Avd. Matematisk statistik

KTH Teknikvetenskap

Statistical Learning Theory for Chow-Liu Trees

Timo Koski
Institutionen för matematik
Kungliga tekniska högskolan (KTH) , Stockholm

As a starting point, the problem approximation or estimation of high dimensional multivariate distributions by products of lower dimensional distributions is introduced and analyzed. We consider the problem as a reverse I-projection, [3], (defined in terms of a Kullback distance) of the high dimensional multivariate distribution to an approximating structure. We shall fist rewrite the pertinent Kullback distance in terms of a set of mutual informations that only depends on the lower dimensional distributions.

For example, Bayesian networks provide an approximation that is a significant simplification of a joint distribution, as the approximating distribution is factorized according to an acyclic and directed graph. The statistical learning problem is that of finding from i.i.d. samples the structure or topology of the graph. There are a number of approaches to this problem [4]. Many of these invoke mutual information and other concepts of information theory frequently used in all applications of statistical learning [5].

When the approximating structure is a tree, the work in [1] established an effective method for maximum likelihood estimation of the probability distribution from i.i.d. samples. The method hinges upon the interpretation of maximum likelihood estimation as a reverse I-projection of an empirical distribution. In fact the algorithm reduces maximum likelihood estimation to solving a maximum spanning tree problem, with mutual informations as weights, as will be shown.

It will also be shown that there is a property of of almost sure asymptotic consistency for the maximum likelihood estimate. Then we shall study certain recent results from [6] finding the exponential rate of convergence of the

maximum likelihood estimation using the technique of large deviations and error exponents. The possibilities of extending these error exponents to other structures than trees will be touched upon.

Finally, we shall discuss a Bayesian predictive version of the likelihood function in the Chow-Liu theory, and demonstrate its applications to supervised, see [2], and unsupervised learning.

Referenser

[1] C.K. Chow & C.N. Liu: Approximating Discrete Probability Distributions with Dependence Trees. *IEEE Transactions on Information Theory* 1968, 14, 462-467.

[2] J. Corander, M. Gyllenberg & T. Koski: Learning genetic population structures using minimization of stochastic complexity. *Entropy*, 2010, in print.

[3] I. Csiszár & F. Matus: Information projections revisited. *IEEE Transactions on Information Theory* , 49, 1474− 1490, 2003.

[4] T. Koski & J. Noble: *Bayesian Networks. An Introduction*J. Wiley 6Sons, New York, London, 2009.

[5] D.J.C. MacKay: *Information theory, inference, and learning algorithms*, Cambridge Univ Press, 2003

[6] V.Y.F. Tan, A. Anadkumar, L. Tong, A.S. Willsky: A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree. *International Symposium on Information Theory, ISIT 2009*, IEEE Press 2009. London, New York, 2009.