CS 410/510: Topological Methods in Data Analysis and Machine Learning – Final Project

Instructor: Tao Hou

1 Summary

- Due: June 12
- What you need to submit: A report in .pdf (we don't have a presentation). More details on the page limit, format, etc. are provided in later sections.
- Work in groups: You should work in a group of 2-4. Make sure you state clearly the group members on the title page of your report. **Only one** of the group members needs to upload the final report to the Canvas submission folder.
- You could choose to do either of the two types of work for the project (details on each type are provided in later sections):

  1. Paper summary: You summarize some research papers and present your understanding of them.
  2. Experiments: You take some data, analyze it using the approaches learned in class, and report your observation. You don't need to submit your codes.

  You are encouraged to read detailed instructions that follow on **both** types of work no matter you choice (it should give you a better idea).

- Other things you could do:
  - Your summary of "papers" could also be more open-ended: you could choose to survey some other scientific literature.
  - You could do a combination of both: You read some papers, and do some experiments based on your reading.
  - You could also explore some problem on your own: This is like doing research, you are exploring something new.
  - You could also work on some mathematical or algorithmic problems related to topology or geometry.
  - If you have questions regarding whether your proposed work falls in the scope, you can ask.

2 Format of the Report

The report should have an abstract, an introduction, followed by several sections describing your work. Reference needs to be provided if you want to refer to others' work. If your report is very theoretical, you may need a "preliminary" section.

Your report should be in **single column**, reasonably spaced. See this ACM template for an example. There is a "soft" limit of **6 pages**: You are encouraged to put the most essential information in the first 6 pages, and if you have more to write, you can put them in "appendix".

You could use either LaTex or Word for typesetting, although LaTex is slightly encouraged. Here are a few resources for learning LaTex:

- `https://www.overleaf.com`: an online (free) system for preparing LaTeX documents. It saves your efforts on installing and configuring a LaTeX editing/compiling environment on your own computer.
- `https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes`: a short intro to LaTeX.
- `https://www.overleaf.com/learn/latex/Algorithms`: a short tutorial on typesetting pseudocodes using LaTeX.
- `https://youtu.be/Jp0lPj2-DQA?si=7_BBZcQYRgW5TyvH`: a Youtube video introducing how to use LaTeX on Overleaf.

3 Regarding Paper Summary Project

You may want to inspect the common techniques used and compare the difference for papers you summarize. The more insights you could get from theses approaches yourself, the better it is. It's okay that you don't understand some parts of the paper, try to present clearly what you can digest. The number of papers to summarize can vary: 3 would be enough. You could do less. Just try to read to the best you can.

When reading them, try to think about the following:

- What specific approach are they using?
- What type of topological features are they capturing?
- Why are the topological features helpful?
- What are the things that these topological approaches can do that other more traditional approaches cannot?
- If they are using persistent homology, what filtrations are they building? Why do they use the specific way to build the filtration but not others?
- If they are using persistent homology, how do they interpret the persistence diagram?
- Which part of the proposed method you think is challenging?
- Any shortcomings or possible improvement?

4 Regarding Experimental Project

Choose a dataset (see Section 6). Apply techniques in TDA (learned inside or outside of the course) on the dataset. Interpret the topological summary you get and report any topological, qualitative phenomena you can find. Also report your approach. Feel free to integrate TDA techniques into pipelines from other CS domains such as machine learning or visualization. You may need to preprocess the data you have, due to several reasons: the dataset may not be immediately ready for the TDA pipeline or simply is too big to process (either each individual item is of too high a dimension or the number of items is large).

5 Sample Papers

Here are a list of papers in different domains (the list is only supposed to give you an example; you are encouraged to find them on your own by using, e.g., google scholar or just google):

Computer vision:

- Wong CC, Vong CM. Persistent homology based graph convolution network for fine-grained 3d shape segmentation.
- Ver Hoef L, Adams H, King EJ, Ebert-Uphoff I. A primer on topological data analysis to support image analysis tasks in environmental science.
- Qaiser T, Tsang YW, Taniyama D, Sakamoto N, Nakane K, Epstein D, Rajpoot N. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features.

Medical imaging / healthcare:

- Bukkuri A, Andor N, Darcy IK. Applications of topological data analysis in oncology.
- Moon C, Li Q, Xiao G. Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers.
- Rammal A, Assaf R, Goupil A, Kacim M, Vrabie V. Machine learning techniques on homological persistence features for prostate cancer diagnosis.
- Xia K, Wei GW. Persistent homology analysis of protein structure, flexibility, and folding.

Neuroscience:

- Ellis CT, Lesnick M, Henselman-Petrusek G, Keller B, Cohen JD. Feasibility of topological data analysis for event-related fMRI.
- Stolz BJ, Emerson T, Nahkuri S, Porter MA, Harrington HA. Topological data analysis of task-based fMRI data from experiments on schizophrenia.
- Anderson KL, Anderson JS, Palande S, Wang B. Topological data analysis of functional MRI connectivity in time and space domains.

Materials science:

- Nakamura T, Hiraoka Y, Hirata A, Escolar EG, Nishiura Y. Persistent homology and many-body atomic structure for medium-range order in the glass.
- Srensen SS, Biscio CA, Bauchy M, Fajstrup L, Smedskjaer MM. Revealing hidden medium-range order in amorphous materials using topological data analysis.
- Cramer Pedersen M, Robins V, Mortensen K, Kirkensgaard JJ. Evolution of local motifs and topological proximity in self-assembled quasi-crystalline phases.

Misc.

- Herring AL, Robins V, Sheppard AP. Topological persistence for relating microstructure and capillary fluid trapping in sandstones.
- Ravishanker N, Chen R. An introduction to persistent homology for time series. Wiley Interdisciplinary Reviews: Computational Statistics.
- Rieck B, Yates T, Bock C, Borgwardt K, Wolf G, Turk-Browne N, Krishnaswamy S. Uncovering the topology of time-varying fmri data using cubical persistence.

6 Sample Datasets

Here is just a list after a search over the internet; feel free to use one you find on your own:

General

- **UCI Machine Learning Repository**
  `https://archive.ics.uci.edu/`
- **Kaggle Datasets**
  `https://www.kaggle.com/datasets`
- **OpenML**
  `https://www.openml.org/`

Computer Vision

- **ImageNet**
  `http://www.image-net.org/`
- **COCO (Common Objects in Context)**
  `https://cocodataset.org/`
- **MNIST & Fashion-MNIST**
  `http://yann.lecun.com/exdb/mnist/`
  `https://github.com/zalandoresearch/fashion-mnist`
- **CIFAR-10/100**
  `https://www.cs.toronto.edu/~kriz/cifar.html`

Natural Language Processing

- **GLUE/SuperGLUE Benchmarks**
  `https://gluebenchmark.com/`

- **SQuAD (Stanford Question Answering Dataset)**
  `https://rajpurkar.github.io/SQuAD-explorer/`
- **Common Crawl**
  `https://commoncrawl.org/`
- **IMDb / Yelp / Amazon Reviews**
  IMDb: `https://ai.stanford.edu/~amaas/data/sentiment/`

Healthcare / Bioinformatics

- **MIMIC-III / MIMIC-IV**
  `https://physionet.org/content/mimiciii/`
- **PhysioNet Challenge Datasets**
  `https://physionet.org/about/database/`
- **The Cancer Genome Atlas (TCGA)**
  `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`

Scientific

- **OpenNeuro**
  `https://openneuro.org/`
- **Materials Project**
  `https://materialsproject.org/`
- **Protein Data Bank (PDB)**
  `https://www.rcsb.org/`

Other Interesting Niches

- **Google Dataset Search**
  `https://datasetsearch.research.google.com/`
- **Awesome Public Datasets (GitHub)**
  `https://github.com/awesomedata/awesome-public-datasets`
- **FiveThirtyEight Datasets**
  `https://data.fivethirtyeight.com/`