Topological Data Analysis

INF563

Steve Oudot

Course organization:

9 sessions, each split into:

- 2 hours of lecture,
- 2 hours of exercises (PC) or lab (TD).

URL: http://www.enseignement.polytechnique.fr/informatique/INF563/

Course description:

Topological Data Analysis is an emerging trend in exploratory data analysis and data mining. It has known a growing interest and some notable successes in the recent years. The idea is to use topological tools to tackle challenging data sets, in particular data sets for which the observations lie on or close to non-trivial geometric structures that can fool classical techniques. These topological tools are able to extract useful information on those structures from the data, and to exploit that information to enhance the analysis pipeline. The objective of this course is to familiarize the students with these tools lying at the confluence of pure mathematics, applied mathematics, and computer science. Emphasis is put on the methods themselves and on their theoretical guarantees. Meanwhile, the lab sessions focus on challenging data sets, primarily multimedia data sets such as collections of images or 3d shapes.

Course evaluation: midterm TD + final written exam

Session details:

Session 1. Context: splendour and misery of dimensionality reduction:

Lecture:

- quick review of linear methods (PCA, MDS) and non-linear methods (Isomap, LLE)
- limitations: trivial topology, domain convexity
- illustrative examples where the conditions are not met so the methods are fooled by the data

Lab session (TD):

Goal: explore artificial and real data sets using some dimensionality reduction package in R or Python:

- some standard data sets: hand-written digits, swiss roll
- some challenging data sets: natural images, molecular energy landscapes (alanine, cyclohexane)
- understanding the reasons why the methods fail (which assumptions are not met by the data)

Session 2. Clustering and persistence theory:

Lecture:

Hour 1:

- hierarchical clustering: single-linkage
- dendrograms vs ultrametrics
- stability theorem for ultrametrics

Hour 2:

- mode-seeking: graph-based hill-climbing
- merge trees of functions -> barcodes
- application to density functions, ToMATo

Lab session (TD):

Goal: implement and test ToMATo

- implementation in the students' preferred language (C/C++, java, python)
- test against artificial data sets: 2d (cones/craters, spirals), 3d (rings)
- test against real data: images (for segmentation in color space), molecular energy landscapes (alanine, cyclohexane)

Sessions 3-4. Quick introduction to homology theory:

Lectures:

- goal/motivation: extract higher-dimensional topological features (not just connected components, but also holes, voids, etc.) from data
- intuition: cycles for capturing topological features, equivalence between cycles
- simplicial homology: abstract/embedded simplicial complexes, chains, boundary operator, cycles, boundaries, homology groups, morphisms, homotopy invariance
- computation: boundary matrix reduction, cubic complexity, almost linear complexity for 0-dimensional homology
- singular homology: singular simplex, then same process as for simplicial homology, equivalence of simplicial/singular homologies
- functoriality

Exercise sessions (PC):

Goal: solve some classical exercises to get familiar with the concepts (cf. [Hatcher], Chapter 2):

- computing the homology of spheres, tori, projective spaces, Klein bottles, etc.
- well-known counter-examples: house with two rooms, Poincaré's homology sphere, etc.
- Brouwer's fixed point theorem
- degree of map S^n->S^n, application to (non-)existence of nonzero continuous tangent vector fields on S^n

Session 5. General persistence theory:

Lecture:

- filtrations, persistence modules via homology functor
- decomposability and resulting persistence barcodes / diagrams
- computation: basic matrix reduction algorithm
- stability theorem

Lab session (TD):

- implementation of the basic matrix reduction algorithm (with coeffs. in Z_2) in the students' preferred language + verification on the previous simple examples
- experimentation to compute persistence homology of various functions (height, distance to a point, curvature etc.) on triangulated surfaces and triangulated planar domains (e.g. square) -> this is in preparation for the next sessions

Homework (DM):

- simple properties of persistence (cf. [Edelsbrunner, Harer], end of Chapter VII):
- manual application of basic matrix reduction algorithm to simple filtrations
- simple proof of the stability theorem
- proof of the Euler-Poincaré theorem using persistence

Session 6. Topological inference:

Lecture:

- theory of distance functions
- connection to persistence
- from unions of balls to simplices

Lab session (TD):

- implementation of Rips filtrations
- application to the analysis of collections of images

Sessions 7-8. Feature design using topology:

Lectures:

- introduction/motivation for stable descriptors to compare data (3d shapes used as a running example)
- stable global topological descriptors
- stable local topological descriptors
- convergence rates
- from descriptors to feature vectors: the kernel trick applied to persistence diagrams

Lab sessions (TD):

- implementation of descriptors for 3d shapes using height filtrations
- implementation of feature vectors from the descriptors
- application to pose detection among a collection of 3d shapes
- datasets: SHREC 2010 / TOSCA

Session 9. Graph structures for TDA:

Lecture:

- Reeb graphs and extended persistence
- Mapper, MultiNerve Mapper
- Signatures, convergence and stability

Lab session (TD):

- implementation of a basic version of Mapper
- application to the analysis of data sets such as the NBA data.