CS373, Instructor: Jean Honorio (jhonorio@purdue.edu), Spring 2021, PROJECT PLAN

Register your project team in Brightspace (see "Project team & plan"). Fill this project plan form and submit it by email to jhonorio@purdue.edu on or before March 29. CC the email to all team members. Do not send the plans to the TAs.

You should wait for my approval before starting to work on the project. Once your plan is approved, it acts as a contract between you and me. If you implement everything correctly, you will receive a grade of 10/10. Remember that changes are still allowed later (in coordination with me), but it is beneficial to have a good initial plan.

1. Team number (as registered in Brightspace).

   Our team number is 23.

2. Students' names and Purdue e-mails (as registered in Brightspace). The project is to be done in groups of 3 students.

   Our team includes Aidan Abbott (abbott34@purdue.edu), Stephen Everett (everet12@purdue.edu), and Mark Gindling (mgindlin@purdue.edu).

3. [Up to 3 lines] Definition of the problem, possibly relevant to your interests.

   For this project, we are going to build a machine learning algorithm that can accurately predict whether a patient has heart disease.

4. [Up to 5 lines] Which dataset will be used? The dataset should be already publicly available. For possible datasets, see the course website https://www.cs.purdue.edu/homes/jhonorio/21spring-cs37300.html

   Describe how many samples/features the dataset has. Give me some idea of how you plan to transform the dataset into a table of data that a machine learning algorithm can use.

   If the dataset is too large, then choose fewer samples (for instance 1000 or smaller). If the dataset has too many features, then arbitrarily choose few of those features (for instance 100 or smaller). Remember that cross validation takes time, and the larger the dataset, the more time this will take.

   We will be using the Heart Disease UCI dataset. This dataset has 303 samples and 14 features. This allows us to use the dataset without having to cull any samples or features. As the dataset is already in a CSV file, we will be able to read the data directly into a dataframe using a library like Pandas.

5. URL where the above dataset is available.

   The dataset can be found at www.kaggle.com/ronitf/heart-disease-uci.

6. [Up to 5 lines] Which TWO machine learning algorithms are going to be used? (e.g., SVM, classification trees, etc.) **You are allowed to either implement this from scratch or use third-party code, e.g., scikit-learn.**

   We are going to use linear SVM and classification trees. We chose these machine learning algorithms because linear SVM is a parametric model and classification trees are non-parametric, and it may be useful to see which performs better.

7. [Up to 5 lines] Which cross-validation technique(s) is(are) going to be used? (e.g., training/validation/testing, k-fold cross-validation, bootstrapping). **You MUST implement this from scratch.**

   You should specify how you will apply the cross-validation technique(s). For instance, for training/validation/testing, specify which percentage of the samples will be used for training, which percentage for validation, and which percentage for testing (e.g., 40%, 30%, 30%). For k-fold cross validation, specify the value of k (e.g., k=10). For bootstrapping, specify the number of bootstraps B (e.g., B=30).

   We plan to use k-fold cross-validation and bootstrapping for our cross-validation techniques. For k-fold validation, we plan to use a value of k = 15. For bootstrapping, we plan to use a value of B = 30 since we have just over 300 samples in the dataset.

8. [Up to 10 lines] Which hyperparameter(s) is(are) going to be tuned. **You MUST implement this from scratch.**

   For the SVM model, we will be tuning the hyperparameter C. For the classification tree, we will be tuning the Gini threshold.

9. [Up to 10 lines] Which TWO experimental results will you show? (e.g., plots of number of samples versus accuracy using different subsets of the dataset, hyperparameter versus accuracy, ROC curves, etc.) **You MUST implement this from scratch. Plots of the data itself DO NOT count as experimental results.**

   We will show ROC curves, as well as hyperparameter versus accuracy results. We will be using hyperparameters versus accuracy to update the hyperparameter to obtain the highest accuracy. We

will be using the ROC curve to also evaluate the accuracy of our performance in terms of precision and recall because we are determining the presence of heart disease. A false positive has much less consequence than a false negative, so we will want to minimize false negatives as much as possible.

10. Which programming language are you going to use? (Only MATLAB, C++, Java and Python are allowed. **Jupyter notebooks are NOT allowed.**)

    We will use Python for this assignment.

Advice: Do not spend too much time on things such as "understanding the data", "memory problems because your data is too big", etc. Only if you are already familiar with computer vision, brain data, natural language processing, big data, parallelism, etc. then you can make use of those things, but this will not imply that you will get a higher grade just based on that fact. In general, I would recommend using easy-to-understand datasets, and smaller subsets of the data, for instance.