موسسه آموزش عالی آزاد توسعه

برگزار کننده دوره‌های تخصصی علم داده

Homework 4: Winter 2019

Due date: ۳۰ دی

Please email your HWs to y.zerehsaz@gmail.com
Please append all your codes to your response
*****Please hand in your HWs as a word file with your email as the document's name. For instance, I would name my word file as y.zerehsaz@gmail.com.docx.**
*****Make sure to copy and paste the codes that you used for each question. I need to see your plots, results and conclusions but not the long output of your codes.**
*****When asked, please explain your results.**

This HW must be completed in RStudio **not Python**.

Consider the "bank" dataset to answer questions A~E.

In this dataset, we would like to build a predictive model for the bank data. Follow to steps to build this model and test its efficiency.

A) First and foremost, we need to divide our data to test and train datasets.

1. Read your dataset and place it in a variable called "d".
2. Now, we need to select 20% of the dataset, and leave it aside as the test dataset. Selecting 20% means choosing 20% of the rows of the data and pulling them out. To do this, we use the "sample" function to generate numbers between 1 and the total number of rows in our dataset.

```
s=sample(nrow(d),floor(0.2*nrow(d)),replace=F)
dtest=d[s,]
dtrain=d[-s,]
```

3. You should now use dtrain data to fit your model, and then test its accuracy using the dtest data. So, use the backward stepwise regression to predict the variable "deposit". Regress this variable on all variables except the variable "duration" and call this model stb. As discussed in class, we are excluding duration since it is not possible to build a predictive model using this variable. (This is true if we assume that every person decides to make a deposit or not right after the call). **Also, remember that the "day" variable in the dataset is a categorical variable, but it is recorded as a numerical variable. Hence, you must change it to a categorical variable before fitting the model.** (use the factor function)

B) Interpret your coefficient estimates using odds ratio concepts (it is enough to interpret two of the coefficients).

C) After fitting the model in part A, you can obtain the estimated probabilities of making a deposit using the "predict" function.
Hint:
phat=predict(stb,dtest,type='response')# this gives you the estimated probabilities of making deposits. So, phat is a vector with 2125 elements.

D) The class of each observation must be determined based on phat. If the estimated probability of making a deposit for an observation is larger than 0.5, then we classify the observation to the class "yes". Otherwise, we decide that the subject will not make a deposit (its class will be "no"). Use phat to obtain the class for each of the observations in dtest data and call the estimated classes yhat. Note that yhat should be a vector with the same length as phat, and it must show the classes for all subjects in the dtest data. The elements of yhat must be either "yes" or "no".

E) The final step is to compute the testing prediction accuracy. To do this, we need to compare yhat to the true response variable in the test dataset. So, we need to see in how many cases yhat elements are equal to those in dtest$deposit variable, and then divide this by the number of rows in the test dataset.
Hint: use yhat==dtest$deposit