School of Electronics and Computer Science
University of Southampton

MSc 2015/16: Projects in Machine Learning

Mahesan Niranjan
January 2016

1. Integrative Analysis of Transcriptome and Proteome

Transcriptomes and proteomes are high throughput measurements of messenger RNA (mRNA) and protein abundances in cells respectively. Various machine learning algorithms are routinely applied to such data to gain a molecular level understanding of biology. Of particular interest is the combined (integrative) analysis of measurements taken at multiple levels across these "*omic*" scales. Recently, Rogers *et al.* [1] developed a clustering algorithm for this purpose. This is done in a probabilistic setting. In this project, I would like to formulate the coupled clustering problem as a structured matrix approximation problem. We will build on the formulation of K-means clustering as a sparse matrix factorization problem introduced by [2] and take it to a structured matrix approximation formulation similar to Markovsky *et al.* [3]. The challenge in the project is to find a neat formulation of the problem, implment it (making good use of the *cvx* [4] toolbox), make a thorough comparison of your work with that of Rogers *et al.* [1]. We will use publicly available datasets, and if permission can be organized in the time, use a dataset coming from a recent study of a cohort of severe Asthma patients undertaken in the `UBIOPRED` project [5] (`http://www.europeanlung.org/en/projects-and-research/projects/u-biopred/home`), of which Southampton University Medical School was a partner.

2. Making Sense of Outliers in Data

Most machine learning problems are formulated as supervised (classification or regression) or unsupervised learning problems. A slightly different problem relates to the detection of novelty in datasets appropriate for situations in which we are interested in abnormal behaviour. Some interesting early work in the subject in the machine learning literature is due to Tarassenko and colleagues(*e.g.* see [6] and [7]). The same issue is dealt with under the subject *outlier detection* in the statistics literature. In recent work in my group ([8], [9]) we studied the use of outlier modelling in the context of integrative analysis of genomic data (*i.e.* the mapping between transcriptome (mRNA measurements) and proteome (relative concentrations of the different proteins in a cell). In particular, in Gunawardana *et al.* [9] we developed a new regression algorithm that is robust to outliers in the data. The aim of this project is to explore this work further, study the convergence and scaling properties of this algorithm. We will attempt to apply such settings to much larger problems such as intrusion detection for computer security [10], [11], [12].

3. Gap Gene Expression in Drosophila Development

Developmental biology is a fascinating subject in which mathematical and computational modelling plays a significant role, starting from the seminal work of Turing ([13]). One of the best studied examples of morphogenesis is in early Drosophila embryonic development. mRNA molecules of the bicoid morphogen, deposited by the mother at the anterior pole of the embryo is translated into protein and diffuses in the medium establishing a concentration gradient which defines the anterior-posterior boundary in about two hours since egg-laying. Recently, we used machine learning methods to understand how the stability of the maternally deposited

mRNA may be regulated [14] [15]. Subsequent to the establishment of bicoid gradient, a circuit known as the gap-gene circuit of around six strongly coupled transcription factors (proteins that regulate gene expression) are activated and contribute to segmental patterning on the embryo. This circuit has been studies in computational and experimental work [16]. While the proteins involved in the regulation is well understood, complete knowledge of the the parameter values of the regulatory interactions and their dynamics still has open issues. Perkins *et al.* [17] and Fomekong-Nanfack *et al.* [18] are examples of detailed computational studies on this circuit. In this project, we will attempt a novel machine learning method, *Bayesian Optimization*, for parameter estimation of the Drosophila gap gene regulation circuit [19], [20]. Bayesian optimization is a recent development in the machine learning literature, which works by exploring a search space in a systematic manner by guided by progressive modelling of uncertainty. A systematic comparison of this method against a genetic algorithm based method ([18]) is a useful and novel undertaking, with scope to learn much about modern machine learning in the context of an interesting biological domain.

4. Reinforcement Learning for Packet Routing

Reinforcement learning [21], [22]) is an exciting subject within machine learning, aimed at problems of planning, as opposed to static and dynamic pattern recognition we are familiar with in classification, regression and time series analysis. It is based on learning by repeated trials to maximize a future reward. A particularly illustrative demonstration of such learning was the TD-Gammon programme [23] that played backgammon to standards of the best human player, improving by self-play. More recently, the subject has attracted much interest from similar demonstrations on the Atari Learning Environment [24], the acquisition of DeepMind by Google etc.[1] Some years ago, Boyal and Littman [25] demonstrated that reinforcement learning when combined with deep neural networks for function approximation, can be used to learn a routing table for computer networks and that under conditions of high loading, outperforms shortest path routing by diverting packets away from bottlenecks in the network. In this project, I want to re-visit this topic and ask how this relates to the topological properties of networks. Router level networks are not formed by random connections. Instead their growth may happen in a preferential attachment style [26]. Alternate topological features have also been considered to heuristically optimize a network to match the Internet's router level topology [27]. In this project, we will implement Boyan *et al.*'s Q-routing framework and carry out a systematic empirical study its convergence and scaling behaviours. We will ask what topological features of the underlying network might aid the ability to learn paths away from those with bottlenecks.

5. Deep Neural Networks for Speech Enhancement

Speech being a natural medium of human communication, over the last four to five decades, there has been much interest in speech signal processing. Various speech-based communication interfaces are now available in the marketplace. However, the performance of such systems rapidly degrades in noisy environments. Though the speech of a single speaker in noise-free environments can be recognized by machines to very high accuracies, human performance in adverse conditions (*e.g.* the cocktail party effect) far exceeds what we can nowadays do on machines. One aspect of this is that over several decades most automatic speech processing systems relied on statistical spectral analysis of short durations of speech signals. Various parametric spectral features (*e.g.* Mel Frequency Cepstral Coefficients - MFCC) have been explored. A rather different representation of acoustic signals is to more accurately mimic the auditory processing in the mammalian ear. The auditory image model ([29], and references therein) extracts a time-frequency rep-

[1] `http://deepmind.com/`

resentation every 10 ms, and is regarded as a much richer representation of the underlying signal, capturing perceptually important aspects of it. Wang *et al.* [28] (and references therein) demonstrate a framework for speech enhancement which uses neural networks. They show that a network could be trained to predict which parts of a spectrum contain noise and which contain target speech (or predict a signal to noise ratio), using statistical signal analysis techniques. The aim of this project is to design a speech enhancement system that couples the above two ideas: *i.e.* using auditory image modelling as representation from which a neural network is trained to predict signal to noise ratios in regions of the spectra. The challenge in the project will come from the fact that the auditory representation is in the form of time-frequency matrices every 10 ms. Hence we will use techniques from computer vision to extract features from these images. The project will be in collaboration with the Institute for Sound and Vibration Research (ISVR `http://www.southampton.ac.uk/engineering/research/centres/isvr.page`).

6. Sequential Monte Carlo Algorithms

Sequential Monte Carlo methods [30], also known as particle filters, are a powerful subset of algorithms for Bayesian inference. While they are particularly useful for sequential problems (*e.g.* time series analysis, tracking in computer vision etc.), they have also been successfully used to process batch data [31] with computational and performance advantages. More recently, such methods were successfully applied to parameter estimation in models of systems biology [32]. In this project, we will start from this class of algorithms (reviewed in [33]), and examine their role in a class of systems biology models. In particular, we will study the scenareos parameter sensitivities, in which some parameters in models are *sloppy* and others considered *stiff* [34]. The aim of this project is a systematic study of SMC-based parameter estimation algorithms on a large number of carefully curated models of systems biology (see `https://www.ebi.ac.uk/biomodels-main/`).

References

[1] S. Rogers, M. Girolami, W. Kolch, K. Waters, T. Liu, B. Thrall, and H. Wiley, "Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models," *Bioinformatics*, vol. 24, no. 24, pp. 2894–2900, 2008.

[2] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," Tech. Rep. GT-CSE-08-01, Georgia Tech., 2008.

[3] I. Markovsky and M. Niranjan, "Approximate low-rank factorization with structured factors," *Computational Statistics and Data Analysis*, vol. 54, no. 12, pp. 3411 – 3420, 2010.

[4] M. Grant and S. Boyd, "Users guide for `cvx` version 1.21." `http://cvxr.com/cvx`, 2010.

[5] D. E. Shaw, A. R. Sousa, S. J. Fowler, L. J. Fleming, G. Roberts, J. Corfield, I. Pandis, A. T. Bansal, E. H. Bel, C. Auffray, C. H. Compton, H. Bisgaard, E. Bucchioni, M. Caruso, P. Chanez, B. Dahlén, S.-E. Dahlen, K. Dyson, U. Frey, T. Geiser, M. Gerhardsson de Verdier, D. Gibeon, Y.-k. Guo, S. Hashimoto, G. Hedlin, E. Jeyasingham, P.-P. W. Hekking, T. Higenbottam, I. Horváth, A. J. Knox, N. Krug, V. J. Erpenbeck, L. X. Larsson, N. Lazarinis, J. G. Matthews, R. Middelveld, P. Montuschi, J. Musial, D. Myles, L. Pahus, T. Sandström, W. Seibold, F. Singer, K. Strandberg, J. Vestbo, N. Vissing, C. von Garnier, I. M. Adcock, S. Wagers, A. Rowe, P. Howarth, A. H. Wagener, R. Djukanovic, P. J. Sterk, and K. F. Chung, "Clinical and inflammatory characteristics of the european u-biopred adult severe asthma cohort," *European Respiratory Journal*, vol. 46, no. 5, pp. 1308–1321, 2015.

[6] S. Roberts and L. Tarassenko, "A probabilistic resource allocating network for novelty detection," *Neural Computation*, vol. 6, no. 2, pp. 270 – 284, 1994.

[7] P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis, "Support vector novelty detection applied to jet engine vibration spectra," in *NIPS*, pp. 946–952, 2000.

[8] Y. Gunawardana and M. Niranjan, "Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes," *Bioinformatics*, vol. 29, no. 23, pp. 3060–3066, 2013.

[9] Y. Gunawardana, S. Fujiwara, A. Takeda, J. Woo, C. Woelk, and M. Niranjan, "Outlier detection at the transcriptome-proteome interface," *Bioinformatics*, 2015.

[10] P. Laskov, P. Dussel, C. Schafer, and K. Rieck, "Learning intrusion detection: Supervised or unsupervised?," in *Image Analysis and Processing ICIAP 2005* (F. Roli and S. Vitulano, eds.), vol. 3617 of *Lecture Notes in Computer Science*, pp. 50–57, Springer Berlin Heidelberg, 2005.

[11] I. Perona, I. Gurrutxaga, O. Arbelaitz, J. I. Martín, J. Muguerza, and J. M. Pérez, "Service-independent payload analysis to improve intrusion detection in network traffic," in *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pp. 171–178, Australian Computer Society, Inc., 2008.

[12] B. Farran, C. Saunders, and M. Niranjan, "Machine learning for intrusion detection: Modeling the distribution shift," in *IEEE Workshop on Machine Learning for Signal Processing*, pp. 232 – 237, 2010.

[13] A. M. Turing, "The chemical basis of morphogenesis," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 237, no. 641, pp. 37–72, 1952.

[14] W. Liu and M. Niranjan, "Gaussian process modelling for *bicoid* mRNA regulation in spatio-temporal bicoid profile," *Bioinformatics*, vol. 28, no. 3, pp. 366–372, 2012.

[15] W. Liu and M. Niranjan, "The role of regulated mRNA stability in establishing bicoid morphogen gradient in drosophila embryonic development," *PLoS ONE*, vol. 6, no. 9, p. e24896, 2011.

[16] J. Jaeger, M. Blagov, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, S. Surkova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz, "Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster," *Genetics*, vol. 167, no. 4, pp. 1721–1737, 2004.

[17] T. J. Perkins, J. Jaeger, J. Reinitz, and L. Glass, "Reverse engineering the gap gene network of ¡named-content xmlns:xlink="http://www.w3.org/1999/xlink" content-type="genus-species" xlink:type="simple"¿drosophila melanogaster¡/named-content¿," *PLoS Comput Biol*, vol. 2, p. e51, 05 2006.

[18] Y. Fomekong-Nanfack, M. Postma, and J. A. Kaandorp, "Inferring drosophila gap gene regulatory network: a parameter sensitivity and perturbation analysis," *BMC Systems Biology*, vol. 3, no. 1, pp. 1–23, 2009.

[19] M. D. Hoffman, E. Brochu, and N. de Freitas, "Portfolio allocation for bayesian optimization," in *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pp. 327–336, 2011.

[20] D. Jones, M. Schonlau, and W. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[21] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[22] G. Rummery and M. Niranjan, "On-line Q-learning using connectionist systems," Tech. Rep. `CUED/F-INFENG-TR 166`, Cambridge University Engineering Department, 1994.

[23] G. Tesauro, "Temporal difference learning and td-gammon," *Commun. ACM*, vol. 38, pp. 58–68, Mar. 1995.

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, p. 529533, 2015.

[25] J. A. Boyan and M. L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," in *Advances in Neural Information Processing Systems 6*, pp. 671–678, Morgan Kaufmann, 1994.

[26] A. L. Barabasi, J. De, P. Lettres, L. Al, N. Cimento, H. Jeong, H. Jeong, Z. Neda, Z. Nda, and A. L. Barabsi, "Measuring preferential attachment in evolving networks," *Europhysics Letters*, vol. 61, no. 61, pp. 567–572, 2003.

[27] L. Li, D. Alderson, W. Willinger, and J. Doyle, "A first-principles approach to understanding the internet's router-level topology," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 4, pp. 3–14, 2004.

[28] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, pp. 1849–1858, Dec. 2014.

[29] S. Bleeck, T. Ives, and R. D. Patterson, "Aim-mat: the auditory image model in matlab," *Acta Acustica*, vol. 90, no. 4, pp. 781–787, 2004.

[30] A. Doucet, N. de Freitas, , and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[31] J. De Freitas, M. Niranjan, A. Gee, and A. Doucet, "Sequential monte carlo methods for optimisation of neural network models," *Neural Computation*, vol. 12, no. 4, pp. 955–993, 2000.

[32] X. Liu and M. Niranjan, "State and parameter estimation of the heat shock response system using Kalman and particle filters," *Bioinformatics*, vol. 28, no. 11, pp. 1501–1507, 2012.

[33] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, pp. 174–188, Feb 2002.

[34] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput Biol*, vol. 3, p. e189, 10 2007.