IST652 Scripting for Data Analysis
Summer 2020

Final Project Instruction

This is an individual project that you could program using Python and apply the data analytics and machine learning knowledge you learned from this course to solve a real-world problem set.

1. Choose data mining problem and data set

For this project, you must choose your own dataset. It can be one found from an on-line source, one of your own, or one of the ones from the UCI repository (http://archive.ics.uci.edu/ml/) or Kaggle datasets (https://www.kaggle.com/datasets). You are encouraged to choose data challenges on Kaggle that are related to the current COVID-19 pandemic (https://www.kaggle.com/search?q=covid+in%3Adatasets).

Some rules/tips about choosing data sets:

a. Do not choose the datasets that we have already analyzed in class or for homework.
b. It should not be a small or made-up dataset. For this semester, "small" is defined as fewer than 5000 examples in the dataset. There should also be at least 10 features with mixed types (including both categorical and numerical attributes).

2. Machine learning and statistical modeling design

Define a problem on the dataset and describe it in terms of its real-world organizational or business application. The complexity level of the problem should be at least comparable to one homework assignment.

You should implement an end-to-end data mining and machine learning modeling workflow that covers the following stages: problem framing (classification vs. regression vs. others), data understanding, Exploratory Data Analysis (EDA) and visualization, Data Preprocessing and Preparation, Machine Learning and Statistical Modeling (with at least one supervised learning algorithm and one unsupervised learning algorithm), model performance evaluation and algorithm fine-tuning and model interpretation.

3. Project proposal

By **11:59pm EST, Thursday, 06/25/2020**, you should post a brief project proposal to the Week 6 Discussion Forum describing the dataset you plan to use (a summary of the meta data such as number and type of features, etc.), and tentatively what type(s) of data preparation and algorithms you would like to investigate and what is the target outcome you try to study/model. Please also provide brief justifications regarding your choices of

the data mining algorithms for the problem set, and any other relevant information regarding your analysis plan.

4. Final project paper

To complete this project, write a final report that conforms to general research paper format. See (Pang, Lee, and Vaithyanathan, 2002) as an example. Your report should be within 8 pages, 1 inch margin on all sides, and at least 12 point Arial or Times New Roman. Remember that your project paper serves as the tour guide for your readers to be able to repeat your data mining process and discover the same patterns as you did. It is very important to cite and paraphrase relevant work appropriately.

5. Project video presentation

Please record a presentation of your final project that is no shorter than 10 minutes with detailed explanation of your work and upload to YouTube or other video sharing site. Include the URL link for your video presentation at the end of your final project report. Prepare a slide deck to summarize your final project and guide your presentation by walking the audience through your research.

Final Project Submission:

Upload the following files for the final project submissions:

- Final project paper (with the URL link for the video presentation)
- Final project presentation slide deck (used in the video presentation)
- All the source Python codes developed for the final project.

Submit your final project files to the Blackboard by **11:59pm Friday 07/03/2020**.

References

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of EMNLP 2002*, 79-86.