Preface

The methods of conventional statistics were developed in times where both dataset collection and modeling were carried out with paper and pencil. The appearance of computers first displaced the pencil for model calculation. Machine learning gained prominence by exploiting this new opportunity, enabling the construction of efficient high-dimensional models using comparatively small training sets

Another change of similar magnitude is underway. Pervasive and networked computers have reduced the cost of collecting and distributing large-scale datasets. We now need learning algorithms that scale linearly with the volume of the data, while maintaining enough statistical efficiency to outperform algorithms that simply process a random subset of the data.

For the sake of the argument, assume that there are only two categories of computers. The first category, the “makers,” directly mediate human activity. They implement the dialogue between sellers and buyers, run the accounting department, control industrial processes, route telecommunications, etc. The makers generate and collect huge quantities of data. The second category, the “thinkers,” analyze this data, build models, and draw conclusions that direct the activity of the makers. Makers produce datasets whose size grows linearly with their collective computing power. We can roughly write this size as nd where n is a number of examples and d the number of features per example. The computational needs of state-of-the-art learning algorithms typically grow like n^2d or nd^2. If thinkers were to use these algorithms, their collective computing power would have to vastly exceed that of the makers. This is economically infeasible. The Googles of this world cannot deploy vastly more resources than all their customers together: their advertisement income cannot exceed the wealth of those who receive the messages, because the advertisers would never sell enough to recoup their expense.

The title “Large-Scale Kernel Machines” may appear to be a contradiction in terms. Kernel machines are often associated with dual techniques that implement very large parameter spaces at the expense of scalability in the number of examples. However, as was made clear during the NIPS 2005 workshop, kernel machines can scale nicely by cleverly approximating the conventional optimization problem solved during learning. Because we derive these large-scale systems from relatively well understood kernel machines, we can assess more soundly the impact of their increased scalability on their statistical efficiency.

This book offers solutions to researchers and engineers seeking to solve practical learning problems with large-scale datasets. Algorithms are described in detail;

experiments have been carried out on realistically large datasets. Many contributors have made their code and data available online.

This book is also intended for researchers seeking to increase our conceptual understanding of large-scale learning. Large-scale learning research so far has mostly been empirical. Many useful algorithms lack firm theoretical grounds. This book gathers information that can help address the discrepancy between advances in machine learning mathematics and advances in machine learning algorithms.

The first chapter provides a very detailed description of state-of-the-art support vector machine (SVM) technology. It also reviews the essential concepts discussed in the book. The second chapter compares primal and dual optimization techniques. The following chapters progress from well understood techniques to more and more controversial approaches. This is, of course, a very subjective assessment since most chapters contain both aspects. This progression includes:

- Fast implementation of known algorithms, leveraging special kernels, sparse data, or parallel computers. Some chapters describe experimental setups that should be considered as masterpieces of engineering.
- Approximations that are amenable to theoretical guarantees, such as multipole approximations and fast matrix-vector multiplication.
- Algorithms that perform very well in practice but are difficult to analyze theoretically. This part includes three very effective methods to improve the scalability of kernel machines: greedy selection, nonconvex optimization, and selective sampling.

Finally, we invited the authors of the final chapter to rationalize their mistrust in kernel algorithms for large-scale problems. They consider problems that animals perform effortlessly, such as perception and control. They argue convincingly that local kernel representations are inefficient for these problems. Meanwhile, this argument might not apply to other relevant tasks, such as mining transaction logs. This analysis suggests that this class of problems is different enough to justify a specific scientific approach.

Léon Bottou, Olivier Chapelle, Dennis DeCoste, and Jason Weston,
December 1, 2006.