Representation Learning: A Probabilistic Perspective

David M. Blei
Columbia University
Spring, 2020

Description

We will study representation learning from a probabilistic modeling perspective, focusing on its theory, algorithms, and applications. Topics will include tensor factorizations, deep generative models, the variational autoencoder, sequential data, adversarial training, and attention mechanisms. We will study applications involving discrete data, such as text, networks, and user behavior.

Prerequisites

- You have taken STCS6701 *Foundations of Graphical Models*.
- You can derive and implement approximate posterior inference algorithms for new models.
- You have a good high-dimensional dataset in hand.

Organization

The course is open to doctoral students. No auditors and no pass/fail.

Our meetings are based around discussion. The discussion will focus on the readings and on your individual research, including "best practices" for research. I want us to share with each other the tools and techniques that we use to think, code, and write. At a higher level, this is a course about doing research.

During the first ten minutes of each session, students will pair off and discuss their last week's research and reading. The rest of the session is devoted to group discussion.

Coursework and grade

The coursework is the following:

- weekly readings
- weekly responses to the readings
- weekly progress on your final project repository (see below)
- a final project

You are graded on completing the responses, working consistently on research, the final project, and participation in the class community.

Final project

Every student will explore representation learning and document their research. This research culminates in a final project.

The project has two components: a report and a git repository.

The repository documents your research (except files that are too large, like raw data and the outputs of analysis). It is organized like this:

- `readme.md`
- `journal.md`
- `doc/`
- `src/`
- `etc/`

This repository will document your research and exploration through the semester.

The file `readme.md` simply contains an abstract of the project. At first, it is an "aspirational abstract," one that describes the research program you want to complete. You will refine it through the semester.

The file `journal.md` is a diary of your progress. It contains dated entries with a description of what you are doing, what you found, what you are thinking, and so on. It is mainly a resource for you, but I will glance at it too (at the end of the semester). Please update and commit it at least once per week.

The `doc/` directory contains latex documents that you are writing, a subdirectory for each one.

The `src/` directory contains the code you are writing.

The `etc/` directory contains anything else—materials, notes, photos of whiteboards, and so on—that you want to keep track of.

(Feel free to have other directories too, such as `dat/`, `out/`, and `fig/`. But keep the top level directories to three characters.)

Commit often, at least every week. You are graded on the quality of the project and the path that you took to get there.

"Syllabus"

We will try to understand representation learning as inference of local latent variables in a probabilistic model. (They might also be per-dimension latent variables, as in a word embedding.)

Why do we care about learning representations? Low-dimensional representations capture the correlation patterns among components in high-dimensional data. These patterns can help interpret

the data—understanding what kinds of hidden structures are helping to generate it—and can help make predictions. We often hope the uncovered patterns reveal something causal about the data, though this is an ambitious hope.

Notice how the concept of a "representation" is open ended. Its form can be binary, non-negative, real-valued, sparse, dense, hierarchical. And we can also choose how to generate an observation from a representation. My hunch is that this sampling model is important.

We will do research together about this topic. Each of us will have a dataset we care about and will explore probabilistic representation learning on that dataset. Be curious—unlike other projects, we are not yet interested in writing papers or having fancy new results. At first, you will simply explore the ideas of representation learning in your dataset.

We might explore many themes. Some of the more modern themes include the following:

- disentanglement
  (Achille and Soatto, 2017)
- attention mechanisms
  (Vaswani et al., 2017)
- representation in probabilistic deep learning
  (Bengio et al., 2013; Neal, 1992; Saul et al., 1996; MacKay, 2003; Ranganath et al., 2015)
- word embeddings
  (Bengio et al., 2003; Schnabel et al., 2015; Pennington et al., 2014; Mikolov et al., 2013; Rudolph et al., 2016; Ruiz et al., pear)
- invariance
  (Arjovsky et al., 2019)
- variational autoencoders and their extensions
  (Kingma and Welling, 2013; Tomczak and Welling, 2018; Dieng et al., 2019)
- tensor factorizations
  (Schein et al., 2016, 2015; Zhou et al., 2014)
- high-dimensional discrete data
- using representations in downstream tasks

But we might also explore some older themes:

- Bayesian nonparametrics
  (Neal, 2000; Kingman, 1993)
- non-negative matrix factorization
  (Lee and Seung, 1999; Gopalan et al., 2015; Cemgil, 2009)

Finally there are some ideas that will orbit these ideas:

- black-box variational inference
  (Ranganath et al., 2014)
- model checking and diagnostics
  (Box, 1980; Meng, 1994; Gelman et al., 1996; Ranganath and Blei, 2019)
- empirical Bayes
  (Efron, 2017)

We may well dive into them, even for multiple weeks. I'm eager to hear your ideas too.

References

Achille, A. and Soatto, S. (2017). Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv:1907.02893*.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

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

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

Cemgil, A. (2009). Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*, pages 4:1–4:17.

Dieng, A., Kim, Y., Rush, A., and Blei, D. (2019). Avoiding latent variable collapse with generative skip models. In *Artificial Intelligence and Statistics*.

Efron, B. (2017). Bayes, oracle Bayes, and empirical Bayes.

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

Gopalan, P., Hofman, J., and Blei, D. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*, pages 326–335.

Kingma, D. and Welling, M. (2013). Auto-encoding variational bayes. *ArXiv e-prints*.

Kingman, J. (1993). *Poisson Processes*. Oxford University Press, USA.

Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Meng, X. (1994). Posterior predictive p-values. *The Annals of Statistics*, pages 1142–1160.

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

Neal, R. (1992). Connectionist learning of belief networks. *Artificial Intelligence*.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

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

Ranganath, R. and Blei, D. (2019). Population predictive checks. *arXiv:1908.00882*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*.

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*.

Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. (2016). Exponential family embeddings. In *Neural Information Processing Systems*.

Ruiz, F. J., Athey, S., and Blei, D. M. (to appear). Shopper: A probabilistic model of consumer choice with substitutes and complements. *Annals of Applied Statistics*.

Saul, L., Jaakkola, T., and Jordan, M. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76.

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*, pages 1045–1054.

Schein, A., Zhou, M., Blei, D., and Wallach, H. (2016). Bayesian Poisson Tucker decomposition for learning the structure of international relations. In *International Conference on Machine Learning*.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Empirical Methods in Natural Language Processing*.

Tomczak, J. and Welling, M. (2018). VAE with a VampPrior. *Artificial Intelligence and Statistics*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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