CS 6220: Data Mining Techniques
Course Project Description

Yizhou Sun

College of Computer and Information Science
Northeastern University

Spring 2016

General Goal

In this project, you will have an opportunity to apply the data mining algorithms and techniques you learned in the class to some real-world problems.

- You can choose any problem that you are interested in, and formalize it into a data mining task.
- Get some data related to the task.
- Apply some data mining algorithms to your data.
- Evaluate and compare your algorithms.
- Submit a report, together with your data and code.
- Finally, present your project to the whole class.

Detailed Stages and Deadlines: 1. Form Groups

- 3-4 students
- Deadline: Jan. 27 (11:59pm)
- Where to submit: Blackboard
- What to submit: Group name; Group members; Group leader
- Points: 1 point.

Detailed Stages and Deadlines: 2. Project Proposal

- Deadline: Feb. 23 (11:59pm)
- Where to submit: Blackboard
- What to submit: A 2-Page proposal including
  - 2.1 Problem and goal
    - What do you want to solve?
    - Why do you think it is important?
    - What results do you expect?
  - 2.2 Formalization into data mining task
    - Which data type?
    - Which function? E.g., Frequent pattern mining, classification, and clustering.
  - 2.3 Data plan
    - What kind of data?
    - Where and how do you get the data?
    - Make sure get data in time
  - 2.4 Schedule: detailed plan of your project
- Points: 5 points
- Note: We will discuss with every group about your proposals later that week.

Detailed Stages and Deadlines: 3. Midterm Check

- Deadline: Mar. 22 (11:59pm)
- Where to submit: Blackboard
- What to submit: A Temporary report
  - A draft of report
  - Discuss about progress
  - Issues and difficulties you have met
- points: 2 points
- Note: We will discuss with every group about your progress later that week.

Detailed Stages and Deadlines: 4. Final Report

- Deadline: Apr. 27 (11:59pm)
- Where to submit: Blackboard / Course System
- What to submit: A final report, data, and code
    - Problem introduction, formalization, algorithms, experiment results, etc..
- points: 12 + 2 points

Detailed Stages and Deadlines: 5. Final Presentation

- When: Apr. 27 (in class)
- Who to present: whole class
- How long to present: 15 mins (include Q and A)
- In what form: slides (include demo if you like)
    - Motivating your audience, problem introduction, formalization, algorithms, experiment results, demo, etc..
- points: 5 points (peer evaluation)

Grading

Total: 30 points of regular credit and 2 points of extra credit

1. Group formation (1 point)
2. Proposal (5 points)
3. Midterm check (2 points)
4. Data and code (5 points)
    - Any programming language that can run in CCIS environment (Java and Python recommended)
    - Documentation
5. Final report (12 points)
    - At least two algorithms and two evaluation methods
6. Additional features (2 extra points)
    - Novelty of the problem
    - Your own data
    - More than two algorithms/evaluation methods
    - Other innovative features (e.g., new algorithm)
7. Presentation (5 points)

Grading

Collaboration Rules

- Every member in a team gets the same score (encourage teamwork)
  - Exception: the team has the right to claim someone as a free rider, and we will lower his/her score
- A table describing your division
  An example:

Task	People
1. Collecting and preprocessing data	Student A
2. Implementing Algorithm 1	Student B
3. Implementing Algorithm 2	Student C and D
4. Evaluating and comparing algorithms	Student A
5. Writing report	Student B and C
6. Slides, demo, and Presentation	student A, B

- Peer evaluation

Resources and References

Datasets

- UCI Machine Learning Repository
  http://archive.ics.uci.edu/ml/
- DBLP "four-area dataset":
  http://www.ccs.neu.edu/home/yzsun/data/DBLP_four_area.zip

Sample Projects from Previous Semesters

- Face Recognition
- Outlier Detection from Clinical Lab Data
- CCIS COURSE PLANNER
- Stylometry Classification for Authors
- MBTA Arriving Time Prediction
- Price Range Prediction for Boston Real Estate Data
- Student Application Recommendation System.
- ...

A Simple Example: Email Classification

Problem

- Determine whether a given email is spam or not

Data Mining Task

- Binary classification

Data

- UCI spam data set (http://archive.ics.uci.edu/ml/datasets/Spambase)
- Number of instances: 4601
- Number of attributes: 57
- The last column denotes whether the e-mail was spam (1) or not (0)

A Simple Example: Email Classification

Algorithms

- Naive Bayesian classifier
- Artificial neural network
- AdaBoost

Evaluation and Comparison

- Error rate
- ROC and AUC
- Speed

Have fun with your project!

$$\ddot\smile$$