Data Science Final Exam

2020-06-16 7:00 ~ 10:00 PM

In the final exam, you will be working with two different dataset, "online shopper intention" and "real estate valuation" in "**final_test.RData**" file. "online shopper intention" dataset is for classification problem and "real estate valuation" dataset is for regression problem.

After obtaining your best classification and regression model, you need to include their prediction result of 3 different vectors in an RData file named "st21100123.RData – st+your student ID".

You should turn in your RData file on the designated website:

http://jict.handong.edu:3838/ds_final_2020/

and your record will be listed on the chart. You can register your record on the chart **until 10:00 PM**. After that, you can only check your AUC, Balanced Accuracy, RMSE, R^2. The record listed on the chart is mainly considered to evaluate your work in this test.

You should turn in your **reporting document until 12:00 AM** that you have 2 more hours to summarize your work. Turn in your report and RData together to Edmodo. (no need for R script file)

In the report, describe your classification and regression model that yield your final result, briefly. The report with the best record will be shared with your classmate, honorably, to inspire others.

Again Rdata can be registered until 10 PM and the ranking chart will be locked.

기말시험에서는 "**final_test.RData**"에 포함된 2종류의 dataset으로 문제를 풉니다. dataset 하나는 classification, dataset하나는 regression 문제입니다. 최선의 예측모델을 만들어서, test data에 대한 예측결과를 3개의 vector에 담아, RData 로 제출하시오. RData 파일 이름은 "st21100123.RData – st+your student ID"으로 합니다. (형식이 틀리면 제출불가)

RData 파일은 위 링크의 웹사이트에 등록하면 여러분의 기록이 저장됩니다. 웹사이트의 기록을 **등록하는 것은 10시까지만 가능**하며, 10시가 지나면 기록을 확인할 수 있지만 등록할 순 없습니다. 기말시험의 성적은 웹사이트에 등록된 기록들을 중심으로 결정됩니다.

보고서는 12시까지 제출로서, 10시가 지나도 2시간 정도 넉넉하게 여러분이 한 것을 정리할 수 있습니다. 에드모도에 제출하면 됩니다. 보고서는 여러분의 최종모델에 대한 설명을 간단하게 하면 됩니다. 가장 성능이 좋았던 학생의 모델+보고서는 동료들도 함께 배울 수 있도록 공유될 것입니다. **시간 제한을 명심하고 제출 못하는 불상사가 없도록 해주세요 (late 제출받지 않음)**

Problem 1 (Classification) 50pts

The task is to predict behavior of online shopper whether they make some purchase online (positive samples) or spend no money for the session. Each observation represents a session which is a user's visit of e-commerce website. Each session belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period. The dataset consists of 10 numerical and 8 categorical attributes. The **'Revenue'** is the outcome variable we attempt to predict.

Feature	Feature Description
Administrative	Number of pages visited by the visitor about account management
Administrative duration	Total amount of time (in seconds) spent by the visitor on account management related pages
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site
Informational duration	Total amount of time (in seconds) spent by the visitor on informational pages
Product related	Number of pages visited by visitor about product related pages
Product related duration	Total amount of time (in seconds) spent by the visitor on product related pages
Bounce rate	Average bounce rate value of the pages visited by the visitor
Exit rate	Average exit rate value of the pages visited by the visitor
Page value	Average page value of the pages visited by the visitor
Special day	Closeness of the site visiting time to a special day
OperatingSystems	Operating system of the visitor (8 categories)
Browser	Browser of the visitor (13 categories)
Region	Geographic region from which the session has been started by the visitor (9 categories)
TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct) (20 categories)
VisitorType	Visitor type as "New Visitor," "Returning Visitor," and "Other" (3 categories)
Weekend	Boolean value indicating whether the date of the visit is weekend
Month	Month value of the visit date
Revenue	Class label indicating whether the visit has been finalized with a transaction

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site

The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session.

The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

The dataset for training model is named "shopping_train" and "shopping_test" is for testing with no revenue variable. You need to make 2 vectors for estimated probability of being positive (revenue = true) and predicted result of TRUE/FALSE logical vector. Name the vectors "**prob_shopping_test**" and "**pred_shopping_test**" and include them in the RData file. Your record will be evaluated in terms of **AUC** and **Balanced Accuracy**. Balanced Accuracy is calculated as follows:

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

where P and N are the number of positive samples and the number of negative samples, respectively.

중요한 내용만 한글로 적으니 자세한 내용은 위의 설명을 참조하세요. 이 문제는 온라인 쇼핑몰 이용자가 구매를 할 것인지 하지 않을 것인지를 예측하는 모델을 만드는 것입니다. revenue 변수가 구매 여부를 표시합니다. 각 변수에 대한 자세한 설명은 위를 참고하세요. 데이터 프레임의 한 행은 사용자가 한번 웹사이트에 접속해서 접속을 끊고 나갈 때까지의 한 session의 기록이며, 각 session은 서로 다른 이용자입니다. 1년 동안의 기록을 수집한 것입니다.

학습 데이터는 "shopping_train"이며, 테스트 데이터는 "shopping_test"인데 revenue변수를 포함

하고 있지 않습니다. "**prob_shopping_test**"는 사용자의 구매 예측 확률 값을 저장한 numeric vector, "**pred_shopping_test**"는 사용자의 구매 여부 예측(TRUE/FALSE)를 저장한 logical vector로 RData에 저장하여 제출하여야 합니다. 여러분의 예측 결과는 AUC값과 Balanced Accuracy로 평가됩니다. Balanced Accuracy에 대한 자세한 설명은 위를 참조하세요.

Problem 2 (Regression) 50pts

The second task is to build a regression model to estimate the price of real estate in Taiwan. The dataset is collected from Sindian Dist., New Taipei City, Taiwan. The input variables are given as follows:

X1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)

X2=the house age (unit: year)

X3=the distance to the nearest MRT station (unit: meter)

X4=the number of convenience stores in the living circle on foot (integer)

X5=the geographic coordinate, latitude. (unit: degree)

X6=the geographic coordinate, longitude. (unit: degree)

The outcome variables is named Y that implies house price of unit area (10,000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared)

TIP 1: X1 the transaction date is not a single numeric variable but combination of two variable year and month. So you need to find a way to make use of the X1 in you model.

X1은 단일 변수가 아니라, 연도와 월이 합쳐진 변수입니다. 모델에 사용하기 위해서는 적절한 변환이 필요하겠지요.

TIP 2: X5 and X6 indicates geographic coordinate which is not straightforward to use in regression model. They can be well-fit to Euclidean distance and methods using Euclidean distance. Furthermore, it is also possible to introduce more variables based on geographic location such as distance to downtown, distance to highway, and distance to shopping mall, etc. (the location of those places you can find from web.)

위도 경도 정보 그 자체는 모델에 사용하기가 적절하지 않을 수 있습니다. 하지만 Euclidean distance와 아주 잘맞을 수 있고, Euclidean distance를 사용하는 모델에 잘 맞을 수 있겠지요. 또한 위치정보는 근처에 있는 주요 장소(시내중심가, 고속도로, 쇼핑몰 등)과의 거리를 계산하여 새로운 변수를 추가하는 것도 가능합니다. (다른 주요 장소의 위도 경도는 인터넷에서 찾을 수 있겠지요)

The dataset for training model is named "housing_train" and "housing_test" is for testing with no Y variable. You need to make a numeric vector of predicted housing price. Name the vector "**pred_housing_test**" and include it in the RData file. Your record will be evaluated in terms of **RMSE**.

중요한 내용만 적으니, 자세한 내용은 위를 참조하세요. 이 문제는 대만의 부동산 가격을 예측하는 regression model을 만드는 것입니다. 입력 변수는 6개로 위의 변수 설명을 참고해주세요. (팁도 읽어보세요)

예측하고자 하는 변수는 Y로 부동산의 평당 가격입니다. (1평 = 3.3 meter^2). 가격 단위는 10,000 대만 달러입니다. 모델로부터 예측한 결과는 벡터 "**pred_housing_test"**로 만들어 RData파일에 포함하여 제출하세요. 여러분의 결과는 RMSE를 기준으로 평가될 것입니다.