Machine Learning

Homework 1: 10 class Classification

Master in Artificial Intelligence and Robotics

SAPIENZA
UNIVERSITÀ DI ROMA

Professor
Luca Iocchi

Teacher Assistants
Damiano Brunori, Michela Proietti

Homeworks

Each homework will assign up to 2 points that will be added to the final score of the exam in any session within this academic year.

Homework point will remain valid independently of acceptance/failure in exam sessions

Homeworks are not mandatory

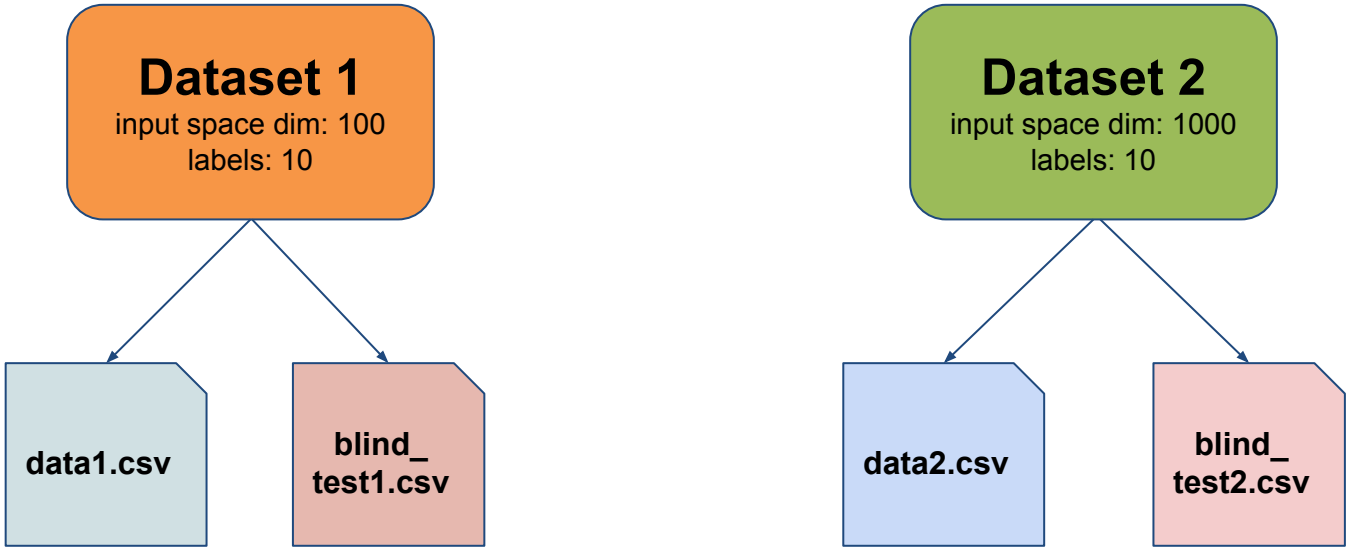
It is not possible to deliver homeworks outside the deadline given during the course

Homework 1 : 10-class classification
Deadline: **3/12/2023 23:59 CET** (STRICT DEADLINE!!!)

Datasets

Two datasets with *N* samples and different input space dimension *d*.

The datasets contain the data for training and a blind test for evaluation.



Training sets

Training set .csv files are structured as follows:

Index	X	Y
0	$[V^0_1, V^0_2, V^0_3, V^0_4, \dots V^0_j, \dots, V^0_d]$	C_k
…	…	…
i	$[V^i_1, V^i_2, V^i_3, V^i_4, \dots V^i_j, \dots, V^i_d]$	C_k
…	…	…
N	$[V^N_1, V^N_2, V^N_3, V^N_4, \dots V^N_j, \dots, V^N_d]$	C_k

Blind test sets

Blind test sets .csv files are structured as follows:

Index	X
0	$[V^0_1, V^0_2, V^0_3, V^0_4, \dots V^0_j, \dots, V^0_d]$
…	…
i	$[V^i_1, V^i_2, V^i_3, V^i_4, \dots V^i_j, \dots, V^i_d]$
…	…
N	$[V^N_1, V^N_2, V^N_3, V^N_4, \dots V^N_j, \dots, V^N_d]$

File format

The elements $\mathbf{v}^i$ belonging to the **X** column are feature vectors that represent your input data. The **Y** column contains the associated labels C_k (k=0,...,9).

Thus, each line *i* in column **X** is a feature vector structured as:

$$[V^i_1, V^i_2, V^i_3, V^i_4, \ldots V^i_j, \ldots, V^i_d]$$

where *d* is the size of the feature vector (100 for Dataset 1 and 1000 for Dataset 2) and *i=1,...,N* where N represents the total number of samples (= number of feature vectors = rows in the csv file).

File *load_data.py* contains a function to load data from .csv files

Assignment through Classroom

Dataset folder

https://drive.google.com/drive/folders/1JHEfzsdQchFzpT9D__iIgYz4q5VwjW1g

you can use any method, any tool, any programming language.
Exception: you cannot use solutions based on neural networks

You can pre-process the input data

Deliver though the assignment:

1) A report (PDF file)

2) A ZIP file with the code you used in the project.

3) [OPTIONAL] CSV files with predictions on the blind test sets

Assignment through Classroom

Report

- PDF file of about 10 pages excluding code, with your name and matricola code
- describe the implemented solutions
  - how data have been preprocessed
  - which methods/algorithms have been used
  - which configurations of the methods have been tried
  - description of the evaluation method used
  - results using appropriate metrics.
- provide comparative solutions for both datasets (hint: start with the easiest one).
- For each dataset
  - compare at least two different solutions, obtained, for example, by using different models, different learning algorithms, different values of some hyper-parameter, by testing different ways of preprocessing, etc.
- Conclusions to discuss the comparative results.
- Computational training time can also be interesting to report and comment.

Assignment through Classroom

Submit the files (PDF report, ZIPped code, and CSV predictions) through this assignment, make sure to turn the assignment in.

NOTES:

1) do ***NOT*** put the PDF report and the CSV files with predictions on the blind test sets into the ZIP file!

2) no other submission mode will be considered (e.g. do ***NOT*** send submissions by email).

This assignment must be **individual** (i.e., one submission for each student) and **original** (i.e., not equal or too similar to other works either from other students in this class or from other sources).

Evaluation will be based on the appropriateness and correctness of the described solution, regardless of the numeric results (as long as they are reasonable). The results on the blind tests also do not affect the evaluation of this homework.

Leaderboards

Submit the predictions of your best models for data in the blind tests
They will be scored and will lead to a leaderboard for each dataset (**not affecting exam votes!!!**)

IMPORTANT

Predictions on blind tests must be delivered as csv files with N rows (1 for each sample in the blind test set) and 1 column with the label of the class (0..9)

The name of the file should be d1_<your_matricola>.csv for the first dataset and d2_<your_matricola>.csv
Example: if your matricola is 1234567, the file names should be d1_1234567.csv and d2_1234567.csv.

Failure to follow these rules will not allow blind test evaluation (it will not affect the evaluation of the report in any case)

Matricola codes will be used to refer to your solution in the leaderboard.