Assignment 2/ Lab 2 – Part 2

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

February 13th, 2020

Lab2 Part 2 consists of

- Regression
  - New multivariate dataset
- kNN
  - New Abalone dataset
- Kmeans
  - Iris dataset
- Need all three for Assignment 2

Go over the lecture in-class exercises first, it will help/guide you to do the Lab2 part 2

The Dataset(s)

- http://aquarius.tw.rpi.edu/html/DA

- Two new ones; dataset_multipleRegression.csv, abalone.csv

- And …. Visit this link:
- http://aquarius.tw.rpi.edu/html/DA/group1

- Code fragments, i.e. **they will not run as-is, on the following slides as Lab2_knn1.R, etc.**

Exercise1: Regression

- Retrieve this dataset: dataset_multipleRegression.csv
- Using the unemployment rate (UNEM) and number of spring high school graduates (HGRAD), predict the fall enrollment (ROLL) for this year by knowing that UNEM=7% and HGRAD=90,000.
- Repeat and add per capita income (INC) to the model. Predict ROLL if INC=$25,000

Exercise 2: Classification

- Retrieve the abalone.csv dataset
- Predicting the age of abalone from physical measurements.
- The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope: a boring and time-consuming task.
- Other measurements, which are easier to obtain, are used to predict the age.
- Perform knn classification to get predictors for Age (Rings). Interpretation not required.

Exercise 3: Clustering

- The Iris dataset (in R use data(“iris”) to load it)
- The 5th column is the species and you want to find how many clusters without using that information
- Create a new data frame and remove the fifth column
- Apply kmeans (you choose k) with 1000 iterations
- Use table(iris[,5],<your clustering>) to assess your results

- Due: **Monday, 17th February 2020 by 11:59pm on LMS**
- Include both Lab 2 Part 1 & Part 2
- Make sure to include Part1 and Part 2 in same document and name it as Lab2_Assignment2
- “YourName_RCSID_Lab2_Assignment2”

REMINDER: NASA IMPACT talk Today

- NASA Research Scientists will be talking to the Data Analytics students to help with their term project tomorrow (Those who are planning to use the NASA dataset for the Data Analytics term project)
- NASA Scientists are willing to mentor the students who are going to use the provided NASA datasets during the Data Analytics term project (*NASA internship opportunities!!!*)
- **Meeting Time: 2 pm - 2:30 pm**
- **Location: Amos Eaton building room 215**
- **Date: 02/13/2020 (Thursday)**
- Please come to Amos Eaton room 215 by 1:55 pm tomorrow (02/13/2020, Thursday) The meeting starts at 2 pm.
- ***(Please come a little early before we start the meeting***).

Project Dataset Inspection:

- Project idea(s) and dataset inspection on Monday, 17th 2020 during the class time, One-on-One with the instructor.
- We will go over your dataset during this time.
- You need to provide the URL of the dataset (if the dataset is obtained from Web) that you are planning to use for the project.