Computational Linguistics 34(3), pages 449--452, 2008

Semisupervised Learning for Computational Linguistics

Steven Abney
(University of Michigan)

Chapman & Hall / CRC, 2007, (Computer science and data analysis series, edited by David Madigan et al.), 2008, xi+308 pp; hardbound, ISBN 978-1-58488-559-7, $79.95, £44.99

Reviewed by
Vincent Ng
University of Texas at Dallas

Semi-supervised learning is by no means an unfamiliar concept to natural language processing researchers. Labeled data has been used to improve unsupervised parameter estimation procedures such as the EM algorithm and its variants since the beginning of the statistical revolution in NLP (e.g., Pereira and Schabes (1992)). Unlabeled data has also been used to improve supervised learning procedures, the most notable examples being the successful applications of self-training and co-training to word sense disambiguation (Yarowsky 1995) and named entity classification (Collins and Singer 1999).

Despite its increasing importance, semi-supervised learning is not a topic that is typically discussed in introductory machine learning texts (e.g., Mitchell (1997), Alpaydin (2004)) or NLP texts (e.g., Manning and Schütze (1999), Jurafsky and Martin (2000)).[1] Consequently, to learn about semi-supervised learning research, one has to consult the machine-learning literature. This can be a daunting task for NLP researchers who have little background in machine learning. Steven Abney's book *Semisupervised Learning for Computational Linguistics* is targeted precisely at such researchers, aiming to provide them with a "broad and accessible presentation" of topics in semi-supervised learning. According to the preamble, the reader is assumed to have taken only an introductory course in NLP "that include statistical methods — concretely the material contained in Jurafsky and Martin (2000) and Manning and Schütze (1999)." Nonetheless, I agree with the author that any NLP researcher who has a solid background in machine learning is ready to "tackle the primary literature on semisupervised learning, and will probably not find this book particularly useful" (page 11).

As the author promises, the book is self-contained and quite accessible to those who have little background in machine learning. In particular, of the 12 chapters in the book, three are devoted to preparatory material, including: a brief introduction to machine learning, basic unconstrained and constrained optimization techniques (e.g., gradient descent and the method of Lagrange multipliers), and relevant linear-algebra concepts (e.g., eigenvalues, eigenvectors, matrix and vector norms, diagonalization). The remaining chapters focus roughly on six types of semi-supervised learning methods:[2]

1 While Manning and Schütze (1999) and Jurafsky and Martin (2000) do discuss self-training, they do so only in the context of Yarowsky's word sense disambiguation algorithm.

2 The presentation of the methods here does not reflect the order in which they are introduced in the book; rather, it is motivated by the book's section 1.3, which gives an overview of the "leading ideas" of the book.

- *Self-training.* After introducing the self-training algorithm and its variants, the author discusses its applications in NLP and its relationship to other semi-supervised learning algorithms.
- *Agreement-based methods.* The co-training algorithm, along with a theoretical analysis of its conditional independence assumption and its applications in NLP, are presented. Additionally, a random field, which penalizes disagreement among neighboring nodes, is introduced as an alternative way of enforcing agreement.
- *Clustering algorithms.* Basic hard clustering algorithms (e.g., k-means, graph mincuts, hierarchical clustering), EM (as a soft clustering algorithm), and their role in semi-supervised learning are discussed.
- *Boundary-oriented methods.* Two discriminative learning algorithms, boosting and support vector machines, are introduced as a means to facilitate the discussion of their semi-supervised counterparts: co-boosting and transductive SVMs.
- *Label propagation in graphs.* In graph-based approaches to semi-supervised learning, the labels of the labeled nodes are propagated to the unlabeled nodes, with the goal of maximizing the agreement of the labels of proximate nodes. The author shows that this goal is equivalent to finding a *harmonic function* given the labeled nodes, and presents several algorithms, including the method of relaxation, for computing this function.
- *Spectral methods.* Spectral methods for semi-supervised learning can be viewed as interpolation across a partially labeled graph as described above using a "standing wave". The author explains the connection between such a wave and the spectrum of a matrix, and establishes the relationship of spectral clustering algorithms to other semi-supervised learners, including graph mincuts, random walks, and label propagation.

The book is rich in theory and algorithms, and although it is targeted at those who lack relevant mathematical background, each theory and algorithm is presented in a rigorous manner.

Another nice feature of the book is that it reveals the connection among seemingly disparate ideas. As mentioned above, it shows that many semi-supervised learners can in fact be viewed as self-training; also, the description of the connection between spectral clustering and other semi-supervised learners is insightful.

In addition, I like the organization of the book. One reason is the presentation of co-training: while the algorithm is presented in chapter 2, its theoretical underpinnings are not described until chapter 9. This enables the reader to see its applications in NLP (in chapter 3) before going through the mathematics, which could be important for researchers who are linguistically but not mathematically oriented. Another reason is that the preparatory material is presented on a need-to-know basis. This allows the discussion of algorithmic ideas as soon as the reader grasps the relevant fundamentals. For instance, function optimization and basic linear algebra concepts are presented in separate chapters, with the latter being deferred to chapter 11, right before the discussion of spectral clustering in chapter 12.

While the discussion of self-training and co-training is complemented by their application to NLP problems, the same is not true for the remaining semi-supervised

learners described in the book. The reader is often left to imagine the potential NLP applications of these learners, and as a consequence is unable to gain an understanding of the state of the art of semi-supervised learning for NLP. In fact, given its scarcity of NLP applications, the book perhaps does not merit its current title. It does have a rich bibliography on semi-supervised learning for NLP, but most of the references are not cited in the text.

The book also lacks a discussion of the practical issues in applying the semi-supervised learners. For instance, the author does not mention that in practice it is not easy to choose k in k-means clustering, merely describing k as a parameter of the clustering algorithm. As another example, when introducing the EM algorithm, the author applies it to a generative model that can be expressed in exponential form, without acknowledging that one of the most difficult issues surrounding the application of EM concerns the design of the right generative model given the data. The lack of NLP applications in the book has unfortunately enabled the author to sidestep these practical issues. On a related note, one can hardly find discussions of the strengths and weaknesses of the semi-supervised learners in the book. This could leave the reader without the ability to choose the best learner for a given NLP problem, and is probably another undesirable consequence of the book's reluctance to discuss NLP applications.

The author's decision to focus exclusively on semi-supervised *classification* problems effectively limits the scope of the book. One consequence of this decision is that the reader may not be able to apply the EM algorithm to train a hidden Markov model for solving *sequence-learning* problems as basic as part-of-speech tagging upon completion of this book. Given the recent surge of interest in structure prediction in the NLP community, and the fact that co-training and semi-supervised EM have been applied to structure-prediction problems such as statistical parsing and part-of-speech tagging, the book's sole focus on classification problem is perhaps one of its weaknesses.

There are a few occasions in which the reader might not get a complete picture of the capability of an algorithm. For instance, the reader might think that spectral methods can be applied only to binary classification tasks, owing to the book's exclusive focus on such tasks in its discussion of spectral clustering. Similarly for the treatment of support vector machines: the reader may get the impression that SVMs cannot be used to learn non-linear functions, as the discussion of kernels is deliberately omitted due to their irrelevance to transductive learning. While it is important to keep the presentation focused, I believe that the author can easily remove potential confusions by explicitly stating the full capability of an algorithm and referring the reader to the relevant papers for details.

Given the rapid growth of semi-supervised learning research in the past decade, there is currently a need for a broad and accessible reference to this area of research. Abney's book serves this purpose in spite of the aforementioned weaknesses, and I believe that it is a useful starting point for any non–machine-learning experts who intend to apply semi-supervised learning techniques to their research. As someone who has some prior knowledge of semi-supervised learning, I still find this book insightful: as mentioned above, it reveals deep connections among apparently disparate ideas. If I were to teach a course on semi-supervised learning for NLP, I would undoubtedly use this book as a primary reference.

References

Alpaydin, Ethem. 2004. *Introduction to Machine Learning*. The MIT Press.

Collins, Michael and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the 1999 Joint Conference on Empirical Methods in Natural*

Language Processing and Very Large Corpora, pages 100–110.
Jurafsky, Daniel and James H. Martin. 2000. *Speech and Language Processing*. Prentice Hall.
Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.
Mitchell, Tom M. 1997. *Machine Learning*. McGraw Hill.
Pereira, Fernando and Yves Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135.
Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

Vincent Ng is an assistant professor in the Department of Computer Science at the University of Texas at Dallas. He is also affiliated with the University's Human Language Technology Research Institute, where he conducts research on statistical natural language processing and teaches undergraduate and graduate courses in machine learning. Ng's address is: Department of Computer Science, University of Texas at Dallas, Richardson, TX 75080-0688; e-mail: vince@hlt.utdallas.edu