CSC 240/440 Data Mining
Fall 2018
Final Course Project
Instructor: T. F. Pawlicki, PhD

Contents

1 Description

The goal of the final project is to develop a data mining solution to predict an outcome, discover patterns or classify events. As discussed in class, you can work individually or in teams. The maximum team size is 2. Each team will choose from one of the provided projects / data sets.

The team's performance will be judged on the following criteria:

- Demonstrate solid understanding of the data (you could demonstrate that in a number of ways including: identify missing data and how you handled it, provide descriptive statistics of attributes, visualizations, description of dimensionality of the data set and features)
- One or more of (in increasing order of complexity and hence merit):
    - Apply an algorithm covered in class
    - Perform a comparative study using different existing algorithms
    - Apply a modified version of an algorithm covered in class
    - Apply a new algorithm
- Extra credit will be given for special achievements as applicable: e.g. getting on a leaderboard, demonstrating significantly improved performance compared to previous publications, obtaining novel results beyond what is provided in a relevant research paper, working with significantly large versions of data sets, etc.

You are welcome to use any programming language of your choice. You are also free to use packaged libraries and modules that are publicly available. However, you would be expected to clearly cite the package in the references/bibliography section of your report.

2 Data Sets

Your team can choose to work on any of the following data sets for the project. You may propose a different data set that you are interested in, but it must be similar to the data sets below in terms of size and complexity. You must get approval of the instructor before proposing an alternate data set. By providing data sets from different domain areas, my intent is that you can choose to work on a data set that aligns with your background and/or interests. Project goals for each data set is provided below as a suggestion but the team is free to develop their own project goals or additional goals if they so desire.

1. "House Prices: Advanced Regression Tech- niques" on kaggle https://www.kaggle.com/c/house-prices-advanced-regression-techniques. See the link for an overview description.

A suggested goal is to predict the final price of each home in the test data set, with a large number of features describing (almost) every aspect of residential homes. A very good detailed analysis visualization of some features can be found at https://www.kaggle.com/pmarcelino/house-prices-advanced-regression- techniques/comprehensive-data-exploration-with-python/run/879421. (Here you may also learn how to read in the data using pandas).

2. Movie Recommendation System using the MovieLens data set (https://grouplens.org/datasets/movielens/

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org). The data sets were collected over various periods of time, depending on the size of the set.

A suggested goal is to build a recommender system that can predict a given user's ratings for movies using different algorithms and compare their performance. Collaborative filtering (CF) is a technique (algorithm) popularly used by recommender systems that could be explored.

Note that a number of different data sets (varying sizes) are available on the GroupLens website. The sizes vary significantly. The 1 MB data set (labeled "small") provided at: https://grouplens.org/datasets/movielens/latest/ could be used for this project. Larger data sets could also be used if the team would like to experience working with large data sets and scale their results. Note that in this case, the team should plan effectively how to handle the larger data sets in terms of required computational resources, techniques etc.

3. Predict the popularity of an online news article (https://archive.ics.uci.edu/ml/datasets/online+news+popularity)

The goal is to develop a multi-class classifier that predicts the labels for a test data set. Observations are online news articles and the goal is to predict the level of popularity of the article, i.e. number of shares in social networks. The data comes from website mashable.com.

4. Predict impact on re-admission rates for patients hospitalized with diabetes (https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008#)

This data set provides the ability to predict the Impact of HbA1c Measurement (i.e. measurement of blood sugar levels) on Hospital Readmission Rates. The research paper: https://www.hindawi.com/journals/bmri/2014/781670/ provides good background information and insight on the topic.

A suggested goal is to build an algorithm that can successfully predict probability of re-admission within a pre-defined period of time, i.e. predict if the class labels (corresponding to time to re-admission) would be "<30", ">30" or "none". The teams are encouraged to work on goals and approaches beyond what is demonstrated in the above referenced research paper. Extra credit will be given for such novel approaches.

3 Grading

Your team is expected to prepare and submit the following items as part of the project. The weightage for each component and grading criteria are listed below.

Deliverable	Weightage	Criteria / Items to include
Proposal	15%	• Describe project choice and rationale • Team composition • What is the goal of the analysis? • Planned technical approach to solving the problem (It is understood approaches evolve during the project, but having a high level vision is important) • Role of team members
Final Report	70%	• This grading component will significantly depend on the performance of your algorithm, and results achieved. • The typical structure of a project report includes the following sections: title, author(s), introduction, related work, methodology, experiment, conclusion, and bibliography (either ACM or AAAI format)
Code	15%	• Commented code with Readme file on how to run the code to obtain the results

4 Schedule

Item	Submission Deadline
Project proposal	November 5, 2018 at 11:59 pm
Project report with code and presentation slides (as a zip file to BB)	December 12th, 2018 at 11:59 pm