**Data Mining, Final Exam**

**Your ID#: [ 2016270202 ]**

**Your Name: [ 강민규 ]**

|  |
| --- |
| **[Instruction] 안내**  **(1)** Please read and answer the questions carefully. Write your answers in Korean or English.  (문제를 잘 읽고 신중하게 답변하십시오. 영어 또는 한국어로 답안을 작성하세요).  **(2)** Please compress (.zip) and submit this report (.docx), and the entire R-code (.R file) you wrote into the BlackBoard.  (이 시험지와 답변에 사용된 R코드를 압축하여 블랙보드에 지정된 시간까지 반드시 제출하십시오.)  Midterm: 16:00 PM ~ 17:30 PM (90 Min.)  \*I highly recommend that you start submitting to the Blackboard at 17:25PM, at least 5 minutes before the end. If you do not submit by the given time, you will receive 0 points. (시험 끝나기 5분전인 5시 25분에는 블랙보드에 제출을 시작하길 권장합니다. 5시 30분에는 제출할 수 없도록 닫히며, 제출을 못했을 시 0점 처리됩니다.)  **(3)** The R-code should be executable when the TA runs. The submitted compressed file (.zip) must be named Final\_YourID\_Yourname.zip.  (R 코드는 TA가 돌렸을 시 깔끔하게 돌아가야 하며, 제출될 압축파일은 반드시 Final\_YourID\_Yourname.zip 로 명명하여 제출하십시오.)  **(4)** There is a partial score. Even if you can't resolve it completely, I hope you can try it as far as you can.  (부분 점수가 있으므로, 완벽하게 해결하지 못하더라도 할 수 있는 만큼 시도해 보길 바랍니다.)  **(5)** Only (Q4) is an open-ended question about theory, and the rest of the quiz includes the process of analyzing data through R programming.  (Q4만 이론에 관한 서술형 문제이며, 나머지 문제는 R프로그래밍을 통해 자료를 분석하는 실습 문제입니다.) |

**(Q1) Load Q1\_data.tsv** to model a KNN classifier (**k=7**) **without** cross-validation (\*Use all data as train and test sets) and **get a confusion matrix** for evaluating the model. \*\*\*If you start modeling without thinking, your computer will crash down. **(20 Points)**

*(\*\*\*Hint: Data Cleaning is not an “option”, it is a “necessity”. Before modeling, all values ​​must be numeric values.) mutate\_if() can be employed….*

*Data can be easily load to your R workspace as follows:*



**(Kor: Q1\_data.tsv 파일을 불러와서 KNN 분류기 (k=7)로 교차검증을 고려하지 않고 (모든 자료를 훈련 및 테스트 셋으로 사용하십시오) 모델링 하십시오. 그 후, 모델을 평가하기 위한 교차행렬을 완성하십시오. \*\*\*만약 생각 없이 모델링을 시작한다면 컴퓨터가 멈출 것입니다) \*힌트: 데이터 클렌징은 선택이 아닌 필수입니다. 당연하게도, 모델링전에 사용되는 모든 자료의 값은 실수 값이어야 합니다. mutate\_if() 가 도움이 될지도…)**

|  |
| --- |
| **[Result Screen-shot]**  **[Summary of Interpretation of Results and Solutions to Problems]** |

**(Q2)** Let's **implement genLTOCV(data) methods** to make it possible to perform modeling projects with leave-**two**-out cross validation (L**T**OCV). This method must return multiple folds (Train and Test sets, respectively) as list() format variable. After that, apply the data used in (Q1) **to compare** **3NN, 5NN, and 7NN models** **in terms of sensitivity**. Then answer the questions in the Answer sheet below. **(30 Points)**

**(Kor: 모델링 프로젝트에서 활용될 수 있는 leave-two-out cross validation (LTOCV) 기법의 genLTOCV(data) 함수를 구현하십시오. 이 함수는 과제와 마찬가지로 각 폴드의 훈련셋과 테스트셋을 묶어 리스트 형식으로 리턴해야 합니다. 그 후 (Q1)문제에 적용하여 3NN, 5NN, 7NN 을 서로 민감도 측면에서 비교하십시오. 마지막으로 아래의 답변 시트에 답변하십시오.)**

|  |
| --- |
| **[Q2-1: How many training sets** are generated in total?**]**  **(Kor: 몇 개의 훈련자료가 생성됩니까?)**  **[Q2-2: Which model** shows the best performance in terms of **sensitivity**?**]**  **(Kor: 민감도를 기준으로 봤을 때 어떤 모델이 최고의 성능을 보여줍니까?)**  **[Q2-3:** **What is the value of k** in k-fold cross validation **equal to** your implemented LTOCV algorithm?**]**  **(Kor: 본인이 구현한 LTOCV알고리즘은 k-FOLD Cross validation 에서 k가 몇일때와 동등합니까?)**  **[Result Screen-shot]** |

**(Q3)** After loading the given "Q3\_data.tsv", perform decision tree modeling using all samples without cross-validation. This material is a revised version of the grade results for students in data-related classes in the previous semester. Then answer the questions in the Answer sheet below. **(30 Points)**

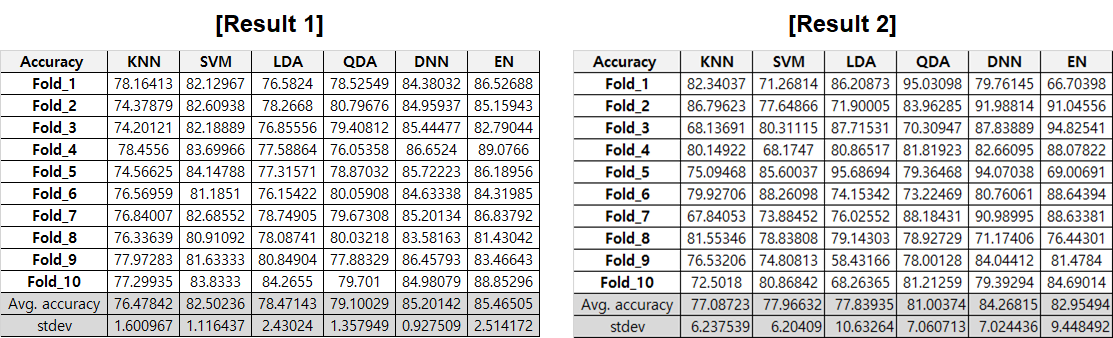
*(\*\*\*Hint: (1) It is normal when modeling is finished within 1 second. (2) As I always emphasize, all data used for modeling must be in numeric format except for class labels. (3) When using rpart.plot(), you must adjust it by referring to the manual, such as the "extra" option.)*

**(Kor: “Q3\_data.txv”자료를 불러온 뒤 의사결정나무 모델링을 교차 검증 없이 실시하십시오. 해당 자료는 직전학기 데이터 과학 관련 과목을 수강한 학생들의 성적을 수정한 자료입니다. 모델링 이후 아래의 답변 시트에 답변하십시오. (힌트: (1) 모델링은 1초만에 끝나야 정상입니다. (2) 항상 강조하지만 모든 자료는 반드시 숫자여야합니다. (3) rpart.plot()함수를 활용할 경우 “extra”등의 옵션을 매뉴얼에 따라 잘 조정해야 합니다.) 마지막으로 아래의 답변 시트에 답변하십시오.)**

|  |
| --- |
| **[Q3-1:** **Draw a decision tree plot** to figure out what is the decisive factor for given grades (A+, B+,..., F).**]**  **(Kor: 의사 결정나무 플랏을 그려서, 각 변수들이 최종성적에 기여하는 바에 대한 알고리즘을 모델링하십시오.)**    ##modeling  model <- rpart(Grade ~ ., data =trainData, method = "class")  rpart.plot(model,extra="auto" ,box.palette = "green")  **현재로써는 해당 코드로 실행했을 때 underfitting 된 모습을 학습이 잘 되지 않는 것으로 보입니다.**  **[Q3-2:** Can you predict your **expected grades (i.e. A+, B+, …, F)** in this course based on the data?**]**  \* Clearly sketch the modeling process to achieve that goal, and briefly state why you think so.  **(Kor: 이 자료를 바탕으로, 본 수업 본인의 최종학점 (A+, B+, …, F)을 예측할 수 있습니까? 만약 그렇다면, 해당 목적을 이루기 위한 전체적인 모델링 과정을 명확하게 스케치하고, 왜 그렇게 생각하는지에 대해서 서술하십시오.)**  **이 자료를 바탕으로는 예측할 수 없습니다. 데이터를 눈으로 봤을때는 final 점수가 높을수록 학점이 높아보이나 의사 결정트리로는 잡아내지 못하는 듯 합니다. 다른 linear model을 쓰면 조금 나은 결과를 얻을 수 있을 듯 합니다.**  **[Q3-3:** Can you predict your expected **final exam score** in this course based on the data?**]**  \* Clearly sketch the modeling process to achieve that goal, and briefly state why you think so.  **(Kor: 이 자료를 바탕으로, 본 수업 본인의 기말고사 성적을 예측할 수 있습니까? 만약 그렇다면, 해당 목적을 이루기 위한 전체적인 모델링 과정을 명확하게 스케치하고, 왜 그렇게 생각하는지에 대해서 서술하십시오.)**  **마찬가지로 현재 모델로는 불가능하고 final 을 결과값으로 두고 나머지 hw 변수들을 바탕으로 기말고사 점수 값을 예측하는 linear regression 모델 정도만 만들어도 예측 가능 할 듯합니다.**  **[Q3-4:** Assume that there is a scholar who argues that **cross-validation is not necessary when modeling only for generating decision-making algorithms** on data, as in (Q3-1). Do you agree with this opinion?**]**  \*Be clear about why you think so.  **(Kor: (Q3-1)과 같이 데이터의 의사 결정 알고리즘 생성만을 위해 모델링을 할 경우 교차 검증이 필요하지 않다고 주장하는 학자가 있다고 가정합시다. 이 주장에 동의합니까? 그 이유를 명확히 서술하십시오).**  **데이터의 구성에 따라 예측결과와 또 의사 결정트리의 구조가 달라지는데 한 번의 검증으로 검증을 하는 것은 동의 할 수 없는 의견입니다.** |

**(Q4)** The *[Result 1]* and *[Result 2]* tables at the bottom are the results of two other 10-fold cross validations on the same data. **Make all possible interpretations** based on this result. **(20 Points)**

**(Kor: 하단의 [결과 1]과 [결과 2]는 동일한 자료에 대해서 두 번의 다른 10-fold 교차 검증을 실시한 결과입니다. 이를 바탕으로 가능한 해석을 모두 이끌어내십시오).**

****

|  |
| --- |
| **[Interpretation] \*within half a page** |

**-At the end-** Well done for one semester!