CS 6170 Computational Topology Lecture 25

Apr 11

Data Science for topologists

- Jennifer Gamble @ AWM Symposium.

→ Data for machine learning : requires cleaning , pre-processing

① How to deal with missing value

② How to define "distance" between samples

③ How to normalize columns

- if one column has range [0,1] and other column has range $[0, 10^6]$ → how do you define "distance" that gives equal importance to both ?

→ Predictive modeling : linear regression / machine learning (classification etc.)

→ When are topological methods useful ?

ⓐ Exploratory data analysis → understanding structure of data.

ⓑ When data is not in form of point cloud but there is some distance / dissimilarity measure (eg. Matrix, Network)

* ← ⓒ Understanding behavior of traditional m.l. methods.

→ Decision tree : recursively divide data along different dimensions — over-fitting training data ? training accuracy vs. test accuracy.

→ Random forests : made up of multiple individual decision trees.

→ Each tree randomized : either use random subset of attributes / random subset of samples for training

→ final prediction is average over all decision tree predictions.

# Key idea : apply mapper to the predictions.

* ← ⓒ

⇒ Understand the behavior of random forests.

→ Network Analysis using topology

: Graph as 1-skeleton of a simplicial complex

→ Build higher dimension simplicial complex ⇒ apply homology

# Node Dominance Collapse : simplify network by reducing number of nodes but still preserving homology.

TDA on large dataset

→ How to handle very large data ?

① spassification : preserve homology / spectral properties etc.

② sampling

③ Parallel / Distributed computing : Scale up computational capability.

④ sketching / approximation

---

Topological Complexity in Protein Structures

- Erica Flapan