Machine Learning Project, 2021/2022

The project should be done in groups of 2 students, in Python language, and comprises 2 parts: i) regression and ii) classification.

Each group should work independently. Consultation of other people, exchange of ideas or software is not allowed and may invalidate the work.

The teacher will help to clarify doubts about Python language or basic Machine Learning tools but should not solve the project. Furthermore, he/she will ask questions during the lab session to assess the student knowledge about these topics.

Part 1 - regression

First problem

Given a training set with 100 examples $\mathcal{T}_r = \left\{(x^{(1)}, y^{(1)}), \ldots, (x^{(100)}, y^{(100)})\right\}$, with $x^{(i)} \in \mathbb{R}^{20}$, $y^{(i)} \in \mathbb{R}$, we wish to train predictors $\hat{y} = f(x)$. The students may train more than one predictor but must select **only one** predictor to be evaluated. Predictor evaluation will be done using a different set of data (test set) that cannot be used to select the predictors.

To evaluate the predictor, the students receive the feature vectors, $x^{(i)}$, associated to the test set and should calculate the corresponding outcomes $\hat{y}^{(i)}$ and send them to the professor.

Second problem

The second problem is identical to the first one but some of the training examples (less then 10%) are not generated by the model used to generate the other data. They can, therefore, be considered as outliers. We wish to train one or more predictors valid for the majority of examples. Again, **only one** predictor will be evaluated using an independent test set without outliers.

Part 2 - classification

For classification a dataset of face images will be provided. This dataset is an adaptation of the well known UTKFace dataset. The version we will use as training set contains grayscale images of approximately 7300 subjects, with 50x50 pixels Two different classification tasks will be performed using these data. For each, the students may train several classifiers but must select **only one** to be evaluated. Classifier evaluation will be done using a different set of data (test set), which can not be used for training.

First classification task

The first classification task is a binary one where we wish to create a model that predicts the gender of each subject. For this task the label is either 0 (male) or 1 (female).

Second classification task

The second classification task is a multiclass problem in which we wish to identify a person's ethnicity. For this task the label is an integer from 0 to 3, denoting Caucasian, African, Asian or Indian.

Results evaluation

For each of the four tasks (first regression problem, second regression problem, first classification task, second classification task) each group will be asked to submit in fenix a zip file containing a vector with the prediction/classification of their 'best' model for an independent test set that will be provided (without labels). The score results of all groups will be compiled and used to create a leaderbord that students can use to check how their models ranks against others. The regression score will be the **mean squared error** and the classification score will be the **balanced accuracy**. The schedule with the deadlines for each submission will be announced in the Labs section of the course webpage in fenix.

Report

Each group must prepare a report with a maximum length of 10 pages. The report should present a detailed account of the two parts, including: goal, methodology, numerical evaluation (figures, statistics) and conclusions.

Assessment

The project will be evaluated by the professor responsible for each laboratory shift. Evaluation includes

- methodology;
- experimental results;
- project report;
- interaction with the professor during classes and presence in the lab sessions.

The professor may ask specific questions to each student during the laboratory classes and take this information into account.