Milestone Progress Report

Work Completed

I have explored multiple machine learning algorithms to classify our dataset, including Decision Tree, Perceptron, and AdaBoost.

- **Decision Tree:** Initially, I implemented trees using both Collision Entropy and Information Gain methods, resulting in a moderate accuracy of 54% on Kaggle. Subsequently, I experimented with a Weighted Gini impurity approach, significantly improving accuracy to 80%.
- **Perceptron:** I evaluated several variations of the Perceptron algorithm—standard, average, decaying, margin, and aggressive versions. Additionally, I incorporated the weighted approach from the Decision Tree experiments. This resulted in a best-performing accuracy of 72%.
- **AdaBoost:** I combined Decision Tree and Perceptron models into an AdaBoost ensemble, achieving the highest accuracy of 84% thus far.

Dataset Descriptive Statistics

The dataset consists of 7,597 samples and 361 integer-valued features. The label distribution is imbalanced, with approximately 33.88% positive cases. Feature analysis highlights significant variation; for example, feature x_1 ranges from 0 to 294 with a mean of approximately 0.27, whereas feature x_2 ranges from 0 to 28,161 with a mean of 305.71. Many features (x_4 through x_{357}) have constant zero values, suggesting potential sparsity or irrelevance. Additionally, features x_{359} and x_{360} exhibit notably large maximum values (540,483 and 806,971, respectively), indicating potential outliers or highly skewed distributions.

Plan Until Next Milestone

Moving forward, I plan to:

1. Explore various data preprocessing methods, such as mRMA and Pearson correlation coefficient analysis.
2. Implement Support Vector Machines (SVM) and evaluate performance using different kernels.
3. Develop and test Fully Connected Neural Network models.

These strategies aim to further enhance model performance and explore more sophisticated classification techniques.