Scoring Rules, Divergences and Information in Bayesian Machine Learning

Ferenc Huszár

Probability distributions and random variables play a central role in Bayesian computation. In order to design effective Bayesian machine learning algorithms one therefore needs rich frameworks for describing, characterising, comparing and manipulating probability distributions and random variables. This thesis is centred around one such framework: information geometry defined by strictly proper scoring rules. Part I of this thesis is devoted to a general introduction to this framework.

The scoring rule framework allows us to define general notions of divergence and information. Divergence functionals quantify the difference or discrepancy between probability distributions. The best known example, the Kullback-Leibler (KL) divergence, has been used for decades in a wide range of Bayesian applications. Divergences are of central importance in approximate Bayesian inference, and the generalisations presented here can be used to construct and analyse new classes of methods. In Part II of this thesis I introduce a general framework called loss-calibrated approximate inference. I will also study the properties of two closely related algorithms, kernel herding and Bayesian quadrature. Both of these methods are special cases of loss-calibrated inference.

Scoring rules also provide meaningful generalisations of Shannon's mutual information which has been used in a variety of applications in statistics and communications. Quantifying the value of information is essential in active machine learning and optimal experiment design, where one seeks to select the most informative measurements to perform in a sequence of experiments. In Part III of this thesis I introduce a unifying framework for Bayesian active learning based on scoring rules. I will show how a wide range of old and recent work can be understood as special cases of this framework. I will also present a technique called Bayesian Active Learning by Disagreement (BALD) and apply it to classification, preference elicitation and quantum tomography.