University of Algiers - 1 Benyoucef Benkhedda
Faculty of Sciences
Department of Computer Science

Academic Year 2021/2022
Module: Machine Learning

ML Mini Project

Project Objective:

The purpose of this project is, in the first place, to familiarize you with the structure, selection, and quality of datasets for supervised learning, including Regression and Classification issues. In the second place, to enable you to deepen your understanding of machine learning, by studying and applying different supervised learning techniques, seen in class, on the selected regression and classification problems.

The work **MUST** be done in teams of five students, which means **5 persons make a '*Team*'** to work together on this project. The team's members can be from different TP groups.

- **You are asked the following**:

a) Go to the *Machine Learning Repository of the University of California at Irvine (UCI)*.
https://archive.ics.uci.edu/ml/datasets.php

b) Start with a quick site scan to understand what it is. On the left side of the page, you can see different tasks including regression and classification.

c) Select from this website:

- Three (3) training data for the regression, and
- Three (3) training data for the classification.

- Whatever problems you select, make sure that they are appropriate for regression and classification techniques, and that they are not very complex or very simple problems. It's up to you to choose the datasets that suit you the most.

d) Introduce on the "*Machine Learning*" course platform (in order of priority of problems):

- The full names of your team members,
- The 3 data choices for the regression task,
- The 3 data choices for the classification task.

(choice N ° 1 is the one you prefer the most, then N ° 2, then choice 3 least preferable)

e) The submission of team names and problem choices on the course platform will be **closed on**

Friday 12/31/2021 at 11. 59p.m (23:59 - Be punctual!).

f) The assignment of a team number, a regression problem and a classification problem will be confirmed for each team by **Sunday 01/02/2022** at the latest to start your work. The first choice is given where possible. If the same dataset is chosen as the first choice by two different teams, a fair and equitable draw will be made.

- **Start the Project** :

1) Do the necessary research on your two assigned problems and apply the necessary data preprocessing to improve its quality, its representation if it is modified to be more appropriate to the techniques considered, etc.

2) Design your solution for each supervised learning technique covered in class and implement it to solve these problems using **Python *(Python 3 is recommended)***. Apply the techniques implemented on your datasets that will be used to train your model:

- *Linear Regression (LnR)*
- *Logistic Regression (LgR)*
- *Neural Networks (NN)*
- *Support Vector Machine (SVM)*
- *Decision Trees (DTs)*
- *Gradient Descent (GD)*
- *Normal Equation (NE)*
- *Regularization, Cross Validation, …*

3) Prepare a **report** of ***no more than 20 pages (without the appendix)***, *Times New Roman*, *11*, *single spaced* (*paragraph format*) in **English language.** This report will contain two main parts, a part for the regression task and a part for the classification task. For each part you must:

-Present the problem in question in detail with its characteristics.

-Present the selected data set, explaining the reasons for this choice. Present the characteristics of the database such as its size, the number and types of attributes with their meanings, the different problems present in the database with examples

-Present a clear explanations of the methods applied to improve its quality

-Any other information that you find useful and necessary related to your problems being treated.

-Explains with all the necessary details and justify your design choices of the best architecture and configuration for each technique and the results obtained, including summary / comparison tables, graphs, analyzes, discussions of results, conclusions, etc.

4) The report must contain the complete code of your implementation in the Appendix and must include:

- Guidelines for using your solution.
- A clear and precise mention in the report of *"**Who did What?**"*. Tasks must be shared equally between members of the same team; **you will lose points if this is not the case**. In addition to the fact that everyone will be questioned and graded on their part, members of the same group ***must all master their problems***.

5) Prepare a **PowerPoint presentation** of the entire Mini Project, including the demo of the solutions. Each team will have **20 minutes** to present the proposed solutions and answer questions. You must have a good grasp of the problems, the techniques and the various details of the solution design.

- **What you must submit for each team:**

a. Each team must submit a report respecting the number of pages and the format indicated above, and clearly explaining what was requested.

b. The report must be sufficiently complete with a written, scientific presentation as clear and short as possible (*it is not about filling in, a good report doesn't necessarily have to be long!*). You are limited to *20 pages* in *Times New Roman font*, *size 11*, not including cover page, references and appendix.

c. An **online submission** via the **ML course platform** of **3 files**:

a- The Report *(Two electronic versions (**pdf** + **word**))*
b- The Presentation
c- A rar / zip file that contains:

- The **report** of your mini project *pdf* and *word* versions
- The **processed datasets** used for both *Regression* and *Classification* problems
- All the *Scripts* of your solutions well organized in directories and having meaningful names. Use meaningful names for variables and functions in your implementation as well as clear and precise comments in the code to clarify your work.
- All screenshots of **figures** from your tests with meaningful names. The report should contain the comparison tables of the results and the figures that you think are important.

d. « ***Non-plagiarism and intellectual dishonesty commitment*** » downloaded from the platform, completed, signed by each member of the team, and inserted in your report at the end. Certifying that your reports as well as your code for the whole project is entirely your work except where references are given (without copy / paste).

Learn to rely on your personal efforts and develop your skills and intellectual honesty!

- **Important dates :**

- Completed Work Submission Deadline: **Wednesday February 9, 2022 11.59 pm**
- Presentation Submission Deadline: **Sunday February 13, 2022 11.59 pm**
- Oral Presentations Date: **from February 14 to 19, 2022.**

- **NB**

1. Work that is not submitted through the Machine Learning course platform will not be accepted.
2. The names of Scripts, Figures and Reports **must contain** the team number « *E.g.:* ***X_Team30.(ext)*** »
3. Only one submission per team, which means, one person will be responsible for submitting the names and team selections for the first phase, as well as the electronic versions for the second phase.
4. Each double submission results in loss of points for the team.
5. Each late submission results in loss of points for the team.
6. **Each plagiarism detection will result in loss of points, or even cancellation of the project, for the team, *except where references are given without copying / pasting* (an anti-plagiarism tool will be used to verify that it is of your work). No form of phrase and / or code plagiarism *(without references)* will be tolerated.**
7. The oral presentation of the work will be made according to a schedule which will be communicated to you later. Attendance at presentations according to the schedule is compulsory for all the teams presenting in the same date. Any question asked to the presenters, its owner will be rewarded according to its relevance**.**
8. The grade for your work will take into consideration the quality of the work; compliance with the requested formats; the respect of deadlines and the good presentation and organization of your project.
9. Strict adherence to the mentioned rules will be noted!
10. Any interesting innovation will be rewarded.
11. The score of each member of the same team will not necessarily be the same, each according to his/her contribution and understanding of the problem, techniques, solutions, presentation, answers to questions during the presentation, etc.…

Your team's score for the project will be relative to that of other teams…!!!

Doors are open for Creativity and Innovation …… EnJoY !

Good luck