CSC-272: Final Project

project proposal due by 10:30 a.m. on Wednesday, March 25, 2020
project writeup due by 10:30 a.m. on Friday, April 24, 2020

Overview

The final project will give you the opportunity to use the techniques covered in the course to:

- organize, experiment with, and analyze a collection of data relating to human behavior that interests you
- draw conclusions based on your analysis
- present your results

Requirements

You must work together on the project in teams of three students. Team members will submit a single final report and receive the same grade for the project, with the possibility of adjustments based on peer review. If you need help finding people to work with, email me and I will try to help you form a team.

Your final project must include all of the following:

- the effective application of at least one (and typically more than one) data-mining strategy to a dataset that you choose, with thorough analysis of its effectiveness
- a clear and compelling presentation of the results that you obtain, both from the data mining and any other analysis that you perform
- at least **two** examples of data visualization. You may want to make use of the tools at Tableau (*http://tableau.com*), Google Charts (*https://developers.google.com/chart/*), or the tools associated with Google's Public Data Explorer (*http://www.google.com/publicdata/directory*) to help you with this, although other visualization tools (e.g., Excel) are also fine.
- a thorough, well-organized write-up (described in detail below)
- a 20 minute final presentation, organized and rehearsed, with appropriate visual displays

Choosing a Dataset and a Problem to Solve

You should begin by selecting a dataset to analyze. Possible sources of datasets include:

- the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html)
- the datasets associated with Google's Public Data Explorer (http://www.google.com/publicdata/directory)
- U.S. Government open data repository (http://www.data.gov/)
- FedStats, which includes a large number of datasets compiled by federal agencies (http://fedstats.sites.usa.gov/)
- datasets available from the U.S. Census Bureau (http://www.census.gov/)
- UNdata, a collection of databases compiled by the United Nations (http://data.un.org/)
- Collection of datasets provided by Weka (http://www.cs.waikato.ac.nz/ml/weka/datasets.html)
- KDnuggets dataset repository (http://www.kdnuggets.com/datasets/index.html)
- Kaggle.com competition datasets (http://www.kaggle.com/competitions)
- The *Journal of Statistics Education* data archive (http://www.amstat.org/publications/jse/jse_data_archive.htm)
- One of numerous open APIs (iTunes, Billboard Top 100, Wikipedia, TMDb, Statistica, Natl. Weather Service, etc.)
- One of numerous auth APIs (Spotify, Twitter, Goodreads, YouTube, Instagram, Google Maps, Tumblr, Fitbit, etc.)
- a dataset that you find using a search engine. Try a search using the keywords "dataset" and whatever subject you are interested in (e.g., "population dataset")
- a dataset that you construct via data scraping from a web site
- a dataset of your own creation

You are welcome to choose any dataset that interests you, relates to human behavior, and that has enough data to enable a meaningful analysis.

In making your choice, you should be sure consider what problem or problems you would be able to solve by employing data mining on the dataset. In other words, you should ask yourself: *How could I use data mining to answer one or more questions about this dataset?*

Although you are welcome to employ association learning as a supplemental component of your analysis, ***you must choose either classification learning or numeric estimation as one of your approaches***. This means that you will need to choose a dataset that is amenable to one of these types of data mining -- i.e., to building a model that will determine, predict, or estimate one of the attributes in the dataset, based on the values of other attributes. The medical diagnosis and credit-card promotion problems that we have covered in lecture are examples of this type of problem.

Important notes:

- You will be able to use the Weka data-mining software to perform the actual data mining. Don't worry if you are uncertain at this point about the particular algorithms that you will use. All that matters for now is that you have a sense of the general data-mining approach that you will employ.
- It may be necessary to transform the dataset in some way before performing data mining. For example, if you are working with data about countries, you may need to combine the records for a given country from multiple tables to create one large record containing all of the information about that country. You may possibly have to discretize some numeric data, or contend in a logical way with missing data, incorrect data, or noisy data. You should consider what transformations will be needed on the datasets that you are considering.
- You will need to make sure that you have enough data for a meaningful analysis. One rough guideline is to aim for a dataset with at least 100 instances and at least 4 attributes (excluding attributes that are merely identifiers such as the name of a country).

Writing a Project Proposal

Before beginning work on your project, you will need to submit a brief proposal outlining what you intend to do. This will allow me to make sure that you are on the right track, and to give you some initial guidance.

Each team only needs to submit one proposal.

Your proposal should include the following:

1. the names of all students working on the project
2. a description of the dataset that you will be analyzing, including information about where it can be obtained (e.g., a URL). Include in your description a list of the key attributes that are present in the dataset.
3. the type(s) of data mining that you intend to perform (classification learning, numeric estimation, or association learning), and an explanation of your choice of approach(es). *Remember that that you must chose either classification learning or numeric estimation as one of your approaches.* You are, of course, free to incorporate more than one approach in your project plan.
4. Explain briefly the goal of the model that you are hoping to build using data mining. For example, if you were working with the medical-diagnosis dataset from lecture, you would say something like this: *We are hoping to use data mining to build a model that will allow us to determine a patient's diagnosis, based on his or her symptoms.*
5. a description of any pre-processing that you will need to perform on the dataset before you perform data mining. If no transformations are needed, you should briefly explain why the current format of the dataset is amenable to the type of data mining that you will perform.
6. how you plan to divide up the work among the members of your team.

Your project proposal should be submitted electronically via e-mail.

Splitting Your Dataset

Your project will include an application of data mining to learn a model that predicts the value of some output or class variable. In order to validate the model or models that you produce, you should separate your dataset into two files. One file -- containing N percent of the examples (where N is typically somewhere between 70 and 90) -- should be used for training, and the remaining 100 - N percent should be used for testing. The testing examples should not be touched until you have developed a model using the training examples and are ready to test it. We will be covering how to use Weka to split your dataset in an upcoming lecture. Note that you should save your dataset as an ARFF file *before* splitting it.

Experimentation

As should be obvious from the discussions we've had so far, data mining is not a perfect science. It requires a sense of curiosity and a willingness to experiment. Accordingly, *experimentation will be at a premium in this project*. That includes fine-tuning the settings of the algorithms that you use (as we've done in lab), and doing research to learn about algorithms and approaches that we haven't covered in class. This principle may also be applied to data collection and cleaning. Acquisition of data via scraping, or via an API, for example, will be worth a lot. So will the use of overlay data taken from different sources. Be sure to highlight such explorations in your final report.

Writing the Report

You should submit a written report on your final project that incorporates all of the items mentioned in the *Requirements* section above.

Your report should include at least the following sections:

1. **Introduction**: a one- or two-paragraph overview of your project that summarizes the key points found in the rest of the report, including the problem or problems that you attempted to solve and a high-level (not too specific) description of the results that you obtained. Think of this section as a brief preview of what the rest of the report will contain.

2. **Dataset description**: information about of the dataset that you analyzed, including its key attributes and details about where it was obtained (e.g., a URL). A table that includes a brief description of each attribute is often helpful.

3. **Data preparation**: a description of any steps that you took to prepare your data for analysis and mining. (A lecture on this topic is upcoming.) ***Discuss the steps at a high level that would make sense to someone who is not familiar with Weka.*** For example, rather than saying "I applied the unsupervised/attribute/Discretize filter to the following attributes:...", it would be better to say "I used Weka to perform equal-height discretization of the following attributes:...".

4. **Data analysis**: a description of the analysis that you performed -- including the data-mining algorithm or algorithms that you employed. ***You should include a brief description of each data-mining algorithm -- enough so that someone who is not already familiar with it can understand what it does.*** You can often find some information about a given algorithm by clicking on its name in Weka and then clicking the *More* button in the window that pops up. You can also try using a search engine to find out more information about the algorithm. Include references to any sources that you use.

5. **Results**: a summary of the results of your analysis. The exact form of this section will depend on the type of analysis that you performed, but make sure that the results are presented in a clear and compelling way that includes at least two examples of data graphics. ***You may include graphics that are drawn from Weka, but they do not count towards the requirement for two data visualizations***. Rather, your data visualization should illustrate relationships between various attributes of your data set. (Data visualization is a topic to be covered in class soon, including a lab session on Tableau.)

   This section should also include the structured model(s) that you produced using data mining. For your classification/estimation models, you should specify how well the model performed on the training examples and how well the model performed on the test examples. If you include a confusion matrix (another upcoming topic), please turn it into a nicely formatted table. Don't just copy the text version of the matrix from the Weka window into your report.

You should also include a brief discussion of each model. Does the model make intuitive sense? Why or why not? How well does the model generalize? ***In discussing your results, beware of making overly confident claims.*** It is better to be realistic and cautious. For example, instead of saying "The results clearly show that attribute A is determined by the value of attribute B", it would be better to say something like "The results suggest that attribute B may have an impact on the value of attribute A."

6. **Conclusions**: a one-paragraph summary of the report, reminding the reader of the key points that you want him or her to remember. This should include overt emphasis on the experimentation that you did.

7. **Appendix**: You may also want to optionally include appendices at the end of the report (e.g., the text version of large models produced by Weka).

Your report does *not* need to be overly long. Just make sure that it is a clear and complete presentation of the steps that you took and the conclusions you reached. Submit your report electronically via e-mail as a Word or PDF file. Also submit your dataset(s) in its final form, after any transformations that you applied. Additionally, submit any supporting files that you created – slide decks, Tableau projects, Python code, etc.

Project Presentation

Your team will be assigned a 20 minute slot for presentation of your work during the last day of class or the last lab period. Plan on 15 minutes for presentation and 5 minutes for questions, with overlapping transition time. The presentation should include:

- an accessible, engaging summary of the key objectives and results of the project
- a visual presentation using an appropriate tool of your choice
- comparable, confident, and rehearsed participation of all team members

Inspiration

For inspiration as you begin thinking about your project, visit the “Hall of Fame” of past projects:

http://cs.furman.edu/~ktreu/csc272/hof.html

Elements of this project assignment are borrowed verbatim and with permission from Dr. David Sullivan of Boston University.