Machine Learning - Aprendizagem de Máquina - Assignment #1

Inês Dutra and Maria Pedroto

Deadline: April 5th, 2025

February 27, 2025

Objectives

Let's look into the assumptions and behavior of different machine learning methods. The general objective of this assignment is to dive a little bit deeper into how different ML methods behave on a dataset.

Task

Generate artificial datasets that illustrate the assumptions and characteristics of different methods. Datasets are ideally **bidimensional**. Among other dataset properties you can experiment with:

- number of instances (rows in your dataset table)
- number of classes (for multi-class classification problems)
- proportion of classes (balanced and imbalanced class problem)
- distribution of points within each class (shape of point clouds)
- shape of the border between the class regions, from linear to whatever
- level of noise
- level of overlap between the classes

Consider the methods: logistic regression, LDA, QDA, Decision Tree without pruning, Decision Tree with a maximum depth of 2, SVM linear, SVM RBF.

Investigate method assumptions

Find for each of the listed methods a dataset where the respective assumptions are met and assumptions of the other methods are not met (if possible). In other words, a dataset where that method is hard to beat using cross validation.

Explain why this dataset is appropriate for the method.

Suggestion: use datasets with 2 predictors and one class variable with only two distinct classes that can be visualized. This is not mandatory, but may facilitate the task.

Bias variance and model capacity

1. Find a dataset where by varying the level of noise (or other dataset properties such as border shape) different levels of tree pruning are ideal, from no pruning to maximum pruning. Produce a plot of level of noise (or other some other property if you prefer) against `ccp_alpha`.

2. Measure bias and variance error decomposition for three versions of this dataset along `ccp_alpha` (decision trees).

3. Interpret the results in terms of model capacity, bias error, variance error and total error.

Ensembles

With the dataset you used to study trees, compare Bagging, RandomForest and AbaBoost and other forms of boosting.

1. Compare the learning curves of the three methods along the number of trees and using Out-of-Bag (OOB). OOB error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging).

2. Compare the final solution of the three methods using cross validation.

3. Discuss the results.

Submit

A notebook with answers to these questions. The notebook is `ipynb` by default. If exporting to `pdf` or `html`, make sure that all contents are exported and not truncated. Each answer must be clearly identified and respecting the order of the questions.

Notes:

- Steps taken must be succintly described.
- Results must be summarised as much as possible.
- Well formatted and clear references to any paper, slides, book or site consulted for this assignment.
- Each group presents the work in-class by the end of the deadline. All elements of the group should participate in the presentation.
- All students should participate as each student will be evaluating one of the groups and feedback will be given to these evaluations.

Formal deadline is April 5th 2025, to be submitted in Moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1. Presentations are made in class (date to be announced).

Suggested structure

- Cover page: authors clearly identified at the top of the notebook, title and date
- Answers
- References
- Appendice (optional - non-essential add ons, longer and more detailed experiments, etc.)

Important

- Only Moodle can be used for submission (no email)
- Guidelines:

  - The answers should be clearly written.
  - The objectives for each experiment and plot you show should be clear.
  - It is not necessary to describe the methods but you should explain your steps when doing experiments and making choices.

Evaluation

- This assignment is worth at most 6 values out of twenty.
- Components
  - Writing 30%
  - Technical 60%
  - Presentation 10%: Clarity, confidence, communication, timely delivery.

Groups

Assignments are submitted by groups of 2 to 3 students. Different elements may have different grades based on the contribution distribution and interactions about the assignment. This distribution can be declared by the members of the group or can be inferred by the lecturers on the basis of individual interaction. Other group sizes will not be considered.

It is advisable that the students from the same group share general tasks but may have specific tasks that can be know by all. Group work is important for learning from other people, but each individual must acquire independent skills.

Ethical principles

When submitting, students commit themselves to follow strong ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole work. The contribution of different members within a group must be declared up front in the header of the report stating clearly a percentage of contribution per member. All the materials used and consulted must be credited in the work as references.