Machine Learning - Aprendizagem de Máquina - Assignment 2

Alipio Jorge and Inês Dutra

15/05/2021

Objectives

Supervised machine learning has many approaches for solving the problem of finding the best approximation for the unknown function $f(x)$. Each dataset may have a different best model, if any, subject to an evaluation metric. However, if we take the average of all algorithms over all possible datasets, each algorithm behaves equally well as any other. This is known as the no free lunch theorem (https://machinelearningmastery.com/no-free-lunch-theorem-for-machine-learning/). In practice, however, we have a limited set of problems of interest, for which a best model exists, which implies that some solutions may be potentially better than others.

The challenge of this assignment is the following. Given a method A from a pool of methods L find a dataset D for which A is the best. A only has to be better than the other methods in L. Additionally, you are invited to explain why this happens.

Datasets

This task has no given data. You are supposed to produce artificial datasets or find existing ones. You can use random generation methods.

Tasks

- Task: Consider the following classification methods (our L):
  - Logistic Regression
  - Linear Discriminant
  - Quadratic Discriminant
  - RandomForest
  - DecisionTrees
  - SVM linear
  - SVM rbf
  - SVM poly
  - MLP tanh
  - MLP relu
- Choose all of them or a **subset of them** and **try to find** an artificial classification dataset where that method+configuration is the best (or gets its best performance) or the less complex (number of parameters) of the best. Are the results statistically significant?
  - Suggestion: explore 2d binary classification datasets with 500 to 1000 points. The datasets can be linearly or non linearly separable, may have regions. You can also try more than two classes and actually more than 2 dimensions (but then explanations may be harder to find)
- For each resulting model, try to obtain **accuracy**, **AUC**, **F1**, **number of parameters/complexity**, **training time**. Use cross validation. Take into account random aspects of the algorithm.
- For each resulting case, try to find explanations for why the best method is the one you found.

- **Produce an essay** in the style of a paper with your results and conclusions. Use the following template: https://www.overleaf.com/latex/templates/ieee-conference-template-example/nsncsyjfmpxy (or a similar one).

Suggested structure

- Introduction
- Objectives
- Methods and hyper parameters
- Dataset generation
- Experimental setup
- Results and limitations
- Related work
- Discussion and conclusions
- References

To submit:

- Important notes (I apologize for the caps but these are recurrent issues):
  - ONLY MOODLE CAN BE USED FOR SUBMISSION. No email submission.
  - DO NOT SUBMIT ALL THE FILES AS A ZIP OR EQUIVALENT.
- Submit:
  - a pdf version of the essay (maximum 8 pages, including references)
  - a fully operational support R Markdown document or a Jupyter notebook.
  - A 10 minutes video presentation of the work, featuring all the elements of the group. Please include a link to the video as a **footnote on the first page of the essay**.
- Guidelines:
  - The document should have a clear narrative interleaved with plots and tables.
  - The objectives for each experiment and plot should be clear so that the reader understands why it is worth to read a particular part.
  - The conclusion should contain the important lessons learned.
  - It is not necessary to describe the methods. It is more important to point out the differences in the methods and relate those with the concepts you have learnt.
  - Presentations are **collective**. Different members of the group present different parts as the group wishes but all should participate.

Evaluation

- This assignment is worth 5 values out of twenty. The previous assignment should be worth 6 values but it has been announced as 3.5. Please get in touch with me if you have a problem with that.
- Components
  - Essay 30%
    - Clear narrative 15% (the reader understands what is going on and why experiments are being made and lead to the results and conclusions)
    - Writing correctness 15% (Avoid typos and grammatical errors. Use clear sentences, to the point.)
  - Technical 60%
    - Correctness 20% (Avoid technical errors)
    - Coverage 20% (The requested tasks were covered)

    * Conclusions and interpretation 10% (good insights and good links to the course)
    * Added value 10% (Out of the box ideas that had not been requested)
- Presentation 10%
    * Clarity, confidence, creativity, communication.

Groups

Assignments are submitted by groups of up to 4 students. Different elements may have different grades based on the contribution distribution. 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 perform overlapping work and only after that, exchange ideas with each other. Group work is important for learning from other people, but each individual must acquire independent skills.

Submissions

Formal deadline is 8th June 2021, to be submitted in moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

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 of the 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.