CS 5821: Machine Learning

Group Project: Exploratory Data Analysis using Machine Learning Techniques

Official Start Date: Tues. Oct 22

Official End Date: Week of Dec. 3

Overview

The primary purpose of this exercise is for students to gain practical experience in data analysis using machine learning techniques: explore data sets and apply machine learning methods to seek answer(s) to question(s) that students find interesting. A secondary purpose is to facilitate information sharing so that students can learn from each other. As such, students are expected to present / demonstrate their work, view others' presentations, and participate in Q&A as appropriate.

Programming platforms / Languages / Tools

Students may choose any set of tools and language to complete the project, but naturally there should be intra-group commonality. Also, there is an expectation that results will be presented graphically using plots, etc.

Procedure

1. Form a group of 2 to 4 people.

2. Register your group (students' names) and topic (or title) no later than Oct. 29th in eLearning.

3. Decide on a topic of interest. The decision should be based on interest and availability of data. (see below)

A. Standard Topics

- Spam / Non-spam email filter
- Web content filter (block objectionable web content): Unit (entire websites or selected parts)? Define "objectionable"…
- An interpretable predictor of house prices for any specific city or county.
- Rank the rankings: among the popular rankings of world universities, which one is the most trustworthy and why?

B. Create Your Own

Typically need a problem statement or hypothesis, and probably some appropriate assumptions. For example, is X R Y? In other words, is X related to Y? This relationship may be causal (cause and effect), correlated (similarity), coincidental (by chance), concurrent (simultaneous), etc. More concrete examples include:

- Does smoking cause lung cancer?
- Do consumers who buy apples also buy bananas and carrots?
- What is a balanced diet?

- Does it make sense to buy an electric vehicle in 2021?
- Will data scientists still be in demand by the time I graduate?
- How will automation impact society in 2030?
- Can we use X to predict stock prices t time units in advance?
- Can we use X to predict box office results t time units in advance?
- What is the true cost of palm oil?
- Will renewable energy sources (collectively) replace fossil fuels as the chief sources of energy by 2030?
- How will the makeup of energy sources in 2030 differ from today's makeup?
- How will a 1% tax reduction affect the economy? What tax (personal income, corporate, etc.)? Is there an optimal rate?
- Is the income / wealth gap widening in _____ (nation, region, city, etc.)?
- What will be the size of the Amazon rainforest in 2030, both in absolute term and relative to its current size?

N.B. Some of the above questions are **still quite vague and will require further refinement.**
Be creative. A surprising outcome is often an interesting one (high information content).

C. Recommended Report Structure

0. Abstract (~200 words) that summarizes the report succinctly.
Suggested format:
Problem Statement / Hypothesis
Methods Used (and why?)
Key Findings (and significance)

1. Introduction
2. Methods and Results (numbers and/or plots with explanation)
3. Key Findings and significance, distilled from 2. Justification of "best method" for the task.
4. Conclusion (may include your personal reflection of what went well or what did not)
and suggestions for possible extensions)
5. List of References

Grading Scheme (out of 20)
I. Report (~5 pages): 15
II. Presentation (approx. 10 minutes) clarity, organization, effectiveness, timekeeping, teamwork): 5
III. Best presentation award: 3 extra points for winning team(s), by popular vote.

Report Specifications
1. Should be about 5 pages excluding a cover page and list of references.

2. A cover page is required. It should include the following: group number or group name, students' names, project title, abstract.

3. A Contents Page is optional; it is useful if there are many sections / subsections in the report.

4. References should include full bibliographic information, including full list of authors (not xxx et al.), title, journal or conference name, pages, year, and other relevant info, e.g. DOI or URL if available.

5. References should be listed in the order that they appear in the main text.

6. Use headings and subheadings appropriately to aid readability.

7. One PDF file of the report is to be submitted per group.

8. The submitted file must be named “GroupX.pdf”, where X is your group number or name.
Code associated with the project should be uploaded as a separate file. Alternatively, use R markdowns, jupyter, github, etc.