Complex Prior Knowledge in Learning

Neil D. Lawrence
Marc Dymetman

27th January 2008

Outline

Outline

Machine Learning History

- Philosophy: Machine learning, Statistics and AI
  - Statistics assumes statistician
  - AI and Machine Learning replace the human.
- Practice:
  - Humans always in the loop.
  - AI and expert systems
  - Machine Learning and Data Driven Systems

The Future

If we are to replace the human in the loop

- Need to understand knowledge transfer.
- Data set shift.
- Semi supervised learning
- Multitask learning and transfer learning.

Less Ambitiously (but still difficult): Incorporate Prior Knowledge with Data

- Mix between expert system approach and neural network "black box".
- Need to incorporate our knowledge of the system with the data.
- How best to do this? Application domain dependent? Need to transfer knowledge between applications.

All these subjects fall within the scope of the thematic programme.

Outline

The Methodologies

- How are we encoding prior knowledge?
  - graphs
  - probability distributions
  - similarity measures (kernels)
  - differential equations.
- How do these technologies inter-relate?

Outline

Common Themes

- Mechanistic models: built from knowledge of physics or human expert knowledge.
- Data driven models: built from data with limited assumptions (such as smoothness)

Common theme in applications is bridging the gap between these two.

- For illustration we will consider two case studies.
  - Computational and Systems Biology
  - Language Modelling
- But there are many more!!

Computational and Systems Biology

- Many different systems: some well studied (cell cycle, signalling pathways), others not.
- Large volume of data large, but sparse relative to complexity of the networks.
  - Some quantities easy to measure (mRNA expression) in high throughput. But still temporal/spatial resoultion problems.
  - Other quantities (protein concentration) difficult to measure in high throughput.
- Quality and quantity of data improving constantly, but biologists want answers now.
- For some systems there is (noisy) data about which constituents react with what. ChIP on chip.
- The underlying mechanism of interactions is (somewhat) known. Chemical reaction kinetics — but parameters etc. are unknown.

Rich playground for interaction between data and models of the interactions!

What's Out There?

- Broadly speaking two large areas:
  - Computational biologists have focused on constructing models (Kalman filters, "Bayesian" networks, linear models, PCA, AR models) from data. *Data driven approach.*
  - Systems biologists have focused on the appropriate differential equation structures for modelling given systems. *Mechanistic approach.*
- Challenge is to combine these approaches. Bridge the gap between Mechanistic and Data driven approach.
- Consider the mechanistic approach as prior knowledge for the data driven approach.
  - Alternative perspective is to use data to "fit" mechanistic models.

Text and Language Processing

- A lot of training data.
- Data is inherently complex in structure
  - Risk of overfit even with very large training data.
  - Requires constraining the space of fitted models.
- Lots of existing expert knowledge (*e.g.* linguistic constraints).
  - Incorporation of this knowledge not always validated by data.

Rich playground to study interplay between prior knowledge and data!

Historic Perspective

Rule Based Systems

- Before c. 1990 mainstream systems (for parsing translation *etc.*) *rules* defined by experts.
  - Could be seen as an *extreme case* of "prior knowledge". No training data to "tune" the encoded knowledge.
  - Standard linguistics is about the discovery of such rules.

Data Driven Systems

- Opposite extreme: data only, no "prior knowledge".
  - *e.g.* basic n-gram language models — collect statistics over 3-grams and use these for computing probability of a test sentence.
  - Suprisingly effective in many cases (widely used in speech recognition).

Implicit Prior Knowledge in Data Driven Systems

Prior knowledge is encoded, but implicitly

- Even in "data driven" systems a lot of prior knowledge is *implicit* in the models/representations.
  - *n*-gram models incorporate "smoothing". Algorithms attempt to exhibit good generalisation properties.
  - In statistical machine translation pioneer IBM systems were generative models that perform operations (*e.g.* fertility, distortion) that have implicit linguistic motivations. *Once* the generative model is defined tuning of parameters is done by looking at the data.
  - In current discriminative models for NLP, models learn to discriminate good outputs from bad on the basis of training data alone, but often the underlying feature functions are complex (*e.g.* in phrase-based SMT), and designed carefully to address the problem at hand.

Towards Explicit Prior Knowledge

- Minimum Discription Length
  - *e.g.* in Grammar inference [Grünwald, A minimum description length approach to grammar inference]
- Bayesian Parametric
  - *e.g.* in Topic Modelling [Blei et al., 2003, Latent Dirichlet Allocation]
  - *e.g.* in Language Model Smoothing [**?**, A Hierarchical Dircilet Language Model]
- Currently hot: Bayesian non-parameteric
  - Dirichlet processes, Chinese restaurant processes, ...
  - Allow automatic selection of model complexity
    - ★ *e.g.* in topic modelling: automatic determination of the number of underlying topics.
    - ★ *e.g.* in parsing: automatic determination of number of non-terminals in PCFG.

Surge of Publications

- Topic Modelling
  - finite Bayesian model; variational [Blei et al., 2003]
  - HDP-based model; sampling [Teh et al., 2006]
- Language Modelling
  - Pitman-Yor $\rightarrow$power-law; sampling

The Challenge of Incorporating Expert Knowledge

- By and large the approaches described concentrate on "generic" models of prior knowledge, not on "specific" expert linguistic knowledge.
- One possible approach to that might be in the line of statistical relational learning
  - *e.g.* Markov logic networks [Richardson and Domingos, 2006]
    - ⋆ Main idea: ask experts to formulate their belifs through first order logical formulas (a form of prior knowledge):
      e.g. "if student is author of publication, professor is co-author ..."
    - ⋆ These formulas become binary features on possible worlds. Can be false in a given world and are associated with weights that are learnt on the basis of data.
    - ⋆ Good results on such tasks as link prediction. Few applications to "core" NLP so far.

Outline

Computational and Systems Biology

Learning in Computational and Systems Biology
Glasgow, March 27th - 28th 2008, *Co-located with MASAMB 2008*
Co-organized by Mark Girolami and Simon Rogers

- Combining data with models in systems biology
  - how to estimate differential equation parameters
  - how to estimate difficult to measure chemical species
- Validation of model structure: which hypothesis is correct?
  - Hypothesis in the form of non-linear differential equations.
  - Sampling approaches to efficient computation of Bayes factors.

Stochastic Processes and Differential Equations

Title TBA

May 2008, Cumberland Lodge, U.K.

Co-organizers: Cedric Archambeau, John Shawe Taylor

- Focus on combining differential equations with data.
  - mainly stochastic (applications in climate, weather, computational biology) but also ordinary differential equations.
- Bring together Machine Learning technologies with statistics, physics, control *etc.*
- Focus on both applications and methodologies.

Bayesian Workshop

Bayesian Research Kitchen
Early June 2008, Lake District, U.K.
Co-organizers: Neil Lawrence, Joaquin Quinoñero Candela

- Small gathering
- Bayesian "reality check".
  - Focus on future directions for Bayesian research.
- May lead to larger Participation Workshop in Second Half of Programme (*c.f.* GPRT/GPIP).

Language Workshop

Prior Knowledge for Text and Language Processing
9th July 2008, Helsinki. *Proposal submitted to ICML/UAI/COLT*
Co-organizers: Guilliaume Bouchard, Hal Daumé III, Marc Dymetman and Yee Whye Teh

- Draft call for papers:
  - Prior knowledge for language modelling, parsing, translation.
  - Topic modelling for document analysis and retrieval.
  - Parametric and non-parametric Bayesian models in NLP.
  - Graphical models embodying structural knowledge of texts.
  - Complex features/kernels that incorporate linguistic knowledge; kernels built from generative models.
  - Limitations of purely data-driven learning techniques for text and language appllications.
  - Typology of different forms of prior knowledge for NLP (knowledge embodied in generative Bayesian models, in MDL models, in ILP/logical models, in linguistic features, in representational frameworks, in grammatical rules ...).
  - Formal principles for combining rule-based and data based approaches

Other Workshops

- Mining and Learning on Graphs
- Machine Learning for Systems Biology
- Computational Biology Summer School (Sami Kaski)
- Follow up events from September 2008 - March 2009.
- **Over to you!!**

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

P. Grünwald. A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, and G. Scheler, editors, *Symbolic, Connectionist and Statistical Approaches to Learning for Natural Language Processing*, volume 1040 of *Lecture Notes in Artificial Intelligence*, pages 203–216. Springer-Verlag, Berlin, Germany.

M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, pages 107–136, 2006.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.