Research Directions to Validate Topological Models of Multi-Dimensional Data

Nello Blaser
nello.blaser@uib.no
Department of Informatics
University of Bergen, Norway

Michaël Aupetit
maupetit@hbku.edu.qa
QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

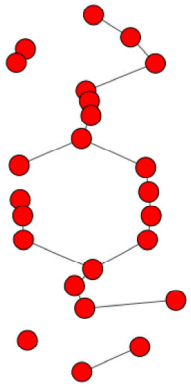
Goal

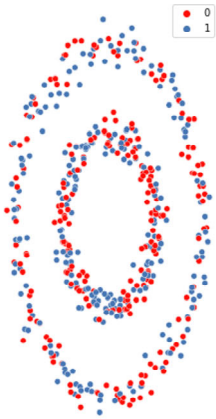
Topological methods in machine learning aim to quantitatively encode shape information from multi-dimensional data points.
Validation relies on defining a validation measure to compare topological models.

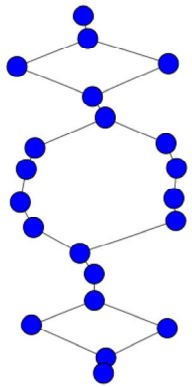
What could be a validation measure relating topological properties of the model and statistical properties of the data for the Mapper [1] and the Generative Simplicial Complex [2,3,4] models?

Research directions for validation

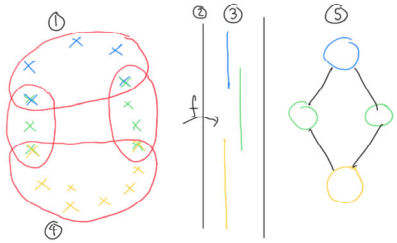
Different samples (blue/red) from the same distribution.
But very different Mapper nerves…





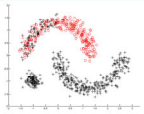


Mapper [1]



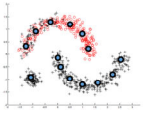
1) Input data
2) Filter function f
3) Cover $Im(f)$
4) Cluster preimages
5) Compute nerve

[2,3,4] Generative Simplicial Complex



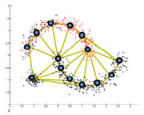
Input

1) Input (labeled) data $x \in \mathbb{X}$

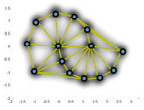


GSC

2) Gaussian Mixture Model
- Parameters θ estimated with Expectation Maximization (EM)
- #centers w selected with Bayesian Information Criterion (BIC)

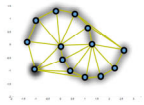


3) Delaunay complex of GMM centers w (Here the case of the Delaunay Graph)

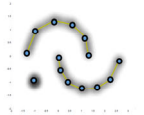


4) Gaussian kernel $g(x, w, \theta)$ convoluted to each simplex W_σ with its own prior weight π_σ

$$p(x, S, W, \theta) = \sum_{\sigma \in S} \frac{\pi_\sigma}{|W_\sigma|} \int_{W_\sigma} g(x, w, \theta) dw$$

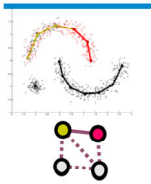


5) EM: prior weights of generative simplices which do not explain data tend towards 0



6) BIC: Simplices with 0 prior get pruned

Output



7) Max A Posteriori gives class label for each simplex

8) Summary graph/simplex based on connected components in initial and pruned Delaunay complex

References

[1] Pek Y. Lum et al. Extracting insights from the shape of complex data using topology. Sci Rep, 3:1236, 2013
[2] Michaël Aupetit. Learning topology with the Generative Gaussian Graph and the EM algorithm. NIPS 2005
[3] M. Maillot, M. Aupetit, G. Govaert. A generative model that learns Betti numbers from a data set. ESANN 2012
[4] P. Gaillard, M. Aupetit, G. Govaert. Learning topology of a labeled data set with the supervised Generative Gaussian Graph. Neurocomputing, 71(7-9): 1283–1299, 2008

Research directions for validation

At step 6), BIC is used to select a « good » simplicial complex based on a statistical criterion on the density $p(x, S, W, \theta)$

- Can we prove it also gives a « good » topological model of the data?
- Or can we find another criterion which does link statistical and topological models properly?
- How the number of data x and the parameters θ of the model impact the « coupling » between density and topology?

From step 4), can a filtration based on the priors π_σ give an interesting topological model?

From steps 2) and 4), can we use multidimensional persistence theory on the number of centers w and the priors π_σ?

Looking for a Post-doc or PhD on these topics, please contact us!