Practical Session. Fundamentals of regression and classification.

Javier Civera, University of Zaragoza
jcivera@unizar.es

Objectives

1. Train regression and classification models
2. Evaluate the performance of the trained models
3. Understand the assumptions and the formulation of some of the basic regression and classification models and relate them to the obtained results

Introduction

In this practical session, we will train and evaluate simple regression and classification models in two public datasets: a simplified version of the Year Prediction - Million Songs Database and the CIFAR-10 Database.

Evaluation

You need to:

- share the Google Colab notebook with jcivera@unizar.es
- **submit it in Moodle max two weeks after the lab session (by Oct. 20th)**. To do this, download the colab document as .ipynb, complete the report in the given format[1] and upload them to moodle.

You are allowed 4 *late-days* in total **for all the labs**. You can use them when you want, i.e., for example you can submit 1 day late each of the four labs, or up to four days late one of the labs.

The following will be evaluated:

- Correctness and extent of the work.
- Correctness in the use of technical terms in the explanations.
- Analysis and discussion of the results.
- Organization and cleanliness of the code.

[1] https://www.overleaf.com/read/nccjycmbcnbj

1 Regression in the Year Prediction - Million Songs Database

The goal of this first part of the practical is to train and evaluate a regression model. We will use the Year Prediction - Million Songs Database[2]. Start your work from the following **notebook**, which loads the data.

1.1 Split the data

First of all, after you have your data, split it into the training, validation and test sets. Make sure you do not look at the test set until you report your final metrics.

1.2 Explore the data

Report first the size of the problem (how many samples are in the dataset? How many dimensions does your data have?).

Plot and/or analyze in more depth the relation of each dimension with the output. For example, for each input variable and the output, 1) compute the correlation (you can use the `numpy` function `corrcoef`) and 2) plot the relation between the two variables in a 2-d graph (you can use the `pyplot` function `scatter`).

1.3 Train several promising models, decide on the best one and report expected performance

Train the regression models we have seen in class (linear regression, polynomial models, ridge/Huber regression, RANSAC, trees, forests...). If training is slow, consider doing quick explorations reducing the dataset size or the data dimensionality. Report several metrics and analyze their values. Extract conclusions and explain the results as much as possible.

2 Classification in the CIFAR-10 dataset

The goal of the second part of the practical is to train and evaluate classification models in the CIFAR-10 dataset, a toy dataset for visual classification[3]. Start your work from the following **notebook**, which loads the data.

2.1 Split the data, explore it and train several promising models

As in the previous case, split your data into training, validation and test sets. Show the images and display the size of the dataset. Finally, train and evaluate several classification models from the ones we learned in class. If training is slow, remember that you can use PCA to reduce the dimensionality of the data or you can try things out with smaller versions of the full dataset.

[2]The dataset was first presented in: Thierry Bertin-Mahieux and Daniel P.W. Ellis and Brian Whitman and Paul Lamere. The Million Song Dataset. Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011). Find an overview here `https://samyzaf.com/ML/song_year/song_year.html`

[3]Find an overview of the dataset in `https://www.cs.toronto.edu/~kriz/cifar.html`

2.2 Transform the data

As you might now from the computer vision course, finding appropriate transformations of the visual data is complex. And as I told you in class, finding appropriate features is crucial for learning methods to work. For you to understand the important of using good features, we will use here the so-called GIST descriptors[4], for which you have code to extract them in the notebook of the practical.

Compute the GIST descriptors of all the CIFAR-10 images, and use them as the feature vectors to train again your models. Select reasonable values for the hyperparameters of your models and report the performance of your classifiers in the test set.

2.3 Show examples of the classification results

Once your classifier is trained, show several examples for which your classifier predicted the correct label, and some others for which your classifier predicted the wrong one.

[4]Originally proposed in Aude Oliva and Antonio Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. International Journal of Computer Vision, Vol. 42(3) (2001) 145-175. It basically convolves the image with Gabor filters at different orientations and scales, and averaging their responses in a grid.