IERG 5130 Readings for Topic 1

Note: This assignment should be completed by groups.

Topic 1 covers the basic concepts of graphical models, practical techniques in graphical model formulation, as well as exponential families. With these knowledges, we are now ready to dive into the world of graphical model formulation.

In this assignment, we will explore an important family of graphical models — **topic models**. Topic models originate from the area of document understanding, and have been later extended to other domains, *e.g.* scene or event understanding in computer vision.

Reading List

Here are several representative papers in this area:

- T. Hofmann. "Probabilistic latent semantic indexing". Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". Journal of Machine Learning Research (JMLR), vol. 3, pp. 993 - 1022, 2003.
- Michael Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. "The Author-Topic Model for Authors and Documents". Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI), 2004.
- David M. Blei, and John D. Lafferty. "Correlated Topic Models". Advances in Neural Information Processing System (NIPS), 2005.
- David M. Blei, and John D. Lafferty. "Dynamic Topic Models". Proceedings of International Conference on Machine Learnning (ICML), 2006.
- David M. Blei, and Jon D. McAuliffe. "Supervised Topic Models". Advances in Neural Information Processing System (NIPS), 2007.

Note: All these papers have already been uploaded to Piazza. You may also directly click the links on the titles above to reach the papers.

Most of these papers include both model formulation parts and inference/estimation algorithm derivation parts. We will discuss algorithms in topic 2 and 3. For now, please focus on the model formulation parts, and try to understand how the models are motivated and formulated.

Questions

Please bear the following questions in mind when you read the papers, and try to answer them when you finish the reading.

1. These models share a common paradigm in document modeling, that is, *bag of words*.
   - What is the key assumption behind this paradigm?
   - Is this assumption realistic? If not, why does it work?

2. What are the key differences between pLSA and LDA (from the standpoint the modeling, not algorithms)? Please specify an important capability of LDA that is not possessed by pLSA.

3. The *author-topic model*, *correlated topic model*, *dynamic topic model*, and *supervised topic model* are all extensions of LDA. Each model incorporates one additional aspect based on LDA. For each of these models, please answer the questions below:

    - What is the additional aspect/relationship that the model takes into account?
    - How is this additional aspect/relationship introduced into the formalism (*e.g.* what variables are introduced and how they are connected to the model)?
    - What assumptions do they make when formulating the additional factors (*e.g.* what are considered as independent under what conditions)?

4. What is the your understanding of the tradeoff between *"sophistication"* (*e.g.* incorporate more structures or factors) and *"simplicity"*. In practice, how would you find a right balance between them?

5. In scientific publications, each paper usually consists of the following aspects — *text document*, *authors*, *year*, *citations*. Try to formulate a graphical model that takes into account all these aspects as well as their relations. When formulating the model, you can make reasonable assumptions by yourselves (but please explicitly articulate them in your discussion). You will have chance to present your formulation during the in-class discussion session for topic 1.