# The soft tree: a new start in constructing tree-shaped predictors

**Antonio CIAMPI**

***Abstract:***
Prediction Trees are an invaluable tool in data analysis. Their strength is, however, less in providing highly accurate predictions than in gaining an economic and interpretable understanding of how predictors are related to outcome. Indeed a tree is a particularly transparent prediction rule based on a hierarchy of simple binary questions concerning an automatically selected number of predictors. On the other hand, the predictive skill of a tree is often inferior to that of other prediction rules.
Much work has been done in the last 15-20 years to 'boost' the predictive power of trees. Several approaches have been devised, such as Boosting, Random Forests, and, more generally, Ensemble Learning. These approaches achieve excellent predictions, but at the price of obscuring interpretation. There is a need to build interpretability in the highly efficient 'black boxes' thus obtained.

In this presentation we will outline a new approach to tree-growing that aims to improve prediction while preserving a measure of interpretability: the Soft Tree.
The Soft Tree originates in work on Symbolic Data Analysis. Symbolic Objects were used, in an earlier work, to model imprecise data. Our work made possible to build trees with 'hard nodes' (decision nodes based on binary question) out of imprecise data. This suggested the idea of replacing a hard node with a soft node in a tree-growing process.
We will describe the development of a soft tree. First, we considered the task of predicting a binary outcome. For continuous predictors, we replace a binary question with a question that has an answer between 0 and 1: thus, at a given node, a subject 'goes to the left branch' with probability $p$, which is a sigmoid function of a predictor $z$, and 'goes to the right' with the complementary probability. This idea is implemented in a tree-growing process that recursively selects predictors and estimates the sigmoid functions at the nodes. The construction is heavily based on the EM algorithm. The result is a tree in which all subjects determine all nodes. We expect that such a tree should be more stable and should yield more accurate predictions than an ordinary 'hard' tree, while preserving interpretability. Simulations and bench mark analyses appear to support this claim

Current and future research is directed at improving the algorithm, extending it to other outcomes (categorical, continuous, multivariate etc.), and gathering further evidence about its predictive skills, as compared to those of other prediction rules. As well, we intend to investigate the possibility of constructing soft trees from symbolic data—defining soft nodes from imprecise data.

***Keywords:***
Prediction trees, CART, Recursive Partition, imprecise data, probabilistic splits

***References:***
"Growing a tree classifier with imprecise data", A. Ciampi, E. Diday, J. Lebbe, E. Perinel, R. Vignes, in Pattern Recognition Letters 21 (2000) pp 787-803.

"Prediction trees with soft nodes for binary outcomes", A. Ciampi, A. Couturier1, Shaolin Li., STATISTICS IN MEDICINE, in Statist. Med. 2002; 21:1145-1165 (DOI: 10.1002/sim.1106)