Artificial Intelligence L2-L3 project

2023-24

Dear all,

Due to the change in the 23-24 curriculum for the years to come, Both L2 and L3 students have attended the same AI course. This has had some impact, notably that the number of Practical Work groups increased from 2 to 4, meaning that it was not possible this year to include courses on Deep Learning in this course (that L2 students will have next year).

But while Deep Learning appears everywhere (making it easy to find information about it), it has a structural problem that impacts AI in general: it is not an explainable model (because once Deep Learning has “learned” something into its millions of neural connexions, it is impossible to know what it has leant.

Therefore, during the course, I have (among other algorithms) presented Decision Trees, which (contrarily to Deep Learning) are a way to do Data Science in an explainable way.

However, Decision Trees also have a problem: they lead to “overfitting” and may result in very deep trees if there are many decision variables. For this, it has been proposed to evolve for an identical dataset **several** “sub-trees”= trees with part of the data missing. The part of the data that is removed to create a “sub-tree” is random, and all of the sub-trees resulting from the elaboration of Decision Trees on the randomly reduced data-set is called… a Random Forest and it is observed that the performance of a Random Forest is much better than the performance of a single Decision Tree on the same data-set.

For this project, I have found 2 great web pages that explain the concept of Random Forests really well:

1. https://medium.com/@brijesh_soni/why-random-forests-outperform-decision-trees-a-powerful-tool-for-complex-data-analysis-47f96d9062e7#:~:text=Random%20forests%20are%20an,subset%20of%20the%20input%20data.
2. https://towardsdatascience.com/interpreting-random-forests-638bca8b49ea

After the lecture courses we have done together, you have ALL the elements to deeply understand the Random Forest concept described in these 2 web pages, so, as part of this project, please read and understand them.

The second link “Interpreting Random Forests: Comprehensive guide on Random Forest algorithms and how to interpret them“ uses the Wine Quality Dataset as an example on the interpretation of a Random Forest.

What I am asking you in this project is to create and interpret a random forest on the same dataset as the Wine Quality Dataset, **but without the information on Alcohol content** that you will be able to access here: https://seafile.unistra.fr/f/04d6b66055e049179c6f/?dl=1

In the beginning of her web-page, Mariya Mansurova says that thanks to her article, you will ***“learn how to find answers to the following questions:***
What features are important, and which ones are redundant and can be removed?
How does each feature value affect our target metric?
What are the factors for each prediction?
How to estimate the confidence of each prediction?”

The project is to be done by a group of 2 or 3 students (please register on this Google Sheet page).

You need to provide:

1. a 5-10 pages report written in LaTeX **presenting your work and answering the questions above** (without the information on alcohol content for red and white wine).
2. a link to the zip file containing the directory of your work.

The deadline for this project is MARCH 24

Please remember that I will be attentive to cheating (plagiarism)

Pierre Collet