ECE 443 (Fall 2022) – Term Project Instructions (100 points)

Last updated: November 5, 2022

The term project has three main phases, with three 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. A student/team failing to deliver on one of the phases will forfeit points for the subsequent phase(s).

1 Phase I (Due: November 18, 2022)

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

1.1. (2 points) **Formation of teams:** Each team must have no more than three and no less than two members, with a designated point-of-contact (POC); the POC is the only one who would submit the files required for grading of the different phases of this term project.

Specific deliverable: Provide a chosen name for the team and a list of names of the team members, with the POC of the team clearly marked on the list, as part of this Phase I submission.

1.2. (5 points) **Declaration of project datasets:** The project needs to revolve around four different datasets for a three-person team and three different datasets for a two-person team, with each dataset being different from the other in terms of nature/modality, as per the following requirements:

- Each one of the datasets can be a tabular dataset, a time-series dataset, an imaging dataset, a video dataset, a text dataset, a sound/speech dataset, a multimodal dataset, etc.
- One of the declared datasets must be a tabular dataset, with the number of samples at least 200; i.e., $n \geq 200$, and the number of raw features (attributes) at least 20; i.e., $p \geq 20$.

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.

1.3. (3 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 declared 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: December 19, 2022)

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

2.1. **Creation and submission of separate notebooks, one for each of the 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:

 waheed.bajwa@rutgers.edu

- 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 of the notebooks as `<TeamName>_Dataset<n>.ipynb`, where you would replace `<TeamName>` with your team's name, and `<n>` with 1, 2, 3, 4 for each dataset. As an example, for a two-person team named Anaconda, the notebooks should be named as `Anaconda_Dataset1.ipynb`, `Anaconda_Dataset2.ipynb`, and `Anaconda_Dataset3.ipynb`.
- You must ensure that your submitted notebooks are fully executed, so that they are not required to be rerun during grading. *Please double- and triple-check this to ensure compliance with this requirement.*
- The following breakdown of points for this part of Phase II of the term project should guide your code development for each notebook.

(a) (3 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) (3 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) (6 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) (32 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 three methods as much as possible.

(e) (8 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.

(f) (5 points) **Discussion on ethical issues for each dataset/task:** Provide a discussion on the ethical aspects of the machine learning tasks that you carried out on the declared datasets. This discussion should be carried out in a markdown cell and should be carefully formatted for readability purposes.

(g) (3 points) **Bibliography for each notebook:** Provide bibliographic references that helped you during the preparation of the notebook. These references, which should be provided in a markdown cell at the end of the notebook, should be referenced within the body markdown cells of each notebook as much as possible.

2.2. (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 different notebooks, and its purpose is to convince the teaching staff that you fully understand the different aspects of the submitted notebooks. The presentation should be uploaded on Canvas on its respective assignment.

 waheed.bajwa@rutgers.edu

3 Phase III (December 22, 2022 and December 23, 2022)

Phase III of the project would be a 10-minutes in-person Question-and-Answer session involving the teaching staff and every team that submitted a video presentation. In other words, teams that do not submit the video presentation would lose points for this phase of the project. Teams would be provided a sign-up opportunity on a first-come, first-served basis to sign-up for the Q&A session slots that will be held on December 22, 2022 and December 23, 2022.

3.1. (15 points) Come prepared to the Q&A session with all information about your project. Based on your video presentation, the teaching staff would ask you three to five questions of technical nature. The answers to these questions will also be used to assess parts of the submitted notebooks.

 waheed.bajwa@rutgers.edu