EE 660 **Project Assignment** Posted: Mon., 10/12/2020
Jenkins See timeline below for due dates

Introduction

For this project you will pick your own topic and design your project. You are encouraged to pick a topic (or dataset) of interest to you, and that is appropriate for a machine learning class project.

You will submit a project proposal (as your Homework 6), a final written report that describes your approach and results, and your computer code. A timeline of due dates and grading criteria are given at the end of this assignment.

Types of Projects

There are two overall types of projects; you may choose either one for your project.

(1) Type 1 project. Solve a machine learning problem by implementing a machine learning system of your own design, that uses real-world data. For this, you will choose one (or more) set(s) of real-world data, and define the goals of your project. For example, the goal of your project might be to use regression or classification techniques to predict the output attribute y as well as possible. You could additionally include other goals, such as understanding what the limitations in your final system are caused by; investigating the attributes that are most predictive, and assessing why; etc. You will typically have other issues to address as well, such as number of data points N not being ideal, missing or noisy data, imbalance of data set, categorical feature values, preprocessing steps, etc. See the "Dataset Tips" document for suggestions of where to find datasets, and criteria for sifting through them to find one appropriate for a class project.

(2) Type 2 project. Perform one or more experiments in machine learning. The experiments would typically use synthetic data, so that the data can be controlled and varied in various ways; synthetic data also allows you to generate "unknowns" to numerically estimate the out-of-sample error directly. It might also be applied to real-world data to assess the effects of realistic data.

This would typically also involve some theory – either to predict what would happen, or to help interpret the results of what did happen. Experimental work would typically have a statement of what will be learned from the experimental results, or a prediction of what is expected; and explanations and interpretation (after the experiment) based on some theory, intuition, or conjecture. Or, a project might start with a theoretical component that develops some predictions, and then run some numerical experiments to test them.

A good example of an experiment is Sec. 4.1.2 of AML, especially Exercise 4.2, including the results shown in Fig. 4.3 and some of its interpretation.

Suggestion: If you're not sure what you want to do, you can try the following.

(1) For a Type 1 project, start by finding a dataset that you're interested in, and develop a project and goals based on that data. Or, you can also browse through Kaggle competitions to get an idea of what kinds of topics could constitute a project.

(2) For a Type 2 project, you can choose some aspect of class material you find interesting, and pose some questions of how some variables would depend on others; especially where it isn't obvious, where we haven't given examples that show the dependence, or where you can think of a lot more to try than in the examples we covered in class.

Guidelines and Ground Rules

Groups: You may do your own individual project, or you may work in a team of 2 students. Teams of 3 students may be allowed in cases that clearly warrant it. Your project will be graded accordingly; that is, 2 students should accomplish about twice the work of one student (or solve a problem that is an appropriate factor more difficult). Note that if you work in a team, you will submit one project final report together. All students should participate in writing the final report. Moreover, the report should clearly state the contributions each student made to the project. Usually all students of a team will receive the same grade for the project, although different grades may be assigned in exceptional cases.

Your course project must be work that you do specifically for this course. If you want to do a project that is on a topic you have worked on previously, or are currently working on (*e.g*., as part of your research, or a project for another class), that is OK. But, you must clearly distinguish between what is done for EE 660 this semester, and what is done for other purposes (*e.g.,* research or other class work). In your proposal and your final report, you must include a brief summary of the other work and describe how the EE 660 project work is distinguished from it. Also, consider how much background information will need to be described in your project report for the project work to be understandable to people that may not have the domain knowledge you have; too much would imply it's not a good topic for a class project.

Code - writing your own vs. using available code from the internet. OK to use code from the internet - be sure to state so in your report. It's also OK to write your own code in the language of your choice*. Keep in mind that your project topic should be focused on machine learning issues. Spending almost all your time coding up a well-known but complicated algorithm will not leave you much time to do anything else. (Likewise for coding a lot of feature extraction.) On the other hand, if your project consists of running lots of different algorithms from the internet without understanding what the algorithms are doing, then you are missing the point of the project.

Suggestion: Best to use only standard libraries, and code up what else you need yourself; and for functions/methods you use from libraries, make the effort to understand what they actually do.

Data: For real-world data, it is recommended to use dataset(s) that are publically available on the internet. You may also acquire your own data. However, be advised that data gathering (and subsequent processing of it to make it usable) can be very time consumptive, so think this through carefully during your planning/proposal stage if you want to acquire your own data. A team effort can make acquiring your own data more feasible.

Suggestion: Try to make the size of your project big enough to be interesting to you or your team, and to not be a trivial project; but small enough to be consistent with the amount of time and resources available. Keep in mind we will also have homework assignments during the project period, although we'll generally keep them shorter than they were in the first half of the semester to help give you time to work on your project. Also consider the computational resources you have, and the likely amount of computation needed for your proposed project (for example, datasets with 1 million data points will likely eat up a lot of computational resources if you use the entire dataset).

Requirements

Your project is required to include the following elements.

Significant machine learning content. This should be the main part of your project, and will include the use of ML concepts, techniques, and algorithms. It will also include some understanding of, or insightful attempts at understanding, results that you are observing (intermediate results as well as final results).

Use of real-world data for Type 1 projects, or use of synthetic data and/or real-world data for Type 2 projects, as described in project types above.

Complexity analysis. Some consideration of complexity of your approach wherever reasonably possible. This could include complexity of the model(s) used and hypothesis set(s), the number of data points, and anything known or relevant about the underlying target function. If it isn't tractable to analyze the complexity mathematically, then a rough estimate using principles like degrees of freedom, perhaps accompanied by some numerical experiments, should be done. Whatever method you use, it should help you make good choices in developing your model(s), managing the number of data points, size of test set, etc.

Estimation of out-of-sample error. Some valid method(s) for estimating the out-of-sample error, or predicted error on unknown (new) data. Ideally, this would include application of some theory as well as some numerical results. A simple example for Type 1 projects based on classification, is to use a true test set, and to use a theoretical error bound to estimate error bars on the true out-of-sample error. A simple example for Type 2 projects using synthetic data, is to numerically estimate the out-of-sample error by drawing new data points; a sample mean and sample standard deviation can be used to estimate the out-of-sample error and its error bar. In this case, it could also be interesting to compute the theoretical out-of-sample error bound, if possible, for comparison.

Reporting and interpretation of intermediate (or multiple) results. For Type 1 projects, this would typically be done using validation set(s), with or without cross-validation. Accumulating a numerical estimate of mean and standard deviation of the (cross-)validation error can give intermediate results to be interpreted or explained. For Type 2 projects, this will depend on the experiments being performed, and could involve results of smaller experiments that together comprise a larger experiment, or merely a set of different results from one overall experiment.

Interpretation and understanding of your methods, results, and procedures. Your report should demonstrate that you have an understanding of what you are doing and discovering. Where the reason behind some results or findings are unclear, state so and try to make a conjecture that could explain it, and/or suggest an experiment that could shed more light on the issue.

Baseline system and comparison with it. For Type 1 projects, clearly describe your baseline system(s), and how you evaluated their performance. Compare with your final system's performance. It is often advisable to have 2 baseline systems: (i) trivial and (ii) non-trivial. Examples of trivial and nontrivial baseline classifiers will be given in Discussion 8.

Description of how the data was used - training sets, validation sets, test sets, any cross-validation loops, etc. You should use your datasets in a valid way. Consider using a diagram or flow chart to make your description clear. This may be included in the next item below rather than a stand-alone description.

Description of the overall procedure (methodology) followed. For example, this could be a list of steps, sequence of paragraphs, or flow chart showing, for example: drawing data samples, choices of hypotheses, preprocessing, separation of data into various sets, training algorithms, model selection, feature selection, choosing parameters and validation, final choices, and final testing.

***Allowed languages** are MATLAB, Python, C/C++. If you want to use other languages, check with the TAs or instructor first.

Methods and techniques you can use. A minimum of 50% of your project work should use methods and techniques covered in EE 660. This includes topics already covered in class, as well as topics we haven't yet covered (refer to the course outline and Discussions 7 and 8 for upcoming topics). You can also include methods and techniques from EE 559, and from outside of both classes; but these (combined) should constitute less than 50% of your project.

Citation of others where appropriate. This applies to both your project final report and your code. In the final report, any statements taken from other sources must be cited and referenced as such. Similarly, any results of others that are stated in your report must also be cited and referenced. Instructions for doing this will be included with the Project Final Report Instructions (to be posted later). Any code that is taken from elsewhere and used in your project, must be commented as such in your code. *Failure to cite other*

sources where appropriate amounts to plagiarism, and will result in deduction from your project score. In egregious cases, your final course grade will be lowered directly, as a penalty.

Comment: Details and instructions for the final report will be posted later.

Grading Criteria

Criteria used to grade the projects will include: workload (difficulty of problem, amount of work), technical approach and execution, data handling (correctness and appropriateness), performance (correctly estimated or evaluated; comparison with baseline system(s) and work of other people if available), analysis (understanding and interpretation), and write up (clarity, completeness, conciseness).

Timeline

Item	**Date**
H6W8 posted (Dataset Information Form(s) and Project Proposal Form)	Fri. 10/16
H6W8 due (Dataset Information Form(s) and Project Proposal Form)	Fri 10/23, 5:00 PM PDT
Final project reports and computer code due	Thur., 12/3, 3:00 PM PST