PM8: ISMB2006 Tutorial

Title: Bayesian networks for bioinformatics: an introduction to inference and learning

Topic Area: Machine Learning and Artificial Intelligence

Main Presenter:

- Dr Chris J Needham
- School of Computing, University of Leeds, Leeds, LS2 9JT, UK.
- chrisn@comp.leeds.ac.uk
- Phone work: +44 113 343 5767, cell: +447711572384
- Fax: +44 113 343 5468
- http://www.comp.leeds.ac.uk/chrisn/
- Statistics to undergraduate Sport and Exercise Science Students.

Second presenter:

- Dr James R Bradford
- Institute of Molecular and Cellular Biology, University of Leeds, Leeds, LS2 9JT, UK.
- j.r.bradford@leeds.ac.uk
- Phone work: +44 113 343 3072, cell: +447855442646
- Fax: +44 113 343 5468
- http://www.bioinformatics.leeds.ac.uk/
- Bioinformatics to post-graduate and undergraduate students

Other contributors to the tutorial presentation:

Dr Andrew Bulpitt and Dr David Westhead.

50-word abstract: Bayesian networks provide a neat compact representation for expressing joint probability distributions and for inference. They are becoming increasingly important in biology for inferring cellular networks and pathways, biological data integration and genetics. This tutorial introduces the Bayesian approach to inference and learning parameters and structures for Bayesian networks.

Tutorial level: Intermediate

Prior knowledge required: An introductory knowledge of statistics.

Suitability of this tutorial for ISMB:

Many applications in computational biology have taken advantage of Bayesian networks or more generally, probabilistic graphical models. These include: protein modelling, systems biology; gene expression analysis, inferring cellular networks and pathway modelling; biological data integration; protein protein interaction and functional annotation; DNA sequence analysis; genetics and phylogeny linkage analysis.

With this growing use of Bayesian networks and Bayesian methodologies, there has been a lack of suitable introductory information about Bayesian networks which is accessible to an audience without significant mathematical and statistical backgrounds.

This tutorial will fit the multi-disciplinary ISMB audience, both students and researchers, since it will be based around biological examples and begin at an introductory level with numerous examples to demonstrate how to use Bayesian networks. In the second half, the focus will be on the higher level concepts, rather than becoming involved in the complicated mathematics behind the learning methods. This will provide the audience with an understanding how and why Bayesian networks works and a time when they are becoming the machine learning method of choice.

Profile of Presenters

Dr Chris Needham and Dr James Bradford are working alongside Dr Andrew Bulpitt and Dr David Westhead as researchers on a bioinformatics project 'Protein function prediction using uncertainty' which is beginning to look at predicting gene ontology classifications of proteins by integrating information from multiple sources. This project between the Institute of Molecular and Cellular Biology and the School of Computing at the University of Leeds has heavily used Bayesian networks and other machine learning tools over the last year. Dr Andrew Bulpitt is a lecturer with five years experience in teaching and producing learning materials in the School of Computing. Dr David Westhead is a lecturer with seven years experience teaching bioinformatics in the Institute of Molecular and Cellular Biology, and author of the book 'Instant Notes in Bioinformatics'. Their input to the presentation will ensure it is suitable and accessible to the ISMB audience.

Over the past year, we have begun to develop an 'introduction to Bayesian networks talk' which has been well received in Computing, Bioinformatics, and at the Leeds Annual Statistics Workshop 2005. We have written a primer on 'Inference in Bayesian networks' to appear in Nature Biotechnology (January 2006).

Profile of Presenter 1 (Dr Chris Needham)

- I am interested in machine learning and artificial intelligence. In a previous post-doc position, on a 'Cognitive Vision' project, I used numerous continuous machine learning and symbolic AI methods to learn autonomously from noisy audio-visual data. Part of this work, on 'Learning protocols from perceptual observation' won the British Computer Society Machine Intelligence Award in Dec 2004, as the result of a live demonstration of the computer vision and artificially intelligent system.
- I have experience of teaching in higher education. In 2002/03 I was employed as Teaching Fellow in Sport and Exercise Science at the University of Leeds and taught Statistics to over a hundred first year undergraduates. Since then, I have given a few lectures in the School of Computing, on courses in image and signal processing, computer vision, and on the masters course perceptual sensory systems.

Profile of Presenter 2 (Dr James Bradford)

- I am currently a research associate in the Leeds Bioinformatics research group. My main research interest is prediction of protein function and protein-protein binding sites using machine learning. I have three years experience in machine learning methods such as support vector machines and Bayesian networks.
- My background is in Biochemistry and Molecular Biology thus I will provide expert input into the biological examples we will use throughout the tutorial. My teaching experience includes three years of demonstrating experience at both post graduate and undergraduate level in Bioinformatics.

Relevant publications

Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2006) Inference in Bayesian networks. *Nature Biotechnology* (in press)

Bradford JR, Needham CJ, Bulpitt AJ, Westhead DR (2006) Insights into protein-protein interfaces using a Bayesian network prediction method (submitted)

Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR (2006) Using Bayesian networks to model the effect of missense mutations on protein function (submitted)

Bradford JR, Dobson P, Doig AJ, Siepen JA, Westhead DR (2006) Support vector machines in bioinformatics (submitted).

Bradford, J.R. and Westhead, D.R. (2005) Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics* **21**: 1487-1494

Tutorial Outline:

SECTION 1 (25 mins)

- Intro – learning from data - principles
- Machine learning (other approaches – recap on decision trees/SVMs)
- Probability: classical vs Bayesian. Methods for probability assignment
- Probability theory. Sum, product and Bayes' rule.
- Bayesian inference. Example.
- Summary highlighting benefits of Bayesian statistics.

SECTION 2 (50 mins)

- Bayesian networks. What are they?
  [A cell signalling pathway example will be used throughout this section]
- Conditional independence – incorporating prior knowledge
- Conditional probability distributions
- Joint probability distributions
- Compact representation – parameter reduction
- Inference in Bayesian networks
- Calculating posterior probabilities
- Summary highlighting benefits of Bayesian network representation and inference

SECTION 3 (30 mins)

- Learning parameters from data
- Parameter priors
- Continuous variables as well as discrete (Gaussian parameters)
- Point estimates: maximum likelihood (ML), maximum a posteriori (MAP) estimates
- Bayesian learning – model averaging
- Summary highlighting benefits of learning model parameters from data

(break 30mins)

SECTION 4 (25 mins)

- Structure learning
- Structure priors
- Markov Chain Monte Carlo methods
- Causality
- Summary highlighting benefits of structure learning and learning causal relationships

SECTION 5 (50 mins)

- Examples

SECTION 6 (30 mins)

- Discussion