Topic Modeling Research

Melody Jiang

May 2019

1 Possible Research Directions

1.1 Two main approaches to clustering

(i) **Distance-based** clustering. Using this approach, we analyze matrix of pairwise distances. Namely, suppose y_i and y_j are data points, then D is a distance matrix whose d_{ij} entry represents distance between y_i and y_j.

(ii) **Model-based** clustering. For example, $y_i \sim \sum_{h=1}^{k} \pi_h \mathcal{K}(\theta_h)$.

1.2 Two main problems in clustering

(i) sensitivity to kernel

(ii) issues in high dimensions (large p)

1.3 Semi-solutions

1. **C-Bayes**. All derivations from assumed models (e.g. kernel misspecification). See [Miller and Dunson, 2018].
2. **Model plus distance-based clustering**. See [Duan and Dunson, 2018].
3. **Calculating better distances**. E.g., geodesic or intrinsic distace (Didong Li & Dunson, in preparation).
4. **To address issues in high dimensions**, cluster on the latent variable level or varational autoencoder (VAE).

2 Literature Review

Three main tasks in textmining are clustering, classification, and information extraction [Allahyari et al., 2017]. Topic modeling can be applied to all of these tasks [Lu et al., 2011]. We would like to focus on the most commonly used topic model, Latent Dirichlet Allocation (LDA), and LDA's robustness in the documeng clustering task. Lu et al. investigated LDA's task performance in document clustering and found LDA's performance is quite sensitive to the setting of its hyper-parameter and parameter [Lu et al., 2011]. In terms of robustness, Wang et al. proposed a model-based approach to make LDA more robust by using localization and empirical Bayes [Wang et al., 2018].

2.1 Latent Dirichlet Allocation (LDA)

Here, we reproduce LDA introduced in [Wang et al., 2018].

3 Pilot Study

4 Annotated Bibliography

[Wang et al., 2018] Title: A general method for robust Bayesian Modeling This apaper proposes a general model-based approach to robustify Bayesian models.

[Doyle and Elkan, 2009] Bursty Bayesian models.

[Blei et al., 2003] Title: Latent dirichlet allocation An introduction to latent dirichlet allocation.

References

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Doyle and Elkan, 2009] Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288. ACM.

[Duan and Dunson, 2018] Duan, L. L. and Dunson, D. B. (2018). Bayesian distance clustering. *arXiv preprint arXiv:1810.08537*.

[Lu et al., 2011] Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203.

[Miller and Dunson, 2018] Miller, J. W. and Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13.

[Wang et al., 2018] Wang, C., Blei, D. M., et al. (2018). A general method for robust bayesian modeling. *Bayesian Analysis*, 13(4):1159–1187.