Data Science

Homework 3

여러분은 아래 주어진 task들을 수행하여 그 과정과 결과 및 설명을 담은 **보고서(1)**와 task를 수행하기 위해서 사용한 R code를 담은 **R script 파일(2)**을 제출해야 합니다. 과제를 수행하기 위해 필요한 데이터는 각각 **Bikes.RData**와 **income.RData** 파일에 포함되어있습니다.

Single Variable Model for Regression

For this task, you will build a single-variable model to predict the number of bikes rented in an hour using either one variable of weather condition, the type of day (holiday, working day, or weekend), or the time of day. You will train the model on data from the month of July (bikeJuly data frame).

The data frame has the columns:

cnt: the number of bikes rented in that hour (the outcome, target variable)

hr: the hour of the day (0-23, as a factor)

holiday: TRUE/FALSE

workingday: TRUE if neither a holiday nor a weekend, else FALSE

weathersit: categorical, "Clear to partly cloudy"/"Light Precipitation"/"Misty"

temp: normalized temperature in Celsius

atemp: normalized "feeling" temperature in Celsius

hum: normalized humidity

windspeed: normalized windspeed

instant: the time index -- number of hours since beginning of data set (not a variable)

mnth and yr: month and year indices (not variables)

1. bikeJuly data frame을 사용하여 목적 변수(cnt)를 가장 잘 예측하는 (단일 변수 모델) single variable model(Best Model)을 찾으시오(**RMSE 기준**). 목적 변수가 연속형이므로 regression model이고 단일 변수 모델이므로 하나의 입력 변수만을 모델에 사용해야 합니다. 가장 성능이 좋은 모델을 찾아가는 과정을 보고서에 함께 기술하시오. (최선의 모델을 만들기 위해서 기존의 변수 외에 새로운 변수를 추가하는 것도 가능합니다.)
2. 1번에서 찾은 모델의 RMSE와 R square 값을 계산하시오.
3. 1번에서 찾은 모델을 bikesAugust data frame에 대해서 테스트하시오. 테스트 RMSE와 R square의 값은 어떻게 계산되나요?
4. 찾은 모델은 overfitting(과적합)이라고 할 수 있나요? 할 수 있거나 혹은 할 수 없거나 왜 그런지 이유를 설명하시오.
5. 여러분이 찾은 모델의 성능을 Test 데이터 기준으로 다른 동료와 비교하시오. 여러분의 모델이 동료의 모델의 성능보다 낮다면 왜 그런지 생각해보고, 여러분의 모델을 개선하시오. 비교 과정/결과와 개선 방법/개선 결과를 보고서에 기술하시오.
6. 학습 데이터와 테스트 데이터에서, 각각 y축에 목적변수, x축에 목적변수에 대한 모델의 예측 값을 그린 산점도(scatter plot)을 그리시오. 그래프로부터 알 수 있는 것이 있다면 설명해 보시오.

Single Variable Model for Classification

For this task, you will work with a medium sized dataset about the income of people given a set of features like education, race, sex, and so on. Each observation is labeled with 1 or 0: 1 means the observation has annual income equal or above $50,000, 0 means the observation has an annual income lower than $50,000 (Source: UCIMLR). This label information is stored in the income variable. The data sets are two, one for training a prediction model and one for testing the model.

1. 학습데이터(**train** data frame)를 사용하여 목적 변수를 가장 잘 예측하는 (단일 변수 모델) single variable model을 찾으시오 (**AUC 기준**). 목적변수가 범주형이므로 classification model이며 단일 변수 모델이므로 하나의 입력 변수만을 모델에 사용해야 합니다. 가장 성능이 좋은 모델을 찾아가는 과정을 보고서에 함께 기술하시오. (최선의 모델을 만들기 위해서 기존의 변수 외에 새로운 변수를 추가하는 것도 가능합니다.) 어떤 변수를 사용하

였나요? AUC는 얼마가 나오나요?

2. 1번에서 찾은 모델을 **test** data frame에 대해서 test해서 AUC를 계산하시오.

3. 찾은 모델은 overfitting(과적합)이라고 할 수 있나요? 할 수 있거나 혹은 할 수 없거나 왜 그런지 이유를 설명하시오.

4. Best model에서 Threshold를 변화시킬 때, precision과 recall 값이 어떻게 변화하는지 그래프로 표현하시오. 그래프를 바탕으로 어떤 threshold을 선택하는 것이 좋을지 여러분의 의견을 이유와 함께 제시하시오. (학습 데이터 테스트 데이터 각각에 대해)

5. Trade-off 관계에 있는 precision과 recall을 하나의 measure로 보기 위해서 F_1 Measure라는 것을 사용하기도 합니다. F_1 Measure를 계산하는 수식은 아래와 같습니다. F_1 값이 가장 크게 되는 threshold는 얼마인가요? (학습 데이터 테스트 데이터 각각에 대해)

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

6. 학습데이터와 테스트 데이터 각각에 대해서 여러분 모델의 ROC 커브를 그려보시오. 그래프를 통해서 알 수 있는 점이 있다면 설명해보시오.

7. Accuracy가 가장 크게 되는 threshold는 얼마인가요? (학습 데이터 테스트 데이터 각각에 대해)

8. 여러분이 찾은 모델의 성능을 Test 데이터 기준으로 다른 동료와 비교하시오. 여러분의 모델이 동료의 모델의 성능보다 낮다면 왜 그런지 생각해보고, 여러분의 모델을 개선하시오. 비교 과정/결과와 개선 방법/개선 결과를 보고서에 기술하시오.