Machine learning assignment nº 2

Introduction

Data availability has been exponentially increasing worldwide, compelling institutions and corporations to implement effective data management protocols as to exploit them and propel novel breakthroughs and the improvement of existing technologies or services. In such purpose’s view, machine learning is a fast-growing field of study that provides means for the development of systematic methodologies to data handling and data-driven decision making, ever more prevalent *e.g.* in the biomedical and customer services sectors.

A diverse set of machine learning algorithms is at the public disposal, each posing advantages and disadvantages depending on the task and available data at hand. Therefore, understanding the basic principles behind them and their pitfalls is paramount to make a grounded selection of a suitable model for a given problem.

Sample datasets act as proxies to reality. There is the assumption of being a reliable representation of a given phenomenon, governed by underlying periodic patterns, so that predictions can be inferred from them, in a process called learning. Provided that real world datasets are coupled with some degree of randomness or noise, outliers, sparse and missing data, each prediction is associated to an error, which is fractioned into irreducible error, bias, and variance; an error not only introduced by the dataset itself (noise, outliers, size), but also the principles behind each model (complexity and parameter numerosity). To produce good results on a task, an estimator must be able to make generalized predictions of reality from a sample as, that is, a compromise between *bias* and *variance* must be had. By common practice, estimator performance assessment is done *via* error monitoring given by defined loss functions, such as the residual sum of squares (RSS), cross-entropy and the 0-1 loss function, or *via* scoring metrics: Accuracy, F1-score, Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) score. Thusly, the suitableness of a candidate model for a given dataset is reflected by these metrics as well as training times, computational requirements, and model complexity.

In view of acquiring an intuition on what models are the most adequate for a given binary classification, machine learning problem, ten classifiers: Logistic Regression, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest, Support Vector Machines with linear, polynomial, and radial basis kernels, Multi-Layer Perceptron with ReLu and Hyperbolic tangent activation function, were trained and tested on several bidimensional artificial datasets of different topologies, sizes and noise degree, designed to create a performance discrepancy where a certain method or a subset of methods from the pool produce significantly better results than the others. When possible, hyperparameter tuning is used in order to select the optimal hyperparameter values. The experimental setup along with the results and their respective discussion will be further presented in this report.