Analytical Project

Overview

The **analytical project** is the "other" project in the course (technically viewed as an equivalent to a midterm exam. The project is to be performed in teams of four people (feel free to stay in the teams you have formed for the **design project**.)

Due Date: ***June 4***, Thursday, in-class.

The hard copies of the project deliverables (see below) are to be submitted before the end of the last class/lab period. The soft copies should be posted to the course wiki around the same time.

Assignment

The project consists of two parts: **mandatory** and **extra credit**. Each is described below.

Mandatory Part

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/csc466-2009/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 (and will) be added to this page during the next week. (there will be at least six datasets offered to you).

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).

Extra Credit Part

For **extra credit** (assessed at 10–20% of the full credit for the assignment *per dataset*), you can do the following:

- **Create** from scratch or from publically available data a **new dataset**, **or select** an existing dataset that is not on the list for mandatory 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 **mandatory part** of the assignment.
- **Perform** the **mandatory part** of the assignment on more than four (4) datasets made available to you.

You get 20% extra credit for (correct and complete) work with a newly created by you dataset provided you (a) submit it to the course wiki and (b) grant me permission to use the dataset in future courses.

You get 15% extra credit for (correct and complete) work with an existing dataset.

You get 10% extra credit for (correct and complete) work with more than four datasets from the list of datasets in the **mandatory part**.

Deliverables and Submission

Each team shall produce the following artifacts.

For Mandatory Part.

- A written report for each dataset selected by the team. The report shall, at the very least, contain the following:

- 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.

The **soft copy** of the written report (or reports, if you use separate documents for each dataset) shall be made available on the team's course wiki page. The **hard copy** of the written report shall be submitted to the instructor on **June 4** during the classtime or lab period.

For Extra Credit Part.

- For extra credit involving datasets made avaialble to you by the instructor, you need to submit the same style report as for the datasets used in the mandatory part.
- For extra credit involving existing datasets (that you discovered yourselves), submit a report similar to the one for the mandatory part, adding to it, a thorough description of the dataset used. Also, include the link to the data on the team's wiki page.
- For extra credit involving the dataset(s) you created, submit a report similar to the one for the mandatory part, adding to it a thorough description of the dataset you created.

  Additionally, if you want to receive the full 20% credit, submit the full dataset to the CSC 466 course wiki. The submission shall contain any actual data files you have created, a README or a data dictionary file containing the explanation of the attrbitues used in the data set and a (non-exclusive) permission to me (and, hopefully, others) to use the dataset for research and educational purposes. The data dictionary/README shall contain the names of everyone who participated in creating a dataset[1]

Submit hard copies of all extra credit documents together with your mandatory part submission.

Note: Extra credit is assessed on a per-dataset basis. That is, for example, if you submit analysis of three extra brand new, created by you, datasets, you can get 60% of extra credit. All extra credit is cumulative, and you can submit analyses of different types of datsets. E.g., if you submit analysis of one existing dataset (not supplied by me) and one dataset you have created, you may be subject to 35% extra credit.

Theoretically, the amount of extra credit you can receive is unlimited.

GOOD LUCK!

[1] Joint participation of members of different group in **creation** of a dataset is allowed!