15.095 Machine Learning Under a Modern Optimization Lens
Project Proposal: Multimodal AI-Text Detection

Team Members: Elie Juvenspan (ejuven@mit.edu), Hindy Rossignol (hindyros@mit.edu)

Problem Summary

The rapid growth of large language models has made distinguishing AI-generated text from human writing increasingly difficult. Our current XGBoost model already achieves an out-of-sample precision of about 90%, but the remaining 10% of ambiguous cases are critical for Mercor's mission to ensure reliable content verification. Our project aims to close this gap by developing a multimodal classifier that combines linguistic features and text embeddings within an optimization-based framework.

Dataset

We will use the **Mercor AI Detection Dataset** from link to data, which provides labeled samples of human and AI-generated text. Each entry includes the full text and metadata about its source. We will extract auxiliary linguistic features (sentence length, punctuation frequency, lexical richness) to complement embedding representations provided in the tabular dataset. We may also augment the dataset by generating additional samples from large language models (e.g., Mistral, Gemini, LLaMA) and curated human-written text to enhance robustness.

Methodology

Our approach views each stage of the project through the optimization lenses introduced in class—convex, robust, and mixed-integer optimization.

We will begin with **feature engineering as an optimization problem**, treating feature selection as a subset selection task to maximize predictive power while controlling model complexity, using LASSO regularization and feature importance analysis to foster sparsity and interpretability.

We will then perform **model and hyperparameter optimization** on a focused set of algorithms—Logistic Regression, Random Forests, XGBoost, and Optimal Classification Trees with Hyperplanes (OCT-H).

Cross-validation and regularization will guide model selection, with OCT-H providing a link to mixed-integer optimization and interpretability.

Finally, we will develop a **multimodal fusion** layer combining structured linguistic features with transformer-based embeddings (e.g., BERT), optimizing the fusion weights to leverage both modalities effectively.

Challenges and Approach

The main challenges involve limited data, class imbalance, and generalization. With only 269 samples, overfitting is a concern; we will use extensive cross-validation, strong regularization, and paraphrasing-based data augmentation.

Class imbalance will be addressed through class weighting, and evaluation will emphasize ROC-AUC over accuracy. Because test samples may include unseen topics, we will emphasize topic-agnostic features and lightweight domain adaptation.

To balance model complexity and interpretability, we will use automated feature selection and sparse models. Transparency will be enhanced through interpretable methods such as logistic regression and OCT-H, complemented by SHAP value analysis.

Finally, to improve robustness across writing styles, we will explore domain-invariant features and robust optimization objectives.