Bocconi

Università Commerciale Luigi Bocconi

Department of Decision Sciences

Statistics Seminar

Bayes and Big Data: The Consensus Monte Carlo Algorithm

Steven Scott

Google

Thursday, 9th October

12:30pm Room 3-E4-SR03 Via Rontgen 1 Milano

Abstract

A useful definition of “big data” is data that is too big to comfortably process on a single machine, either because of processor, memory, or disk bottlenecks. Graphics processing units can alleviate the processor bottleneck, but memory or disk bottlenecks can only be eliminated by splitting data across multiple machines. Communication between large numbers of machines is expensive (regardless of the amount of data being communicated), so there is a need for algorithms that perform distributed approximate Bayesian analyses with minimal communication. Consensus Monte Carlo operates by running a separate Monte Carlo algorithm on each machine, and then averaging individual Monte Carlo draws across machines. Depending on the model, the resulting draws can be nearly indistinguishable from the draws that would have been obtained by running a single machine algorithm for a very long time. Examples of consensus Monte Carlo are shown for simple models where single-machine solutions are available, for large single-layer hierarchical models, and for Bayesian additive regression trees (BART).

Department of Decision Sciences

Via Röntgen 1 - 20136Milano

Tel. 02 5836.5632
Fax 02 5836.5630