Multi-Class Classification with Standard and Group Cross-Validation

Najib Attar
Computer and Informatics Engineering
Istanbul Technical University
Istanbul, Turkey
Student ID: 150220905
attar22@itu.edu.tr

***Abstract*—We address a multi-class classification problem using a labeled training dataset and an unlabeled test dataset. Model performance is evaluated using the macro-F1 score under both standard 5-fold cross-validation and Group 5-fold cross-validation based on subject identifiers. RidgeClassifier and Logistic Regression models are explored in this study. The final model is selected according to GroupKFold performance and is used to generate predictions for the Kaggle test set.**

***Index Terms*—Multi-class classification, cross-validation, GroupKFold, macro-F1 score, machine learning**

I. Introduction

Machine learning methods are widely used for classification problems in a variety of real-world applications. In this project, we address a multi-class classification task in which each sample belongs to one of four possible classes. The available data consists of a labeled training dataset and an unlabeled test dataset, and the objective is to build a model that generalizes well to unseen data while providing a reliable performance estimate during training.

A common approach to model evaluation is standard k-fold cross-validation, where the dataset is randomly divided into k disjoint folds. Although this method is effective in many scenarios, it may produce overly optimistic performance estimates when the dataset contains correlated samples. In particular, when multiple samples originate from the same subject, randomly splitting the data can lead to information leakage between training and validation sets [3].

To address this issue, this study employs both standard 5-fold cross-validation and Group 5-fold cross-validation. In the group-based setting, samples belonging to the same subject are assigned to the same fold, preventing subject-level data leakage and providing a more realistic evaluation of model performance. Model performance is assessed using the macro-F1 score, which equally weights all classes and is suitable for multi-class classification problems.

Several linear classification models are explored in this work, including RidgeClassifier and Logistic Regression. The final model is selected based on GroupKFold cross-validation performance and is subsequently trained on the full training dataset to generate predictions for the unseen test set. Our final Kaggle submission achieved a public macro-F1 score of **0.41616**, ranking **26** on the public leaderboard under the Kaggle username **Najib Attar - 150220905**.

II. Dataset and Preprocessing

A. Dataset Description

The provided dataset consists of a labeled training set and an unlabeled test set. The training dataset contains a total of 22,496 samples, where each sample is represented by 1,793 numerical feature columns. In addition to the feature columns, the training data includes a label column indicating the class of each sample and a *person_id* column identifying the subject from which the sample was collected.

The classification task is a multi-class problem with four possible class labels, namely $\{0, 1, 2, 3\}$. The presence of multiple samples per subject makes the dataset susceptible to subject-level data leakage if not handled carefully during model evaluation.

The test dataset contains 10,656 samples and includes the same 1,793 feature columns as the training dataset, but does not include label information. The objective is to predict the class label for each test sample and generate a submission file in the required format for evaluation.

B. Preprocessing

Prior to model training, several preprocessing steps are applied to the data to ensure robust and fair evaluation. All preprocessing operations are performed within a unified pipeline to avoid information leakage between training and validation sets during cross-validation.

C. Missing Value Handling

The dataset contains missing values in some feature columns. To address this issue, missing values are imputed using the median value of each feature, computed from the training data only. Median imputation is chosen due to its robustness to outliers and its suitability for numerical features.

D. Feature Scaling

After imputation, feature values are standardized to have zero mean and unit variance. Feature scaling is necessary for linear models to ensure that all features contribute equally to the learning process and to improve optimization stability.

E. *Pipeline-Based Processing*

All preprocessing steps are integrated into a single machine learning pipeline together with the classification model. This pipeline-based approach ensures that imputation and scaling parameters are learned exclusively from the training folds and then applied to the corresponding validation folds, thereby preventing data leakage during cross-validation.

III. Methods

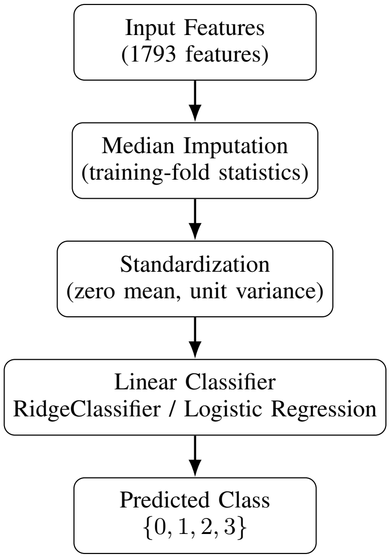

Fig. 1. Learning pipeline used in training and inference. All preprocessing steps are executed within a single pipeline to prevent data leakage during cross-validation.

A. *Models*

In this study, linear classification models are employed due to their efficiency, interpretability, and suitability for high-dimensional data. Linear models are particularly appropriate for this task, as the number of feature dimensions is relatively large compared to the number of samples.

B. *RidgeClassifier*

The RidgeClassifier is a linear classification model that incorporates ℓ_2 regularization to penalize large model coefficients. This regularization helps reduce overfitting and improves generalization performance, especially in high-dimensional settings. RidgeClassifier is computationally efficient and integrates naturally with standardized features, making it well suited for repeated cross-validation experiments.

C. *Logistic Regression*

Logistic Regression is another widely used linear classification method that models class probabilities directly. In this work, a multinomial formulation of Logistic Regression is considered to handle the multi-class nature of the problem. Regularization is applied to control model complexity and prevent overfitting.

D. *Model Selection*

Both RidgeClassifier and Logistic Regression are evaluated using the same cross-validation strategies to ensure a fair comparison. Model selection is based on Group 5-fold cross-validation performance, as this evaluation scheme provides a more realistic estimate of generalization when samples are grouped by subject. Based on the obtained results, RidgeClassifier demonstrates superior and more stable performance under the group-based evaluation and is therefore selected as the final model.

The regularization parameter of RidgeClassifier was selected through Group 5-fold cross-validation. Multiple values of the regularization strength α were evaluated, and the value yielding the highest mean macro-F1 score under group-based validation was selected. This tuning strategy ensures that the chosen model generalizes best to unseen subjects.

E. *Cross-Validation Strategy*

To evaluate model performance and ensure reliable generalization estimates, two different cross-validation strategies are employed in this study: standard 5-fold cross-validation and Group 5-fold cross-validation. Both strategies use the macro-F1 score as the evaluation metric.

F. *Training and Testing Procedure*

During training, preprocessing steps and model parameters are learned exclusively from the training folds within each cross-validation split. Model performance is evaluated on the corresponding validation folds using the macro-F1 score. After model selection, the final classifier is trained on the full training dataset and then applied to the unlabeled test dataset to generate predictions for submission.

G. *Standard 5-Fold Cross-Validation*

In standard 5-fold cross-validation, the training dataset is randomly partitioned into five disjoint folds of approximately equal size. During each iteration, four folds are used for training while the remaining fold is used for validation. This process is repeated until each fold has served as the validation set once.

Standard 5-fold cross-validation provides a baseline estimate of model performance and is commonly used when samples are assumed to be independent and identically distributed. However, this approach does not account for potential correlations between samples originating from the same subject.

H. *Group 5-Fold Cross-Validation*

In Group 5-fold cross-validation, samples are split into folds based on their associated *person_id*. All samples belonging to the same subject are assigned to the same fold, ensuring that no subject appears simultaneously in both the training and validation sets.

This group-based evaluation strategy prevents subject-level data leakage and yields a more realistic assessment of model performance in scenarios where multiple samples are collected from each subject. As a result, Group 5-fold cross-validation is used as the primary criterion for model selection in this work.

IV. Results and Conclusions

The performance of the evaluated models is assessed using macro-F1 score under both standard 5-fold cross-validation and Group 5-fold cross-validation. Table I summarizes the mean macro-F1 scores obtained for each model and evaluation strategy.

TABLE I
Cross-Validation Performance (Mean Macro-F1)

Model	CV Strategy	Mean Macro-F1
RidgeClassifier	Standard 5-Fold	0.7550
RidgeClassifier	Group 5-Fold	0.2562
Logistic Regression	Standard 5-Fold	0.7573
Logistic Regression	Group 5-Fold	0.2414

The results indicate that both models achieve relatively high macro-F1 scores under standard 5-fold cross-validation. However, a substantial performance decrease is observed when using Group 5-fold cross-validation, highlighting the impact of subject-level data leakage in the standard evaluation setting. Under the group-based evaluation, RidgeClassifier demonstrates higher and more stable performance compared to Logistic Regression, and is therefore selected as the final model. In addition to cross-validation results, the final submission achieved a public Kaggle macro-F1 score of **0.41616**, ranking **26** on the public leaderboard (Kaggle username: **Najib Attar - 150220905**).

Based on the Group 5-fold cross-validation results, RidgeClassifier is selected as the final model due to its superior performance compared to Logistic Regression under the group-based evaluation. The selected RidgeClassifier uses a regularization parameter of $\alpha = 10.0$, which achieved the highest mean macro-F1 score among the evaluated models.

After model selection, the RidgeClassifier is trained on the full training dataset using all available samples. The trained model is then applied to the unlabeled test dataset to generate class predictions for each test instance.

The predictions are saved in a file named `predictions.csv`, following the required submission format. The file contains two columns: `ID`, which ranges from 0 to 10655, and `Predicted`, which contains the predicted class labels belonging to the set $\{0, 1, 2, 3\}$. This submission file is subsequently uploaded to the Kaggle evaluation platform for final assessment [1].

In this project, a multi-class classification task was addressed using a labeled training dataset and an unlabeled test dataset. Linear classification models were evaluated using both standard 5-fold cross-validation and Group 5-fold cross-validation, with performance measured by the macro-F1 score.

The results demonstrate that standard cross-validation can lead to overly optimistic performance estimates when samples are correlated at the subject level. By contrast, Group 5-fold cross-validation prevents subject-level data leakage by ensuring that samples from the same person are not split across training and validation sets, providing a more realistic assessment of model generalization.

Based on the group-based evaluation results, RidgeClassifier was selected as the final model and trained on the full training dataset. The trained model was then used to generate predictions for the test set, which were submitted in the required format for final evaluation.

Acknowledgment

The author would like to thank the course instructors for their guidance and support throughout this project.

References

[1] Kaggle, "Kaggle Competitions Documentation." [Online]. Available: https://www.kaggle.com/docs/competitions

[2] Y. Zhao, X. Cao, J. Lin, D. Yu, and X. Cao, "Multimodal Affective States Recognition Based on Multiscale CNNs and Biologically Inspired Decision Fusion Model," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1391–1403, Apr.–Jun. 2023.

[3] M. Ali, F. Al Machot, A. Haj Mosa, M. Jdeed, E. Al Machot, and K. Kyamakya, "A Globally Generalized Emotion Recognition System Involving Different Physiological Signals," *Sensors*, vol. 18, no. 6, p. 1905, 2018.