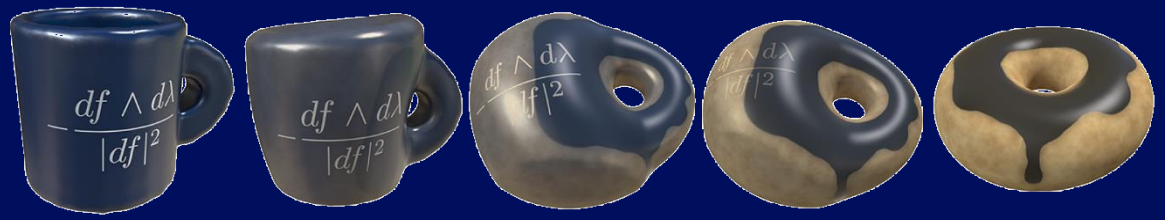
Topological Data Analysis of Decision Boundaries with Application to Model Selection

Karthikeyan Natesan Ramamurthy, Kush R. Varshney, and Krishnan Mody

IBM

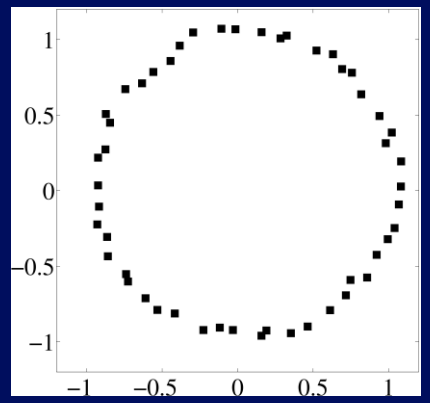
Topology

- Study of shape
  - Connected components and holes (cavities) of various dimensions
  - 
- Betti numbers denote the number of k-dimensional holes
  - Circle: $(\beta_0 = 1, \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \dots)$
  - Torus: $(\beta_0 = 1, \beta_1 = 2, \beta_2 = 1, \beta_3 = 0, \beta_4 = 0, \dots)$
- Statistics of Betti numbers characterize topological complexity

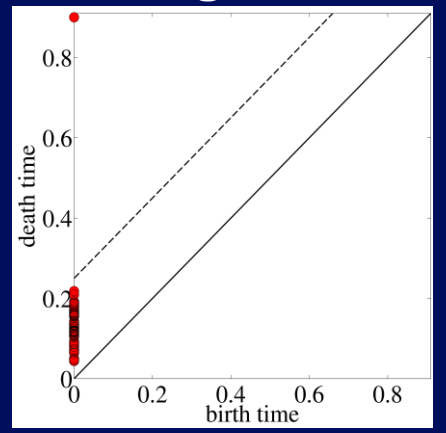
Topological Data Analysis

- Analyze shape of structures from which point clouds of data are sampled
- Persistent homology
  - Set of tools to recover Betti numbers from data point clouds
  - Construct simplicial complexes at different scales
  - See which ones persist across scales
  - Report Betti numbers of those persistent features

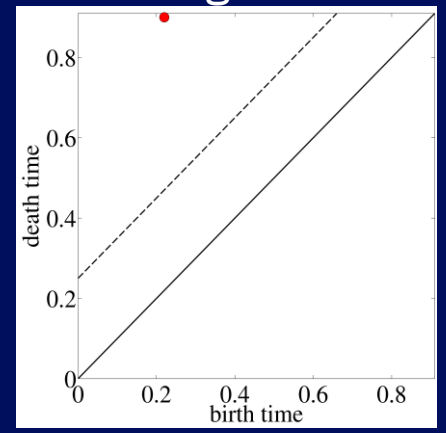
data points



β_0 persistence diagram

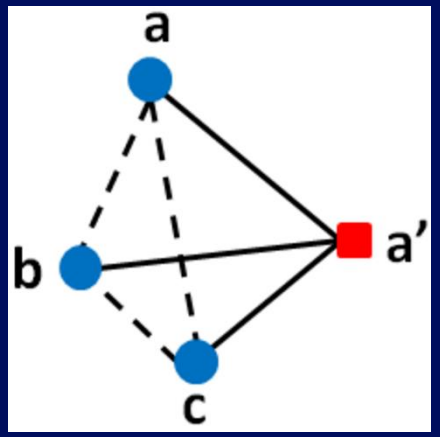


β_1 persistence diagram

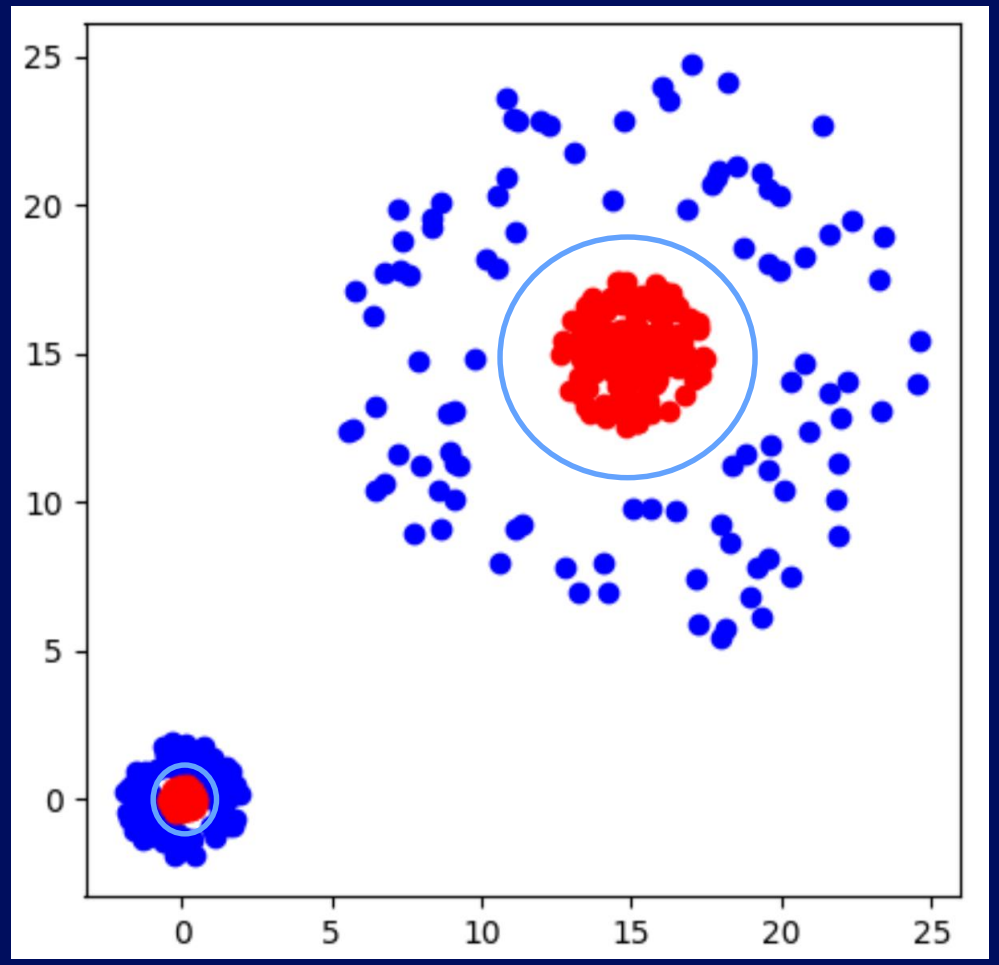


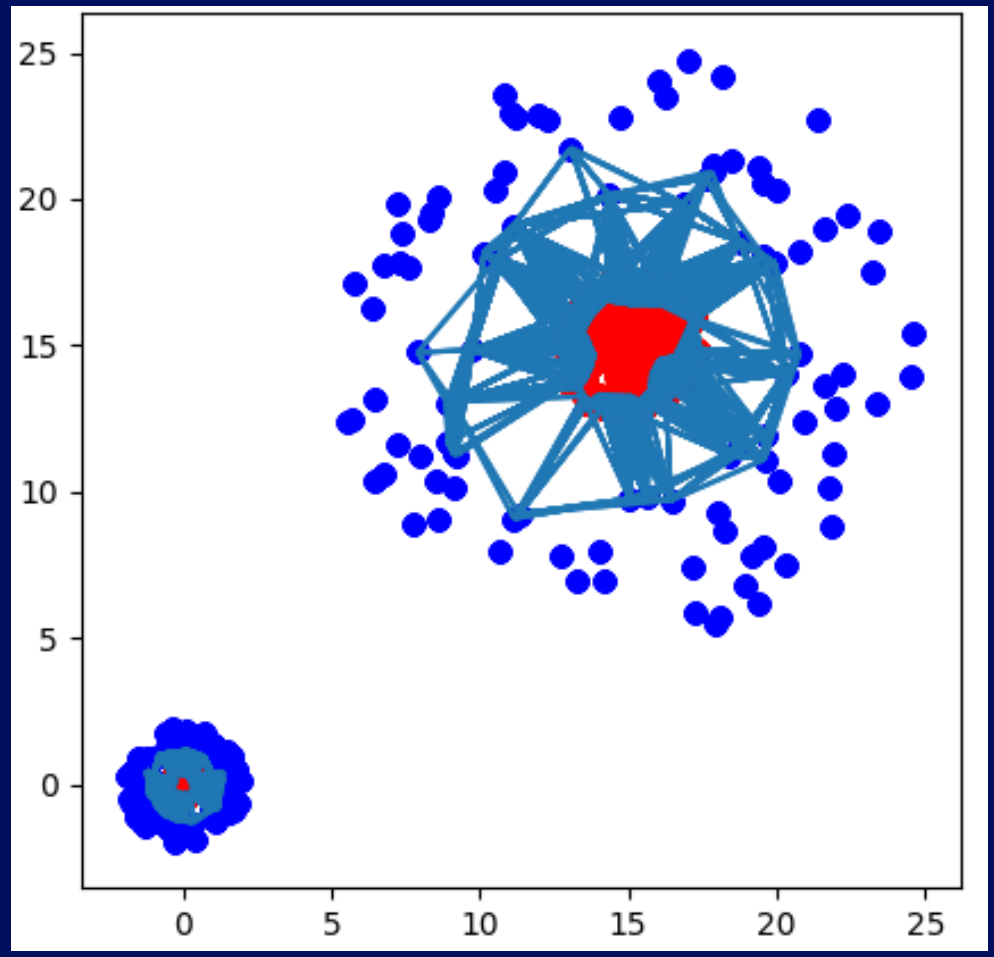
Great for Unlabeled Data, But What About Labeled Data?

- Labeled data is common in supervised classification problems
- The decision boundary is the most interesting structure in supervised classification problems
- In this work we propose to characterize the topology of the decision boundary by extending persistent homology
- Connect data points with opposite class labels
  - Complete simplices by length two graph walk
  - Labeled Vietoris-Rips complex



Local Scaling





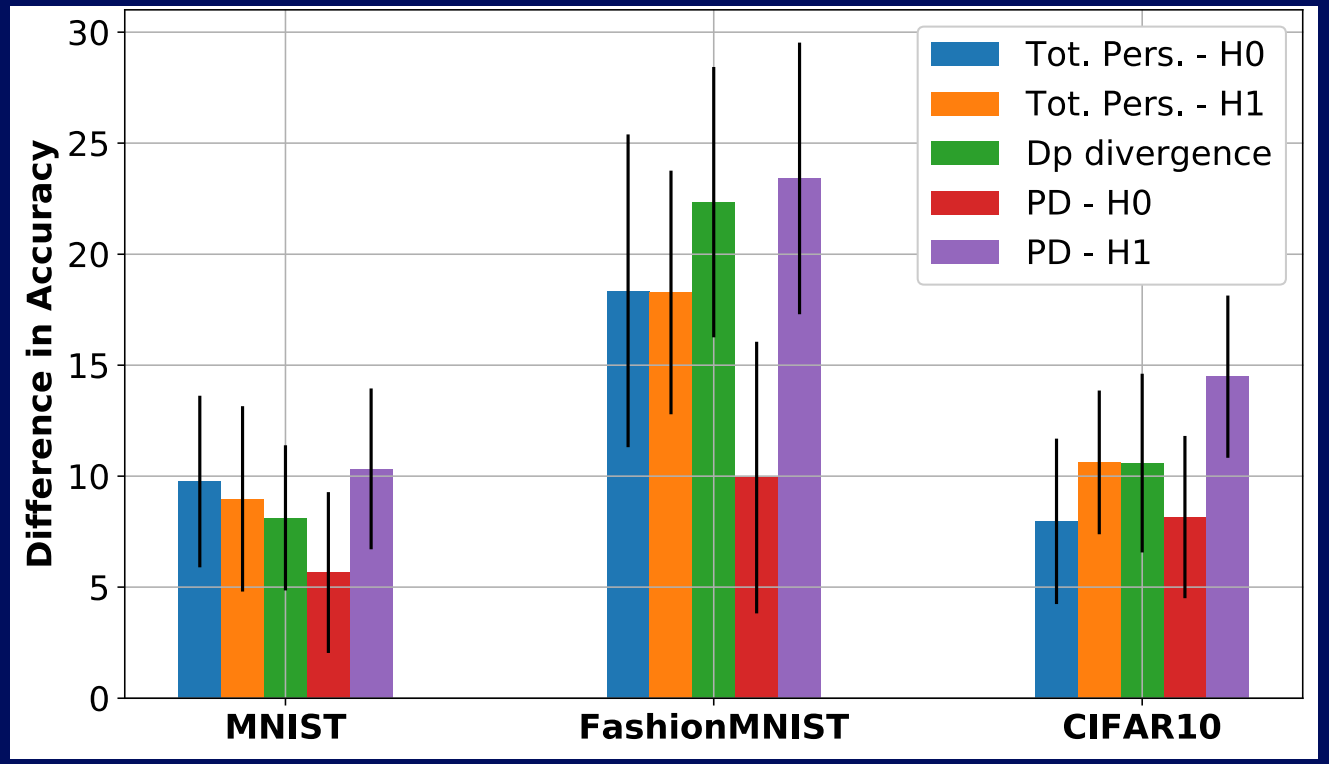
Application to Model Marketplaces

- Growing trend of model marketplaces to buy pre-trained models
- Vendor unwilling to let user download model for free trial
- Consumer unwilling to send their data
- Still need to do some matching to figure out which model is appropriate for which dataset

- Match on topological complexity

Results

- When choosing a pre-trained classifier for a novel dataset, one whose decision boundary topological complexity matches that of the dataset yields good generalization



What Else is in the Paper?

- Theoretical results on labeled Čech complexes

Come See Us This Evening

- Poster #124