XGBoost for Classification

Midterm Presentation 2 - Classification

MSSC 6250 - Spring 2025
Praful Aggarwal / John Schliesmann / Jeremy Buss

In our industry roles, we often classify diverse sets of data, necessitating a broad modeling toolkit. This project will enhance our ability to discern the usage of tree-based classification models specifically XGBoost (Extreme Gradient Boosting) and its extension of the gradient boosting toolset.

Project Objectives

The primary goal of this project is to explore XGBoost for classification of structured data, focusing on understanding its strengths, weaknesses, and practical applications compared to other classification algorithms. Specifically, we aim to:

- Understand the theoretical foundation of gradient boosting and how XGBoost enhances it.
- Investigate and analyze different datasets for classification with XGBoost.
- Compare XGBoost with other classification algorithms, such as logistic regression, decision trees, and random forests, identifying the conditions under which each method is appropriate.
- Explore hyperparameter tuning and its impact on model performance and feature importance.

Methodology

We will implement with Python, leveraging libraries like scikit-learn and XGBoost with the following steps:

- Data: Find publicly available or simulated dataset to illustrate key points below.
- Model Fitting: Apply logistic regression, decision trees (CART), random forests/boosting and XGBoost.
- Model Performance Evaluation: Assess and compare the performance of the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Hyperparameter Tuning: Conduct a systematic exploration of XGBoost hyperparameters using techniques like cross-validation and grid search.
- Feature Importance: Analyze feature importance via XGBoost to understand the relative contribution of different features to the classification task.

Conclusion

By the end of this project, we expect to have a comprehensive understanding of XGBoost for classification, its capabilities, and its limitations. This knowledge will directly benefit our work by enabling us to build high-performing classification models for a variety of real-world problems. We also expect to provide our classmates with the intuition to apply XGBoost and analyze classification problems effectively.