Aix-Marseille University

Machine Learning Project

General Overview

The objective of the project is to predict sentiments in finance-related tweets. The database contains two variables:

- The *text* variable: contains the tweet text
- The *label* variable: contains three categories corresponding to the associated sentiment (0 = negative; 1 = positive; 2 = neutral)

For grading criteria 1 to 3, you must implement a binary classification method where you will try to predict negative sentiment vs. the rest.

For grading criterion 4, you will be evaluated on your ability to implement a multi-class classification method and cross-validation.

Presentation Date: Thursday, April 10, 2025

Important:

- You will work by groups of 3 or 4 and choose a name for your group.
- You must send your code the day before the presentation.
- DO NOT FORGET TO WRITE DOWN YOUR NAMES ON THE REPORT. YOU SHOULD ALSO SAVE THE REPORT AS NAME1NAME2NAME3_MACHINELEARNING.pdf

1 Grading Criterion 1: Implementing Algorithms (/3)

You must compare the performance of the following algorithms using the same explanatory variables:

- Logistic Regression
- K-Nearest Neighbors
- Random Forest
- Neural Networks
- Two models not covered in class

For each model, indicate the precision, recall, weighted f1-score, time needed to obtain results, and main hyperparameters used.

2 Grading Criterion 2: Model Optimization

2.1 Selection of Explanatory Variables (/2)

For the model of your choice from the list in Section 1, you must:

- Vary the explanatory variables (X). These changes should be made to reduce the test error.
- Show how the test error evolves with these choices.

2.2 Selection of Hyperparameters (/2)

For two models of your choice from the list in Section 1, you must:

- Optimize the hyperparameters
- Show how the test error varies with the hyperparameters

3 Grading Criterion 3: Selection and Evaluation of the Best Performing Model (/4)

For the model achieving the best weighted f1-score (your flagship model), you will need to:

- Explain why you chose this model (algorithm, hyperparameter, explanatory variables)
- Indicate the precision, recall, weighted f1-score, and correct classification rate
- Plot the ROC curve
- Try to identify the weaknesses of this model

4 Grading Criterion 4: Ability to Learn Coding Independently (/3)

For this grading criterion, you must predict the label variable using:

- Multi-class classification methods (hence keeping the label variable with three classes)
- Cross-validation

5 Grading Criterion 5: Presentation and Summary Report (/6)

For the presentation

1. Content
    - A response to each of the sections above
    - Presentation of the approach and potential limitations
    - Brief presentation of the code structure and content
2. Format

- Strict respect of allotted time (10 minutes **maximum**)
- Presentation without written notes
- The quality of the presentation will be taken into account in the final grade

For the summary report:

- 2-page report **maximum**
- Response to grading criteria
- The quality of the report (readability, clarity, conciseness) will be taken into account in the final grade