CS 395
Project Part 1
30 Points Total
Due in Canvas by 11:59 PM on Wednesday, February 27, 2019

This is the first of three parts of your course project.

This first part asks you to show your data sources and a brief description of how you plan to proceed. This is intended to get you thinking about the project in detail, identify your data source(s) and describe what problem you plan to address with this data, and a give a high-level idea of how to proceed.

The second part will be due in April and will require you to provide some preliminary results of your project.

The third (and final) part will be due at the end of the semester and provide/present your results. This final project deliverable will incorporate what is being asked in this assignment (subject to refinements after receiving feedback) so when you prepare this first part, keep this in mind.

You can work on the project alone or with one other person. In other words, there can be a maximum of 2 people in a single group.

Requirements for this first project assignment are given below.

1. (5 points) Choose and describe your dataset. The dataset you choose should be publicly available. It helps if it is labeled so you can verify the metrics (e.g., accuracy).

   Here are some publicly available datasets. This is not an exhaustive list, so if you are not sure if your dataset is a good one, please email me. See me also if you need more suggestions.

   - https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/
   - https://archive.ics.uci.edu/ml/datasets.html
   - https://www.datasciencelearner.com/datasets-for-machine-learning-projects-data-scientist/
   - https://www.dataquest.io/blog/free-datasets-for-projects/
   - https://medium.com/datadriveninvestor/the-50-best-public-datasets-for-machine-learning-d80e9f030279

   All things being considered, choose one with more instances (rows) than fewer instances.

2. (10 points) Describe your data. You will probably need to have it in a single file for processing, so if it is in multiple files now, you need to do this first.

   a. What type of data is it?
   b. Is it nominal, ordinal, interval or ratio?
   c. What are the ranges or possible values?

3. (5 points) Provide a few sample rows of the dataset.

4. (5 points) Transforming the data.

   a. Describe any transformations (e.g., one hot encoding) or normalizations you plan to do to prepare the dataset.
   a. How will you break the data into test and training sets? At what ratio?

5. (5 points) Describe what problem you are trying to solve in one paragraph.

Make sure if you are working with another person, both names appear on your submission.

Ideally, provide the above in a single pdf or docx document (if you need to have a second document to show the sample rows, that is okay, but don't zip up your work, just provide two documents)