An Empirical Analysis of Topological Persistence as a Supplementary Measure of Dataset Drift

Christopher Shultz

Introduction

Background

* What is dataset drift? What are the typical mechanisms for detecting it? What are their pros/cons?
* What is TDA? Simplicial Complexes? Persistnet Homology? Persistence Entropy?
* What new information does this view provide that is otherwise unobservable? [a] captures topological changes that may be missed with traditional methods; [b] PE considers features across multiple scales, making it sensitive to both local and global changes in the data; [c] TDA methods like PE are robust to noise, ensuring that drift detection isn’t overly sensitive to minor fluctuations; [d] combining PE with traditional metrics provides a more comprehensive understanding of drift.

Method

* Compute PE by generating persistence diagrams for the data. Calculate PE from these diagrams.
* PE method can be integrated into existing drift monitoring systems alongside traditional metrics.
* Thresholds for acceptable levels can be established based on historical data and domain knowledge.
* Empirical studies can demonstrate the effectiveness of PE in detecting drift – one below.
* Simulation
  + Use the Adult dataset. Build a logit model on it.
  + Inject drift at various intensities.
  + Demonstrate model performance degradation.
  + Quantify drift using PE and show spike/drop that scales with drift amount.