UNIVERSIDADE DE AVEIRO
Physics Department

Foundations of Artificial Intelligence

PROJECT 1 Instructions

Each group of two students is supposed to work on one project topic.
You are strongly encouraged to propose a machine learning problem you would prefer to work, not listed below, that may reflect better your interests. Please, discuss your idea with the instructor.

I. PROJECT GOALS

The goal of this project is to apply suitable machine learning algorithms learned in class or self-learned to solve a specific data science problem (classification or regression). Represent the results in graphical/table formats and make analysis and conclusions.

II. PROJECT PROPOSALS

Project proposal 1: Breast Cancer Wisconsin (Diagnostic) Data Set

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

Project proposal 2: Car Evaluation Data Set

https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Project proposal 3: Wine Quality Data Set

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Project proposal 4: Machine Learning for cybersecurity
https://github.com/PacktPublishing/Hands-on-Machine-Learning-for-Cyber-Security

Project proposal 5: Mammographic Mass Data Set

This data set can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The aim is to discriminate benign from malignant cases assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases are benign.
Data source: http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Project proposal 6: Heart Disease Data Set

This dataset contains 4 heart disease related datasets. For the present project you will use the Cleveland database and the referred subset of 14 features form a total of 76 attributes. The goal is to distinguish presence (values 1,2,3,4) from absence (value 0) of heart disease in the patient.
Data source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Project proposal 7: Bank Marketing Data Set

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe or not a term deposit.
Data source: https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Recommended Data Repositories:

- Kaggle Data Repository : https://www.kaggle.com/datasets

- UCI Machine Learning Repository : https://archive.ics.uci.edu/ml/index.php

III. PROJECT ASSESMENT

1. **Report.** The project is evaluated based on a submitted report (IEEE Latex format). The work done by each student has to be explicitly specified. All project's files (pdf and Latex files of the report, the presentation slides and the code implementing the algorithms) are sent to the course instructor (petia@ua.pt) in a compressed format having the following name: P1_FIA2022_XXXXX_YYYYY (where XXXXX and YYYYY are substituted by the academic (mechanographic) number of each student. If the file is too big to email as an attached document, feel free to use any big file transfer option you may know (we transfer, dropbox, link in a cloud. etc.)

2. **Oral presentation** of the report in class (about 10-15 min.).

IV. Evaluation criteria (total score 20)

1. Report content (10):

- Data description and preprocessing (if necessary normalization, feature selection, transformation, etc.). Motivation for choosing the particular problem.
- Data visualization (histograms, box plots, other plots).
- Short description of the implemented ML models.
- Model training (data splitting – train, validate, test, k-fold Cross validation). Visualize graphically the cost function trajectory over iterations. Training with regularized and non-regularized cost function.
- Model hyper-parameter selection - regularization parameter *lambda*, number of NN hidden layer units, number of hidden layers (if necessary), etc.. Systematic approach instead of just one or several randomly chosen values.
- For a classification problem, you need to present the confusion Matrix (accuracy, precision, recall, F1 score, etc.).
- Performance comparison between the models.
- Results in graphical or table formats.
- Conclusions.
- Problem complexity.

2. Report formatting (3) :

- IEEE Latex format, affiliation (Department, University, subject, course instructor), abstract, keywords, work load per student.
- Sufficiently detailed report.
- References, reference citation in the report.
- Clear figures (title, legends, axis labels) and tables referred in the text.

3. Oral presentation (4)

- Slide Organization, slide numbers, affiliation.
- Clear and convincing presentation by both students.

4. Novelty and contributions (3)

- Compare your solution with the works of other authors (published references), try to propose a better solution, e.g. improve the performance of the ML model in solving the problem you work with.