Scalable Inference for Gaussian Processes

Résumé du projet de recherche (Langue 1)

In many areas of sciences involving the studying of complex phenomena, a fundamental scientific question is how to reliably quantify the level of confidence in any predictions and in any conclusions made on the functioning of systems of interest. A well-calibrated quantification of the level of uncertainty in the analysis of these systems makes it possible to reliably assess the level of risk associated with some configurations of the system and balance the cost of decisions. This project employs the expressive language of probabilities to characterize the degree of belief about systems’ variables, as probability theory is the pillar on which a principled theory for decision-making is built. Applications that motivate the importance of being able to reliably quantify uncertainty include: Systemic diseases – Systemic diseases, such as diabetes, obesity, cardiovascular diseases, Alzheimer's disease, colorectal cancer, hepatocellular carcinoma (HCC), breast cancer, melanoma to name a few, represent some of the main causes of motor impairment and death in the western countries at present. Even though considerable steps have been made in the direction of understanding risk factors associated with these, we are still far from knowing definitive markers that can be targeted to stop or reverse the progression of systemic diseases. Natural disasters – Accurately and timely predicting the occurrence of extreme weather conditions, the effect of tsunamis on coastal reefs, and the incidence and magnitude of earthquakes is important in order to raise alerts in cases when there is a concrete danger for the population. Even though short-term predictions can reasonably be made, the extreme complexity and stochasticity of these systems poses formidable computational challenges in producing accurate predictions and well-calibrated confidence levels, so that natural disasters, as of today, still account for several deaths and billions of dollars of damages to cities every year. In many domains, models offer mathematical abstractions encoding any understanding on the way systems evolve and respond to certain inputs. Models are usually characterized by a number of parameters that may have an interpretation as unobserved physical quantities. In many fields, such as climatology and geo-sciences, given the complexity of the phenomena under investigation, there are several competing models that are adopted to make predictions and they are usually computationally intensive to run.

Résumé du projet de recherche (Langue 2)
The question that forms the motivation for this project is how to meaningfully exploit the information given by data, possibly in combination with a model, to accurately characterize uncertainty in predictions and in the analysis of the behavior of complex phenomena. The advantage of taking a probabilistic approach is that mathematical properties of probabilities are well understood and offer a principled mathematical implant - known as Bayes theorem - to update prior beliefs about some variables of interest in light of observed data. By taking a probabilistic approach, it is possible to (i) infer, rather than optimize, parameters, namely consider all model parameters that are consistent with observed data when making predictions, and (ii) carry out model selection, that is, obtain a quantitative criterion to assess whether a model is more suitable than another in modeling data. Furthermore, in cases where a mathematical formulation of a model is lacking, it is possible to infer functional relationships among variables. This can be done by placing a prior distribution over functions directly, and make data update this prior belief, thus effectively making data determine the suitable relationships existing among variables. Even in cases where a mathematical abstraction of the system exists but is computationally expensive to simulate, a widely adopted framework involves the emulation of the response of the model using a surrogate function. Again, this amounts in updating prior beliefs about functions that emulate the response of models based on simulations for some configurations of parameters, making it possible to account for the uncertainty in the emulation phase when making predictions. Probabilistic nonparametric modeling based on Gaussian processes offers a suitable framework for carrying out quantification of uncertainty in the aforementioned scenarios, and is the one that will be considered in this project. Although their name implies otherwise, nonparametric models are richly parameterized in a way that makes them flexible to capture the functional relationships that exist among data and increase their capacity as more data are processed. This is exactly what is needed in the “Big Data” era where we would like flexible models to learn from data, without imposing restrictive assumptions on the relationships among variables, and that can automatically tune their complexity to the kind and amount of data available. Although extremely appealing from the theoretical point of view, employing these models is computationally expensive, as they typically require storage that is quadratic in the number of data and numbers of operations that are cubic in the number of data. A further complication is that inference of these models requires repeating these computations several (sometimes millions of) times to yield accurate quantification of uncertainty, and most of the basic calculations, e.g., matrix factorizations and Markov chain Monte Carlo, are inherently serial. Moore's law suggests that the capacity of computers' memory and the number of operations that computers can make per second roughly double every other year. While this law has been holding since the birth of computing units, the development of modern computer architectures has changed considerably in the last decade. In particular, the trend is now to increase computational speed by exploiting parallel architectures. Even though the development of novel architectures is followed by the development of scientific libraries that exploit this, developing a scalable inference framework for nonparametric probabilistic modeling that can take advantage of the trend in the development of hardware requires the development of novel syntheses of statistical inference and computing approaches.

Informations complémentaires (Langue 1)

The probabilistic nonparametric modeling framework is extremely powerful as it offers a principled way to quantify uncertainty in the analysis and predictions of complex systems. For most systems of interest, however, employing this framework when dealing with the variety and today's sheer amount of data poses novel computational and statistical challenges, and calls for the development of novel scalable paradigms of computations. The aim of this project is to develop novel techniques at the interface between Mathematics, Statistics, Computer Science and Physics to scale computations for probabilistic nonparametric models, and demonstrate that this will ultimately lead to a better quantification of uncertainty compared to current approaches. Objective 1: Develop novel techniques to scale and distribute computations for Gaussian process models. Objective 2: Demonstrate the capability of probabilistic nonparametric models to accurately quantify uncertainty in the analysis of data outperforming current state of the art approaches. The two objectives contain an exciting mix of theoretical and applied work so that the impact of the theoretical work will be maximized.