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

References

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

[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.