An unprecedented wealth of data is being generated by genome sequencing projects and other experimental efforts to determine the structure and function of biological molecules. The demands and opportunities for interpreting these data are expanding rapidly. Bioinformatics is the development and application of computer methods for management, analysis, interpretation, and prediction, as well as for the design of experiments. Machine learning approaches (e.g., neural networks, hidden Markov models, and belief networks) are ideally suited for areas where there is a lot of data but little theory, which is the situation in molecular biology. The goal in machine learning is to extract useful information from a body of data by building good probabilistic models -- and to automate the process as much as possible. In this book Pierre Baldi and Søren Brunak present the key machine learning approaches and apply them to the computational problems encountered in the analysis of biological data. The book is aimed both at biologists and biochemists who need to understand new data-driven algorithms and at those with a primary background in physics, mathematics, statistics, or computer science who need to know more about applications in molecular biology. This new second edition contains expanded coverage of probabilistic graphical models and of the applications of neural networks, as well as a new chapter on microarrays and gene expression. The entire text has been extensively revised.
This book is decidedly a mix: some very good information, combined with some very puzzling omissions and uneven editing. First, the good. The description of stochastic context free grammars is the best I’ve seen. I don’t know any other reference that even hint at how to use generative grammars to evaluate likelihoods. Once they caught my interest, though, the authors did not carry through with training and evaluation algorithms I could really use. I suspect that parts of the information are there, but I’ll have to go back over their opaque notation again to work out just what they’ve given and just what’s been left out. This same pattern - an interesting introduction with missing or mysterious development - recurs throughout the book. The discussion on clustering and phylogeny goes the same way: a number of techniques are mentioned but not developed. The authors mention a tree drawing problem, not just building the tree’s topology, but ordering the branches for the most informative rendering. Again, a critical topic and one that most authors miss - in the end, these authors miss it, too, by mentioning but not filling in the idea. Their discussion of neural nets suffers badly from the authors’ partial presentation. Evaluation of network output for a given input is relatively straightforward, and they present it in some detail. Training the net is the real problem, though, and is given less than a page. Baldi and Brunak give more of the fundamentals than most authors. For example, they explain the maximum entropy principle well enough that I’ll use it in lots of other areas. They give some coverage to topics of intermediate complexity, such as the forward and backward algorithms for HMM training.

Download to continue reading...


Dmca