Data Science

Homework 4

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 "**income_hw4.RData"** 파일에 포함되어있습니다.

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.

이번 숙제는 숙제3에서 사용했던 인구조사 데이터를 활용해서, 해당 사람이 연봉이 5만불이 넘을 것인지 넘지 않을 것인지를 예측하는 모델을 작성하는 것입니다. 데이터는 학습용과 테스트용으로 나누어져 있습니다.

아래 문제를 풀고, r script와 보고서를 edmodo에 제출하시오.

1. Decision Tree

Question 1.

Build the best decision tree model that predict ***income*** variable as accurately as you can. Explain how you come up with that model and explain your model's performance in terms of confusion matrix, accuracy, precision, and recall.

(Korean translation) decision tree를 사용해서 ***income*** 변수를 가장 잘 예측하는 예측 모델을 만드시오. 어떻게 그 모델을 찾게 되었는지 설명하시오. confusion matrix, accuracy, precision, recall을 사용해 만든 모델의 성능을 설명하시오.

Question 2.

Visualize your decision tree model. you may choose to use R function, excel spreadsheets, or any software for visualization. or you can even hand-draw the model. What kind of information or insight can you see from your visualization?

(Korean translation) 앞서 만든 decision tree모델을 그려봅시다. R 함수를 사용해도 되고 엑셀을 써도 되고, 편한 프로그램을 써도 됩니다. 혹 손으로 그려도 상관없습니다. 그린 모델(시각화)을 보고 무엇을 알 수 있는지 설명하시오.

Question 3.

How could you know that your decision tree model is not over-fitted?

(Korean translation) 당신의 decision tree 모델이 overfitting 되지 않았다는 것을 어떻게 알 수 있나요?

2. kNN

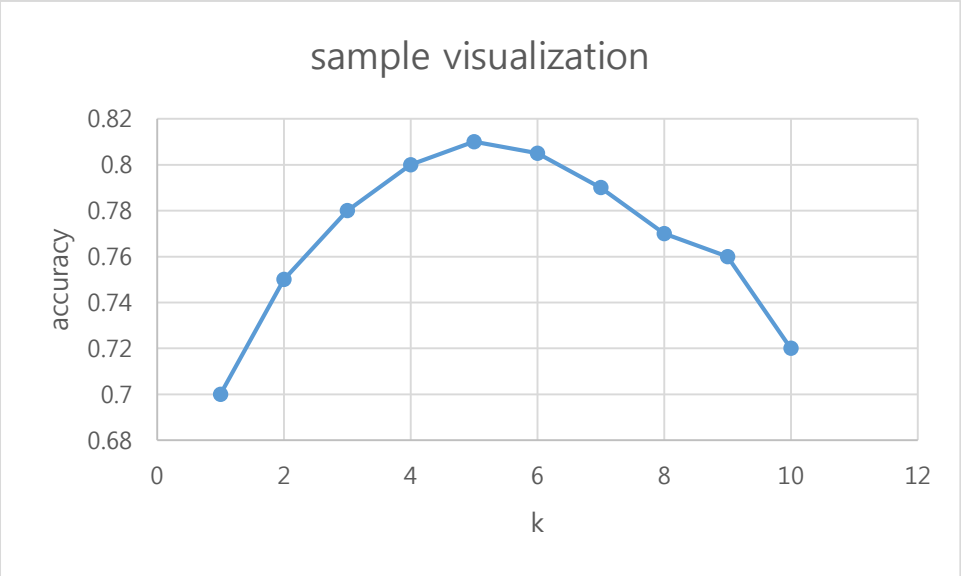
Question 4

Use kNN method to predict ***income*** variable in the test dataset. Explain confusion matrix, accuracy, precision, and recall of your model.

(Korean translation) kNN 방법으로 test data에 있는 ***income*** 변수를 예측하시오. confusion matrix, accuracy, precision, recall를 설명하시오.

Question 5

Try different k to find the best performing k option. Draws graphs show how AUC, accuracy, precision, and recall changes over k values. You can use R, Excel spread sheets, or any other SW to draw a graph. Or you can just hand-draw the graph as well. what is the best k?



(Korean translation) 다양한 k를 시도해보고 가장 성능이 좋은 k를 찾아보시오. k를 변화하면서 accuracy, precision, recall, AUC가 변화하는 것을 보기 위해 그래프를 그리시오. R이나 엑셀이나 어떤 SW 사용해도 되고, 손으로 그려도 됩니다. 가장 성능이 좋은 k는 무엇입니까(AUC 기준으로)?

Question 6.

Draw ROC curve and find the AUC when you use the "best k" for kNN that you find in Question 4.

(Korean translation) best K를 사용한 kNN에 대해서 ROC 커브를 그리고 AUC를 계산해보시오.

Question 7.

In order to apply kNN to the dataset, what kind of data transformation or preparation have you performed? Explain why and how.

(Korean translation) 데이터에 kNN을 적용하기 위해서 데이터를 어떻게 변형하였나요? 이유와 방법을 설명하시오.

Question 8.

Compare three modeling methods of decision tree, kNN, and Naïve Bayes. Which model outperforms others?

(Korean translation) 해당 데이터 셋에서 decision tree, kNN 두가지 모델을 비교했을 때 어떤 것이 성능이 더 좋은가요? 그 이유는 무엇이라고 생각하나요?

Quest 9.

여러분이 만든 모델의 성능을 여러분의 Group Member들과 비교해보세요. 차이점을 파악하고 여러분의 모델에서 개선해야할 점이 있다면 개선해봅시다.