LCPB 21-22 Exercise 4, XGBoost

1.

Consider the dataset generated for exercise 03 on convolutional neural networks (CNN), namely the samples of the stochastic time series with labels 0,1,2 depending on the eventual addition of another transient signal.

a) Compare the accuracy of a CNN with that of an *XGBoost* model trained with the features extracted by *tsfresh* from same data, in the limit of small datasets. For instance, try values of N=20, 50, 100, 150, 200, 250, 300, 400, 500. In all cases show also the standard deviation of the accuracy, obtained from several independent training and test procedures on different datasets.

b) For task a) we have seen during the lesson that XGBoost finds some features more relevant than others. Find the description of those features in the documentation and try to provide an explanation of why they are relevant for that problem.

c) OPTIONAL: with the features extracted by *tsfresh,* train a standard (non-convolutional) feed forward neural network (FFNN) and compare the performances with those of XGBoost. Than keep only the most relevant features from XGBoost and train another FFNN with this smaller set. Is the new FFNN working better than the one trained with all features?

2.

For the labeling of simple two dimensional data (as the one generated during the lesson), try different parameters (gamma, lambda, n_estimators, …), aiming to find the simplest yet effective XGBoost model that keeps a good accuracy.