Lab 2

Task A: Classification

Alejandro Silva Rodríguez

Marta Cuevas Rodríguez

Aprendizaje Computacional
Universidad de Málaga

Septiembre 2024

Contents

1 Introduction

In computational learning, applying classification algorithms to predict disease progression is critical for deriving meaningful insights from complex biomedical data. The primary focus of this project is to evaluate and compare the performance of various classification methods, utilizing established metrics like precision, recall, specificity, and accuracy. These metrics provide a quantitative basis to assess each model's strengths and limitations, enabling us to identify the most suitable algorithm for predicting disease outcomes.

This project will analyze a dataset relevant to disease classification, covering essential aspects such as the dataset's class distribution, balance, and overall characteristics. By exploring different methods through detailed metric calculations, we aim to determine the best-performing model for predicting outcomes in the dataset. Additionally, graphical representations will aid in visualizing and comparing each method's effectiveness, making it easier to assess their respective advantages. Overall, this project provides a systematic approach to classification model evaluation and supports informed decision-making in selecting models suited to biomedical data analysis.

2 Objectives

- **Implement an algorithm for performance metric calculations**: Develop an algorithm that calculates a range of classification performance metrics, including precision, recall, specificity, accuracy, and others, providing a comprehensive comparison of each method's effectiveness.
- **Analyze the dataset characteristics**: Examine key properties of the dataset, such as the number of samples, class distribution, and class balance. Understanding these characteristics will help us assess the dataset's impact on classification performance and ensure a fair comparison of methods.
- **Provide detailed definitions for each metric**: Define each performance metric in detail, including its mathematical formula, what it measures, its range, and an interpretation of whether higher or lower values are preferable for classification tasks.
- **Visualize results for clearer comparisons**: Generate visualizations, such as plots comparing False Positives and False Negatives, Precision and Recall, and Accuracy versus F-measure. These graphs will facilitate a better understanding of each method's strengths and weaknesses.
- **Identify the best classification method based on performance**: Conduct a thorough analysis of each metric to identify the method with the best overall performance, providing a rationale for why this model is most suitable for predicting disease outcomes in this dataset.

3 Methodology and Results

1. Information about the dataset. Number of samples, number of classes, number of samples per class, if the dataset is balanced or unbalanced...

2. Information about the metrics you will use to compare the performance of the methods. Name, what it represents, its definition, range of the value provided by that metric, if lower/higher values are better...

3. A table with the yielded performance of each method for each metric. You can distribute the methods along the columns and the metrics along the rows. Highlight best results in bold.

4. Analysis of the performance. For each metric, describe briefly which method is the best and the worst, and why.

5. Conclusion. According to the analysis, determine which method, in general, is the best and why

following figures: - FN against FP - PR against RC - ACC against Fm

4 Conclusion

5 Repository Access

All additional information, including the source code and full documentation of this project, is available in the GitHub repository [1].

References

[1] Marta Cuevas. Lab2_computational_learning. `https://github.com/martacuevasr/Lab2_Computational_learning`, 2024. Último acceso: 1 noviembre 2024.