Final Project for DASC 5420

Your Name

April 1, 2025

Introduction

- Brief overview of the project
- Objectives and goals
- Importance of the topic

Data Preparation

- Data collection and preprocessing
- Algorithms and techniques used
- Tools and frameworks

Model Fitting

- Model selection and design
- Details in logistic regression
- etc

Selected Models

- **Logistic Regression**: Designed custom algorithms to fit the log-odds regression from scratch.
- **K-Nearest Neighbors (KNN)**: Utilized grid-search to determine the optimal value of K.
- **Random Forest**: Applied ensemble methods to improve prediction accuracy and interpretability.

Customized Logistic Regression

- Logistic regression models the probability of a binary outcome using the log-odds (logit) function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

  where p is the probability of the positive class, and $x_1, x_2, \ldots, x_n$ are the input features.
- The model parameters $(\beta_0, \beta_1, \ldots, \beta_n)$ are estimated by maximizing the likelihood function.
- In this project, we designed custom algorithms to solve the logistic regression problem from scratch, focusing on optimizing the log-likelihood function.

Logistic Regression: Model Fitting

- TO DO.

Best Hyperparameters of RF upon different datasets

In Random Forest, key hyperparameters such as *n_estimators*, *max_features*, *max_depth*, *max_leaf_nodes*, and *min_samples_leaf* need tuning to optimize model performance. We used grid search with cross-validation to explore combinations. Parameters were tested over selected ranges (e.g., *n_estimators* $\in \{50, 100, 200, 300, 500, 1000\}$, *max_features* $\in \{1, 2, 5, 10\}$, etc.). The table below shows the best settings for each dataset based on validation accuracy. Results indicate that different balancing strategies affect the ideal model complexity and tree structure.

Dataset	n_estimators	max_features	max_depth	max_leaf_nodes	min_samples_split
Raw	50	4	15	None (Unlimited)	5
Random Oversampling	100	1	30	None (Unlimited)	1
SMOTE-balancing	100	1	30	None	1
Advanced-balancing	300	3	10	None (Unlimited)	1

Table: Optimal Hyperparameter Combinations for Random Forest Across Datasets

Results

- Key findings
- Visualizations and analysis
- Interpretation of results

Conclusion

- Summary of findings
- Future work
- Questions