CSYE 7245 - Big-Data Systems and Intelligence Analytics

Assignment 1 - Machine Learning (INFO 7390) Review

Due Friday, February 2 2018

Submission: Put the data and Jupyter notebook files in a folder. Make sure all links to data are relative to the folder so the TAs can run the notebooks.

Individual Assignments

These are individual assignments. They cannot be done in groups.

Machine Learning Review

Find a public dataset or machine learning competition and use machine learning techniques to analyze the data. This should be REVIEW of what what learned in INFO 7390 - Advances in Data Sciences and Architecture. You cannot use a INFO 7390 project for this assignment.

No two students can analyze the same data so you MUST e-mail the TA's for approval.

Part A - Get Some Data (25 points)

- Data cleaning
  - Are there missing values? (10 %)
  - Are there inappropraite values? (10 %)
  - Remove or impute any bad data. (10 %)
- Answer the following questions for the data in each column:
  - How is the data distributed? (10 %)
  - What are the summary statistics? (10 %)
  - Are there anomalies/outliers? (10 %)
- Plot each colmun as appropriate for the data type:
  - Write a summary of what the plot tells you. (10 %)
- Are any of the columns correlated? (10 %)
- Write a clear summary of what the EDA tells you (20 %)

Part B - Analyze Some Data (50 points)

What is expected?

a. A clear description of the question being asked. (10 %) b. Background research of related work. (10 %) c. Data sources? (10 %) d. What algorithms are being used and code sources. (10 %) e. References. (10 %) f. Analysis (50 %)

These assignment will provide practice in real-world analysis and application of machine learning algorithms. The research can take one of the following forms:

i. Tweaking an existing machine learning algorithm.

ii. Applying an existing machine learning algorithm in a novel context.

iii. Validating an existing machine learning algorithm in real-world contexts.

iv. Creating a novel machine learning algorithm.

v. Competing in a compeition like Kaggle https://www.kaggle.com/

vi. Student suggested.

Part C - Write a Report (25 points)

The report must have:

a. Abstract (10 %)

b. Introduction (5 %)

c. Code with Documentation (50%)

d. Results (20 %)

e. Discussion (10 %)

f. References (5 %)

List of datasets for machine learning research

- List of datasets for machine learning research
- UC Irvine Machine Learning Repository
- Public Data Sets : Amazon Web Services
- freebase
- Google Public Data Explorer
- datahub
- data.gov