Ph.D. in Information Technology
Thesis Defense

June 13th, 2024
at 14:30 p.m.
Sala Conferenze Emilio Gatti

Antonio CONSOLO – XXXVI Cycle

SPARSE SOFT DECISION TREES AND KERNEL LOGISTIC REGRESSION: OPTIMIZATION MODELS AND ALGORITHMS

Supervisor: Prof. Edoardo Amaldi

Abstract:

Machine Learning (ML) methods have recently achieved remarkable performance in several tasks arising in a variety of fields, ranging from medical diagnosis to natural language processing. Unfortunately, most of the underlying models are black boxes (e.g., deep neural networks) and this may substantially restrict their applicability. Model interpretability is of particular importance in applications where the ML models support the decisions of the domain experts and justifiable predictions are required. In fields such as healthcare, finance, education and criminal justice, ML models should not only provide accurate predicted responses for new input vectors but also valuable explanations concerning the derivation of the response in terms of the features. For some ML models, the interpretability can be improved by increasing the sparsity, that is, by reducing the number of nonzero parameters. According to Occam's razor principle, sparser models may also lead to lower generalization error.

In this thesis, we investigate and improve two ML methods, namely, soft decision trees for classification and regression tasks as well as kernel logistic regression for binary classification. In particular, we devise improved model variants, establish theoretical properties and develop decomposition algorithms for training them.

Decision trees are supervised ML methods widely used for classification and regression tasks. They are inherently interpretable since, for every input vector, they reveal to the domain experts the clear sequence of decisions (explanation) leading to the tree response. A decision tree is a binary directed graph where at each branch (internal) node a splitting rule is applied to the input space and at each leaf node is associated either a class label or a continuous response. During the past decade, growing attention has been devoted to globally optimized decision trees with deterministic or soft splitting rules at each branch nodes, which are trained by optimizing the error function over all the tree parameters.

Concerning soft classification trees, we first propose alternative sparsification methods based on concave approximation of the l_0 norm. The results reported for 24 benchmark datasets from UCI and KEEL repositories indicate that our approximate l_0 norm performs better than the original l_1 and l_∞ regularizations and leads to sparser trees, namely, with a smaller number input features per branch node. Then, we determine bounds on the Vapnik-Chervonenkis (VC) dimension of such models, which plays a key role in statistical learning to derive bounds on testing (generalization) error. Finally, we present a general node-based decomposition scheme for training soft classification trees and a practical proximal version. Experiments on larger datasets show that the proposed decomposition method is able to significantly reduce the training times without compromising the testing accuracy.

As to soft regression trees, we propose a model variant where, for every data point, a potential linear prediction is available at each leaf node but the actual tree prediction is the one associated to a particular leaf node. Our nonlinear optimization formulation for training such soft trees is well-suited to decomposition and to impose fairness constraints. After investigating the universal approximation property of our model variant, we present a convergent node-based decomposition algorithm which includes a heuristic for the reassignment of the data points along the tree and a specific initialization procedure. Experiments on 15 benchmark datasets from different types of applications show that our model trained with the decomposition algorithm outperforms two state-of-the-art soft and deterministic regression tree models based on continuous nonlinear and discrete optimization approaches in terms of both accuracy and/or computational time. Other experiments on the same datasets show that our SRT approach is significantly more robust than classical Hierarchical Mixture of Experts trained with the Expectation-Maximization algorithm and provides single soft trees whose testing accuracy is comparable to that of Random Forest. We also present l_0-based sparsification method to tackle sparsity on the branch nodes of soft regression tree. Moreover, we focus on enhancing the flexibility of such trees to consider two distinct group fairness measures, addressing potential algorithmic bias. Finally, we apply our model and decomposition algorithm to investigate a real-world case on the evolution of multidimensional inequality of opportunity in Australian HILDA survey dataset.

Logistic regression is a well-known probabilistic classification model that provides in addition to the class membership of the input vector also an estimate of the corresponding conditional probability. While the use of kernels allows to capture complex nonlinear relationships within the data, the resulting kernel logistic regression models are generally not sparse with respect to data points. Nevertheless, inducing sparsity by involving only a subset of data points during training can positively affect the model's generalization capability.

We consider sparse kernel logistic regression for binary classification. In particular, we present an exact sparsity-inducing formulation and devise a decomposition algorithm of sequential minimal optimization type, which exploits second order information, and for which we establish global convergence. Numerical experiments conducted on 12 datasets from the literature show that the proposed binary kernel logistic regression approach achieves a competitive trade-off between accuracy and sparsity with respect to two well-known alternative approaches, while retaining the advantages of providing informative estimates of the class probabilities.

PhD Committee

Prof. Pietro Luigi Belotti, **Politecnico di Milano**
Prof.Vicenzina Messina, **Università Degli Studi Di Milano- Bicocca**
Prof. Laura Palagi, **Università degli Studi di Roma “La Sapienza”**