STAT 37790 / CMSC 35425: TOPICS IN STATISTICAL MACHINE LEARNING

Syllabus, Spring 2016

Topics in Statistical Machine Learning is a second course in machine learning, with students assumed to have previous exposure to machine learning and intermediate statistical theory. Topics cover selected areas of current interest and activity in the research community. The spring 2016 course will focus on word embedding algorithms, network models, and probabilistic models for discrete data.

Schedule

CLASSES	Mondays and Wednesdays 1:00-2:20 pm	Jones 226

Contact Information

Instructor:

John Lafferty	Jones 121	lafferty@galton.uchicago.edu
Office hours:	Thursday 5:00 pm	

Prerequisites

Class members are assumed to be conversant in standard statistical machine learning ideas and the underlying statistical principles, at the level of Bishop's *Pattern Recognition and Machine Learning*, or Hastie et al.'s *Elements of Statistical Learning*. Background reading will be provided as needed.

Course Structure

The course will be a research seminar, emphasizing informal presentations and discussion of recent research. In a typical week, we will present background material and context for a particular topic. One of the class members will present a research paper on that topic, and lead a discussion. Topics will be chosen from among the following (subject to change):

1. Language models
2. Word embeddings

3. Probabilistic models of discrete data
4. Matrix and tensor factorizations
5. Models of graphs and networks

Requirements

Students taking the course for a letter grade will be required to complete the following work:

1. complete three problem sets;
2. present one paper in class;
3. complete a project on a chosen topic.

Course Organization

Each class will involve a discussion of a research paper. Students will sign up for presentations on specific topic/papers. A project on a research-oriented topic, selected by the student, must also be completed. Projects can be done individually, or in groups of two students. Each student/group will give a presentation on their research project. A project proposal is due at the end of the fifth week. Start thinking about project ideas right away!

Projects

Each student/group should prepare a project report that analyzes a problem involving discrete data, preferably (but not necessarily) using some of the ideas and methods discussed in the class. The report should include the following components:

1. ***Description of the data***. Describe the data, including where it came from, how it was collected and the meaning of the variables. Clearly state the objective of analyzing the data, and any preprocessing that was carried out on the data. Give plots that summarize and visualize the data.
2. ***Probabilistic model***. The main part of the project should describe a model of the data. Typically this will be a generative probabilistic model. What assumptions are made by the model? How is the model fit to the data? How is uncertainty in the model expressed? Include descriptive plots.

3. ***Discussion***. Discuss your results, comparing different modeling approaches and natural baselines. What conclusions can you draw? Is your model effective? How can it be improved or extended? Describe additional work that could be carried out.

Course Calendar

The course calendar and other organizational material will be maintained on the course Piazza site, https://piazza.com/uchicago/spring2016/stat37790/home

Week	Date	Topic	Exams and Assignments
1	March 28 March 30	language models	
2	April 4 April 6	word embeddings	assn1 out
3	April 11 April 13	word embeddings	
4	April 18 April 20	models for discrete data	assn 1 due; assn 2 out
5	April 25 April 27	models for discrete data	project proposals due
6	May 2 May 4	matrices and tensors	assn 2 due, assn 3 out
7	May 9 May 11	matrices and tensors	
8	May 16 May 18	networks and graphs	assn 3 due
9	May 23 May 25	networks and graphs	
10	May 30 June 1	project reports	projects due

Readings

A preliminary list of readings for each of the four topics is given below. (The starting point for this list is a similar course offered at Columbia last semester.) This list will be augmented and modified as the quarter goes along.

1. Word embeddings: Arora et al. (2015, 2016); Brown et al. (1992); Bengio et al. (2003); Hashimoto et al. (2015); Le and Mikolov (2014); Levy and Goldberg (2014); Mikolov et al. (2013); Nalisnick and Ravi (2015); Pennington et al. (2014); Vilnis and McCallum (2015)

2. Probabilistic models of discrete data: Adams et al. (2009); Bhattacharya and Dunson (2012); Falish et al. (2003); Gopalan et al. (2015); Hoffman and Blei (2015); Hoffman et al. (2014); Mimno et al. (2015); Paisley et al. (2011); Ranganath and Blei (2015); Taddy (2013, 2015)

3. Matrix and tensor factorizations: Hoff (2014); Johndrow et al. (2014); Schein et al. (2015); Yang and Dunson (2015); Zhou et al. (2014); Zhou and Carin (2015)

4. Models of graphs and networks: Airoldi et al. (2008, 2012); Gao et al. (2014); Linderman and Adams (2014); Wolfe and Olhede (2013)

References

Adams, R., Murray, I., and MacKay, D. (2009). Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *International Conference on Machine Learning*.

Airoldi, E., Blei, D., Erosheva, E., and Fienberg, S. (2012). Handbook of mixed membership models and their applications. *CRC Press*.

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2015). Rand-walk: A latent variable model approach to word embeddings. arxiv:1502.03520v5.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2016). Linear algebraic structure of word senses, with applications to polysemy. arxiv:1601.03764v1.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Bhattacharya, A. and Dunson, D. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, 107(497):362–377.

Brown, P. F., Pietra, V. D., de Souza, P., Lai, J., and Mercer, R. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Falish, D., Stephens, M., and Pritchard, J. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, 164:1567–1587.

Gao, C., Lu, Y., and Zhou, H. (2014). Rate-optimal graphon estimation. arXiv:1410.5837.

Gopalan, P., Hofman, J., and Blei, D. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*.

Hashimoto, T., Alvarez-Melis, D., and Jaakkola, T. (2015). Word, graph and manifold embedding from Markov processes. arXiv:1509.05808.

Hoff, P. (2014). Multilinear tensor regression for longitudinal relational data. arXiv:1412.0048.

Hoffman, M. and Blei, D. (2015). Structured stochastic variational inference. In *Artificial Intelligence and Statistics*.

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2014). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Johndrow, J., Battacharya, A., and Dunson, D. (2014). Tensor decompositions and sparse log-linear models. arXiv:1404.0396.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. arXiv:1405.4053.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Linderman, S. and Adams, R. (2014). Discovering latent network structure in point process data. arXiv:1402.0914.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mimno, D., Blei, D., and Engelhardt, B. (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences*.

Nalisnick, E. and Ravi, S. (2015). Infinite dimensional word embeddings. arXiv:1511.05392.

Paisley, J., Wang, C., and Blei, D. (2011). The discrete infinite logistic normal distribution for mixed-membership modeling. In *Artificial Intelligence and Statistics*.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12, pages 1532–1543.

Ranganath, R. and Blei, D. (2015). Correlated random measures. arXiv:1507.00720.

Schein, A., Paisley, J., Blei, D., and Wallach, H. (2015). Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In *Knowledge Discovery and Data Mining*.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*.

Taddy, M. (2015). Distributed multinomial regression. *The Annals of Applied Statistics*, 9(3):1394–1414.

Vilnis, L. and McCallum, A. (2015). Word representations via Gaussian embedding. In *International Conference on Learning Representations*.

Wolfe, P. and Olhede, S. (2013). Nonparametric graphon estimation. arXiv:1309.5936.

Yang, Y. and Dunson, D. (2015). Bayesian conditional tensor factorizations for high-dimensional classification. *Journal of the American Statistical Association*.

Zhou, J., Bhattacharya, A., Herring, A., and Dunson, D. (2014). Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*.

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320.