67-364 / 95-885 Data Science and Big Data, Spring 2017

Project 2
Exploring Machine Learning

Learning Objectives

In the first project, you 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 provides a variety of tools for machine learning. It is preferred that you use scikit-learn + a Jupyter notebook, but you are welcome to use a tool of choice. If choosing an alternative to scikit-learn please check and discuss with the instructor.

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.

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 April 17 (both Tue/Thu and Fri sections)
   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:

67-364 / 95-885 Data Science and Big Data, Spring 2017

- Stanford, 2013: http://cs229.stanford.edu/projects2013.html
- Stanford, 2012: http://cs229.stanford.edu/projects2012.html
- UT Austin: http://www.cs.utexas.edu/~mooney/cs391L/project-topics.html
- Oklahoma: http://www.cs.ou.edu/~amy/courses/cs5033_fall2014/index.html

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 5pm Friday, April 28 (both Tue/Thu and Fri sections)
   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.

3. *Class Presentation (30% of project credit)*
   Tue/Thu section: May 11 8:30-11:30 am
   Fri section: May 12, 1:00-3:40 pm

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.
Presentation time will be strictly limited to 9 minutes each (presentation + Q&A).

4. *Report (30% of project credit)*
   Due: 11:59 pm May 11 (Tue/Thu section)
   11:59pm May 12 (Fri section)
   How: Submit to Canvas

   If convenient report your work as an Jupyter notebook. But you are welcome to submit a report as a pdf document also. 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.

67-364 / 95-885 Data Science and Big Data, Spring 2017

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.