Final Project

First Draft Due Date: Dec 2, 2022 at 11:59pm EST

Final Due Date: Dec 17, 2022 at 5pm EST

Instructions: The final is an open-ended project to be completed individually, not as group work. Though you may discuss the broad scope of your problem, analysis techniques, and software with your classmates, you must perform the analysis and write the the project report independently. As stated in the syllabus, copying code from a classmate or the internet (even with minor changes) constitutes plagiarism. You are required to submit your project writeup in pdf form (use LaTeX) in a file called `<Williams-ID>-project.pdf` to Gradescope under "Project Review" for the peer review submission and under "Project Final" for the final submission. Please remember that you cannot use late days for these submissions, and late submissions will be assigned a score of 0%.

The total assignment is worth 100 points.

Note on visualization: For portions of the project that require drawing trees, DAGs, or plots about the data, I encourage you to use software packages like `graphviz` (for trees and DAGs) and `matplotlib` or `seaborn` (for plots.) Hand drawn graphs and plots are **not** permitted for the final project.

Note on off-the-shelf software: You are allowed to use off-the-shelf machine learning packages like `sklearn` or `pytorch` for your final project. However, I will expect you to understand how the model works, and report all the important hyperparameter settings used to fit the model.

Overview

The project report must be 6-8 pages (without references) using the template provided on Piazza. This is an open-ended data analysis project where you can gain experience applying machine learning methods to real data. You should define your own analysis objective: What is the goal of applying machine learning to this data, and how might this improve over current baselines? You can use any methods that fall under the broad umbrella of machine learning. That is, you don't need to limit yourself to methods we've directly discussed, but whatever methods you use should be clearly related to the course material. To ensure that you do not use off-the-shelf packages as black boxes, the project also requires a preliminaries section providing a brief formal description of the algorithms you use. So it's important that whatever algorithms you pick are also ones that you would feel comfortable reading up on and explaining in a paragraph or two. You can, for example, explore the use of methods like support vector machines, recurrent neural networks, or agglomerative clustering, but I also expect you to briefly explain how these methods work in your reports.

Requirements:

1. **Clearly state your goal/hypothesis for the project (6 points).**

2. **Create and fill a section labeled Preliminaries (12 points).** In this section, provide a short but clear description of the machine learning algorithms that you will use and any background knowledge required to describe/understand your analysis. That is, provide enough background so the reader can familiarize themselves with the concepts needed to understand your analysis, but don't overload them with extra information that won't come up later in your writeup. E.g., if you are using linear regression with regularization, a minimal working example of how you would describe it might be as follows.

   Linear regression produces predictions for an outcome variable Y as a linear combination of learned parameters θ and input features X. That is, for a given example, we have the following equation that determines how we obtain a prediction $\widehat{y}$ as a function of inputs $x_1, \ldots, x_d$,

$$\widehat{y} = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_3 x_d.$$

   The parameters θ can be learned by minimizing a suitable loss function L using gradient descent. I use the mean squared error, which is defined for n samples of data as follows: $L(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. To avoid overfitting, it is common to apply a regularization penalty in conjunction with the loss: I use the L2 penalty. So, the overall function being minimized is then $L(\theta) + \lambda \sum_{j=1}^{d} \theta_d^2$, where λ is a hyperparameter that will be tuned using a validation set.

   If you are using decision trees/random forests/neural nets you would provide a similar blurb about how these algorithms produce predictions, how they are trained, and how they can be regularized (you only need to describe the algorithms and regularization techniques you will actually use.) This would also be a good place to mention and cite if you are using external software packages for these algorithms, or running your own implementations of them.

3. **Provide a brief description of the data and pre-processing steps (8 points).** In a separate section, provide some background on your data and how you pre-process it for your downstream analysis. This should include details like standardization (if applicable) and also how you generate the training, validation, and test splits. Some other basic information you might want to cover: where does the data come from, how many rows/samples does it have, how many features does it have and which of these will be retained for the final analysis (if your data is of a specific modality, like pictures or text, describe how you process it into a machine learnable form.) There should be sufficient detail for a reader to understand and reproduce your pre-processing of the data.

4. **Identify and set up a suitable baseline (10 points).** Set up a baseline model that you will use to determine "success" or "failure" of your machine learning models. This could be a baseline based on expert opinions/heuristics (like we saw in the protein binding case study), or if this information is hard to obtain, a simple unregularized model, e.g., linear or logistic regression.

5. **Select a suitable machine learning model using training and validation data (15 points).** Describe and provide some visuals (plots/figures) of how you selected the optimal machine learning model from a set of candidate models using just the training and validation data. If you are considering different algorithms, this could be a search over possible algorithms and a search over hyperparameters of each algorithm (if some hyperparameters, such as learning rate or number of bootstraps, are held constant, briefly justify why.) When considering different hyperparameters, e.g., if you are doing a sweep over possible settings of λ for regularization, you could generate a plot showing how validation accuracy changes as a function of the regularization parameter, and use this to justify the final choice of model. By the end of this section it should be clear to the reader what the final model moving into the final testing/production phase is, what its hyperparameters are, and how they were tuned.

6. **Interpret and discuss your results (12 points).** Write a section on Discussion and Results: How does the chosen model perform on the test data relative to the established baselines and what does this

entail? Negative results are perfectly ok – you are not being graded on being able to build a state-of-the-art model, but rather your ability to reason about how to get there. Provide some concrete reasons and evidence for why the model performs well or poorly. This could include generating figures that show the inner workings of the model e.g., figure of the decision tree, or some notion of feature importance, e.g., magnitude of coefficients in a linear/logistic regression, or changes in predictions in a random forest/neural network upon manipulation of some key features. This could also include showing plots demonstrating distribution shift between training and testing data.

7. **Sensitivity analysis/ablation study (10 points)**. A sensitivity analysis/ablation study usually tweaks a small aspect of the training/validation process in an adversarial way to see how this impacts the learning process. This could include adding some small random noise to the validation and test data to mimic distribution shift, downsampling the number of rows of training data so that the model has less data to learn from, or throwing away important features that the model previously relied on. Describe and perform a small ablation study, and report your findings regarding the final model and its training, validation, and testing accuracy relative to the original model/task.

8. **Summarize your thoughts in a short closing section (6 points).** In this section, provide some thoughts on lessons learned and/or alternative analysis techniques that you may have thought of and would have been interesting to pursue and/or any limitations of your current proposal that you would like to address in the future.

9. **A bibliography with appropriate references (1 point)**. In the bibliography you must provide appropriate references and credit to any methods/software/sources you used in completing your project. The project template contains an example of appropriate citation style.

10. **A zip file containing code for your final project (0 points)**. The project code is not assigned any points and is only used to assist in grading and evaluating the project. In the event that it is determined that there is a serious mismatch between the code and the results in the written report amounting to "made-up" results, there will be a significant penalty applied to the final grade.

Grading Rubric

I will grade the project on the requirements, as well as other holistic factors such as the breadth, clarity, creativity of the work.

- Does the project meet the requirements? (80 points, see point distribution in previous section)
- Does the project adequately cover a range of topics in the course material? (5 points)
- Is the writing clear and understandable? (10 points)
- Is the project novel and/or creative? (5 points)

Total: 100 points