Fundamentos de Data Science

LAB N. 3

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. The second one will be using histograms.

Once the model will have been trained, your system will be able to take in input a new point (that was not used for training) and predict the class to which it belongs.

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 error on the training and testing dataset (which error measure should you use?);
- measure your error for different values of the loss matrix;
- experiment with different bin sizes for the histogram-based classifier;
- make your code more modular, factoring it into functions;
- generate some plots (with title, variable names on the axes, etc).

Have fun ! ☺