CS-1390-1: Introduction to Machine Learning

Project Proposal — **Santripta Sharma**

1. Classifier Classification/Meta Classification

Compare and contrast a few methods of classification of data, including logistic regression, k-nn, (shallow) FFNs, decision trees, and some ensemble variants (random forests, ensemble logistic regression, etc.). Collect metrics on accuracy & latency of these models on 10 different classification datasets.

Finally, try to see if there's any correlation between certain meta-features (metadata) of the datasets themselves (size of dataset, number of features/classes, correlation between features, more) and the (accuracy, latency) using rudimentary (hopefully) data analysis. If such a correlation exists, create a new dataset, with the meta-features of each dataset as the data (so 10ish data-points), and build a classifier to predict which method probably works the best given features of a new dataset.

The intuition is that maybe certain pieces of metadata about the dataset, like its features having strong linear correlation, or having a high number of categorical features favour certain models over others.

Caveats:

1. I'm only going to be able to run each classifier on a finite number of datasets, possibly a very limited number of datasets (<15), which might result in the final meta-classifier being overfit to the trained datasets.
2. Such a neat correlation (linear or otherwise) probably does not exist, or may not be identifiable with simple classifiers.
3. It's unclear to me what features of a dataset might be relevant to the problem, or even what a valid 'feature' of a dataset is. Further, structuring the model may be hard with features like 'correlation between features', which would vary in size for different datasets (for m features, mth triangular number = $m(m + 1)/2$). Here, for instance, there would be a maximum number of features M the model could handle, and there would have to be some sort of null equivalent value that the correlations for datasets with < M features could be represented by.

How:

- I will try to implement the bigger models (decision trees, random forest, ffn) from scratch, while using module primitives for simpler concepts or those covered in the course, like logistic regression, kNNs.
- I will be primarily using python and jupyter notebooks for the implementation.

Deliverables:

- Datasets used & preprocessing code (if any)

- Model code
- pdf Report, summarizing results.