Experimental Projects

Besides complying with the project's specifications, it is extremely important that students follow a sound methodology both in the data pre-processing phase and when running the experiments. In particular, no data manipulation should depend on test set information. Moreover, hyperparameters tuning should focus on regions of values where performance trade-offs are explicit.

Neural Networks

Use Keras to train a neural network for the binary classification of muffins and Chihuahuas based on images from this dataset.
Images must be transformed from JPG to RGB (or grayscale) pixel values and scaled down. The student is asked to:

- experiment with different network architectures (at least 3) and training hyperparameters,
- use 5-fold cross validation to compute your risk estimates,
- thoroughly discuss the obtained results, documenting the influence of the choice of the network architecture and the tuning of the hyperparameters on the final cross-validated risk estimate.

While the training loss can be chosen freely, the reported cross-validated estimates must be computed according to the zero-one loss.

(Kernel) Ridge Regression

Download the Spotify Tracks Dataset and perform ridge regression to predict the tracks' popularity. Note that this dataset contains both numerical and categorical features. The student is thus required to follow these guidelines:

- first, train the model using only the numerical features,
- second, appropriately handle the categorical features (for example, with one-hot encoding or other techniques) and use them together with the numerical ones to train the model,
- in both cases, experiment with different training parameters,
- use 5-fold cross validation to compute your risk estimates,
- thoroughly discuss and compare the performance of the model

The student is required to implement from scratch (without using libraries, such as Scikit-learn) the code for the ridge regression, while it is not mandatory to do so for the implementation of the 5-fold cross-validation.

Optional: Instead of regular ridge regression, implement kernel ridge regression using a Gaussian kernel.

Document Classification

Note that the following project relies on a dataset with text features containing italian terms.

The iBC company provides its data relating to tax documents and other types of documents available (eg: identity documents, "CU" certifications, tax receipts, F24). The files are in PDF

format. For each file, the relative annotation is available which defines the macro and micro-category (eg: "identity document" as a macro-category and "paper identity card" as a micro-category). The amount of data is in the order of thousands.
The aim of the work is to define a classification technique that identifies the macro-category and the micro-category of each document. The student can propose one or more techniques to solve the problem and must use a solution (proposed by the student) based on convolutional networks as a benchmark. The student must produce reliability metrics of the proposed techniques (e.g., balanced accuracy, F1, etc...) calculated with cross-validation techniques. The student must also produce confusion matrices for each proposed technique.
Optionally, the following problem can also be faced: some micro-categories of documents can be further divided into various groups according to the template that is followed to create the document. Ex: for “CU” certifications there are templates from various suppliers (e.g.: "Zucchetti"). The objective of the optional work is to define an unsupervised classification technique that subdivides these documents into various classes according to the template used.
The company will provide a remote machine on which the student can do the work. The student will not be able to copy files to other machines.
For further information or clarifications regarding the project, please contact Prof. Sergio Mascetti.

Theory Projects

Students who want to work on a theory project must write an email to the instructor indicating a topic (typically chosen among those covered in class) they would like to focus their project on. The instructor will then suggest one or two papers in that area.
Keep in mind that theory projects are specifically addressed to students who have a good disposition towards mathematics. Do not choose a theory project only because you are not good at coding.
Here is an example of a good report for a theory project.