QAC385-01 Machine Learning--Data Mining

Course Project

Overview

The goal of the class project is to identify a problem that interests you and that can be approached using machine learning. This can be any techniques covered in the course. Use ML best practices in attempt to solve the problem. Projects will be completed in teams of 3-4. You may form your own teams. Anyone who is not on a team will be placed by the instructor.

Each team will present their findings during the last week of class (in 10-minute sessions). This is a chance to show off what you have learned. A written report is due during finals week.

Data Sources

You may use any data source you wish. You are looking for a large dataset that you can apply machine-learning techniques (including training/testing or cross validation) to. Here are some potential sources. You are *not* limited to these.

Source	URL
Kaggle	https://www.kaggle.com/datasets
Data World	https://data.world/
15 Free Public Data Sets	https://www.springboard.com/blog/free-public-data-sets-data-science-project/
Awesome Public Datasets	https://github.com/caesar0301/awesome-public-datasets
Awesome Data Science	https://github.com/bulutyazilim/awesome-datascience#data-sets
Quandl	https://www.quandl.com/
Enigma Public	https://public.enigma.com/
Data.gov	https://www.data.gov/
Pew Research Center	http://www.pewresearch.org/data/download-datasets/
ICPSR Social Science Data	https://www.icpsr.umich.edu/icpsrweb/
KDnuggets	http://www.kdnuggets.com/datasets/index.html
Registry of Research Data Repositories	http://www.re3data.org/
US Stock Fundamentals Data Archive	http://www.usfundamentals.com/download/
UN Data	http://data.un.org/Explorer.aspx
Archive.org	https://archive.org/details/datasets
Google Collection of Public Datasets	https://www.google.com/publicdata/directory
Microsoft Data Science for Research	https://www.microsoft.com/en-us/research/academic-program/data-science-microsoft-research/

Choosing a dataset

1. Select a topic that interests you – you will be studying it for an entire semester, and you want to make sure the topic will sustain your interest.
2. The data set should be relatively large (> 800 cases) and have a reasonable number of variables (10+). These are only suggested guidelines.
3. Avoid datasets that are primarily text data. In this project, you will be using at least three machine learning approaches and the difficulty of the task will be much greater if at least some of the data is not numeric.
4. The goal of the project is to effectively predict an outcome (categorical or numerical). We do not cover recommender systems in this course (see QAC305), so avoid datasets that would be used exclusively to predict a list of recommended products for users.

The Task

Ask a meaningful question and attempt to answer it with machine learning techniques and publicly available data. Apply at least three (3) approaches to a training dataset and evaluate their performance appropriately. Select the best approach given these results and evaluate its performance on a hold-out test dataset. Interpret your findings, including a discussion of the effectiveness of the final model, what you've learned about the importance of the variables initially considered as predictors, and recommendations going forward.

Deliverables

Each deliverable needs only to be uploaded **once** (i.e., by one member of each team).

1. **Project proposal**
   Upload the project proposal. Identify the problem you want to tackle and the data source to be used. Provide a problem statement, description of the data, and data source to instructor. The description of the data should include details and summary statistics for each variable.

2. **Presentation slides**
   Upload your slide presentation to Moodle. This can be in PowerPoint, Google Slide, or PDF format.

3. **Class Presentation**
   Present the results of your project to the class. Each group will have 10 minutes to present.

4. **Final Report**
   Upload the completed final report to Moodle.

Class Presentation

Your group will have 10 minutes to present your final project to the class. Each group should create an attractive slide show to support their presentation (max 10 slides). The presentation should include the following:

(1) The problem you are trying to solve
(2) A brief description of the data used to solve it
(3) A brief description of the methodology employed, including any difficulties that had to be overcome.
(4) The findings (use graphs here!)
(5) The implications and limitations of the study.

Since time is short, practice your presentation to get out the kinks.

Project Report

Include the following sections in your project report

Title and Authors

Include the title of your project and the names of the team members.

Problem Statement

What is the problem you are trying to solve? Why is it important or useful to solve this problem?

Data Description

Describe the dataset in detail. What is the source? How was it collected? What are the variables? What are the characteristics of the variables?

Data Preprocessing

Describe any variable transformations, treatment of missing values, recoding and any other data manipulations completed prior to applying machine learning techniques.

Machine Learning Approach

Describe your analytic techniques in detail. Assume that your audience is *not* familiar with the techniques that you are using.

Results

Describe the results in detail. What did you find? This the section for tables and graphs. You want to communicate your results as clearly and compellingly as possible.

Discussion

How well were you able to solve the problem? What are the implications? What suggestions do you have for other researchers who want to take your work further?

References

Standard reference section (similar to any term paper)

You will be provided with an RMarkdown template to use when writing your report.

Uploading the finished product

Upload the finished product as 3 separate files:

(1) The .Rmd file that generated your report and includes all R code.
(2) A .csv file containing the cleaned data
(3) The well formatted project report as an HTML document

Grading (26% of course grade)

The project will be graded as a whole. Each team member will receive the same grade. Late submissions will not be accepted except under extenuating circumstances.