FGV EMAp
ESCOLA DE MATEMÁTICA APLICADA

Machine Learning 1

LAB EXERCISE 3

This lab exercise must be submitted by February 7nd, 2024 at 10:00 am.
Late submissions will not be accepted and will be marked as zero.

This lab exercise is mandatory and could be assessed.
It is worth 1% of your total final grade for this course.

Submission Instructions

Submit your files through EClass. The files you submit cannot be read by any other students. You can replace your submission as many times as you like by resubmitting it, although only the last version sent is kept.

If you have last-minute problems with your EClass submission, email your assignment as an attachment to alberto.paccanaro@fgv.br with the subject "URGENT – LAB 1 SUBMISSION". In the body of the message, explain the reason for not submitting it through EClass.

All the work you submit should be solely your own work.
Coursework submissions will be checked for this.

Note that the instructions below are less "prescriptive" on the exact steps that you should carry out in this lab. You are getting better at Matlab and I believe it will be useful for you to start thinking by yourself about what you should do in order to perform a meaningful analysis of the data.

A piece of advice: ***always start by looking at the data…***

In this lab, you will build systems for classification that model the density of the different classes using single gaussians and histograms, as we saw in class.

You will be using 2 different datasets, that you will find on the Eclass page for this lab:

1) DensityEstimationDataset.mat . This dataset is constituted by *15000* points in 2 dimensions (first 2 columns), belonging to 3 classes, represented by the integers 0, 1, 2 in the last column.
2) GlassClassification.csv . This dataset is constituted by *198* points in 4 dimensions (first 4 columns), belonging to 2 classes, represented by the integers 1, 2 in the last column.

Using these dataset, you will build two systems for classifying points. These systems will model the density of each class separately and then use Bayes theorem for the classification.

The first system will model the densities using single gaussians, here you need to learn the parameters of the gaussians and create a function that is able to assign a class to a new point given those parameters.

The second one will be using histograms, this is a non-parametric model, you will decide the best way to design such system. (hint: your predictor function can use as input the whole training set instead of parameters).

Today you should aim at finishing a few scripts that implement the above descriptions.

During next week you should try to:

- divide your dataset into training and testing;
- measure the overall and per class accuracy on the training and testing dataset;
- measure your error for different values of the loss matrix;
- experiment with different bin sizes for the histogram-based classifier;
- Generate plots that show your results (with the proper descriptors).

Have fun ! ☺