ML PROJECT

Team Members:

Tanushri Mohan Chougale[PES2UG24CS826]
Neela G[PES2UG23CS376]

Project Title:
Improving One-vs-Rest Multiclass Classification with Unsupervised Structure (Digits Dataset)

Problem Statement

Traditional supervised learning models for multiclass classification, such as One-vs-Rest logistic regression, rely entirely on predefined labels. However, these labels may obscure latent patterns or natural groupings within the data that could improve classification accuracy. The challenge is to determine whether unsupervised learning can reveal underlying structures in the dataset and whether these insights can be used to design a better-performing multiclass classification system based on binary classifiers.

High-Level Architecture

1. **Data Loading & Preprocessing:** Use the Digits dataset (8×8 grayscale images, 10 classes). Normalize features and split into train/test sets.
2. **Baseline Model:** Train a standard One-vs-Rest Logistic Regression classifier as a benchmark.
3. **Unsupervised Structure Discovery:** Apply K-Means, Gaussian Mixture Model (GMM), and Agglomerative Clustering to uncover latent patterns. Evaluate clusters using ARI and NMI.
4. **Integration Strategies:** (A) Add one-hot cluster labels as features. (B) Train cluster-specific classifiers ("local experts"). (C) Use GMM soft probabilities as stacked features.
5. **Evaluation & Visualization:** Compare performance via accuracy, confusion matrices, and PCA plots.

Perform stability checks across multiple seeds.

Results

- Baseline One-vs-Rest accuracy: ≈93–94%.
- Cluster feature augmentation improved accuracy to ≈95–96%.
- GMM soft-feature integration provided the most consistent and robust results.
- Visualizations showed clearer class separation after incorporating unsupervised features.

Conclusion

Integrating unsupervised learning insights into a supervised One-vs-Rest framework enhances multiclass classification performance. This hybrid approach demonstrates that revealing latent data structures can significantly improve accuracy and generalization compared to traditional label-driven methods.