1. Read data first

code for read the data first: theURL="https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data"

data=read.table(theURL, header=FALSE)

2. Data cleaning

1) What’s the percentage of ‘question marks’?

2) If less than 5 %, we can just discard those cells

3. Use the first half for the training data(1:16281)

4. Exploratory data analysis with the training data

5. Explore application of various classification methods, models and algorithms that we had studied in class. Find the test errors and training errors.

Linear Regression

Logistic regression

QDA, LDA, Combination of both

Ridge, Lasso, Combination of both

KNN

PCA (Dimension reduction)

6. Write a final report (make a coherent story.)

Note: members and crosscheck work later. We will be using github for version control and as soon as we have all your aliases, we will give you permission for the appropriate repository. We will later check your usage of it, we suggest you keep all your work in it, including all intermediate steps.

- No more than 8 pages (including code and graphs) -if it doesn't fit, use smaller graphs or par mfrow

- only code essential in showing how you achieved your solution should be included in the report

- graphs should have title, meaningful axis labels, and captions but try to make the graphs as self explanatory as possible

- points will be deducted for poor formatting and convoluted explanations (it's a report, try to demonstrate the concepts you are working with as clearly as possible)

Divide Task

Tuesday: Data Cleaning - Ke, Li, Shilun- meeting at 9 pm after TSA mid term

(Everyone makes sure everyone knows how to github)

Wed: Try different models - Everyone - meeting at 9pm

Thurs: Graphs

Friday: Write report, Latex - Rebecca, Jiun