Mixtures of tree-structured probabilistic graphical models for density estimation in high dimensional spaces

François Schnitzler (ULg)

September 24, 2012

A widely used strategy to derive a model of a stochastic system from observational data is the estimation of a multivariate probability density over the variables of the problem. Such a density can then be used to study the underlying stochastic phenomenon.

Probabilistic graphical models reduce the number of parameters necessary to encode a joint probability distribution by exploiting independence relationships between variables. However, using those models is challenging when there are thousands of variables or more. First, both learning these models from a set of observations and exploiting them is computationally problematic. Second, the number of recorded occurrences of the problem may be quite low with respect to the number of variables. The model constructed may therefore be influenced by the particular sampling of the realisations of the problem, and generalize badly on new, unseen realisations. This source of error is called the variance of the learning algorithm.

Within this context, the problem considered in this thesis is to study and improve the scaling of probabilistic graphical models in terms of the number of variables. The approach selected is to use mixtures of Markov trees.

Markov trees are a class of probabilistic graphical models limited in the probability distributions they can model. However, both learning and answering queries with such a model is considered to be computationally tractable. A mixture or an ensemble model is a weighted average of models. Such a mixture can be constructed to reduce the variance of a learning algorithm. In particular, the present thesis explores the possibility to build mixtures of Markov trees by using the perturb and combine framework. This approach consists in randomizing a learning algorithm and combining the outputs resulting from a repeated application of the randomized algorithm on a given learning set.

This thesis present several new algorithms for learning such mixtures of Markov trees. These algorithms are studied empirically on synthetic and more realistic problems from the literature.