Practice 8: Decision Tree

Loading data into R

```
prsa <- read.csv('PRSA_data.csv')
```

PRSA dataset

- 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.

```
## Rows: 43,824
## Columns: 13
## $ No    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18...
## $ year  <int> 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2...
## $ month <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ day   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ hour  <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,...
## $ pm2.5 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DEWP  <int> -21, -21, -21, -21, -20, -19, -19, -19, -19, -20, -19, -18, -...
## $ TEMP  <dbl> -11, -12, -11, -14, -12, -10, -9, -9, -9, -8, -7, -5, -5, -3,...
## $ PRES  <dbl> 1021, 1020, 1019, 1019, 1018, 1017, 1017, 1017, 1017, 1017, 1...
## $ cbwd  <chr> "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "NW", "...
## $ Iws   <dbl> 1.79, 4.92, 6.71, 9.84, 12.97, 16.10, 19.23, 21.02, 24.15, 27...
## $ Is    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Ir    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

variable	description
No	Row Index
year	관측연도
month	관측월
day	관츨일
hour	관측시간 0h ~ 23h
pm2.5	미세먼지 농도 Fine dust concentration (ug/m^3)
DEWP	Dew Point (이슬점)
TEMP	Temperature (기온)
PRES	Air pressure (기압) hPa
cbwd	Wind Direction (풍향)
Iws	Cumulated wind speed (m/s) (누적 풍속)
Is	Snowfall per hour (시간당 누적 강설량)
Ir	precipitation per hour(시간당 누적 강수량)

Question 1

미세먼지 예보 기준에 따르면 PM2.5가 75(ug/mˆ3)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 삭제한다.

2010년부터 2013년까지의 데이터를 학습 데이터(train data), 2014년데이터를 테스트 데이터로 설정하자.

English

If fine dust concentration **PM2.5** exceeds 75(ug/mˆ3), it is announced as “very bad” air condition and be harmful to human body.

- Let us add a new column names **bad_air** that has TRUE/FALSE where TRUE is for bad air condition and FALSE otherwise.
- Remove rows with missing PM2.5 from the data frame.
- Partition the dataset from year of 2010 to 2013 for train, the rest of year 2014 for testing.

Question 2

학습데이터를 사용하여 미세먼지 나쁨여부(bad_air)를 예측하는 decision tree model(Best Model)을 만드시오.

학습데이터에서 사용가능한 모든 변수를 입력변수로 사용하여 미세먼지 나쁨여부(bad_air) 예측 decision tree를 학습하시오.

학습에 사용가능한 변수와 사용할 수 없는 (혹은 사용하면 안되는) 변수가 있다면 무엇인지 설명하고, 제외하여 모델을 학습하시오.

English

Train a decision model that predicts whether air condition is “very bad” or not in the **bad_air** column.

Use all available variables in the dataset for training and prediction.

If there are some variables (or columns) that are not proper to include in the model as input variable, state what they are and why they are not good to use.

Question 3

문제 2에서 학습한 모델의 Accuracy, Precision, Recall, F1 값을 계산하여보라.

Train data와 Test data 둘다에 대해서 계산한 후 비교하여보라.

이 모델은 과적합인가 이유와 함께 설명하여라.

English

Find out Accuracy, Precision, Recall and F1-score for both train and test dataset of your decision tree model.

Considering performance on both datasets, do you find your model is overfitting?

Explain why you think so.

예시

```
## [1] "accuracy for train dataset: 0.893"
## [1] "accuracy for test  dataset: 0.728"
```

Question 4

pre-pruning 방식을 사용하여 과적합을 해소한 모델을 학습하여보라.

새롭게 학습한 모델의 Accuracy, Precision, Recall, F1 값을 계산하여보라.

과적합이 얼마나 해소되었는지, 성능은 어떻게 변화하였는지 설명하여라

English

Try to resolve overfitting applying pre-pruning approach to your decision tree.

Calculate Accuracy, Precision, Recall and F1-score for your modified decision tree.

State whether overfitting has been resolved and how the model performance has been changed after modification.

Question 5

post-pruning 방식을 사용하여 과적합을 해소한 모델을 학습하여보라.

새롭게 학습한 모델의 Accuracy, Precision, Recall, F1 값을 계산하여보라.

과적합이 얼마나 해소되었는지, 성능은 어떻게 변화하였는지 설명하여라

English

Try to resolve overfitting applying post-pruning approach to your decision tree.

Calculate Accuracy, Precision, Recall and F1-score for your modified decision tree.

State whether overfitting has been resolved and how the model performance has been changed after modification.

Question 6

설명력이 높은 변수를 추가하거나, 중복되거나 의미없는 변수를 제거하는 과정을 feature engineering이라고 한다. 이러한 과정을 통해서 모델의 성능을 개선할 수 있다.

입력변수의 조정을 통해서 성능이 더 높은 모델을 학습하여보라.

(이 때, 문제 4,5에서 적용했던 pruning 방식을 함께 적용해도 됨)

English

You may try to add new more predictive variables or remove variables irrelevant to pm2.5 which is called "feature engineering" process and generally effective to improve the prediction model.

Find a decision tree model better than you found from previous questions by feature engineering process.

(You may also apply pre and post-pruning to your model to make it better)

Question 7

문제 2에서 6까지 얻은 모델 중 가장 성능이 높은 모델을 선정하여.

ROC커브를 그리고 AUC를 계산하시오.

- Train 데이터, Test 데이터 모두에 대해서
- Hint, decision tree model의 predict 명령에서 type = 'prob'를 하면 확률 값을 얻을 수 있습니다.

English

Pick the best decision tree model from Q2 ~ Q6 and draw ROC curve and find AUC for both train and test datasets for the best model.

- Hint, you may obtain estimated probability by giving **type = 'prob'** option when you train a decision tree model.

Question 8

문제 7에서 선정한 모델에서 threshold를 0에서 1까지 변경할 때,

Accuracy, Precision, Recall, F1 값의 변화를 확인하여라.

threshold를 얼마로 선정하는 것이 가장 적절할 것인지 판단하여 설명하여라.

- 모델의 예측 결과에 따라 우리는 미세먼지 농도를 "매우 나쁨"으로 예보하거나 "매우 나쁨은 아님"으로 예보하게 된다고 하자. False Positive와 False Negative의 비용에 대해서 생각해보고, 이런 맥락에서 Precision과 Recall을 어떤 수준으로 조정해야하는지, 그에 따른 threshold를 생각해보자.

English

For the best model from Q7, show how Accuracy, Precision, Recall, and F1-score change over threshold value 0 ~ 1.

What would be the best choice of threshold to choose for the model to forecast the air condition?

Hint, imagine the cost of two different misclassification(error) of false positive and false negative and how we should control threshold to have proper value of precision and recall.