Probabilistic Models of Discrete Data

David M. Blei
Columbia University

January 22, 2016

1 Motivation

This is a course around probabilistic models of high-dimensional discrete data. High-dimensional discrete data pops up in many domains. We take special care when analyzing it. Here are some examples.

Purchase behavior. Users purchase (or click or "like" or read) items; the dimension is the number of items. On Amazon, for example, this is massive. These data come up in econometrics, marketing, and computer science.

Goals:

- To characterize purhase behavior
- To form recommendations
- To answer counterfactual questions about the economy

Neuroscience. Neurons fire over time; the dimension is the number of neurons. There are many neurons.

Goals:

- To characterize patters of firing
- To understand links between neurons and input/output

Documents. Documents are collections of words; the dimension is the number of terms in the vocabulary.

Goals:

- To characterize patterns of word use
- To understand connection between words and connected data (e.g., images)
- To form predictions based on text
- To answer causal questions around language, e.g., in the social sciences.

Genetics. People are collections of genes; the dimension is the locations of alleles on the genome.

Goals:

  - To discover patterns of gene occurrence, i.e., population structure
  - To discover causal connections between genes and diseases.
  - (Note: Disease occurrence is also high-dimensional discrete data.)

Others

- *Social networks.* People are collections of links; The dimension is the number of people we can link to. In some data, such as event data, there is data on the links as well.

- *Surveys.* Similar to genes, but each "location" is a survey question.

- *Music*

- Others?

Combinations. Many data combine high-dimensional discrete data with other types of data—e.g., text and images, text and social networks, data connected to space / time. These come with their own challenges.

2 Our goals for this class

Our goal is to study the state of the art in analyzing these types of data. We will try to identify common problems and possible solutions. Hopefully, we will see opportunities to apply techniques from one domain to another.

We focus on the probabilistic/Bayesian perspective. This is not because it's superior—it's not—but because it gives us a common language. We'll be reading all kinds of papers. We will be trying to understand each one within this perspective.

Side goals. These might even be more important.

- You will get used to reading academic papers at the cutting edge of statistics & machine learning.

- You will make progress on a research project of your choosing. At the end of the semester, you will be well on your way to a research paper. (These can be collaborative, alone, with people outside of the class, whatever.)

- We will become a community of researchers doing projects together, sharing our practices and pitfalls.

3 Prerequisites

- You have to have taken *Foundations of Graphical Models* and are comfortable with its material (Bishop, 2006; Murphy, 2013).

  What this means: You can understand a new probability model from the research literature. If it's conditionally conjugate, you can derive and implement a Gibbs sampler or variational inference algorithm.

- You are here with a research topic that involves discrete data. Your hope is to make progress on that research in the context of this class. (We are going to go around the room today and find out what those projects are.)

4 Readings

The readings will be from statistics, computer science, and other related fields. We will read from among these topics:

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

- Matrix and tensor factorization (Bhattacharya and Dunson, 2012; Hoff, 2014; Zhou et al., 2014; Johndrow et al., 2014; Yang and Dunson, 2015; Gopalan et al., 2015; Schein et al., 2015)

- Multinomial regression (Taddy, 2013, 2015)

- Mixed-membership models and Bayesian nonparametrics (Falish et al., 2003; Airoldi et al., 2008; Teh and Jordan, 2008; E. Airoldi, 2014; Broderick et al., 2015; Zhou and Carin, 2015; Paisley et al., 2011; Ranganath and Blei, 2015)

- Point processes (Adams et al., 2009; Linderman and Adams, 2014)

- Approximate posterior inference (Hoffman et al., 2013; Wang and Blei, 2013; Hoffman and Blei, 2015)

- Model checking and diagnostics (Box, 1980; Rubin, 1984; Gelman et al., 1996; Mimno and Blei, 2011; Gelman and Shalizi, 2012; Mimno et al., 2015)

This plan may change as the interests and priorities of the class develop. More topics we can consider include empirical Bayes, latent space models, robust modeling, and others. Please feel free to contact me with suggestions.

5 Requirements

Each participant is responsible for all the work.

(There are no auditors. If you bureaucratically cannot take the class, e.g., because you are a postdoc, then you are still responsible to do the work.)

In class:

- We are slotted for 1h50m. That's too short. We will usually go for longer, up to 2h30m. Is that a problem for anyone?
- Each week, I will lead the discussion. Note it is a discussion, not a lecture. We will all work together to understand the reading and the ideas.
- Each week, someone is the "scribe"—she takes notes about the discussion and writes a document about the meeting.
- Each week, one or two people will stand up and tell us about their projects.

Coursework:

- Each week you will write a reader response to the reading. This can be one paragraph or more. (There's no limit.) It does not need to be edited or polished. Please include discussion points.
- Each week (at least) you will update your project diary. I suggest briefly updating this each time you sit down to work on the project.
- You are responsible for a final report by the end of the semester. (There is no limit on length.) In addition, there will be several project milestones throughout the semester.

Your repository:

- Each of you will make a git repository `XXX_DDSeminar/` on bitbucket or github; it can be a private or public repository. (I personally prefer bitbucket because the `edu` email entitles you to unlimited private repositories.)
- It should have the following high-level directory structure. (Note, you will likely use subdirectories under these directories.)

`./doc/`	Documents that you write
`./doc/report/`	The final report
`./doc/reading_log.md`	Reader responses
`./doc/project_log.md`	Project progress
`./src/`	Code
`./dat/`	Data (if not too large)
`./ref/`	Reference materials, e.g., PDFs of papers
`./aux/`	Auxiliary (other stuff, e.g., pictures of whiteboards, random notes)

- The reading log contains your weeky response. Please use a `#` header before each one, and give the date. E.g. `# 2016-01-29 Bengio et al., 2003`.

- The project log contains your project progress. Precede each entry with the date, e.g., `# 2016-02-01`.

6 The final project

The main coursework is a final project. The project can be anything that touches on probabilistic models of discrete data. My hope is that you will be pursuing your PhD research here and/or working on a research paper.

The project does not have to be about papers we read. The idea is for you to do deep research on a project while continuing to broaden your knowledge of probabilistic models.

There will be several milestones throughout the semester, and I expect weekly progress on your project. Here is an example plan:

- Identify data and project idea; articulate a concrete problem
  (1 week)

- Clean and summarize data; visualize the data
  (1 week)

- Fit a simple model to the data; identify related work
  (2 weeks)

- Understand the ways the simple model falls short; propose an alternative
  (2 weeks)

- Develop and implement an inference algorithm for the new model
  (2 weeks)

- Run the new model on your data; understand the ways it falls short
  (2 weeks)

- Find, analyze, and assess another data set
  (2 weeks)

- Write the report
  (1 week)

7 Introductions

Let's go around the room:

- Name

- Department

- What type of data are you are interested in?
  What is one problem you want to solve with these data?
  Do you already have a specific project in mind?

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., 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.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Box, G. (1980). Sampling and Bayes' inference in scientific modeling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4):383–430.

Broderick, T., Mackey, L., Paisley, J., and Jordan, M. (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):290–306.

E. Airoldi, D. Blei, E. E. S. F., editor (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

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.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A. and Shalizi, C. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38.

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 preprint arXiv:1509.05808*.

Hoff, P. D. (2014). Multilinear tensor regression for longitudinal relational data. *arXiv preprint 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. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

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

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint 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 preprint arXiv:1402.0914*.

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

Mimno, D. and Blei, D. (2011). Bayesian checking for topic models. In *Empirical Methods in Natural Language Processing*.

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*.

Murphy, K. (2013). *Machine Learning: A Probabilistic Approach*. MIT Press.

Nalisnick, E. and Ravi, S. (2015). Infinite dimensional word embeddings. *arXiv preprint 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, D. (2014). Glove: Global vectors for word representation. *Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

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

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

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.

Teh, Y. and Jordan, M. (2008). Hierarchical Bayesian nonparametric models with applications.

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

Wang, C. and Blei, D. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:1005–1031.

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.