EE 660 **Project Assignment** Posted: Tues., 10/22/2019
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 HW9), 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 types of projects to choose from:

(1) Project of your own design, based on 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) Project based on exploring some Machine Learning issues experimentally or theoretically. Experimental work would typically use synthetic data, so that the data can be controlled and varied in various ways. It might also (or instead) be applied to real-world data to assess the effects of realistic data. For example, you might investigate the effects of violating one or more of our assumptions, which might include: test data or validation data being completely separate from training data; iid assumption of samples drawn. 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 and intuition. Theoretical work would develop some theoretical predictions, and then run some numerical experiments to test them.

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

Guidelines and Ground Rules

Groups: You may do your own individual project, or you may work in a team of 2 or 3 students. 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: 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 or less frequent 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 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, or synthetic data for numerical experiments 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 (number of learning variables), 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 is to use a true test set, and to use an error bound to estimate error bars on the true out-of-sample error, and/or numerical techniques to get an error bar on the test-set (or validation set) error.

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.

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 (see the course outline 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.

Suggestion: Browse through the rest of the course outline before completing your project proposal; feel free to include some techniques and methods we have yet to cover. If you choose to change your plans later based on what you learn, that is OK. Do avoid topics near the end of the semester (e.g., unsupervised learning), unless you are willing to read ahead and learn them on your own to include them in your project work.

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 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 work of other people if available), analysis (understanding and interpretation), and write up (clarity, completeness, conciseness).

Timeline

Item	**Due**
Project proposals (HW9) due (Dataset Information Form(s) and Project Proposal Form)	Wednesday, 10/30, 2:00 PM
Graded project proposals (with comments) returned	Wednesday, 11/6
Final project reports and computer code due	Friday, 12/6, 2:00 PM