Machine Learning Sentiment analysis JUNE
Introduction to the Machine Learning

Machine learning is a subfield of computer science and statistics that deals with the construction and study of systems that can learn from data, rather than follow only explicitly programmed instructions. Besides CS and Statistics, it has strong ties to artificial intelligence and optimization, which deliver both methods and theory to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines and computer vision. Machine learning, data mining, and pattern recognition are sometimes conflated.

Machine learning tasks can be of several forms. In **supervised learning**, the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. Spam filtering is an example of supervised learning, in particular classification, where the learning algorithm is presented with email (or other) messages labeled beforehand as "spam" or "not spam", to produce a computer program that labels unseen messages as either spam or not.

In **unsupervised learning**, no labels are given to the learning algorithm, leaving it on its own to groups of similar inputs (clustering), density estimates or projections of high-dimensional data that can be visualised effectively.: Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end. Topic modeling is an example of unsupervised learning, where a program is given a list of human language documents and is tasked to find out which documents cover similar topics.

In **reinforcement learning**, a computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not.

Working Methodology

As has been discussed in class all data mining process comprising different processes, these are:

- Identification of the problem and resolve it by supervised algorithms, unsupervised or reinforced, or by incorporating several of these learning techniques.
- Pre-processing the data, cleaning, reduction, generation of new variables (review theory).

- Selecting the most appropriated algorithm (compare several algorithms) based on a process of evaluation or estimation error based on **N-fold cross validation techniques**, or other methods.
- Generation of model.
- Prediction or classification with new entries.

Please use the examples of the PDU:

https://pdu.usj.es/mod/resource/view.php?id=32215

Description of the Problem

The problem to resolve is based on defined in the following url:

http://archive.ics.uci.edu/dataset/33/dermatology

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal.

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

Topic Objective

The main objective of this topic is the acquisition of the following concepts and skills:

- Understand the concept of Machine Learning
- Facing a real problem.

- Implement the system in Python

Questions to Resolve

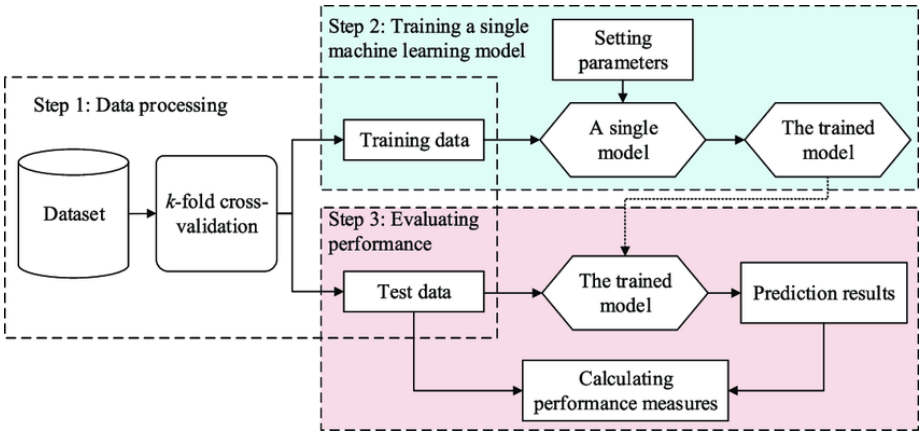
Practical Part (5 points) Evaluation Criteria

Implement (and describe) the machine learning process to resolve the problem. Describe the results obtained. Use two or more different machine learning algorithms to resolve the problem.

- First understand and describe as much as possible the flow of data mining generated.
- You SHOULD describe your code and clearly define all your steps. You should describe the different steps as much as possible, Saying WHAT you have done and WHY (1,25 points) (-0.25 each step missing). The evaluation should be done with Cross validation.
- 0,25 point by each algorithm tested (MAX 0.75). Compare results with all features or removing laboratory analysis.
- Implement almost one bagging algorithm, why in this case improve the results? Or why not? (2 points)
- Compare the different algorithms with the different metrics and explain the confusion matrix.

Theoretical Part (5 points) Evaluation Criteria

(2 points) Explain the following diagram.



- Description of the algorithm. Explain with your words the diagram , use references, and link within the description in course slides . (1,25p)

- Could you highlights the differences between this diagram and the one presented in the course slides in detail, and also elaborate on how this difference enhances the overall process?" (0,75)

(2 points) Explain ID3 Algorithm steps by steps making the calculations for the following example.

- Description of the algorithm. Explain with your words ID3 Algorithm algorithm, use references within the text and generate a diagram with the sequence of the algorithm. (1,25p)

- making the calculations for the following example subset of titanic dataset(0,75)

Survived	Pclass	Sex	SibSp	Parch	Embarked
0	3	male	1	0	S
1	1	female	1	0	C
1	3	female	0	0	S
1	1	female	1	0	S
0	3	male	0	0	S
0	3	male	0	0	Q
0	1	male	0	0	S
0	3	male	3	1	S
1	3	female	0	2	S
1	2	female	1	0	C
1	3	female	1	1	S
1	1	female	0	0	S
0	3	male	0	0	S
0	3	male	1	5	S
0	3	female	0	0	S

(1point) Identify an original process from the University of San Jorge (that you are familiar with), provide concrete examples from the University of San Jorge, and explain how it could be improved using Machine Learning. Define how you would solve and apply it. Explain it step by step, using the CRISP-DM methodology. Make the example specific and elaborate on how to obtain the data for the University of San Jorge's use case.

Materials Needed

The required equipment will be the Colab tool and Python language.

Evaluation

For the evaluation of this activity will be delivered the following documents:

- Activity report (described in detail below)

- Python / Jupyter File pdfs.

The report must be delivered in PDF format or inside Jupyter Notebook and must include the following:

- Responding to questions, written in the form of small 7-9 pages and a description of the process developed
- Indicate if you have used ChatGPT o similar tool, where and how.
- Conclusion: summary of what they learned, identifying initial shortcomings that have been (or not) covered by this study, acquired new knowledge, new skills (ie, things that have been made during the activity that had never been made indicating its level of acquisition (from 0 to 3, to use all the same scale), already had skills but have been worked and reinforced this activity, assessment of individual work (hours, job performance, planning and implementation of task) and collaboration with other peers or the teacher. References, modeled standard citation, with each critic reviews.

Bibliography

Course bibliography

Course Slides