Analytical Project

Overview

The **analytical course project** is viewed as an equivalent to a midterm exam. The project is to be performed in teams of four people (one team of five people is ok).

Due Date: ***May 31/June 4***.

The posters are due *May 31* (the CENG poster day). Complete write-ups are due *June 4*.

Assignment

The project consists of two parts: (a) exploration of provided datasets and (b) creation and exploration of a new dataset. Each is described below.

Exploration of provided datasets.

This part of the project is worth about 65–70%.

The project uses a number of existing, and specially prepared datasets. The datasets are made available to you via the following course wiki page:

http://wiki.csc.calpoly.edu/csc466-2010/wiki/DataToAnalyze

The page contains links to a number of datasets (some reside on the datasets wiki, some other — existing datasets, reside in other repositories). Links to additional datasets may be added to this page throughout the course. during the next week. There are currently 12 datasets listed there.

The Task. Each team will select a number of datasets from the list offered. At least three datasets must be selected. For each selected dataset, the team will perform the following tasks:

1. Make an effort to understand the nature of the dataset. Some datasets (e.g., the `OS-PREFERENCES` dataset) are going to be easier to comprehend than some others.
2. Formulate one or more analytical questions about the data in the dataset.
3. Determine what type of analysis is needed to answer the formulated question.
4. Perform the desired analysis. This may involve using the code you wrote for one or more labs, writing new code, or using existing analytical resources.
5. Determine the answer to your analytical question and record it.

Analytical Questions and Analytical Methods

It is expected that the analytical questions you ask involve use of the KDD methodolgy discussed in the course.

As part of your solution approaches you can conduct any statistical analyses of the data you seem fit, as well as any KDD tasks discussed in the course, or discovered by you independently.

The ground rules for what you can and cannot do are set below.

Allowed Activities

As part of your preparatory and analytical activities you are allowed to do the following:

- Use any programs you (members of the team) created during this course.
- Use any programs other students (outside of your team) created during this course, **with the explicit permission of the authors of the programs**.
- Use any existing code for "menial" tasks (parsing data, reporting) as well as for tasks such as visualization of output. You **must be allowed to use the code by the licensing agreement of the code**.
- Use any existing code for KDD methods both covered and not covered in class, subject to the following two conditions:
  1. You must be allowed to use the code by the licensing agreement of the code.

2. You **must gain sufficient understanding of the methodology implemented by the code**.

   For example, if you decide to use some open source software for learning neural networks from data, I will expect at least one member of the team to be able to coherently explain to me what neural networks are, and what specific types of networks are being constructed by the software used.

- Study new (not covered in class) methods for solving KDD problems discussed in class.
- Study new (not covered in class) KDD problems and methods for addressing them.
- Write new code.
- Enhance code created earlier during this course.
- Use any supporting architectural solutions (e.g., Oracle DBMS, or math/stats packages like SPSS or MatLab) and use any analytical and KDD techniques available through them, subject to the same condition:

  > You **must gain sufficient understanding of the methodology being used**.

Disallowed Activities

The following is a list of **no-no**s for this project. Any of the activities below conducted as part of the project **are considered equivalent to academic cheating!**

You may not:

- Use ANY code you have not been authorized to use (by the authors, or by the licensing agreements).
- Use ANY KDD/analytical techniques (or their implementations), when you did not gain sufficient understanding of the technique.
- Actively seek, and peruse information about the datasets, that contains the answers to your analytical questions.

  Note: some of the datasets are well-known data mining/machine learning datasets, which have been used by many different research teams to test their methods. KDD models developed for such datasets may be discoverable via some targeted web search.

  Note: Some of the datasets are featured in multiple publications. Typically, it is safe to peruse such publications in your work on the project. If a paper publishes, in addition to the evaluation results, the actual models built by the KDD methods for the dataset, you are still allowed to use the paper on the following two conditions:

  - You explicitly acknowledge the source of the model.
  - If the model addresses your analytical questions, you still use tools available to you to generate it.

  (I do not want this assignment to turn into a hunt for existing models. I want you to build your own.)

- Solicit help with your analysis from anyone outside of this class. (In particular, do not ask dataset owners or researchers who used the dataset in their work for help.) If you believe you need to get in touch with the data owners/other researchers because you have a bona fide question or concern, **bring your question(s)/concern(s) to me**, and let me initiate the contact. (this, among other things, will increase the probability and timeliness of the response).

Create your own dataset part

This part of the assignment is worth about 30–35% with potential for extra credit.

- **Create** from scratch or from publically available data a **new dataset**, **or select** an existing dataset that is not on the list for the first part of the assignment.
- **Perform** the same activities (understand the nature, formulate questions, determine and perform analyses, write up results) on the new dataset as you are asked to perform on the datasets from the **first part** of the assignment.
- If you have the rights to the dataset you used for this part of the assignment, consider submitting the dataset to the instructor's datasets wiki to be used in future courses/assignments.

Notes

The rule-of-thumb for this assignment is "one dataset per person". If your team is larger than four people, your team is expected to analyze more than four datasets before the extra credit rules apply.

Your team must analyze **at least** one new dataset. However, you may chose to build/use more new datasets. In this case, you are allowed to replace an instructor's dataset with one of your own. In fact, you will receive extra credit as described below.

As an example, a team of four people may choose to use two datasets from the list provided by the instructor, and may construct and analyze two more datasets that were obtained from other sources. This is considered to be a one dataset replacement. This earns the team 10% extra credit.

Extra Credit

You get 10% extra credit for performing analytical tasks for any of the datasets provided by the instructor above the mandatory three datasets.

You get 10% extra credit for substituting a new dataset for a dataset provided by the instructor.

You get 20% extra credit for each new (i.e., created or discovered by you) dataset for which you perform the analytical tasks, beyond the required dataset (i.e., if a team of four students works with five datasets total, two of which are new, you get 20% extra credit for the extra new dataset).

You get 5% extra credit for each dataset you submit to the instructor's datasets wiki. This includes the submission of the first "create your own" dataset. Note, that in order to submit a dataset:

- You must have the rights to it. That is, you must be the creator/designer/owner (whichever is more appropriate) of the collected data.
- You have to give the instructor the (non-exclusive) rights to use the dataset for research and instructional activities. (The act of submitting the dataset to the wiki serves as such a grant).

Deliverables and Submission

Each team shall produce the following artifacts.

- A poster for the CENG Poster Session, which will be held on May 31. We may actually use the lab period during that day to demo the posters. Each team must create at least one poster – more if you need more space to put your result on.
- A written report for each dataset selected by the team. The report shall, at the very least, contain the following:
  - Description of the dataset used. If this one of datasets provided by the instructor, you need to simply name it. If this is a dataset you creates/selected, provide a full description of the dataset.
  - Description of the analytical question(s) your team studied.
  - A narrative explaining which analytical methods your team used.
  - Results of the use of the methods visualized where possible.
  - Conclusions you team drew from the results.

  Please, merge all reports into a single PDF document. Submit the soft copy to the wiki by the **June 4** deadline. An optional hard copy can be given to me any time during the finals week.

GOOD LUCK!