REPORT 2: OPTIMIZED GRADIENT BOOSTING FOR AUDIO SCORE PREDICTION

NAME: UDAYAN SAHA

COLLEGE: KAZI NAZRUL UNIVERSITY [A state govt, university, Asansol, West Bengal]

CURRENT STATUS: 4TH YEAR B. TECH COMPUTER SCIENCE STUDENT

SPECIALIZATION: DATA SCIENCE

Abstract

This paper presents an enhanced approach to predicting human-evaluated scores from audio samples using machine learning. We optimize the Gradient Boosting Regressor (GBR) by performing hyperparameter tuning to achieve better performance metrics. Key audio features contributing to the predictions are also analysed. The model demonstrates measurable improvements in prediction accuracy, highlighting the potential of gradient boosting for audio score regression tasks.

Introduction

Automated evaluation of audio samples is an emerging area in machine learning with applications in education, entertainment, and health. In this study, we explore the utility of the Gradient Boosting Regressor (GBR) in predicting speaker proficiency scores based on extracted audio features. Building upon baseline models, we aim to improve the model's performance using *hyperparameter tuning*.

Dataset and Preprocessing

The dataset consists of audio feature vectors derived from spoken audio samples. Each sample is associated with a human-evaluated score. Preprocessing steps included:

- Null value handling
- Feature scaling using StandardScaler
- Splitting into training and testing sets (80:20 ratio)

Methodology

We employ the Gradient Boosting Regressor, optimizing its performance using *RandomizedSearchCV* with 3-fold cross-validation. The following hyperparameters were tuned:

- n_estimators: [100, 200, 300]
- learning_rate: [0.01, 0.05, 0.1]
- max_depth: [3, 5, 7]
- min_samples_split: [2, 5, 10]
- subsample: [0.6, 0.8, 1.0]

The search was configured to maximize the R2R^2 score and find the best model among 20 sampled combinations.

Feature Importance Analysis

The top 10 most influential features identified using the model's feature_importances_ were:

1. zcr_mean
2. rmse_std
3. spectral_centroid_mean
4. spectral_bandwidth_mean
5. rolloff_mean
6. mfcc1_mean
7. mfcc2_mean
8. chroma_stft_mean
9. spectral_contrast_mean
10. tonnetz_mean

These features represent both temporal and spectral characteristics of the audio, indicating that fluency and prosody-related traits are predictive of performance scores.

Conclusion

Gradient Boosting with hyperparameter tuning proved to be an effective method for predicting human-assigned audio scores. The model not only improved the accuracy but also provided interpretable insights into which audio features are most significant.

Future Scope

- Multimodal Fusion: Combining audio features with text transcripts or phoneme-level features.
- Deep Learning Models: Implementing CNNs or LSTMs to capture more complex audio patterns.
- Real-time Scoring System: Deploying the model in an API or web app for live assessment.
- Cross-lingual Generalization: Testing the model’s adaptability to other languages and accents.
- Explainable AI: Integrating SHAP or LIME for better interpretability.

This study lays the groundwork for further development in automated audio evaluation systems, emphasizing the role of feature engineering and model tuning in boosting performance.