Geometric and topological methods in machine learning

Frederic.Cazals@inria.fr, Mathieu.Carriere@inria.fr

Academic year: 2021-22

Contents

Procedure to select the projects:

- Groups of two students must register on the following poll `https://framadate.org/3xSNH5dKHbNHht27`

  NB: Options are chosen in a first-come first-served basis: do not vote for a project that has alreaby been taken!
- When choosing a project, make sure to write the two last names.
- **Deadline to fill the doodle (first come first serve basis): March 24th, 2022**
- **Deadline to return your work: April 17th, 2022**

4 k-means versus non negative factorization techniques for data clustering

As studied in class, `k-means++` is a clustering algorithm performing a smart seeding for `k-means`, see [4]. In the sequel, the `k-means` functional (i.e. the function being optimized) is denoted Φ_K.

Non negative matrix factorization (2-NMTF and 3-NMTF) are matrix algorithms which can also be used for clustering [6, 7, 8]. For example, it is well known that 2-factor NMTF and `k-means` are equivalent [7, Thm1].

The goal of this project is to get a deeper understanding of the commonalities/differences between the two classes of methods.

1. Prepare a generic data set defined as a mixture of say five 2D Gaussian, parameterized by the concentration parameters of the Gaussians, and the relative distance d between the centers of the Gaussians. See Fig. 1 for an illustration.

2. In `k-means++`, the selection of the pivots is carried out using a quadratic exponent on distances. Present an experimental study by varying this exponent. Make sure to pay a special attention to

   - the average value obtained versus the extreme ones.
   - the results obtained in the context of the theorem proved in [4] on the expectation for Φ_K.

3. Classical NMTF algorithms do not have the notion of smart seeding. Using various instances of the dataset designed above and/or datasets you are familiar with:

   - perform a comparison between `k-means++` and 2-factor NMTF,
   - perform a comparison between `k-means++` and 3-factor NMTF [8],
   - compare the merits of 2-factor NMTF, 3-factor NMTF, and `k-means++`.

4. As a final investigation, we wish to challenge the algorithms using data of low intrinsic dimension. As seen during the class, embed the data used so fart into a say $D = 50$ dimensional space. Repeat the comparison between `k-means++`, 2-factor NMTF, and 3-factor NMTF.

   Discuss the results in the context of concentration phenomena studied in the class.

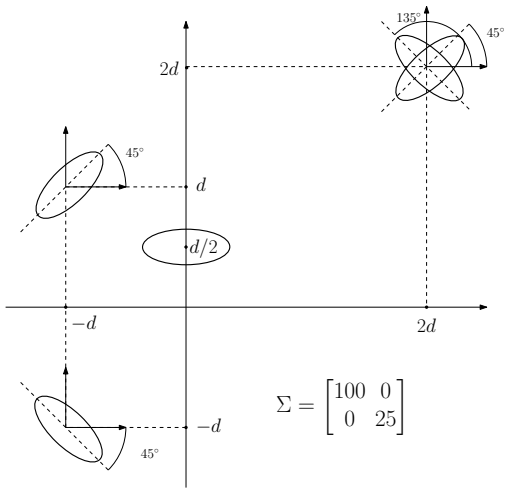

Figure 1: **A mixture of Gaussian distribution parameterized by the distance d.**

Contact. Frederic Cazals: frederic.cazals@inria.fr

References

[1] Peter N Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *ACM SODA*, volume 93, pages 311–321, 1993.

[2] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[3] S. Dasgupta and Y. Freund. Random projection trees for vector quantization. *Information Theory, IEEE Transactions on*, 55(7):3229–3242, 2009.

[4] D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *ACM-SODA*, page 1035. Society for Industrial and Applied Mathematics, 2007.

[5] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

[6] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[7] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[8] Nicoletta Del Buono and Gianvito Pio. Non-negative matrix tri-factorization for co-clustering: an analysis of the block matrix. *Information Sciences*, 301:13–26, 2015.

[9] J. Chodera, W. Swope, J. Pitera, and K. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Modeling & Simulation*, 5(4):1214–1226, 2006.

[10] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in riemannian manifolds. *J. ACM*, 60(6):1–38, 2013.

[11] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li. Hicrep: assessing the reproducibility of hi-c data using a stratum-adjusted correlation coefficient. *Genome Research*, 27(11):1939–1949, 2017.

[12] Marian Gidea and Yuri Katz. Topological data analysis of financial time series: Landscapes of crashes. *Physica A*, 491:820–834, 2018.