ECE443 Machine Learning for Engineers

Term Project Instructions
March 21, 2022
100 points

The term project needs to be completed on an individual basis. It has two phases, with two distinct deadlines, as described below. All submission times, unless otherwise stated, are 11:59 PM Eastern Standard Time and there is no possibility of an extension.

1 Phase I: Due April 4, 2022

Phase I of the term project involves the following deliverables in a single PDF file.

1. *(5 points)* **Declaration of project datasets**: The project needs to revolve around two different datasets, with each one being different from the other in terms of nature/modality, as per the following requirements:

   - Each one of the two datasets can be an imaging dataset, a video dataset, a time-series dataset, a text dataset, a sound/speech dataset, a dataset with a mix of independent variables, etc.
   - At least one of the declared datasets must have categorical independent variables within it.
   - The minimum number of samples per dataset must be at least 500
   - The minimum number of raw features (attributes) per dataset must be at least 50

   Specific deliverable: Provide a brief summary of each dataset, as well as the source URL for each dataset, as part of this Phase I submission.

2. *(15 points)* **Declaration of project tasks**: Specify the machine learning tasks that will be performed in relation to each dataset (one, and exactly one, task per dataset). Collectively, when looking at the two datasets together, you must tackle two out of three tasks of classification, regression, and clustering. That is, all classification (or regression or clustering) tasks are not acceptable. Note that you do not need to finalize (or declare) the methods you will use to solve the tasks.

   Specific deliverable: Briefly discuss what motivated you to select the declared datasets and corresponding tasks.

   Some online resources for datasets:

   - `https://www.kaggle.com/datasets`
   - `https://archive.ics.uci.edu/ml/datasets.php`
   - `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
   - `https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research`

2 Phase II (Due: May 2, 2022)

Phase II of the term project involves the following set of deliverables.

1. **Creation and submission of two separate notebooks, one for each of the two datasets/tasks:** The declared machine learning task for each dataset must be carried out in a well-commented Jupyter notebook, with careful discussions/explanations in markdown cells of the notebook. Additional guidelines for this part of Phase II of the term project include:

   - While you are allowed to use packages such as `sklearn` for this phase of the project, you must only use those modules/functions that you fully understand and you must explain the usage of those modules/ functions in markdown cells.
   - You must name each one of your notebooks as `<LastName><FirstName>_Dataset<n>.ipynb`, where you should replace `<LastName>` with your last name, `<FirstName>` with your first name, and `<n>` with 1 and 2 for each one of the two datasets. As an example, if I were to submit these files, I would name them as `ShirinJalali_Dataset1.ipynb`, and `ShirinJalali_Dataset2.ipynb`.
   - You must ensure that your submitted notebooks are fully executed, so that they are not required to be rerun during grading.
   - The point breakdown for this part of Phase II of the term project in the following should guide your code development for each notebook.

   (a) *(5 points)* **Brief exploration of each dataset**: Carry out a brief exploration of the dataset, such as the number of samples, the number of raw features, the fraction of missing values (if any), the number of categorical variables (if any), histograms of different variables, etc. This exploration should be accompanied with detailed commentary in markdown cells.

   (b) *(5 points)* **Pre-processing of each dataset**: Carry out preprocessing of each dataset, which should be guided by your exploration of the dataset as well as your forthcoming plans for the datasets. This preprocessing could involve, e.g., replacement of invalid entries with plausible values, centering of the data, standardization of the data, encoding of categorical variables, etc. All of the preprocessing steps should be fully motivated and justified in markdown cells.

   (c) *(5 points)* **Feature extraction / feature learning from each dataset**: Depending on the dataset, engage in either feature engineering or feature learning for that dataset. In the case of text dataset, e.g., this would involve transforming the raw text into numerical features. In the case of large images or correlated numerical variables, e.g., this could involve using something like principal component analysis (PCA) to reduce the dimensionality of images or to decorrelate different variables. All of the steps involved in this feature extraction / feature learning component should be fully motivated and justified in markdown cells.

   (d) *(40 points)* **Processing of each dataset using two different machine learning methods**: Carry out the declared task on each dataset using two different machine learning methods, with the parameters for each method (where applicable) carefully tuned using cross-validation, the results averaged over multiple validation folds, and the final results presented in an aesthetically pleasing manner. In addition, use markdown cells to justify different steps in your implementations and explain different aspects of the two methods as much as possible.

(e) *(10 points)* **Comparative analysis of the two methods on each dataset**: Provide a comparison between the two machine learning methods for each dataset across dimensions such as computational complexity, performance, etc., and a final recommendation on the method that should go into production for each dataset. This comparison should include both coding cells (e.g., overlayed plots, side-by-side confusion matrices, etc.) and markdown cells for discussion.

2. *(15 points)* **Video Presentation**. Prepare a 10 minutes (or shorter) video presentation that summarizes your efforts as part of the term project. The presentation can be a mix of slides and snippets of code and markdown cells from the two notebooks, and its purpose is to convince the audience that you fully understand the different aspects of the submitted notebooks. The presentation should be uploaded on Canvas on its respective assignment.