Automated Learning and Data Analysis: CSC 422

Project Reports Guidelines

Thomas Price
Department of Computer Science
North Carolina State University
Spring 2019

Course Project Guidelines

Your class project is an opportunity for you to explore an interesting data mining problem of your choice in the context of a real-world data set. Below, you will find some project ideas, but the best idea would be to combine data mining algorithms with problems in your own research interest. Your class project must be about new things you have done this semester; you can't use results you have developed in previous semesters or in other classes.

Projects can be done in teams of three to four students or individually. For a four-person group, group members are responsible for dividing up the work equally and making sure that each member contributes. For each project, the TAs will be your a project consultant/mentor. Please consult with the TAs before finalizing the project proposal. The final responsibility to define and execute an interesting piece of work is yours. You are strongly urged to consult the TAs or the instructor early on if your project will rely purely on simulated data or if you intend to do a learning theory related project.

Your project will be worth 33% of your final class grade, and will have 4 deliverables:

- Proposal, Due Date: Feb 19th at 11:45PM: 1 page (1%)
- Midterm Report, Due Date: Mar 28th at 11:45PM: 4-5 pages (2%)
- Project Presentation (Date: Apr 23-25th Class Time) (10%)
- Project Final Report (Date: Apr 26th at 11:45PM): 8 pages (20%)

Note that all write-ups in the form of a NIPS paper: (Link here to the latex template). If you are not familiar with latex, please get familiar with it because latex format is becoming the only template accepted by many conferences. The page limits are strict! Papers over the limit will not be considered. Each deliverable of your project will be evaluated based on several factors:

- The novelty of the project ideas and applications. The groups are encouraged to come up with original ideas and novel applications for the projects. A project with new ideas (algorithms, methods, theory) on data mining or new, interesting applications of existing algorithms is scored higher than a project without much new idea/application.
- The extensiveness of the study and experiments. A project that produces a more intelligent system by combining several ML techniques together, or a project that involves extensive experiments and thorough analysis of the experimental results, or a project that nicely incorporates various real world applications, are scored higher.

- The writing style and the clarity of the written paper.

Project Proposal (Due Date: Feb. 19th)

A list of suggested projects and data sets are posted below. Read the list carefully. You are encouraged to use one of the suggested data sets, because we know that they have been successfully used for data mining and machine learning in the past. If you prefer to use a different data set, we will consider your proposal, but you must have access to this data already, and present a clear proposal for what you would do with it.

Page limit: Proposals should be one page maximum.

Suggest to include the following information:

- Project title
- Data set
- Project idea. This should be approximately two paragraphs.
- Software you will need to write.
- Papers to read.Include 1-3 relevant papers.You will probably want to read at least one of them before submitting your proposal
- Teammate (if any) and work division. We expect projects done in a group to be more substantial than projects done individually.
- Midterm milestone: What will you complete by Midway report? Experimental results of some kind are expected here.

Midway Report (Due Date: Mar. 28th)

This should be a 4-5 pages short report in the form of a NIPS paper, and it serves as a checkpoint. It should consist of the same sections as your final report (background, method, experiment, conclusion and references), with a few sections 'under construction'. Specifically, the introduction and related work sections should be in their almost final form; the section on the proposed method should be almost finished; the sections on the experiments and conclusions will have whatever results you have obtained, as well as 'place-holders' for the results you plan/hope to obtain.

Page limit: Midway report should be 4-5 pages (5 pages maximum).

Grading scheme for the project report:

- 70% for proposed method (should be almost finished) and experiments done so far
- 25% for the design of upcoming experiments
- 5% for plan of activities (in an appendix, please show the old plan and the revised one, along with the activities of each group member)

Project Presentation (Date: Apr. 23-25th Class Time)

All project members should be present. The session will be open to everybody.

Project Final Report (Date: Apr. 26th, 2018)

See separate documents

Project Suggestions

Here are some project ideas and datasets for this year:

kaggle.com

NBA statistics data

Click here to download the dataset: contains 2004-2005 NBA and ABA stats for:

- Player regular season stats
- Player regular season career totals
- Player playoff stats
- Player playoff career totals
- Player allstar game stats
- Team regular season stats
- Complete draft history
- coaches_season.txt: nba coaching records by season
- coaches_career.txt: nba career coaching records
- Currently all of the regular season

Project idea: * outlier detection on the players; find out who are the outstanding players. * predict the game outcome.

Netflix Prize Dataset

The Netflix Prize data set gives 100 million records of the form ”user X rated movie Y a 4.0 on 2/12/05”. The data is available: here Netflix Prize and KDD-2007.

Project idea:

1. Can you predict the rating a user will give on a movie from the movies that user has rated in the past, as well as the ratings similar users have given similar movies?
2. Can you discover clusters of similar movies or users?
3. Can you predict which users rated which movies in 2006? In other words, your task is to predict the probability that each pair was rated in 2006. Note that the actual rating is irrelevant, and we just want whether the movie was rated by that user sometime in 2006. The date in 2006 when the rating was given is also irrelevant. The test data can be found at this website.

Enron E-mail Dataset

The Enron E-mail data set contains about 500,000 e-mails from about 150 users.

The data set is available here

Project ideas: * Can you classify the text of an e-mail message to decide who sent it?

Precipitation data

This dataset has includes 45 years of daily precipitation data from the Northwest of the US:

Download Dataset:

Project ideas: Weather prediction: Learn a probabilistic model to predict rain levels. Sensor selection: Where should you place sensor to best predict rain.

Image Segmentation Dataset

The goal is to segment images in a meaningful way. Berkeley collected three hundred images and paid students to hand-segment each one (usually each image has multiple hand-segmentations). Two-hundred of these images are training images, and the remaining 100 are test images. The dataset includes code for reading the images and ground-truth labels, computing the benchmark scores, and some other utility functions. It also includes code for a segmentation example. This dataset is new and the problem unsolved, so there is a chance that you could come up with the leading algorithm for your project.

Download Dataset

WebKB

This dataset contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages.

Download Dataset

Project ideas: * Learning classifiers to predict the type of webpage from the text. * Can you improve accuracy by exploiting correlations between pages that point to each other using graphical models?

Email Annotation

The datasets provided below are sets of emails. The goal is to identify which parts of the email refer to a person name. This task is an example of the general problem area of Information Extraction.

Download Dataset

Project Ideas: * Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of ”person-name” or ”not-a-person-name”.

Object Recognition

The Caltech 256 dataset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds.

Download Dataset

Project ideas: * You can try to create an object recognition system which can identify which object category is the best match for a given test image. * Apply clustering to learn object categories without supervision.

More data

There are many other datasets out there. UC Irvine has a repository that could be useful for you project:

http://www.ics.uci.edu/ mlearn/MLRepository.html

Kaggle Competitions: https://www.kaggle.com/

Sam Roweis also has a link to several datasets out there:

http://www.cs.toronto.edu/ roweis/data.html