95-885 Data Science and Big Data, Fall 2019

Project 2: Adding Value via Machine Learning

Learning Objectives

In the first project, we looked at exploratory data analysis. The goal of that project was to gain experience in understanding the "lay of the (data) land" and to perform descriptive analytics on your data set.

In this second and complementary project you will explore a theme in machine learning. Consider a task T, experience E in performing that task, and a performance metric P measuring how well we perform task T. Learning (in machines and humans) aims to provide a boost in performance with more experience ΔE i.e.,
P(T, E+ΔE) > P(T,E).

Recall that machine learning tasks are categorized into three broad categories (i) supervised (ii) unsupervised (iii) reinforcement learning. Within these categories, for this project, you are welcome to explore any machine learning task of your preference e.g., classification, recommendation, prediction, clustering, association etc. Scikit-learn and Spark MLlib provides a variety of tools for machine learning. You will perform and report your work with a Jupyter Notebook.

At the highest level, in this project, you will have the following intertwined activities (a) pick a data set of your choice (b) in tandem identify a machine learning task (c) perform experiments using the data set and library of tools (d) report your experience and observations. As has been the theme of this course, you are expected to go beyond the just running your data set on a range of ML algorithms. Be sure to focus on the value of the ML work you do. In what way or form would this benefit an organization? As an example, if you are going to predict the arrival delay of various airlines, what value does this add? Can you translate the delayed arrivals to loss revenue? Customer dissatisfaction? Airline reputation?

Related Projects: If you are doing a project in another course and would like to extend that work as the project for this course, please check with the instructor. You are welcome to work on such related projects provided (1) the two projects are sufficiently different to provide different learning experiences (2) you check with the instructor of the other course that doing the related projects is acceptable.

Project Stages

1. *Project Proposal (20% of project credit) /* 1-2 pages.
   Due: 5:00 pm Monday, November 18 (13th week)
   How: Submit to Canvas

As with all projects, the proposal will lay the foundation of work to follow. Provide an overview of the project goals. What type of learning do you plan to investigate? What is the domain? With reference to, P(T, E+ ΔE) > P(T,E), clearly identify the T, P, and E for your domain. What is your motivation for this work? Did a news article, blog post, web site, something we discussed in class trigger your interest? Briefly discuss the data you plan to use. Where did you find it? What is its structure? It is important to discuss in detail how you plan to assess your work (the P part).

A couple of hours search on the web will help you identify a range of machine learning problems that could be potential projects. To seed the search, following are some websites of projects from courses on machine learning:

- Stanford, 2018: http://cs229.stanford.edu/proj2018
  Stanford, 2016: http://cs229.stanford.edu/projects2016.html
- UT Austin: http://www.cs.utexas.edu/~mooney/cs391L/project-topics.html
- Oklahoma: http://www.mcgovern-fagg.org/amy/courses/cs5033_fall2017

Please keep in mind that these are projects from courses exclusively devoted to machine learning. Our course has had a much broader *Data Science* perspective. The course instructor and TAs will provide feedback on your proposal and assist with right-sizing it.

2. *Intermediate Demonstration of Progress (20% of project credit)*
   Due: Before 11:59 pm Monday, November 25 (14th week)
   How: 10-15 minute in person meeting with TAs or Instructor

Setup a time to have a brief meeting with the TAs or course instructor to discuss and demonstrate what has been done to date and what remains to be done. The exact details of this will vary from project to project. We will be looking for a good faith effort demonstrating progress towards project completion. Before you meet with your TA, you will submit the current state of your work in the form of a Jupyter Notebook so that the TA can review project status before meeting in person.

3. *Project Notebook(s) (30% of project credit)*
   Due: Before Friday, December 6, *11*:59 pm (15th week)
   How: Submit to Canvas

   Report your full work as an Jupyter notebook. You should detail your problem domain, datasets, methods, assumptions and approaches you have used in your analysis. Your report should also detail your findings in appropriate technical and 'business' language. Include all useful supporting code, charts, graphs, or summaries. There should be enough detail so that a reviewer can clearly understand, recreate your analyses and evaluate the credibility and soundness of your approach.

4. *Video of project presentation (20% of project credit)*
   Due: Before Sunday, December 8, 11:59 pm (16th week)
   How: Submit to Canvas

   Rather than an in-class presentation, you will submit a video of your project work. You will discuss your domain problem, the type of machine learning, what you did, results. Present your project in a way that those unfamiliar with the domain or data will understand. The video length will be strictly limited to 9 minutes.

5. *Project Video Peer Reviews (10% of project credit)*
   Due: Saturday, December 14, 11:59 pm (EDT – Pittsburgh time)

   Every student will be assigned 5 videos to watch and review. For each of the videos you watch you will submit a review (based on a form we will provide). Your reviews will be assessed for depth, critical assessment, and useful feedback.

Assessment:

These following terms are used to describe your work on this project:

A. *Outstanding*. Deliverables exceed requirements in all respects. Quality of work / reports are outstanding in terms of content, analysis, thoroughness, clarity of thought and expression, as well as quality and depth of insights. Notebook, presentation, screencast are all very well prepared and clearly presented.
B. *Good*. Deliverables meets requirements in all respects and may exceed requirements in some respects. Content, analysis, clarity of thought are good and reports and demonstrations provide some insights into subject matter. Deliverables are well organized, well written and presentation is clear.
C. *Satisfactory*. Deliverables meets requirements in some respects but may be inadequate in some respects. Reports and demonstrations demonstrate basic effort in terms of thought, expression, or analysis. Quality and depth of insight or research is acceptable, but results or analysis are apparently thin or minimal. Conclusions,

details of project plan or supporting documentation and argumentation may be questionable or not well supported. Appearance, lines of argument and/or mechanical details are adequate, but attention to detail is needed.

D. *Unsatisfactory*. Project work generally does not meet requirements. Deliverables are shallow, unconvincing and/or poorly written or presented and there is little to commend it.

Peer Evaluations:

This is a team project. It is expected that each member will contribute equally and effectively towards all aspects of the project (research, development, deliverables, presentation preparation etc.). Peer evaluations will be used to adjust for individual contributions. Please see the course instructors early if there are any unresolvable concerns.