 1

Use fread function to do that.

## Step 2

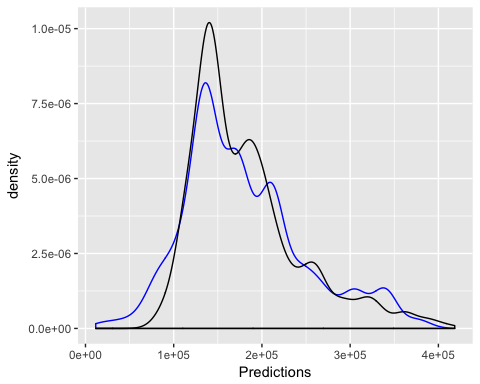
We use substr function to get the first two digits of post code, and then we count them and manage them in a nicer form.

## Step 3

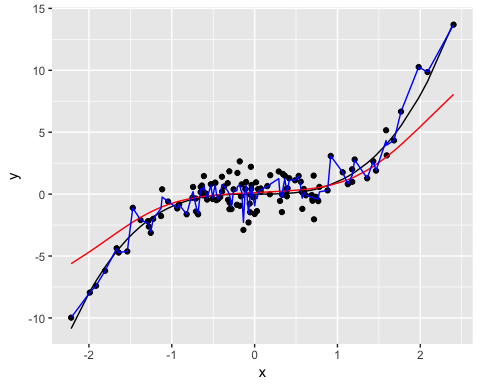
**Before importing the SIREN large dataset, please download and extract it.** To import the large dataset, we split it into 4 parts, and import them separately, and subsequently merge them with the initial dataset. After completing every step, we remove the big dataset and keep the merged ones and then merge all the 4 datasets into 1. Finally, we work on this merged dataset and select the most up-to-date rows for the same siren.  
For each dataset(with around 3000000 observations), the importing needs about 60s. For the whole step it requires less than 10 mins to run. The final dataset we need is *siren\_fin*.

## Step 4

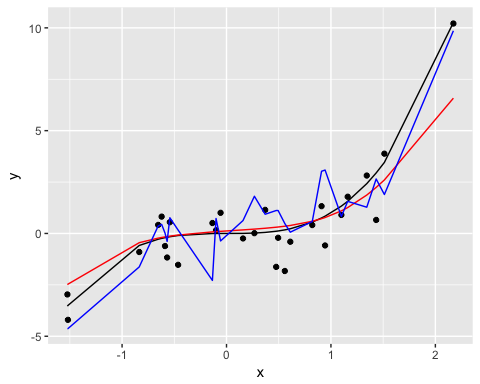
In terms of the number of salaried employees, the distribution of company size is left skewed, which means that there are more companies with less than 50 employees.



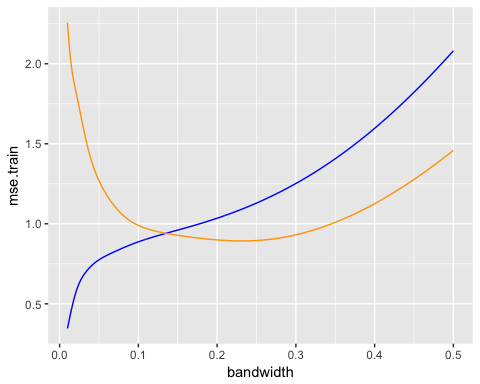
Task1 Step3 - Density Functions of two predictions



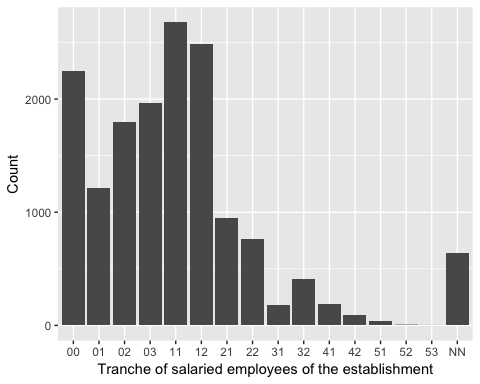
Task2 Step3 - Predictions of ll.fit.lowflex and ll.fit.highflex on training data.



Task2 Step5 - Predictions of ll.fit.lowflex and ll.fit.highflex on test data.



Task2 Step 10 - MSE on training and test data for different bandwidth - local linear regression



Task3 Step 4 - The histogram of the size of the companies that nominated a CIL.

## represeantation meaning  
## 1 NN Non-employer units  
## 2 00 0 employees  
## 3 01 1 or 2 employees  
## 4 02 3 to 5 employees  
## 5 03 6 to 9 employees  
## 6 11 10 to 19 employees  
## 7 12 20 to 49 employees  
## 8 21 50 to 99 employees  
## 9 22 100 to 199 employees  
## 10 31 200 to 249 employees  
## 11 32 250 to 499 employees  
## 12 41 500 to 999 employees  
## 13 42 1,000 to 1,999 employees  
## 14 51 2,000 to 4,999 employees  
## 15 52 5,000 to 9,999 employees  
## 16 53 10,000 employees and more