# CSE4/587 Data-intensive Computing Spring 2017

## LAB3: ALGORITHMS AND MODELS FOR DATA ANALYSIS, LEARNING AND PREDICTION: B. RAMAMURTHY

### OVERVIEW:

The hands-on practical learning components of the course comprises two types of activities: labs covering one or two knowledge units (skills, competencies) of data-intensive computing and a single term project serving as a capstone covering the entire data pipeline. General pedagogical pattern for the labs is: one or two vignettes for learning the main concepts (or theme) of the lab, followed by 2 or 3 activities that apply the concepts.

In Lab1 we wrote a data client and very simple information server. In Lab2 we worked on data cleaning and data munging. In this lab (Lab 3) will apply machine learning algorithms and statistical models to data with the ultimate goal of being able to predict the outcome for a certain query or classify data according to certain criteria. More specifically, we will explore algorithms discussed in Chapter 3 of Doing Data Science textbook [1]: linear regression, k-nearest neighbors, k-means.

## GOALS:

Major goals of the lab3 are to:

1. **Decide on the algorithms** that will be used in solving the problem based on the data characteristics and the attributes of the problems.
2. **Learn to apply algorithms**: linear regression, K-NN and K-means, which algorithm to use and why and when.
3. **Make valid and reasonable assumptions** about the variables (features) of a problem, goodness of a model fit, metrics for evaluation of errors, outliers, scaling, and choosing appropriate data ranges for the computation.
4. **Choose the parameters for prediction** based on experimentation (repeated trials) and acceptable values of error rates. Document the experimentation process and the rationale.
5. **Plot the outcomes** for easy visualization of data analysis.
6. **Interpret the results** to enable decision making.

## OBJECTIVES:

The lab goals will be accomplished through these specific objectives:

1. You will be working in R language environment: Jupyter or R Studio.
2. You will follow the pedagogical pattern: (i) work with classical R data package (ii) a R vignette to understand how to interpret results for the algorithms we will be working with (iii) three original problems one for each of the three algorithms: lm, K-nn, and k-means.

# LAB DESCRIPTION:

Introduction: An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov) have their data available to the public for various purposes. Social network applications such as Twitter and Facebook collect enormous amount of data contributed by their numerous and prolific user. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make subset of their data available for use by registered users for free. Some of them as downloadable data files (.csv, .xlsx) as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as a web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrap the data from these web pages and extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products. Very often data collected is not in the format required for the downstream processes such as EDA, analysis, prediction and visualization. The data needs to be cleaned, curated and munged before modeling and algorithmic processing.

In Lab1 we acquired data from twitter using its REST API and processed it. In Lab2 we prepared the data for (i) question answering (ii) change format to accommodate EDA (iii) understand data by plotting it, and then standardized and attempted to normalize data for supporting further analysis. In Lab3 we will analyze the data with specific algorithms (linear regression, k-nn, k-means), evaluate the error rates of the model (fit) and interpret the results.

Also we will follow the pedagogical pattern

- Preparation before lab (pre-lab)
- One or more R vignette on a specific concept using classic package and/or classic data
- One or two simple activities related to core topic of the lab (in Lab3 it is data analysis)
- Visualize the outcome of the analysis
- Select the parameters for prediction
- Interpret and document the results.

PREPARATION: Here are the preliminary requirements for the lab. **Time needed: 3 to 4 hours (Day 1)**

1. Work environment: You can work in our "Learning Environment" Jupyter [12] or on "Development Environment" in RStudio for R language [13]; (Production Environment will employ a robust programming language such as Java or C++).
2. Make sure you have the data sets (see links [2], [3]): (i) NHL top 100 players obtained from [2] (ii) German credit data from Kaggle [3] and (iii) Data set from Pew Research Center [4] that we worked with in Lab2.
3. Read Chapter 3 and understand the three algorithms described in that chapter.
4. Read about the person who created ggplot2: Hadley Wickham here [5].

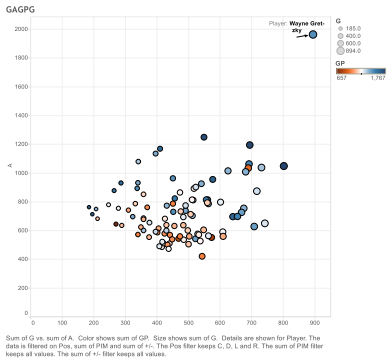
# LAB 3: WHAT TO DO?

1. (5 points) **Classical package**: **H. Wickham's ggplot2 *Vignette* (Time needed 2-3 hours: Day 2)**
   The R package ggplot2 by Hadley Wickham [6] and illustrated through a tutorial from Harvard [7]. Why? You can use ggplot2 for elegantly visualizing your lm, k-means and k-nn results. This tutorial has details on visualizing the output of these methods. **Submit the source code only**: R code is good enough.

2. (5 points) **Classical data**: **Edgar Anderson's Iris data ( Time needed: 2-3 hours: Day 3)**
   Iris data is a famous data in R and we will understand it through a worked out example that is provided in [8]. We will work on a vignette that analysis this data using k-means clustering that we will use later. Submit the source code. **Input**: Iris data **Output**: iris clusters **Process**: K-means as explained in the vignette.

3. (5 points) **Clustering around multiple dimensions: (Time needed: 2-3 hours: Day 4)**
   Here we will be working with real customer data from Berkley UCI Machine Learning Repository [9]. This analysis uses k-means clustering around 5 variables. Pay particular attention to how they handle the outliers and how they interpret the output clusters. **Input:** UCI ML data about customer grocery purchases. **Output**: Clusters around various items bought **Process**: K-means clustering.

4. (15 points) **Featured Activity 1: Linear Model. (2-4 hours each day: Day 5,6)**
   In this activity we will work with one more sports: National Hockey League data. We will work with NHL Top 100 players (not including the recent ones: we will do that). There are many interesting observations. The two variables we will relate are Goals and Assists that each of these players. Fit a linear (regression) model to this data visualized below. Assists on Y-axis and Goals on X-axis. Data is available on UBbox [2].




   a. After the initial fitting, evaluate the fit by noting the $R^2$ and p values of the fit.
   b. Change the model so that it line is forced through Wayne Gretzky. Note the $R^2$ and p values of the fit. This will another line or if you prefer another plot.
   c. Now add the data for another player Patrick Kane (he is a current player). You need to get the data for form online sources. Create a new chart with this addition and force the line through Kane but not Gretzky. Note $R^2$ and p values of the fit.
   d. Make all the model go through (0,0) and create new models. Note $R^2$ and p values of the fit.
   e. Create a table of model names (reference), $R^2$ and p values of the fit. Interpret your results.

**Input**: NHL Top 100 players **Output**: Fitted Linear models, **Process**: "lm" model of R, summary of metrics in a table and interpretation of the results.
**Get the data from**: https://buffalo.box.com/s/jm1hki9hbnlm4kzfkaimkzilys3dm8dg

5. (10 points) **Activity 2: K-nn classification with credit data (Time needed: 3-4 hours: Day 7)**
This data is similar to the class activity and also the one described in the text book Doing Data Science Chapter 3 for K-nn. Repeat it for the German Data given in UbBox [2].
https://buffalo.box.com/s/jm1hki9hbnlm4kzfkaimkzilys3dm8dg
Tabulate your results for various K. Also try it for various sizes for test and training set (20-80%, 30-70%, etc.) Get gata from Kaggle or from here:
https://buffalo.box.com/s/jm1hki9hbnlm4kzfkaimkzilys3dm8dg

6. (10 points) **Activity 3: K-means clustering with Pew Data (Time needed: 3-4 hours: Day 8)**
Pew Research Center [10] has been collecting data for a long time about social issues using survey approach. Study their home page here and the types of questions they try to answer using the data sets collected [4] and analyzed using scientific methods. We are especially interested in look at the data PRC collected about Gaming, Jobs and Broadband. We cleaned the data set in Lab 2. In the lab we will use at least 5 variables (columns) to cluster the data and understand the data. Interpret the cluster characteristics for various K = 3, 5.

I. Here is some information about the data:
**JUNE 10-JULY 12, 2015 – GAMING, JOBS AND BROADBAND**

"This dataset contains questions about video games and gaming; job seeking and the internet; workforce automation; online dating; and home broadband, cable and smartphone use among Americans." The dataset has been downloaded by me and is available as a resource on UBox at [2]. You don't have download it from Pew Research Center. **Please understand this data is for use only in this course. Do not republish it anywhere else as yours.**

II. Download and unzip the file from UBox link given. Here is the list of files and contents.

| File Name | Contents | Our Reference |
|---|---|---|
| June 10-July 12, 2015 – Gaming, Jobs and Broadband - csv | CSV raw data file | File1 |
| June 10-July 12, 2015 – Gaming, Jobs and Broadband - Questionnaire | Survey questions | File2 |
| June 10-July 12, 2015 – Gaming, Jobs and Broadband - Topline | Results per question | File3 |
| June 10-July 12, 2015 – Gaming, Jobs and Broadband - Crosstab | Tabulated results | File4 |
| June 10-July 12, 2015 – Gaming, Jobs and Broadband - sav | SPSS file | We will not use this file |

III. Study the csv file (File 1). Both RStudio and Jupyter load the File1 well and you can see the 2001 observations of 140 variables. Study the variables. Understand the ***independent*** variables: very interesting ones for categorization and factorization: Gender, Income, Age, Race.

7. **You have to work on your own. This is an individual lab. You will get an F for the course if you plagiarize or copy somebody else's work or share your work with somebody.**

DUE DATE: 4/1/2017 BY 11.59PM. ONLINE SUBMISSION.

REFERENCES:


[1] Doing data Science Text book.

[2] UBBox data set for Lab3: https://buffalo.box.com/s/jm1hki9hbnlm4kzfkaimkzilys3dm8dg

[3] Kaggle Data science Platform, https://www.kaggle.com/, last viewed 2017.

[4] Pew Research Center Datasets, http://www.pewinternet.org/datasets/, last viewed 2017.

[5] The Man who revolutionized R. https://priceonomics.com/hadley-wickham-the-man-who-revolutionized-r/, Last viewed 201.

[6] ggplot2, R-package. http://ggplot2.org/, last viewed 2017.

[7] Tutorial on RGraphics, http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html, last viewed 2017.

[8] K-nn classification of Iris data; http://rischanlab.github.io/Kmeans.html , last viewed 2017.

[9] K-means clustering with UCI data set: http://www.learnbymarketing.com/tutorials/k-means-clustering-in-r-example/, last viewed 2017.

[10] Pew Research Center. http://www.pewresearch.org/, last viewed 2017.

[11] Pew Research Center Datasets, http://www.pewinternet.org/datasets/, last viewed 2017.

[12] Jupyter. http://jupyter.org/, last viewed 2017.

[13] The R Language. https://cran.r-project.org/, last viewed 2017.