Northeastern University
CS6220 – Data Mining Techniques
Fall 2017, Derbinsky

Project

The purpose of the project is for your group to grapple with data mining on a dataset of your choosing:

1. Select a dataset / prediction problem
2. Apply one or more algorithms
3. Critically evaluate results
4. Submit a report (including a recorded presentation)

1 Project Scope

1.1 Minimum Viable Project

- Take one or more well-known datasets
- Figure out what pre-processing (if any) is needed
- Compare different algorithms for clustering / recommendation / etc.
- Explain which algorithms work well on which datasets and why

1.2 A More Interesting Approach

- Go find a publicly released dataset (e.g. all of Wikipedia text, Arnetminer, Yelp data), or collect your own
- Identify some questions you would like to answer (e.g. what types of food preferences correlate)
- Use multiple techniques (dimensionality reduction, frequent itemsets, clustering, recommender systems)
- Provide detailed discussion of results

An great example: `http://colah.github.io/posts/2014-10-Visualizing-MNIST/`

2 Deliverables

Each project will comprise of ...

1. a **proposal**, detailing the work to be completed (*2 weeks*);
2. **update 1**, showing preliminary results (*3 weeks*);
3. **update 2**, showing data mining results & a draft report (*3 weeks*);
4. and a final **packet**, including a written document and a link to a YouTube presentation (*3 weeks*).

The list above provides estimates of when each item is due; see the syllabus for specific deadlines. Additionally, each team member will individually submit a **peer evaluation** at the end of the semester.

2.1 Proposal

The proposal lays out the work that the team will do for the project. This document is to be at most **3 pages** and must include at least the following information:

A. The dataset(s) you will analyze

B. Questions you intend to answer

C. Algorithms you intend to apply

D. How you plan to divide up the work in your team

Depending on the quality of the proposal, and this discussion, the team may have to submit an amended proposal.

2.2 Update # 1

It is recommended that all exploratory analysis is complete at this point. The group should submit a summary of work completed, results obtained, and next steps (at most **5 pages**, including plots).

2.3 Update # 2

It is recommended that the group have a rough draft of the final packet to submit, including introduction/background, exploratory analysis, and preliminary data mining analysis (**5-10 pages**, plots in appendices if needed). This is an excellent opportunity for feedback – so the more work submitted, the better.

2.4 Packet

The final submitted packet must have the following items:

- A 5 – 10 page report (introduction/background, exploratory analysis, data mining analysis, discussion)
- A link to a source-code repository with source code, documentation of how to compile/run to reproduce the results, and links to datasets
- A link to a YouTube video with a group presentation (5 – 10 minutes, all members must participate, covers the same sections as the report)

The packet is subject to all rules of academic work, including the requirement for in text citations as well as the use of quotations for any material copied from another source.

2.5 Peer Evaluation

Group projects are sometimes looked upon as being "unfair." To combat contribution inequity, each team member's perception of the quantity of work that s/he performed and that of each team member will be analyzed against the perceptions of the team member(s). Through this process, hopefully equity will be achieved.

Each team member will submit a report rating the relative contributions of each team member (including her/himself) using a single number, as well as optional commentary. The aggregate rating for each student will determine the grade that individual receives, relative to the group grade. In order for this process to work effectively there is the need for each group member to be honest and objective; these ratings and comments will be kept confidential.

3 Grading

The team project grade is based upon ...

Proposal (10%) Includes on-time submission, as well as sufficient coverage of each required item.

Update #1 (10%) Includes on-time submission, as well as sufficient evidence of project progress.

Update #2 (10%) Includes on-time submission, as well as sufficient evidence of project progress and path to conclusion.

Presentation (20%) Includes presentation length, clarity, professionalism (e.g. attire, rehearsed content/distribution), the connection between the data and your discussion/conclusions, as well as a clear explanation of your data mining process and methods.

Packet (50%) Includes on-time submission, as well as the quality and completeness of the writing and other materials. In particular, clarity (is the writing clear), technical merit (are the methods valid), reproducibility (is it clear how results were obtained), and discussion (are the results interpretable).

As described above, each team member will submit a peer evaluation report. These evaluations are a serious statement and are used to re-distribute up to 50% of the grade on the project: if all group members agree they put in equal share, each individual grade will be equivalent to the team grade; however, those who did less will receive up to a 50% lower individual grade, and those who did more will receive up to 50% greater. If you do not submit an evaluation it will be assumed that you did not perform your fair share of the work and your grade will suffer as a result.

No late submissions will be accepted for the peer evaluation, nor the final packet.

4 Sources of Data

Here are a few websites, which have numerous datasets and/or links to other resources:

- `http://archive.ics.uci.edu/ml/`
- `http://ntucsu.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
- `http://lib.stat.cmu.edu/DASL/`
- `http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/00Index.html`
- `http://www.statsci.org/datasets.html`
- `http://www.drivendata.org`
- `https://www.kaggle.com`
- `http://www.itl.nist.gov/div898/strd/`
- `http://www.data.gov`
- `http://aws.amazon.com/public-data-sets/` (warning: really big!)

You are also welcome to seek out your own sources (e.g. via APIs, sensors).