Learning on the Web

Fernando C.N. Pereira

University of Pennsylvania, USA

It is commonplace to say that the Web has changed everything. Machine learning researchers often say that their projects and results respond to that change with better methods for finding and organizing Web information. However, not much of the theory, or even the current practice, of machine learning take the Web seriously. We continue to devote much effort to refining supervised learning, but the Web reality is that labeled data is hard to obtain, while unlabeled data is inexhaustible. We cling to the iid assumption, while all the Web data generation processes drift rapidly and involve many hidden correlations. Many of our theory and algorithms assume data representations of fixed dimension, while in fact the dimensionality of data, for example the number of distinct words in text, grows with data size. While there has been much work recently on learning with sparse representations, the actual patterns of sparsity on the Web are not paid much attention. Those patterns might be very relevant to the communication costs of distributed learning algorithms, which are necessary at Web scale, but little work has been done on this.

Nevertheless, practical machine learning is thriving on the Web. Statistical machine translation has developed non-parametric algorithms that learn how to translate by mining the ever-growing volume of source documents and their translations that are created on the Web. Unsupervised learning methods infer useful latent semantic structure from the statistics of term co-occurrences in Web documents. Image search achieves improved ranking by learning from user responses to search results. In all those cases, Web scale demanded distributed algorithms.

I will review some of those practical successes to try to convince you that they are not just engineering feats, but also rich sources of new fundamental questions that we should be investigating.

J. Gama et al. (Eds.): DS 2009, LNAI 5808, p. 31, 2009.