10/23/2017

CSE 5243

Instructor: Jason Van Hulse

Final Project

Final Due Date: Friday, 12/1/2017 11:59 pm. Submit both the report and code to Carmen. *The weighting of the final project is 20% of your grade.*

Intermediate Deliverable: Please submit, via Carmen, a 1 page proposal by 11/11 at 11:59 pm. The proposal should include an outline of your proposed final project; include the names of all collaborators is you are working in a group.

The proposal counts as 5% of your grade for this assignment. Your grade will be based on turning the proposal in on time and how well thought-out your proposal is. By this point, you should have a good idea what you want your project to be, though you still have not finalized all of the coding, experiments or details. You should provide project background, justification/rationale, description of the data and software you plan to use. Provide a list of the students you will be collaborating with (max of 3 people per team), and what the expected contribution of each team member is. You will receive feedback on your proposal in a week, and may be asked to make modifications. If you deviate significantly from your proposal in the final project, please check with the TA for feedback.

Overview of the Assignment:

This homework assignment will give you an opportunity for in-depth exploration on a topic of personal interest in the area of data mining. You will apply techniques and methods learned in the course to a real-world problem domain. This project will consist of a comprehensive, end-to-end analysis of data, driven by questions that you want to answer using data mining. The goal of your project is to give you hands-on experience applying data mining techniques to one or more real-world datasets, going through the following steps:

- Identifying a problem domain and dataset(s)
- Determining what questions you want to answer via data mining
- Choosing appropriate techniques and algorithms
- Implementing and testing your methods
- Evaluating your techniques on your datasets(s)

- Reporting conclusions

You have the flexibility to choose a topic of interest to you, however you should NOT recycle work from previous classes or assignments. There are two general approaches to picking a topic:

1) Pick an application domain that is of particular interest to you. Define a specific problem of interest in this domain, and gather relevant and useful data from publicly available data sources. Apply data mining principles and methods to solve the stated problem. An example of this type of problem might be in the text mining domain, where you might want to build classification models for text categorization.

2) Pick a data mining algorithm that is of particular interest to you. You might want to implement the algorithm (for example, the Adaboost algorithm) and apply it to a set of benchmark datasets to measure performance.

Further details:

- You should include both the analysis of real-world data and some level of coding to work with or mine this data.
- You are free to use any combination of off-the-shelf and custom software, however your work **must** include some level of coding that you have implemented. This does not mean that you must code all algorithms you use from scratch, but you should choose to implement some data mining techniques/ algorithms, or perform a significant amount of data management/transformation/ cleaning, as part of the project.
- In your report, you should discuss the tools, methods and technology stack used, and what your coding contribution was (as opposed to only using off-the-shelf software).
- Simply pulling a dataset off of the web and applying a suite of off-the-shelf classification algorithms would not receive a high mark because there is limited custom coding involved, unless you had to perform a significant amount of data preprocessing or transformations of that data.
- You should hand-in both your code as well as a well-written report (12-15 pages max for single person, 18 pages maximum for teams).
- Supply appropriate references in a bibliography, if appropriate.

Sample Topics you could focus on (not necessarily a complete list - feel free to review course reference material for additional ideas, or simply ask):

a) Feature Selection

b) Cost sensitive learning
c) Graph mining
d) Web mining
e) Mining social network graphs
f) Semi-supervised learning
g) Collaborative filtering
h) Mining data streams

Group work notice:

It is **preferred** (though not required) that you work on this project as part of a team (max of 3 people per team). Note that as the number of people in the group increases, expectations for the scope and scale of the tackled problem should grow significantly. **Please include a statement in the report stating the contributions of each member of the team.** Make clear the contributions of each individual on the team; it is possible that team members may received different grades depending on their unique contribution.

Evaluation: In addition to the notes listed above, the following dimensions will be considered in grading:

1) Quality of the report: your report should be well written, clear and concise. Clearly explain your problem statement, domain and approach. Justify any design choices that you make.
2) Level of ambition for your project: this is particularly important with 2 or 3 person teams - your project should be relative ambitious in scope. Your project grade will *not necessarily* be correlated with the performance you report.
3) How closely your work adheres to standard data mining practices: A simple example of this would be the use of appropriate model validation practices. If appropriate, you should be following along with the knowledge discovery process we've discussed.
4) Significance of your custom coding (compared to use of off-the-shelf tools), as well as how well you **explain** your contribution. Please add comments to the source code itself to assist the grader in this regard. The grader may review your source code directly to ascertain/validate your contribution.

Please be sure to turn in both your report and code by the due date.