ML 310 Advanced Machine Learning

Homework Assignment#1 Assigned: 1/12/17 Due: 1/27/17 [100 points]

- Problem Set 1 **Probability**

Study the 30 page monograph *A Short Introduction to Probability and Related Concepts* by Harald Goldstein.

Work out in detail solutions to the following problems:

2.2 and 2.5 (Page 11) [5 points]

3.2 and 3.4 (Page 21-22) [5 points]

- Problem Set 2 **Statistics** [10 points]

Study Statistics and Data Analysis A. Abebe et al. Chapter 1 upto Page 26 ONLY
We will skip Section 1.8 and rest of the chapter.

Work out in detail solutions to the following problems. Use R software where appropriate.

Page 3. Exercise 1.2.1 #1
Page 6. Exercise 1.3.1 #5
Page 9. Exercise 1.3.2 #1
Page 17. Exercise 1.6.1 #1
Page 25. Exercise 1.7.1 #6

- Problem Set 3 **Statistical Analyses using R**

Study/Practice R Basics from the R Primer [10 points]
Load the Wine Data Set available here as a CSV file into R and conduct 10 significant statistical operations as below, on any (combination) of continuous variables (i. e., Numerical) from the dataset.
https://archive.ics.uci.edu/ml/datasets/Wine

1. Cumulative Frequency Distribution
2. Cumulative Relative Frequency Graph
3. Stem-and-Leaf Plot
4. Box Plot
5. Standard Deviation
6. Covariance
7. Correlation Coefficient

Report your results with a brief explanation (2 sentences) of what they mean.

Problem Set 4 **Motor Cars (mtcars) Case Study** [45 points]

GOAL: Study a Case Study in depth using R. Apply the skills to analyzing a different dataset. Learn about Regression and Regression Trees algorithms and sue them in a case study of your own.

Please do the following and provide a report (PDF type written).

Your work should include:

1. Some understanding and explanation of the Data set
2. Explanation of Results and Conclusions

Please do not submit just R output without any explanation.

1. Algal Blooms Case Study and Motorcars application

https://cran.r-project.org/web/packages/DMwR/index.html
see Chapter 2 of "Data Mining with R, learning with case studies" by Luis Torgo, CRC Press 2010.

Follow the case study with the algae data set as is.
The operations of this cases study are shown here as one long R file:
http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR/code2.html

Then conduct all of the analysis from tis cases study on the "mtcars" data set, natively available in R. Repeat all of the case study on this dataset including the below:

1. Loading the Data into R
2. Data Visualization and Summarization
    a. IGNORE the Unknown Values code block
3. Multiple Linear Regression
4. Regression Trees
5. Model Evaluation and Selection
    a. IGNORE the Predictions for the 7 algae

Please explain the method and results/conclusions from each of the above 5 Code Sections. What is your overall conclusion about the mtcars data set from this study?

NOTE: you can get motorcars dataset natively included with R, using
>data (mtcars)

Problem Set 5 **Implementation of Resrvoir Sampling** [25 points]

Please submit (1) Working code which is fully commented to describe your solution and (2) Output file showing answers clearly. Please use Java or Python or C++ for the implementation.

Write code for a **Reservoir Sampling** algorithm. It should randomly sample n data values from among a population of size N, which is not known (It is a data stream).

Using your Reservoir Sampling code, sample a random sample of 20 data items from this data set, and print their values to your output file.

Use Mike Stanley's Gyroscope data set as input (Nudge.txt at the below link). Sample a random sample of 20 data items for (MagX, MagY) from this data set, and print their values.
http://www.nxp.com/products/sensors/sample-data-sets-for-inertial-and-magnetic-sensors:SENSORDATA