Persistent Homology in Data Analytics

For students of Computer Science

September 28, 2020

Seminar: Persistent Homology in Data Analytics
Teacher: Luciano Melodia, M.A.

Weekly meetings: Thu. 14:00–15:45 c.t.
Room: 08.130 (most probably).

1 Organizational details

- **Participation:** If you want to participate in the seminar, please be present in all seminar units, especially in the first one, and actively participate in the discussions.
- **Date and room:** So far our seminar room 08.130 of the Chair of Data Management (Computer Science 6) has been selected, but this could change. I will inform you by e-mail if there are any organizational changes.
- **Preparation**: Every seminarist is obliged to make an appointment with me in the week before his or her presentation, in which we check possible difficulties and the quality and correctness of the notes and slides.
- **Examination:** A seminar paper (8–10 pages) must be written on the presentations topic (60–80 minutes including questions during the presentation) and at the latest be submitted to me in electronic form the week after the lecture. The seminar papers are provided on the StudOn page of the seminar.

  The module grade is composed of: the evaluation of the presentation (50%) and the evaluation of the seminar paper (50%).

2 Content

Topological data analysis describes a field in computational topology that uses methods of algebraic topology to investigate sets of points using computer algebra. A particularly prominent tool is persistent homology. Algebraic topology describes, characterises and distinguishes topological spaces. In this seminar we

concentrate on tools for the calculation of homology groups on a filtration of the space. A filtration is a set of subsets up to a given space that induce an exact sequence of homomorphisms.
If we have a discrete set of points of a tame function $f : U \subset \mathbb{R} \to \mathbb{R}$, we want to specify the changes of their sublevel sets along this function. These sublevel sets have the form $L_c^-(f, c) = \{x \mid f(x) \leq c\}$. Starting with the smallest value, we want to show changes in the function. We know from calculus that the number of locally connected components changes only by extreme values. For a local minimum, a new connected component appears. For a local maximum, two connected components are merged. Thus we iterate through the values of the function $\{y_0, y_1, \cdots\}$, considering a finite set and defining a step size to obtain a discrete countable finite set, so that $||x_j - x_i|| = \epsilon$ for $j > i$. In each local maximum, we have two connected components that have the creation values c and c' for the time of their creation. We assume that $c \leq c'$. So the adjacent component with the value c is younger than the component with the value c', since the latter was created at a later function value. We merge the younger component with the older one and store this event as tuple $(c, d) \in \mathbb{R}^2$, where d is the local maximum of the function. If we run the whole function this way, we have a family of tuples describing all the changes of f. Since the very first connected component (the one created with the smallest function value) cannot be merged with another, this connected component is persistent throughout the run and will be persistent in all future. Therefore, d receives the value ∞. We now refer to the space of the tuples as extended Euclidean space $\overline{\mathbb{R}^2} := \mathbb{R}^2 \cup \infty$. Since we consider the tuples as points in extended Euclidean space, we obtain a diagram, the so-called persistence diagram.
The seminar introduces which information can be read from a persistence diagram, how persistence diagrams are efficiently calculated and how this diagram is related to the so-called persistent homology. Furthermore, to generalize the example, it will be worked out how persistent homology can be calculated on multivariate data. You guess right, that persistence diagrams encode a ton of interesting features of data.

3 Organization of the lectures

- Each presentation should last a maximum of 80 minutes and should include at least 60 minutes (10-15 minutes time for questions during the lecture). Projector, Powerpoint, overhead slides and blackboard writings are all permitted.

- The most important aspects are that (1) you have understood exactly what you are talking about and (2) your presentation communicates the content to your fellow students in an understandable and educative way.

- Try to design your presentation in such a way that you want to hear it yourself and that you have real added value over reading the underlying text in a textbook. Of course you can and should add or omit details or structure statements differently. This also includes explaining or precalculating geometric examples.

- Many of the concepts discussed are not directly accessible. You will probably have to edit your texts several times to gain some understanding of the topic and to add many steps that are not described in detail in the literature. If you do not understand a particular concept, please make an appointment with me in time.
- Begin preparing your presentation in time. Late preparation regularly turns out to be a problem in seminars. You probably need much more time for preparation than you think.
- Plan your slides carefully. A disorganized presentation will reduce the quality of your talk. You should not only think about what you want to say, but also how and in what form you want to introduce and illustrate the concepts. Think about what is really essential and avoid too detailed text in your slides.
- Practice your presentation alone or with other students before you give it in the seminar. If you decide to use the blackboard, practice your presentation aloud with a fellow student or alone. Think about what exactly you want to show on the blackboard and write down your ideas beforehand.
- Your seminar paper is a scientific paper. You must therefore list all sources used, both in the presentation and in the written version prepared. The citation standard is up to you, but must be consistent throughout.
- All books can be borrowed from the university library. Please make your own efforts to obtain the copies. All articles can be downloaded free of charge from the university network. If there are any problems, please check first if the article is available for FAU. If not, please write to me in time and I will send you the corresponding copy.

4 Talks

4.1 Topological spaces and groups

Name: **Date:**
Task: Define and illustrate the concepts topological space, metric space, topology of a metric space, bases and subbases of topologies, continuous mappings and homeomorphisms, compactness, groups, homomorphisms of groups and the most important vector spaces, Hilbert spaces, Banach spaces and Fréchet spaces.
Literature: Jänich, Klaus. Topologie. Springer-Verlag, 2013. Chap. 1 and 2.

4.2 Simplicial complexes

Name: **Date:**
Prerequisites: 4.1
Task: Define simplicial complexes and simplicial maps. Explain the simplicial approximation theorem and the nerve theorem. You do not need to show the proofs in their entirety here, but you should have worked through them. Sketch the core concepts and ideas.
Literature: James Munkres: Elements of algebraic topology, CRC Press, 2018. Sections $1-3$ and sections 14 and 16.

4.3 Simplicial complexes associated with point clouds

Name: **Date:**
Prerequisites: 4.1
Task: Define and show different simplicial complexes on a set of points in a metric space. In particular, show the Voronoi diagrams, the Delaunay triangulation, the duality of the two as well as Alpha complexes and Witness complexes.
Literature: Boissonnat, Jean-Daniel, Frédéric Chazal, and Mariette Yvinec. Geometric and topological inference. Vol. 57. Cambridge University Press, 2018. Chap. 4.2, 4.3, 6.1, 6.2.

4.4 Simplicial homology groups

Name: **Date:**
Prerequisites: 4.1
Task: Recall simplices and simplicial complexes. Explain the concept of (free) abelian groups and whether homology groups are free. What are normal subgroups? Define chain groups, cycle groups and boundary groups and lead to the concept of homology groups. Show graphical examples and calculations of homology groups.
Literature: Hatcher, Allen. Algebraic topology. Cambridge University Press, 2005. S. 102-106.

4.5 Persistent homology

Name: **Date:**
Prerequisites: 4.1, 4.3 or 4.4
Task: Define persistent homology, show persistence diagrams and explain persistence diagrams using the relationship between persistent homology classes and the critical values of one-dimensional tame functions. Also explain persistent homology on the filtration of point sets in higher-dimensional spaces.
Literature: Edelsbrunner, Herbert, and John Harer. ”Persistent homology-a survey.” Contemporary mathematics 453 (2008): 257-282.
Boissonnat, Jean-Daniel, Frédéric Chazal, and Mariette Yvinec. Geometric and topological inference. Vol. 57. Cambridge University Press, 2018. Chap. 11.5.

4.6 Computation of persistent homology

Name: **Date:**
Prerequisites: 4.1, 4.3 or 4.4
Task: Specify the algorithm to calculate persistence diagrams and describe the matrix reduction techniques used. Give examples and make exemplary calculations of the triangulation of basic compact geometric objects.
Literature: Afra Zomorodian, and Gunnar Carlsson. "Computing persistent homology." Discrete & Computational Geometry 33.2 (2005): 249-274.
Otter, Nina, et al. "A roadmap for the computation of persistent homology." EPJ Data Science 6.1 (2017): 17.

4.7 Persistent homology and cohomology*

Name: **Date:**
Prerequisites: 4.1, 4.3 or 4.4, 4.5
Task: Remind the audience of the definition of persistent homology. Specify the dual concept of persistent cohomology and explain the proof that both persistent homology and persistent cohomology generate the same barcodes. Why is persistent cohomology more efficient to compute?
Literature: De Silva, Vin, Dmitriy Morozov, and Mikael Vejdemo-Johansson. "Dualities in persistent (co) homology." Inverse Problems 27.12 (2011): 124003.

4.8 Distances and the stability theorem

Name: **Date:**
Prerequisites: 4.1, 4.3 or 4.4, 4.5
Task: Describe the stability of the persistence diagrams in relation to the Hausdorff distance, the bottleneck distance and the Wasserstein distance. Give all three metrics and explain them with illustrative examples. Explain the quadrant lemma without formal proof, so that the listener gets a good understanding of the properties of persistence diagrams.
Literature: David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. "Stability of persistence diagrams." Discrete & Computational Geometry 37.1 (2007): 103-120.

4.9 Statistics in persistent homology

Name: **Date:**
Prerequisites: 4.1, 4.3 or 4.4, 4.5
Task: Define persistence landscapes and make it clear how to obtain them from the persistence diagrams. Explain why the ordinary persistence diagram, unlike the persistence landscape, does not lie in a vector space and explain why this property is important for statistical data analysis.
Literature: Bubenik, Peter. "Statistical topological data analysis using persistence landscapes." The Journal of Machine Learning Research 16.1 (2015): 77-102.