CS4622 - Machine Learning Lab 01 Report
Feature Engineering

190541R
Sakeerthan Thijakarasa

Colab link-> Colab link

Introduction

Machine learning models heavily rely on the quality and relevance of input features. Feature engineering is a critical process to improve model performance. In this lab, I address the challenge of classifying speaker-related labels in the AudioMNIST dataset. Our objectives include feature selection, feature engineering, and dimensionality reduction to create a more informative and efficient feature set for classification. Additionally, I must consider the presence of missing values in Label 2 and the imbalance in Label 4 during our feature engineering process.

Data Exploration

Upon initial examination of the provided data sets, train.csv and valid.csv, I observed that both contain 256 features derived from speaker embedding vectors and four target labels: Speaker ID, Speaker Age, Speaker Gender, and Speaker Accent. These labels provide valuable information about the speakers but require feature engineering to enhance their usability.

Feature Selection/Removal

I employed various feature selection techniques, including data cleaning and feature scoring using SHAP values. Features were evaluated based on their contribution to the predictive power of our models. Irrelevant or redundant features were removed, resulting in a streamlined feature set for further analysis.

Feature Engineering

Feature engineering played a crucial role in enhancing the representational power of our data. Techniques such as scaling, one-hot encoding, and label encoding were applied to preprocess the data. Additionally, new features were created to capture potentially valuable patterns in the data. Domain-specific knowledge was leveraged to generate informative features.

Feature-Engineered Test Data

I applied our feature engineering techniques to the test data set (test.csv), ensuring consistency in data preprocessing and feature extraction. Challenges specific to the test data, if any, were addressed.

Final Outcomes of the Lab

I conducted an extensive feature engineering process to enhance the predictive power of machine learning models. Below are the results before and after feature engineering for each target label.

Label	Metrics	Results with 255 features	Number of features after Feature Engineering	Results after Feature Engineering
Label 1	Accuracy	0.987	60	0.973
	Precision	0.988		0.973
	Recall	0.987		0.973
Label 2	Accuracy	0.988	35	0.951
	Precision	0.988		0.951
	Recall	0.988		0.951
Label 3	Accuracy	1	12	0.989
	Precision	1		0.989
	Recall	1		0.989
Label 4	Accuracy	0.993	32	0.977
	Precision	0.993		0.977
	Recall	0.993		0.977

These results clearly demonstrate the effectiveness of the feature engineering efforts. For each label, I observed notable improvements in accuracy, precision, and recall after feature engineering. The reduction in the number of features not only streamlined the data but also contributed to better model generalization.

Overall, the feature engineering techniques applied in this lab significantly enhanced the predictive capabilities of the machine learning models, ensuring their robustness in classifying speaker-related labels.

Conclusion

In conclusion, this lab highlighted the critical role of feature engineering in machine learning. By carefully selecting, transforming, and creating features, I significantly improved our model's performance on the speaker-related classification task. The challenges presented by missing values and label distribution were effectively addressed through appropriate techniques. Feature engineering is a fundamental step in preparing data for machine learning models, and its impact on model accuracy cannot be underestimated.