ML 4375 – Intro to Machine Learning – Spring 2018 – Mazidi

Project 1

* Select 2 data sets for the project. One data set should be suitable for regression and the other for classification.
* For *each* data set:
  + Decide which column is the target: what you want to learn.
  + Use at least 5 R functions for examining each data set (head, summary, names, cor, str, etc.)
* Produce at least 2 informative graphs for each data set.
* Run *at least* 2 ML algorithms on each data set, using at least 3 algorithms in all.
  + Run summary() on each model.
  + Select an appropriate metric to evaluate how well each algorithm performed on the test set. Provide some commentary and analysis of which algorithm did best and why you think that is the case.
  + Algorithms we have learned so far:
    - regression: linear regression
    - classification: logistic regression, hierarchical clustering
    - regression or classification: knn and k-means
* Upload the Rmd file to eLearning.
* Peer Evaluation
  + see next page
  + The eLearning system will randomly assign 3 peers to evaluate your work. These scores will be averaged by the TA/instructor. The TA/instructor reserves the right to raise or lower the grade in the event of poor grading by your peers.

Top and/or interesting projects will be chosen by the instructors to present to class.

Sources for datasets:

* <https://www.kaggle.com/datasets> (free but you have to join)
* <http://archive.ics.uci.edu/ml/>
* <https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/>

Project Evaluation

Data Exploration (40 points)

* Data Set 1
  + \_\_\_\_\_ (0-5 points) Columns were described; a reasonable target column was identified; data cleaning (if needed) was described; a link to find the data was provided.
  + \_\_\_\_\_ (0-5 points) At least 5 R functions were used for data exploration.
  + \_\_\_\_\_ (0-10 points) At least 2 meaningful graphs were included. Graphs should have labels and be easy to interpret.
* Data Set 2
  + \_\_\_\_\_ (0-5 points) Columns were described; a reasonable target column was identified; data cleaning (if needed) was described; a link to find the data was provided.
  + \_\_\_\_\_ (0-5 points) At least 5 R functions were used for data exploration.
  + \_\_\_\_\_ (0-10 points) At least 2 meaningful graphs were included. Graphs should have labels and be easy to interpret.

ML Algorithms and Evaluation (50 points)

* Data Set 1
  + \_\_\_\_\_ (0-5 points) Meaningful metrics were chosen for the algorithms.
  + \_\_\_\_\_ (0-10 points) Two appropriate algorithms were run on the data set.
  + \_\_\_\_\_ (0-5 points) Summary() was run on the models.
  + \_\_\_\_\_ (0-5 points) Analysis of algorithms and why one might have outperformed others.
* Data Set 2
  + \_\_\_\_\_ (0-5 points) Meaningful metrics were chosen for the algorithms.
  + \_\_\_\_\_ (0-10 points) Two appropriate algorithms were run on the data set.
  + \_\_\_\_\_ (0-5 points) Summary() was run on the models.
  + \_\_\_\_\_ (0-5 points) Analysis of algorithms and why one might have outperformed others.

Project Depth (10 points) {Projects should get lower points for trivial data sets.}

* \_\_\_\_\_ (0-3 points) This project did not meet or minimally met requirements.
* \_\_\_\_\_ (4-7 points) Project met requirements and a little more.
* \_\_\_\_\_ (8-10 points) Project went well beyond the requirements.

Justification and Comments regarding Project Depth: