Machine Learning I

Ricardo Figueiredo up202105430
Sérgio Cardoso up202107918

May 2025

Executive Summary

- **Goals:** use a classification algorithm taught in classes to obtain the accuracy of a set of benchmark datasets
- **Outline of the Approach:** modify an already made implementation of the algorithm on existing open-source code in order to make it robust to the data characteristic in question.
- **Summary of Results:**

Selected Algorithm and Data Characteristic

We chose the CART algorithm due to its easy learning curve. CART stands for Classification and Regression Tree which, in itself, uses a Decision Tree to determine the accuracy of a dataset. Decision Trees are among the most popular models in ML, as they are used not only by CART, but also by other algorithms, such as ID3 (Iteractive Dichotomiser 3) or C4.5, the latter being an extension of CART.

Gini index: $Gini(D) = 1 - \sum_{i=1}^{c} {p_i}^2$, where p_i is the probability of class i.

Selected Algorithm and Data Characteristic

Many classification algorithms are sensitive to certain data. For instance, k-NN is sensitive to the presence of outliers, SVM is sensitive to hyperparameter values.
For this assignment, we were given the following three set of benchmark datasets:

1. Noise or Outliers
2. Class Imbalance in Binary Classification
3. Multiclass Classification

Given that we chose CART for this assignment and its only

Proposal

Emprical Study

- **Experimental Setup:**

Emprical Study

Conclusions

References