CSE 5095 – Bayesian Machine Learning

Derek Aguiar

The probabilistic (or Bayesian) machine learning paradigm provides a unifying methodology for reasoning about uncertainty in modeling complex data. In this class, we will cover the three fundamental components of this paradigm: probabilistic modeling, inference algorithms, and model checking. We consider graphical models as the unifying framework for probabilistic modelling and study several examples from the current literature. Some graphical models, like HMMs or simple mixture models, yield efficient and exact posterior inference, but this is not always the case. In these more difficult settings, we will consider sampling based methods and approximate inference algorithms, such as variational Bayes. Finally, we will consider methods for determining model fitness and advanced topics like Bayesian nonparametrics.

We will reinforce class lectures by developing a research-quality project (ideally connected to your research) on real data that implements this probabilistic framework from problem specification to model checking. To realize this, you should be comfortable with (or willing to learn):

- basic probability, statistics, and calculus
- writing software to analyze large data and some statistical supportive programming language like Python
- scientific writing

1 Logistics

The course instructor is Derek Aguiar (derek.aguiar@uconn.edu); office is ITE 267.

Class times are Tuesdays and Thursdays at 3:30pm - 4:45pm in CAST 201.

Office hours are Wednesdays 9:30am-11:30am, Thursdays 4:45pm-6pm (after class), or by appointment in ITE 267.

The course website is hosted on HuskyCT which is used for course material distribution, announcements, and out-of-class discussions.

Each student will be allowed 3 late days (includes Saturday and Sunday) to be used for homeworks without penalty (*does not include final project presentation/handin or class presentations*).

Please fill out the course questionnaire at https://goo.gl/forms/CrqVdyTz63QA81PK2

1.1 Course requirements

- Required reading and class participation – 25%
- A short demonstration on something you have learned related to the class that you believe would be interesting to share. Examples include: using a machine learning or visualization package, implementing models in a probabilistic programming language (e.g. STAN, Edward, TensorFlow), data preparation with shell scripts, etc. 5%
- 2-3 homework/project progress reports – 20%

- An original research project **or** an extensive review on a subject relevant to the course. This includes a final presentation at the end of the semester. – 50%

We strongly encourage LaTeX for typesetting work.

2 Schedule

A tentative schedule of topics and sample readings (not necessarily required) are included below. A current schedule with readings will be available on the website.

1. Foundations of graphical models
   - Michael I Jordan. *An introduction to probabilistic graphical models.* 2003. URL: https://people.eecs.berkeley.edu/~jordan/prelims/
   - Kevin Murphy. "An introduction to graphical models". In: *Rap. tech* (2001), pp. 1–19
   - Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009
2. Inference
   - Radford M Neal. "Probabilistic inference using Markov chain Monte Carlo methods". In: (1993)
   - Philip Resnik and Eric Hardisty. *Gibbs sampling for the uninitiated.* Tech. rep. 2010
   - David M Blei, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877
3. Models
   - David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022
   - Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959
4. Posterior Checking
   - Andrew Gelman, Xiao-Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies". In: *Statistica sinica* (1996), pp. 733–760
   - Andrew Gelman. "A Bayesian formulation of exploratory data analysis and goodness-of-fit testing". In: *International Statistical Review* 71.2 (2003), pp. 369–382
5. Implementation
   - Dustin Tran et al. "Edward: A library for probabilistic modeling, inference, and criticism". In: *arXiv preprint arXiv:1610.09787* (2016)
   - Bob Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017)
6. Bayesian nonparametrics
   - Samuel J Gershman and David M Blei. "A tutorial on Bayesian nonparametric models". In: *Journal of Mathematical Psychology* 56.1 (2012), pp. 1–12
   - Yee W Teh et al. "Sharing clusters among related groups: Hierarchical Dirichlet processes". In: *Advances in neural information processing systems.* 2005, pp. 1385–1392

3 Resources

3.1 Textbook

There is no required textbook for the course, although, there is assigned required reading, optional reading, and implicit reading required for your final projects. Readings will come from Michael I Jordan. *An introduction to probabilistic graphical models*. 2003. URL: https://people.eecs.berkeley.edu/~jordan/prelims/ and research articles; they will be linked to from the website and hard copies can be made available. Another potentially useful (and free) resources is David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012. URL: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf;

The following are also good references for various topics covered in this course.

- Christopher M Bishop. *Pattern recognition and machine learning*. Tech. rep. 2006
- Kevin P Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012
- Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995
- Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009

3.2 Software

The ability to reproduce results and extend or reuse existing code are essential to research. We highly recommend using Github or some other version control system to document your progress on your final project.

3.3 Writing help

Learning how to write a scientific article is an arduous process. At UConn, the writing center can be a great resource. Other good resources include:

- Paul J Silvia. *How to write a lot: A practical guide to productive academic writing*. American Psychological Association, 2007
- Wendy Laura Belcher. *Writing your journal article in twelve weeks: A guide to academic publishing success*. Sage, 2009

References

Barber, David. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012. URL: http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf.

Belcher, Wendy Laura. *Writing your journal article in twelve weeks: A guide to academic publishing success*. Sage, 2009.

Bishop, Christopher M. *Pattern recognition and machine learning*. Tech. rep. 2006.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe. "Variational inference: A review for statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

Blei, David M, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

Carpenter, Bob et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76.1 (2017).

Gelman, Andrew. "A Bayesian formulation of exploratory data analysis and goodness-of-fit testing". In: *International Statistical Review* 71.2 (2003), pp. 369–382.

Gelman, Andrew, Xiao-Li Meng, and Hal Stern. "Posterior predictive assessment of model fitness via realized discrepancies". In: *Statistica sinica* (1996), pp. 733–760.

Gelman, Andrew et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Gershman, Samuel J and David M Blei. "A tutorial on Bayesian nonparametric models". In: *Journal of Mathematical Psychology* 56.1 (2012), pp. 1–12.

Jordan, Michael I. *An introduction to probabilistic graphical models*. 2003. URL: https://people.eecs.berkeley.edu/~jordan/prelims/.

Koller, Daphne and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Murphy, Kevin. "An introduction to graphical models". In: *Rap. tech* (2001), pp. 1–19.

Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Neal, Radford M. "Probabilistic inference using Markov chain Monte Carlo methods". In: (1993).

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data". In: *Genetics* 155.2 (2000), pp. 945–959.

Resnik, Philip and Eric Hardisty. *Gibbs sampling for the uninitiated*. Tech. rep. 2010.

Silvia, Paul J. *How to write a lot: A practical guide to productive academic writing*. American Psychological Association, 2007.

Teh, Yee W et al. "Sharing clusters among related groups: Hierarchical Dirichlet processes". In: *Advances in neural information processing systems*. 2005, pp. 1385–1392.

Tran, Dustin et al. "Edward: A library for probabilistic modeling, inference, and criticism". In: *arXiv preprint arXiv:1610.09787* (2016).