Data Mining
Georgia State University
Professor: Jaya Krishna Mandivarapu

Home Work 3

Due Date: Oct 26th 2020 at 11.59 PM

Submission Requirements

You must turn work at the SPECIFIED TIME so you can receive credit for Homework!

You can use HW2 jupyter notebook as an initial sample template

Files Required for submission : One Jupyter Notebook and HTML file (Can be download from Jupyter notebook you are working with)

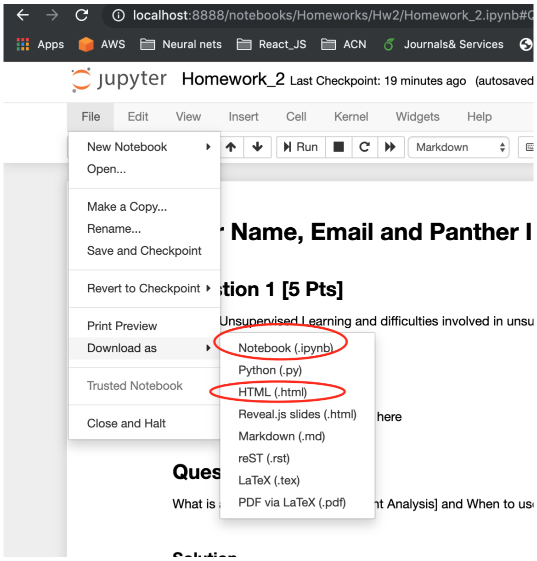


Figure 1: Download as Jupyter Notebook and HTML file

Homework 2 must be **submitted on icollege** by the due date and time. Late homework will be subject to a penalty of 50 percent for 1 day and 80 percent for two days and after 3 days no submission allowed, as stated in the course grading policy. No email or hard copies of homework will be accepted.

You may discuss the assignments with other students in the class, but (as stated in the academic honesty policy) your written answers **must be your own**, and you must list the names of other students you discussed the assignment with.

How to Submit

Log into **iCollege(iCollege)**, select the class to view its drop box folders, select the correct folder for the given assignment and upload the file there.

You will get a confirmation email. Please save the conformation email in the event something goes wrong, for example work was submitted to the wrong folder etc..

Data Mining
Georgia State University
Professor: Jaya Krishna Mandivarapu **Home Work 3**

1. Answer the following? [50 pts]

a) Write and explain about the Linear Regression and it's equation [3 pts]

b) Explain in detail about the loss function of linear regression, R^2, Adjusted R^2 used in the Linear Regression and what is the need for Adjusted R^2? [12 pts]

c) Plot X vs Y in Scatter plot from data in Table 1 and comment on the relation of X vs Y using Covariance,Corelation. Please comment on Covariance and corelation values [5 Pts]

d) Perform Linear regression on the following data using Python? and print β_0, β_1 values in equation y= β_0+ β_1*x. Please write down what is your understanding from those values. [10 Pts]

e) What are different evaluation metrics available for predicting the performance of the Linear Regression? Evaluate all those methods on the given dataset in Table 1 and also please print out the accuracy, R^2, Adjusted R^2 [10 pts]

f) Print ANOVA (Analysis of Variance) table and Parameter Estimates for the given data and explain your understanding of all the variables present in the table.[See hints and explanation for what I am looking for] [10 pts]

X	Y
6	526
3	421
6	581
9	630
3	412
9	560
6	434
3	443
9	590
6	570
3	346
9	672

Table 1: X(No of Weeks) vs Y(Avg Sales)

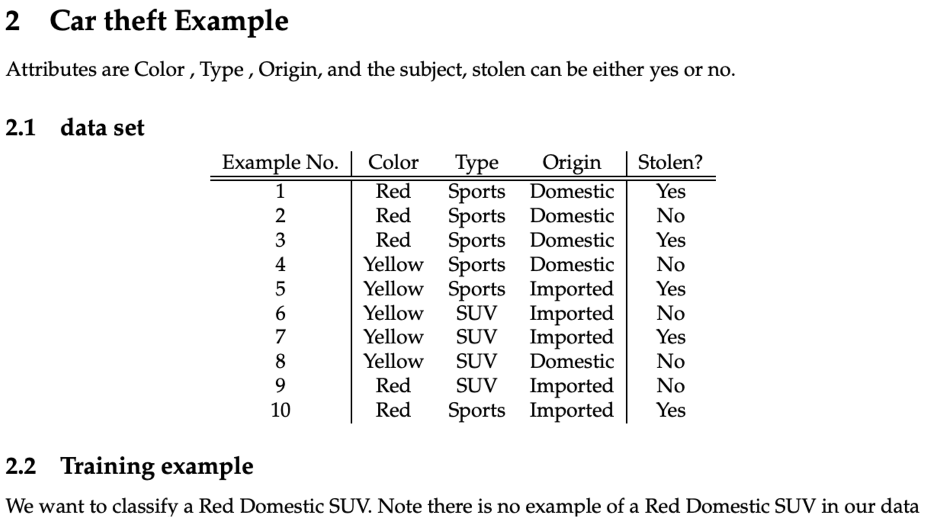
2. Answer the following [30 Pts]

a) What is Conditional probability, Marginal probability and Joint probability? Write their mathematical formulas and give one example each. [5 pts]

b) Explain what is Baye's rule with the formula and what is prior, posterior, likelihood and marginal probability in the Baye's rule. [10 pts]

c) What is Naive Bayes algorithm and how is related or derived or inspired from Bayes rule? [5 pts]

d) Perfom Naive Bayes algorithm on the below dataset in python in which you can classify wheather a **Red Domestic SUV** is stolen or not as shown in 2.2. [10 pts]



2 Car theft Example

Attributes are Color , Type , Origin, and the subject, stolen can be either yes or no.

2.1 data set

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

2.2 Training example

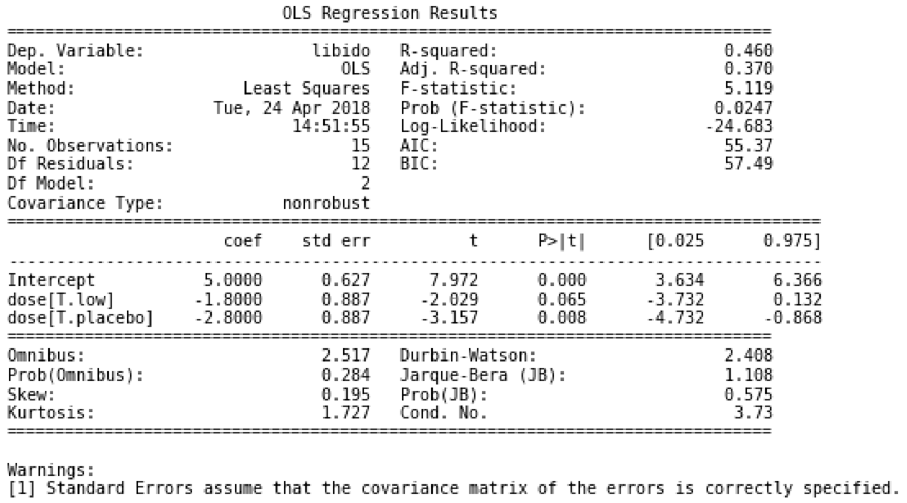
We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data

Hints and Explanation:

convert 1(Table 1),2.1 into a csv then load a csv file or you can prepare your dataframe from 2.1 dataset. [As i posted the text for 2.1,Table 1]

Then Apply both algorithms one after the other then plot the outputs. For question one regression line and question 2 the output predictions

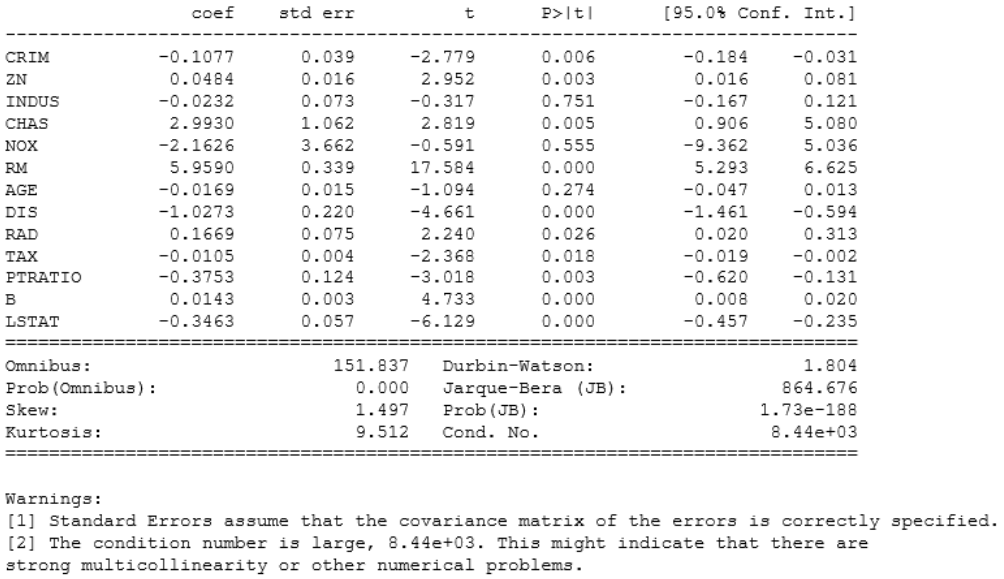
You can use either use sklearn packages or write your own code to do the questions.
ANOVA TABLE EXAMPLE



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 libido   R-squared:                       0.460
Model:                            OLS   Adj. R-squared:                  0.370
Method:                 Least Squares   F-statistic:                     5.119
Date:                Tue, 24 Apr 2018   Prob (F-statistic):             0.0247
Time:                        14:51:55   Log-Likelihood:                -24.683
No. Observations:                  15   AIC:                             55.37
Df Residuals:                      12   BIC:                             57.49
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept           5.0000      0.627      7.972      0.000       3.634       6.366
dose[T.low]        -1.8000      0.887     -2.029      0.065      -3.732       0.132
dose[T.placebo]    -2.8000      0.887     -3.157      0.008      -4.732      -0.868
==============================================================================
Omnibus:                        2.517   Durbin-Watson:                   2.408
Prob(Omnibus):                  0.284   Jarque-Bera (JB):                1.108
Skew:                           0.195   Prob(JB):                        0.575
Kurtosis:                       1.727   Cond. No.                         3.73
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

PARAMETER ESTIMATION



```
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
CRIM          -0.1077      0.039     -2.779      0.006        -0.184    -0.031
ZN             0.0484      0.016      2.952      0.003         0.016     0.081
INDUS         -0.0232      0.073     -0.317      0.751        -0.167     0.121
CHAS           2.9930      1.062      2.819      0.005         0.906     5.080
NOX           -2.1626      3.662     -0.591      0.555        -9.362     5.036
RM             5.9590      0.339     17.584      0.000         5.293     6.625
AGE           -0.0169      0.015     -1.094      0.274        -0.047     0.013
DIS           -1.0273      0.220     -4.661      0.000        -1.461    -0.594
RAD            0.1669      0.075      2.240      0.026         0.020     0.313
TAX           -0.0105      0.004     -2.368      0.018        -0.019    -0.002
PTRATIO       -0.3753      0.124     -3.018      0.003        -0.620    -0.131
B              0.0143      0.003      4.733      0.000         0.008     0.020
LSTAT         -0.3463      0.057     -6.129      0.000        -0.457    -0.235
==============================================================================
Omnibus:                      151.837   Durbin-Watson:                   1.804
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              864.676
Skew:                           1.497   Prob(JB):                    1.73e-188
Kurtosis:                       9.512   Cond. No.                     8.44e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.44e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```