ML 310 Applied Machine Learning

Homework #4 Assigned: 2/23/17 Due: 3/13/17 [120 points]

GOAL: Please become conversant with concepts and implementation using R for the below two topics we discussed in class. Then complete the below exercises.

Section I. Text mining case study

Study Chapters 1 through 4 of the NLP Book: http://www.nltk.org/book/

Analyzing Text with the Natural Language Toolkit
Steven Bird, Ewan Klein, and Edward Loper

Then answer the questions from these chapters as below. Use NLTK and Py programming as needed.

CHAPTER 1: [6 points]
Getting Started with NLTK
Install Python and NLTK
Following the example on Page 5-6, Pick a pair of words and compare their usage in two different texts, using the similar() and common_contexts() functions. Explain your results.

CHAPTER 2: 2.8 Exercises (Page 74) [14 points]
4. ○ Read in the texts of the State of the Union addresses, using the state_union corpus reader. Count occurrences of men, women, and people in each document. What has happened to the usage of these words over time?

CHAPTER 3: 3.12 Exercises (Page 124) [20 points]
9. ○ Save some text into a file corpus.txt. Define a function load(f) that reads from the file named in its sole argument, and returns a string containing the text of the file.

Use nltk.regexp_tokenize() to create a tokenizer that tokenizes the following kinds of expressions: monetary amounts; dates; names of people and organizations.

Section II. Large Datasets analysis using DBI/SQLite [15 points]

Study R and SQLite via DBI
Study tutorial http://www.r-bloggers.com/r-and-sqlite-part-1/
Study ***How to Use DBI: Connecting to Databases with R***

Then create a SQLite DB and load any example .csv data file from the above CMS data sets into the DB.

Conduct any 2 search or modify SQL operations on the DB and report results (e. g. find average; find …)
Show and explain your results.

Section III. Large Datasets analysis using Bigmemory Package [5 points]

Obtain the Netflix Prize Data Set.

Study the paper on **R Bigmemory package**.
Apply any 2 basic R operations on the data set using Bigmemory and report results.

Section IV. Neural Networks

1. Study the Neural network Algorithm basics Tutorial:
   http://gekkoquant.com/2012/05/26/neural-networks-with-r-simple-example/
   http://www.r-bloggers.com/using-neural-networks-for-credit-scoring-a-simple-example/
   http://www.r-bloggers.com/r-code-example-for-neural-networks/
   http://hodgett.co.uk/get-started-with-neural-networks-in-r/

   Then, conduct your own NN Analysis on the Wine classification dataset
   http://archive.ics.uci.edu/ml/datasets/Wine [20 points]
   (Useful Ref (Not essential):
   http://neuroph.sourceforge.net/tutorials/wines1/WineClassificationUsingNeuralNetworks.html

Section V. Deep Learning [20 points]

A. Study the Nature paper on DL – Yann LeCun et al. **Deep Learning** and answer the below question in 2 paragraphs:

   How is Deep learning different from the Neural Networks (ANN) learning and algorithms we studied?

B. MXNetR is one of the 5 key DL packages in R. Study the first 2 sections of this implementation, which we discussed in class, and the Hand-writing Recognition 2 pager. [20 points]

   https://www.r-bloggers.com/deep-learning-with-mxnetr/

   Then, explain in your own words (1-2 para) about how MXNetR was used to implement hand-writing recognition in this DL application.