Scalability and uncertainty of Gaussian processes

Hadji, M.A.

Summary

This thesis investigates the frequentist properties of Bayesian procedures, specifically the use of Gaussian process priors in Bayesian nonparametric statistics. Typically, in Bayesian statistics, there is no fixed unique parameter of our model, but rather a realization of a random variable which will act as a parameter. To this end, an prior distribution on the parameters is assumed. After observing the data, this leads to a posteriori distribution on the parameters, which is used to compute estimates of the parameters. Although this paradigm differs significantly from frequentist statistics, it is still interesting to see how the posterior distribution behaves as a random measure which depends on the "true" parameter.

One, very appealing advantage of the Bayesian framework is that it readily provides built-in uncertainty quantification. Indeed, since it is possible to sample from the posterior, the construction of credible sets containing a fraction of the posterior mass is relatively simple. In Chapter 2, we study the coverage of those credible sets resulting in from Gaussian process priors with squared exponential covariance kernel. As the sample paths of the process are infinitely smooth, the common practice is to rescale it in order to recover the underlying functional parameter. The optimal scaling depends on the smoothness of this parameter, which is generally unknown. The scaling hyper-parameter is thus generally learnt from the data using either hierarchical Bayes or Empirical Bayes techniques. Unfortunately, both methods initially lead to overconfident, unreliable uncertainty statements for a large class of parameters in the context of the Gaussian white noise model. However, blowing up the credible sets with a logarithmic factor or modifying the estimated hyper-parameter with a logarithmic term can get good frequentist coverage while maintaining a reasonable adaptive size.

The following chapters focus on the scalability of Bayesian methods in the context of Gaussian process nonparametric regression. Coming from the Bayesian paradigm, Gaussian process regression allows to make probabilistic statements the regression function based on the data. Moreover, when the noise in the model is Gaussian, it is possible to make use of conjugacy to obtain a closed form solution for the posterior distribution of the regression function. Nonetheless, this model is extremely greedy as its computational complexity scales cubicly with the number of observations. Distributed methods allow to divide the data across different machines which will all perform a local Bayesian nonparametric regression. The local solutions will then be collected by a global machine and aggregated into a global distribution for the regression function.

The naive approach is to simply perform a Gaussian process nonparametric regression with a random subset of the data in each machine and average all the local distributions into a global one. This approach quickly shows its frequentist limits as the convergence rate of the posterior will depends on the number of machines. Other methods are possible: down-scaling the prior locally and then averaging the results,

or up-scaling the likelihood and find the Wasserstein-barycenter of the resulting local distributions for instance. If the smoothness parameter of the Gaussian process prior matches the true regularity of the regression function, then both methods lead to near-optimal recovery and good uncertainty quantification, provided the number of machines does not increase too fast compared to the number of observations.

Another approach would be to partition the design space of the regression such that each machine performs a regression on one of the non-overlapping resulting sub-regions. Though the final posterior distribution contains discontinuities at the borders of each partitions, it will contract optimally to the "true" regression function when the rescaling hyper-parameter of the prior is well chosen. Furthermore, this approach can also lead to adaptive optimal contraction rates. Even when the "true" regularity is unknown, it is possible to learn the hyper-parameter using a hierarchical framework. This will also lead to optimal recovery of the regression function.

This approach can be seen as an aggregation of the posterior samples obtained by the different machines combined with weight functions. Indeed, the weight functions would be indicator functions of a sub-region of the design space. The discontinuities of the latter are then the result of the discontinuities of the weights. A thorough simulation study suggests that by choosing appropriate, data-driven weights, it is possible to achieve adaptive near-optimal recovery and coverage of the underlying regression function.