Data Mining, CSCI 347, Fall 2019
Project Deliverables

Updated: Nov. 24, 2019

Submissions:
Topic Exploration Sept. 23 (~~1%~~ 3%)
Area and Major Question Sept. 30 (~~2%~~ 3%)
Data set(s) Loaded into a Tool and Analyzed Oct. 21 (~~3%~~ 4%)
~~Preliminary Results Nov. 4 (4%)~~

Draft Final Presentation (4%)
Nov. 4 - Carson, Dalton, Keelie and Hunter
Nov. 22 - Kaleb, Xaavan, Amanda and Joseph

Draft Report ~~and Presentation~~ Nov. 22 (~~10%~~ 6%)
Final Report and Presentation ~~Dec. 2nd and 4th~~ (20%) Dec. 4, 3-5pm

This project is an opportunity for you to experiment using data mining tools to explore a question of your choice. This is a chance to:

- explore the challenges of forming a question amiable to data mining,
- get to know an application area
- iterate through the following:
    - formulate a question,
    - locate data to mine,
    - prepare the data for analysis: load the data into a tool, clean it, modify attributes, integrate supplemental data sets,
    - explore the data, visualize it, gain a complete understanding of it (possibly contacting the collectors of the data to get questions answered)
    - extract most useful features
    - mine the data and determining the accuracy of the results, and finally,
- write the results.

Your question may be to predict an outcome, find relations amongst attributes, or determining groupings of data instances. A successful project may or may not answer the chosen question, but will provide insight into the area.

Treat data preparation and analysis as a lab experiment. Document what you have done to the data set what mining you have tried, the results, and what you plan to try next. Document the process sufficiently so that the experiment is repeatable. You may publish your results and it needs to be possible for other data scientists to achieve the same results. Jupyter Notebook is an excellent tool for this.

When submitting a deliverable, submit all previous graded work at the same time. If you updated earlier documents, include both the original graded version and the updated version.

Present each deliverable to the class and include the visual aids with your submissions.

Topic Exploration

Begin by brainstorming possible questions. List questions that you think might be amenable to analysis by data mining. Consider areas in which you may work. Locate datasets which may be helpful.

Select an area in which you will work. Tell why you chose the area you did. List several questions that you might explore and datasets which you may mine. List alternate areas which you considered.

Deliverables:

- A half to full page write-up.
- Share the information with the class in 1-2 minutes (at your seat).

Area and Major Question

Settle on the area in which you will apply data mining. Also determine one major questions, and 3-5 alternative questions. Tell why you have chosen this question. Describe the state of data mining research in this area – what has been tried and what has been learned. Anticipate challenges in this area.

Locate and describe at least one major dataset, and 1-3 supplemental datasets. Tell how this data was collected, characteristics or the data and its completeness.

Deliverables:

- A 3-7 page write-up.
- Share the information with the class in a 6 minute presentation that includes visual aids.
- Visual aids used for the presentation (4 to a page).
- Previous materials.

Dataset(s) Loaded into a Tool and Analyzed

Load the major dataset, and supplemental datasets if relevant, into a tool. For the dataset(s) as a whole, present any characteristics of the dataset that you notice, such as:

- whether it has many or few missing values,
- if it seems consistent (for datasets with a class value, the same evidence generally gives the same prediction),
- if it is balanced (for datasets with a class value, the number of instances is approximately equivalent for each of the class values)
- any other interesting information about the dataset.

Create visualizations of the data.

For major attributes, tell the type of attribute (nominal, numeric), example values (possibly a histogram, or average and standard deviation).

Deliverables:

- A 2-3 page write-up.
- Visualizations shared with the class in a 5 minute presentation.
- Visual aids used for the presentation (4 to a page).
- Previous materials.

~~Preliminary Results~~

~~Try out different data mining techniques in an effort to answer your major question. In each case, document your expectations, describe your process (including evaluation procedures), what you learned and where you will go next. Include the learning output from the trials, along with the evaluations.~~

~~Deliverables:~~

- ~~A 4-5 page write-up.~~
- ~~Preliminary results shared with the class in a 7 minute presentation.~~
- ~~Visual aids used for the presentation (4 to a page).~~
- ~~Previous materials.~~

Draft ~~Project and~~ Final Presentation (Nov. 4 and 22nd)

Update the class on how your project is going. Try out different data mining techniques in an effort to answer your major question. In each case, document your expectations, describe your process (including evaluation procedures), what you learned and where you will go next. Include the learning output from the trials, along with the evaluations.
By this time, some interesting results are expected. (Note that a null result is a result.) Begin drafting your final report.

Deliverables:

- ~~An 8-9 page write-up.~~
- ~~Share with the class your progress and what you are expecting to go into your final report.~~
- Visual aids used for the presentation (4 to a page).
- Previous materials.

Draft Project ~~and Presentation~~

Deliverables:

- An 8-9 page write-up.
- Previous materials.

Final Report and Presentation

Present your conclusions. Write a final paper summarizing what you did and what you conclude.

Deliverables:

- ~~A 10-12 page write-up~~. Final write-up in pages needed
  References in APA format
  Formal tone, appropriate for publication
  Paper abstract, introduction and conclusion included
  Include: name of dataset, where dataset can be found, # instances, # attributes, goal of what you wanted to learn, characteristics that helped/hindered that learning
- Final presentation, time needed
- Visual aids used for the presentation (4 to a page).
- Previous materials.