Final Exam

Introduction to Data Science

2018-06-13

7:00 ~ 10:00 PM

You need to turn in an RData file named "st21100123.RData – st+your student ID" that contains result of following two questions. Explain how you build your prediction model in your reporting document (MS word, HWP, PDF, or any other text format) and submit your document r script file as well. You can choose to write your report in your proficient language from English and Korean. Any of your missing submission or incorrect form of submission may penalize your work.

You must submit your work to Edmodo in time (~10:00PM). The submission page will be locked after 10PM. Only the work that are submitted to Edmodo successfully in time will be graded. You can submit your work several times during the exam (7PM ~ 10PM) if necessary.

(Korean Translation) 아래 2개의 문제를 풀어 결과를 RData 파일에 담아서 제출하시오. RData 파일은 st학번.Rdata 형식으로 제출해야합니다. 예측 모델을 만드는 과정 및 결과를 보고서에 작성해서 제출하고, r script 파일 또한 제출해야 합니다. 제출 형식이 틀리거나 빠진 제출물이 있는 경우 감점이 있을 수 있습니다.

(주의) 10시까지가 시험이고 이 시간을 지나면 (10시 정각이 되면) 답안을 제출하지 못하게 됩니다. 답안이 제출되지 않은 경우 자동 0점 처리되고, 어떤 예외도 허용되지 않기 때문에 제시간에 제출하는 것을 주의하세요. 한번 제출한 이후에 시간 내라면 여러 번 제출하는 것은 가능합니다.

Problem 1

Learn your best classification model to estimate probability that the client subscribe a bank term deposit (variable y = 'yes'). The dataset contains following variables:

(Korean Translation) 은행 고객이 정기 예금 구좌를 개설할 확률(P(y = 'yes'))를 예측하는 classification model을 만드시오. 아래는 변수설명.

Input variables:

bank client data:

1 - age (numeric)

2 - job : type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")

3 - marital : marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)

4 - education (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")

5 - default: has credit in default? (categorical: "no","yes","unknown")

6 - housing: has housing loan? (categorical: "no","yes","unknown")

7 - loan: has personal loan? (categorical: "no","yes","unknown")

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: "cellular","telephone")

9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: "failure", "nonexistent",

"success")

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: "yes","no")

Once you obtain your best classification model based on the training dataset (**bank.train**), save your prediction result on the test dataset (**bank.test.nolabel**) with variable name "**pred_c**". Please make sure that your **pred_c** variables is a numeric vector that contains 8200 predicted **probability** to be **y** variable = "**yes**" for the client in the test dataset. Put your vector variable into your RData file and submit to the Edmodo.

Your classification model will be evaluated based on AUC(area under curve) value.

(Korean Translation) classification model을 만든 후에는 test 데이터에 대한 예측 결과를 pred_c라는 변수에 저장하여 제출하시오. pred_c는 y변수가 yes일 **확률**을 저장한 numeric vector여야 합니다. pred_c 변수를 RData에 넣어서 에드모도에 제출하시오. pred_c는 값이 8200개 들어있어야 합니다. 예측 모델의 성능은 AUC 값으로 평가될 것입니다.

Problem 2

The goal of this problem is build a model that predict final grade of students who take a course of Math and Portuguese. The dataset contains following variables:

(Korean Translation) 수학과 포르투갈어 과목을 들은 학생들의 최종 성적(G3)를 예측하는 모델을 만드시오. 아래는 변수 설명

Variables

1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)

2 sex - student's sex (binary: "F" - female or "M" - male)

3 age - student's age (numeric: from 15 to 22)

4 address - student's home address type (binary: "U" - urban or "R" - rural)

5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)

6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)

7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")

12 guardian - student's guardian (nominal: "mother", "father" or "other")

13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

16 schoolsup - extra educational support (binary: yes or no)

17 famsup - family educational support (binary: yes or no)

18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)

19 activities - extra-curricular activities (binary: yes or no)

20 nursery - attended nursery school (binary: yes or no)

21 higher - wants to take higher education (binary: yes or no)

22 internet - Internet access at home (binary: yes or no)

23 romantic - with a romantic relationship (binary: yes or no)

24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

31 class – course subject: Math or Portuguese

these grades are related with the course subject, Math or Portuguese:

32 G3 - final grade (numeric: from 0 to 20, output target)

Once you obtain your best regression model based on the training dataset (**student.train**), save your prediction result on the test dataset (**student.test.nolabel**) with variable name "**pred_r**". Please make sure that your **pred_r** variables is a numeric vector that contains 183 predicted grade for the student in the test dataset. Put your vector variable into your RData file and submit to the Edmodo.

Your regression model will be evaluated based on R squared.

(Korean Translation) regression model을 만든 후에는 test 데이터에 대한 예측 결과를 pred_r이라는 변수에 저장하여 제출하시오. pred_r는 예측된 성적 값을 가지는 numeric vector여야 합니다.

pred_r 변수를 RData에 포함해서 에드모도에 제출하시오. pred_r는 값이 183개 들어있어야 합니다. 예측 모델의 성능은 R squared(R^2) 값으로 평가될 것입니다.