ICS 485: Machine Learning

Term Project-251

Kaggle Competition!!

Please make sure to join the Kaggle competition link for the project and get a chance to win a Bonus up to 10%!

Please don't share the link with others.

https://www.kaggle.com/t/b22997912b72482abbce03115ac84235

Instructions

- Each student must join the competition individually using their real name.
- After all team members joined, they must form a team of 2 from 'Team' tab.
- Only one team member needs to make submissions on behalf of the team (after the team is formed on Kaggle).

Part-A (70%)

You task is to perform binary classification on the dataset provided:

Minimum requirements:

- [5 points] Dataset analysis and report on important statistics
- [15 points] Feature selection/transformation/engineering
- [10 points] Dealing with missing values (if applicable)
- [10 points] Dealing with imbalanced data (if applicable)
- [40 points] At least 4 Classifiers (each student works on 2 classifier), out of which:
    - One of linear classifier (logistic regression or SVMs)
    - One of: KNN, Decision Trees
    - One Neural Networks
    - One of Ensemble Learning (Random Forest, Adaboost,…)
  - Proper hyper-parameter tuning based on validation set (or cross-validation)…**Note: you can use part of train set to take our validation set or perform cross validation.**

    - List of appropriate evaluation measures with justifications (**we will use F1-score of the minority class as the main metric**)
- [10 points] Error analysis and possible improvements
- [10 points] Final results on the test set

<u>Other possible ideas to try (as examples):</u>

- More classifiers and comparison
- Investigate the concept of margins
- Dimensionality reduction as preprocessing before classification
- Investigate different feature scaling techniques
- Clustering the data in K clusters (K= number of classes) and compare the labels
- Interpreting the learned models (for example by examining the weights of a linear model or by constructing decision rules from the learnt decision tree)
- …

Part-B (30%)-Separate Jupyter Notebook

Implement at least two active learning strategies (e.g., least confidence and entropy, each student works on one) where the dataset is same as part-A EXCEPT the training samples are not labelled. You start with 50 random samples labelled initially and perform active learning on the rest of the training data to achieve comparable results to part-A with minimal number of samples labelled. Each iteration should label between 20 to 50 samples (it is your choice). **Note: Use logistic regression classifier.**

Important Notes:

1. All the documents (code and report) should be submitted in Jupyter notebooks.