Data Science

Homework 4

Perform following tasks and submit your work in html, word, or pdf generated by RMarkdown. (Chance to practice RMarkdown). The dataset you need for this homework is in **"PRSA_data.csv".**

여러분은 아래 주어진 task들을 수행하여 그 과정과 결과 및 설명을 담은 **보고서**를 RMarkdwon으로 작성하여 html, word, 또는 pdf 형식으로 제출하세요. 과제를 수행하기 위해 필요한 데이터는 각각 **"PRSA_data.csv"** 파일에 포함되어있습니다.

PRSA is an hourly data set contains the PM2.5 data of US Embassy in Beijing. Meanwhile, meteorological data from Beijing Capital International Airport are also included.

PRSA 데이터는 2010.1.1 ~ 2014.12.31 기간 동안 중국 베이징의 미세먼지 농도 및 날씨 관련 정보를 기록한 데이터이다. 변수의 설명은 아래와 같다.

No: row number

year: year of data in this row

month: month of data in this row

day: day of data in this row

hour: hour of data in this row

pm2.5: PM2.5 concentration (ug/m^3)

DEWP: Dew Point

TEMP: Temperature

PRES: Pressure (hPa)

cbwd: Combined wind direction

Iws: Cumulated wind speed (m/s)

Is: Cumulated hours of snow

Ir: Cumulated hours of rain

1. When forecasting air condition, "**very bad**" indicates PM2.5 exceeding 75(ug/m^3). Let us add a new variable **bad_air** indicating whether air condition is **"very bad"** = True or not = False. Use the dataset from 2010 to 2013 to train model. To test the model, use subset of data in year of 2014. (remove observation with missing values in PM2.5 Variable)

   미세먼지 예보 기준에 따르면 PM2.5가 75(ug/m^3)를 초과할 때, "매우 나쁨"이 된다. 미세먼지가 매우 나쁨을 의미하는 **bad_air column을 추가하자**. bad_air 변수의 값은 TRUE/FLASE이다. 이 때, PM2.5에 **NA 존재하는 행은 삭제**한다. 2010년부터 2013년까지의 데이터를 **학습 데이터(train data)**, 2014년데이터를 **테스트 데이터로 설정하자**.

(Decision Tree)

2. Find the best decision tree **model A** in terms of "Accuracy" using training dataset. What are the Accuracy, Precision, Recall, and F1 of **model A** for training dataset?

   학습데이터를 사용하여 미세먼지 수치를 예측하는 decision tree model(Best Model)을 만드시오. 학습데이터에 대해서 **Accuracy**가 가장 높은 모델 A를 찾으시오. A 모델의 Training 데이터 기준 Accuracy, Precision, Recall, F1 값을 계산하시오.

3. Find the Accuracy, Precision, Recall, and F1 of **model A** for test dataset. Do you think that model A is overfitting? Explain your argument.

   2번에서 찾은 A 모델의 성능(Accuracy, Precision, Recall, F1)을 Test 데이터에 대해서 계산하시오. 찾은 모델은 overfitting(과적합)이라고 할 수 있는가? 그렇게 생각하는 이유는 무엇인가?

4. If your model A is overfitting, try to fix the issue of overfitting to improve your model. State the process and result of your attempt.

   과적합을 해결하기 위한 다양한 시도를 해보고 그 시도들과 결과를 함께 기술하시오.

5. Do you find that fixing overfitting problem of model improve the performance measures for test dataset? Let us call it **model B** that you have improved **in question 4**

   과적합이 해결됨에 따라서 Test data에 대한 성능이 향상 되었는지 설명하시오. (과적합을 해소하고 Test data에 대한 성능 또한 개선된 모델을 B라고 하자).

6. Assume that we make forecast of air condition to **"very bad"** or "**not very bad"** according to the result of our prediction model. Explain the cost of false positive and false negative

(Cost = Things that can be caused by two different types of errors). In that sense, explain how we should adjust precision and recall since they are in trade off. (Opinion with rationale)

모델의 예측 결과에 따라 우리는 미세먼지 농도를 "매우 나쁨"으로 예보하거나 "매우 나쁨은 아님"으로 예보하게 된다고 하자. False Positive와 False Negative의 비용에 대해서 설명하시오. 이런 맥락에서 Precision과 Recall은 어떻게 조절하는 것이 좋을지 여러분의 의견을 이유와 함께 제시하시오.

7. Draw ROC curves for both **model A** and **model B**, and find AUC for both models. (for both training and test datasets.) (You could get estimated probability by giving **type = 'prob'** option when modeling).

   모델 A와 모델 B에 대해서 ROC 커브를 그리고 AUC를 계산하시오. (Train 데이터, Test 데이터 각각에 대해서). (Hint, decision tree model의 predict 명령에서 type = 'prob'를 하면 확률 값을 얻을 수 있습니다.)

(KNN)

8. This time, we try k-nearest neighbor to predict **bad_air** variable. (classification). Perform all data preparation steps that are needed for knn method on the dataset. Include the preparation process and result in the homework.

   이번에는 k-nearest neighbor 방법을 이용하여 **bad_air**를 예측하고자 한다. knn을 적용하기 위해서 어떠한 전처리 단계가 필요한지 설명하고 전처리를 수행하라. (제출 답안에 전처리 과정 및 결과를 포함)

9. Find out Accuracy, Precision, Recall, F1 value of knn method for test dataset. (You may choose an arbitrary value for k).

   knn을 수행하여 예측 결과의 Accuracy, Precision, Recall, F1을 계산하시오. (k값은 임의로 결정)

10. Show how the performance measures (Accuracy, Precision, Recall, F1, **AUC**) change over k values. Try at least 20 different values for k and draw the result as line graph. Suggest the best k-value for this problem.

    9번에서 사용한 k 값 외에 20개 이상의 다른 k값을 적용하여 Accuracy, Precision, Recall, F1, **AUC** 값의 변화를 측정하시오. 측정된 결과를 아래와 같은 그래프로 표시하고, 최선의 k는 무엇인지 제시하시오.

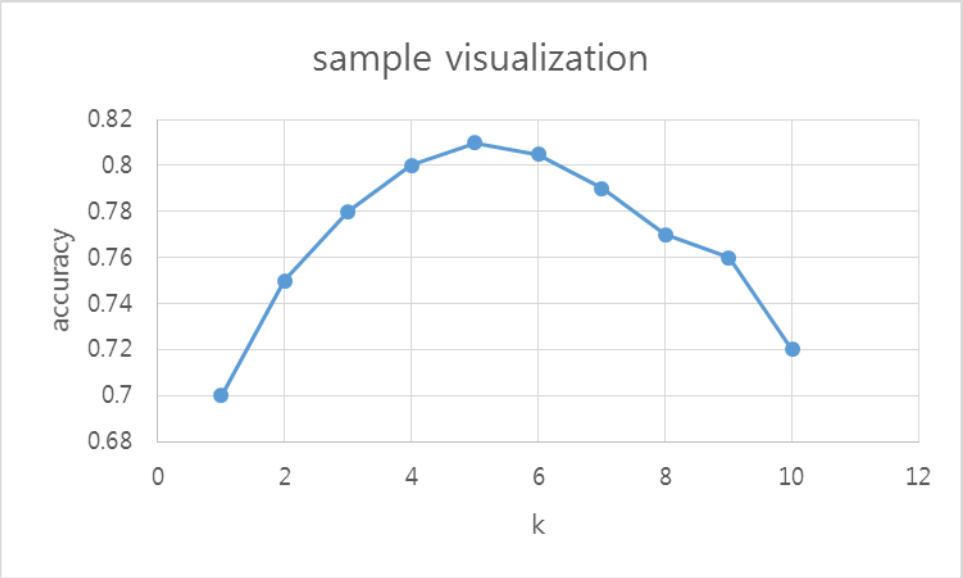

Figure 1 Sample line graph for question 10

11. Precision or Recall is more important than the other as you answered in question 6. According to that, explain how you can improve that measure (Precision or Recall) and apply it to knn method (use the **best k** for the k-value suggested in question 10). Include the process and the result of your improvement in your homework.

    (10번에서 얻은 최선의 k를 이용하시오.) 6번에 답한 바와 같이 precision 또는 recall이 더 중요한 상황이라고 한다면, 더 중요한 지표를 높이기 위해서 어떤 방법을 사용할 수 있는지 설명하고, 더 중요한 지표를 높이는 방법을 적용하여 그 결과를 보이시오. (보고서에 과정과 결과를 포함)