Dados e Aprendizagem Automática
1º Ano, 1º Semestre
Ano letivo 2024/2025

Group Practical Work
October 2024

Theme

Design and optimization of Machine Learning models

Learning Objectives

The aim of this assignment is to sensitize and motivate students to design and develop a Machine Learning project using, among others, the learning models covered throughout the semester.

Work Group

Work groups can be formed by students from different PL class sections but cannot exceed 4 members. A link for students to register in a group will be made available in due course.

Introduction

Mild forgetfulness is often a normal part of ageing. But for some people, memory and thinking issues can become more serious as they get older. Cognitive impairment and dementia are increasingly frequent worldwide, impacting the quality of life of millions of patients and their families. These conditions can be caused by various factors such as genetics, lifestyle, and health conditions. Mild Cognitive Impairment (MCI) is a condition characterized by a cognitive decline that is greater than what would be expected for an individual's age and education level but does not significantly interfere with daily activities. This condition is distinct from dementia, in which cognitive deficits are more severe and widespread and significantly impact daily functioning. However, MCI with memory complaints and deficits has a high risk of progressing to dementia, particularly Alzheimer's disease (AD). This disease is a neurodegenerative disorder characterized by a gradual decline in chronic primary memory and cognitive impairment. The increasing incidence, high rate of disability, and high cost of treatment have made AD one of the most serious diseases affecting humanity. MCI is a transitional state between normal ageing and AD. People with MCI have a higher risk of developing AD. Therefore, in addition to researching this disease, it is also important to predict its onset in time for potential treatment to reduce the number of people with dementia in the long term.

Through the use of Magnetic Resonance Imaging (MRI) of the brain, we can observe differences that are linked to MCI such as shrinkage of the hippocampus, an essential region of the brain responsible for memory, enlargement of the fluid-filled spaces (ventricles) in the brain, and reduced use of glucose.

Radiomics is a quantitative approach to medical imaging that provides textural information through the mathematical extraction of the spatial distribution of signal intensities and pixel interrelationships. After radiomics feature extraction, Machine Learning (ML) or advanced statistical methods are used to analyse the features.

Many studies have been carried out to develop classification models for MCI, AD and diagnostic purposes. More recently, research has focused on predicting the progression of MCI to AD, which is what we aim to do with this

assignment. Fig. 1 illustrates that MCI symptoms can remain stable for years, progress to Alzheimer's disease or another type of dementia, or improve over time.
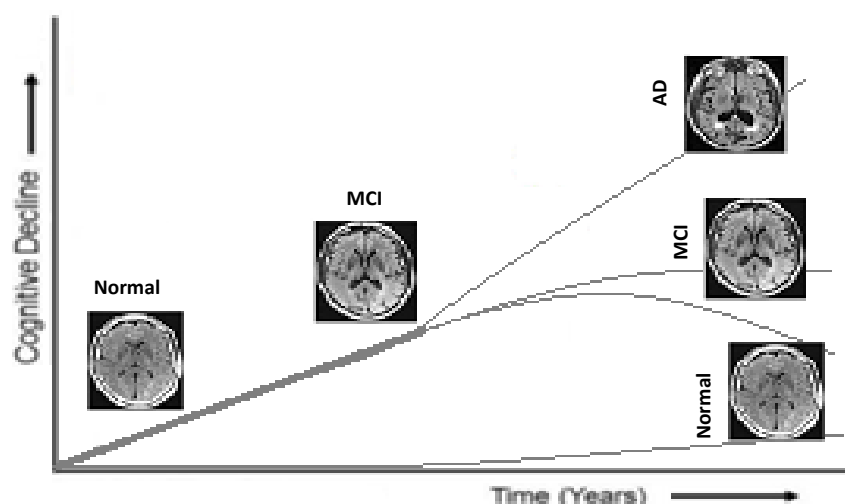

Fig. 1. Progression from normal aging to Alzheimer's disease or another dementia

The data that well be used was obtained from the Alzheimer's Disease Neuroimaging Initiative[1] (ADNI). All the MRI scans were acquired using MPRAGE and had 256 slices with a $1mm^3$ isotropic voxel resolution in NIfTI format. The extraction of radiomic features has been achieved using the open-source package PyRadiomics[2].
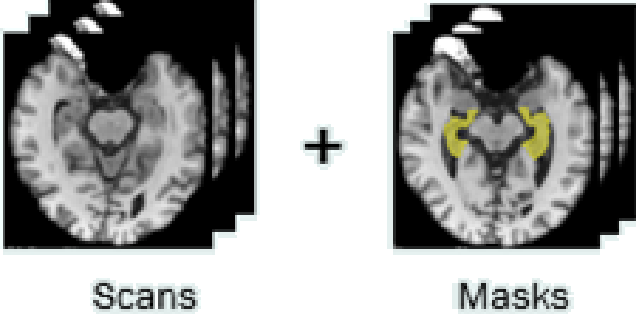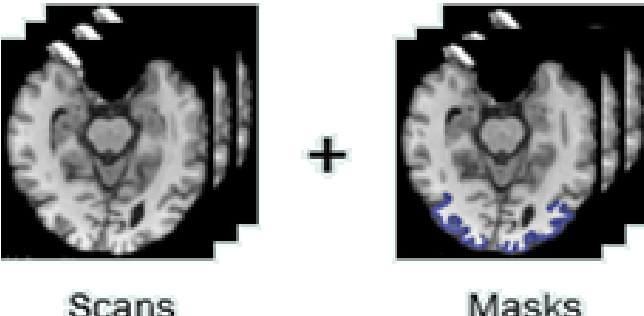
Table 1. Data Description

Patients	426
Exams	852 (two per patient - month 0 and month 24)
Sequence	MPRAGE
Protocols	Acquisition Plane=SAGITTAL; Slice Thickness=1.2; Acquisition Type=3D; Manufacturer=GE MEDICAL SYSTEMS; Weighting=T1
	Acquisition Plane=SAGITTAL; Slice Thickness=1.2; Acquisition Type=3D; Manufacturer=SIEMENS; Weighting=T1
Age Range (month 0)	55.2-91.0 (years old)
Average Age	75.3 (years old)
#Male	241
#Female	185

Only transitions where the patient maintains or worsens their condition were selected: CN-CN, CN-MCI, CN-AD, MCI-MCI, MCI-AD, and AD-AD (CN-AD was excluded Since there was only one exam with that transition).

---

1 http://adni.loni.usc.edu/
2 https://pyradiomics.readthedocs.io/

Table 2. Datasets illustration

Dataset	Ilustration
DS_{hippo}	
DS_{occ}	

For this assignment, two datasets were created (Table 2), each representing the extraction of radiomic features from distinct brain regions: the hippocampus and the occipital lobe. The hippocampus dataset (DS_{hippo}) was selected due to its significant relevance in Alzheimer's research, while the occipital lobe dataset (DS_{occ}) serves as a control, as this region is not typically associated with dementia. The hypothesis is that radiomic features in the postulated region of interest will demonstrate differences that will consequently support delineating patients with MCI that will evolve to AD and those who won't.

Assignment

The assignment includes 2 TASKS.

ANALYSIS AND VALIDATION TASK:

- Explore, analyse and prepare the datasets provided *(DS_{hippo} e DS_{occ})*, seeking to extract relevant knowledge in the context of the problem in question;;

- Use the control dataset *(DS_{occ})* and the models developed for the Competition Task to obtain F1 values (and other metrics) to check whether the hippocampus is indeed relevant in predicting the evolution of dementias and the occipital lobe is not;

- Obtain and critically analyze the results.

COMPETITION TASK (DS_{hippo}):

- The groups will have to work on the dataset *(DS_{hippo})* available at **https://www.kaggle.com/c/sbsppdaa24** :
    - The link above redirects to the Kaggle platform where a competition has been created. The dataset to be used in the competition, as well as all the details and how it works, are available at that link;

- o The first step is to access the Kaggle platform using the following link to register for the competition: **https://www.kaggle.com/t/77fa3e5211a2448ea731fc4d84a4b6f1**

  Students should then form teams with the other members of the working group. The team name should follow the format **GRUPO_<CURSO>_<X>** where **<CURSO>** corresponds to the master's programme (MMC, MEI ou MIEI) and **<X>** to the group number. Submissions cannot be made on the Kaggle platform as long as the group is incomplete;

- Design and optimisation of Machine Learning models for the competition dataset:

  - o Students must submit the results obtained on the Kaggle platform in order to obtain the model's F1 macro score;

  - o There is a **daily limit of 3 valid submissions**, so you should endeavour to start submitting as soon as possible. The competition closes at the end of **20 January 2024**;

- Obtain and critically analyze the results;

- Interpret the results acquired and define their usefulness in the context of the problem underlying the dataset worked on. Determine and explain the most relevant results.

Delivery and Assessment

The results obtained should be the subject of a report, limited to 20 pages, which presents, among other things:

- The areas to be tackled, the objectives and how it is proposed to achieve them;

- The methodology used and how it was applied;

- Detailed description and exploration of both datasets and any processing carried out;

- Description of the models developed, what their characteristics are, how and on what parameters the model was tuned, training characteristics, among other details that should be provided;

- Summary of the results obtained and their critical analysis;

- Presentation of suggestions and recommendations after analysing the results obtained and the models developed.

The whole process must be accompanied by examples and indications that make it possible to reproduce all the steps taken and the results obtained.

During the class period on **27 and 28 November 2024**, a checkpoint will be made on the work carried out by the working groups, with each group using the means it considers most appropriate to demonstrate the results obtained.

On **22 and 23 January 2025** there will be presentation sessions for the work carried out in both TASKS. The working groups will have to choose the slot they want for their presentation, which will be made available in the coming weeks. Each group will have 10 minutes to make their presentation, using whatever means they deem most appropriate.

The report, as well as all other elements produced, must be compressed into a single zip file which must be submitted by a member of the group by **21 January 2025** on the University of Minho's e-learning platform (in "*Conteúdo/Instrumentos de Avaliação em Grupo/Submissão TPG*").

Peer Assessment

Each group should carry out a collective analysis of the contribution and effort that each member has made to the progress of the work. From this analysis they should be able to identify the members who worked above, at and below the average. For this assessment component, 1 value is provided for each student, reflecting their individual contribution to the development of this assessment tool.

So, one member of the group should send an email, with the other members of the group in CC, to valves@di.uminho.pt, filipa.ferraz@di.uminho.pt, dad@di.uminho.pt and bruno.fernandes@algoritmi.uminho.pt. The subject should be "**AP DAA - Avaliação Por Pares**".

In the text of the email, each member of the group should indicate their delta (the amount to be added to the mark for this component). Remember that deltas can be negative, zero or positive and that, in each group, the sum of the deltas must always be equal to 0.00 and, individually, can never exceed one.

Example 1 (corresponds to an equal effort by all):

PG1234 João DELTA = 0
PG5678 António DELTA = 0
PG9123 Maria DELTA = 0
PG4567 Rita DELTA = 0

Example 2 (António receives 1 additional value, Rita keeps her classification, João and Maria are deducted 0.5 values each):

PG1234 João DELTA = -0.5
PG5678 António DELTA = 1
PG9123 Maria DELTA = -0.5
PG4567 Rita DELTA = 0

Code of Conduct

The participants in this academic work declare that they have acted with integrity and confirm that they have not resorted to the practice of plagiarism or any form of misuse or falsification of information or results in any of the stages leading to its preparation. They also declare that they are aware of and have respected the University of Minho's Code of Ethical Conduct.

References

In addition to the material provided in class, it is advisable to consult sources such as:

- Machine Learning. T. Michell, McGraw Hill, ISBN ISBN: 978-1259096952, 2017.

- Introduction to Machine Learning. Alpaydin, E. ISBN: 978-0-262-02818-9. Published by The MIT Press, 2014.

- Computational Intelligence: An Introduction. Engelbrecht A., Wiley & Sons. 2nd Edition, ISBN: 978-0470035610, 2007.

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Hastie, T., R. Tibshirani, J. Friedman, 12nd Edition, Springer, ISBN: 978-0387848570, 2016.

- Machine Learning: A Probabilistic Perspective. K.P. Murphy, 4th Edition, The MIT Press, ISBN: 978-0262018029, 2012.