Finding the Homology of Decision Boundaries with Active Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Accurately and efficiently characterizing the decision boundary of classifiers is important for problems related to model selection and meta-learning. Inspired by topological data analysis, the characterization of decision boundaries using their homology has recently emerged as a general and powerful tool. In this paper, we propose an active learning algorithm to recover the homology of decision boundaries. Our algorithm sequentially and adaptively selects which samples it requires the labels of. We theoretically analyze the proposed framework and show that the query complexity of our active learning algorithm depends naturally on the intrinsic complexity of the underlying manifold. We demonstrate the effectiveness of our framework in selecting best-performing machine learning models for datasets just using their respective homological summaries. Experiments on several standard datasets show the sample complexity improvement in recovering the homology and demonstrate the practical utility of the framework for model selection.

1 Introduction

Meta learning refers to a family of algorithms that assess the complexity of the target data to select an appropriate learning model for solving a problem of interest. In the context of classification problems, the complexity of the data can be characterized by understanding the geometry of the decision boundary; for example, by using topological data analysis (TDA) [1, 2, 3]. This scenario makes sense in settings where large corpora of labeled training data are available to recover the persistent homology of the decision boundary for use in downstream machine learning tasks [3, 4, 5, 6]. However the utility of this family of methods is limited in applications where labeled data is expensive to acquire.

In this paper, we explore the intersection of active learning and topological data analysis for the purposes of efficiently learning the persistent homology of the decision boundary in classification problems. In contrast to the standard paradigm, in active learning, the learner has access to unlabeled data and sequentially selects a set of points for an oracle to label. We propose an efficient active learning framework that adaptively select points for labeling near the decision boundary. A theoretical analysis of the algorithm results in an upper bound on the number of samples required to recover the recover the decision boundary homology. Naturally, this query complexity depends on the intrinsic complexity of the underlying manifold.

There have been several other studies that have explored the use of topological data analysis to characterize the decision boundary in classification problems. In [7], the authors use the persistent homology of the decision boundary to tune hyperparameters in kernel-based learning algorithms. They later extended this work and derived the conditions required to recover the homology of the decision from only samples [3]. Other works have explored the use of other topological features to characterize the difficulty of classification problems [8, 9, 5]. While all previous work assumes full knowledge of data labels, only samples near the decision boundary are used to construct topological features. We directly address this problem in our work by proposing an active approach that adaptively and sequentially labels only samples near the decision boundary, thereby resulting in significantly reduced query complexity. To the best of our knowledge, this is the first work that explores the intersection of active learning and topological data analysis.

Our main contributions are as follows:

- We introduce a new algorithm for actively selecting samples to label in service of finding the persistent homology of the decision boundary. We provide theoretical conditions on the query complexity that lead to the successful recovery of the decision boundary homology.
- We evaluate the proposed algorithm for active homology estimation using synthetic data and compare it's performance to a passive approach that samples data uniformly. In addition, we demonstrate the utility of our approach relative to a passive approach on a stylized model selection problem using real data.

2 Preliminaries

In this section, we define the decision boundary manifold and discuss the labeled Čech Complex [3] which can be used to estimate the homology of this manifold from labeled data. For more background and details, we direct the reader to the appendix.

2.1 The Decision Boundary Manifold and Data

Let $\mathcal{X}$ be a Euclidean space that denotes the domain/feature space of our learning problem and Let μ denote the standard Lebesgue measure on $\mathcal{X}$. While our theory and methods hold more generally, for the sake of clarity, we will restrict ourselves to the binary classification setting in the sequel and let $\mathcal{Y} = \{0, 1\}$ denote the label set. Let p_{XY} denote a joint distribution on $\mathcal{X} \times \mathcal{Y}$. Of particular interest to us in this paper is the so-called Bayes decision boundary $\mathcal{M} = \{\mathbf{x} \in \mathcal{X} | p_{Y|X}(1|\mathbf{x}) = p_{Y|X}(0|\mathbf{x})\}$. Indeed, identifying $\mathcal{M}$ is equivalent to being able to construct the provably optimal binary classifier called the Bayes optimal predictor:

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } p_{Y|X}(1 \mid \mathbf{x}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

Following along the lines of [3], the premise of this paper relies on supposing that the set $\mathcal{M}$ is in fact a reasonably well-behaved manifold[1]. That is, we will make the following assumption.

Assumption 1. *The decision boundary manifold $\mathcal{M}$ has a condition number $1/\tau$.*

The condition number $\frac{1}{\tau}$ is an intrinsic property of $\mathcal{M}$ (assumed to be a submanifold of $\mathbb{R}^N$) and encodes both the local and global curvature of the manifold. The value τ is the largest number such that the open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^N$ for every $r < \tau$. *E.g.*, in Figure 1, where $\mathcal{M}$ is a circle in $\mathbb{R}^2$, τ is its radius. We refer the reader to the appendix (or [11]) for a formal definition.

Now we will suppose that we have access to to N i.i.d samples $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \subset \mathcal{X}$ that are drawn according to the marginal distribution p_X. Notice that in the classical (passive) learning setting, one typically assumes access to N i.i.d samples from the joint distribution p_{XY}. Indeed the goal of this paper is to demonstrate that one may obtain labels for far fewer labels than N feature vectors while achieving similar performance to the passive learning setting if one is allowed to sequentially and adaptively choose the labels observed. Based on the observed data, we define the set $\mathcal{D}^0 = \{\mathbf{x} \in \mathcal{D} : f(\mathbf{x}) = 0\}$, that is the set of all samples with Bayes optimal label of 0; similarly, we let $\mathcal{D}^1 = \{\mathbf{x} \in \mathcal{D} : f(\mathbf{x}) = 1\}$.

2.2 The Labeled Čech Complex

[1] Note that it is conceivable that the decision boundary is not strictly a manifold. While this assumption is critical to the rest of this paper, it is possible to extend thge results here by following the theory in [10]. We will leave a thorough exploration of this for future work

As outlined in Section 1, our goal is to recover the homological summaries of $\mathcal{M}$ from data. Homological summaries such as Betti numbers estimate the number of connected components and the number of holes of various dimensions that are present in $\mathcal{M}$. Since we only have a sample of data points in practice, we first construct a simplicial complex from these points that mimics the shape of $\mathcal{M}$. We can then estimate the rank of any homology group H_i of dimension i from this complex. This rank is called the Betti number β_i and informally denotes the number of holes of dimension i in the complex. The multi-scale estimation of Betti numbers, which involves gradual "thickening" of the complex results in a persistence diagram PD_i. This encodes the *birth* and *death time* of the $i-$dimensional holes in the complex. For more background refer [12].

In the passive learning setting, the authors in [3] consider the same problem of estimating the homology of $\mathcal{M}$ and propose a new topological data analysis tool that is especially well suited called the Labeled Čech (LČ) Complex. For ϵ, let $B_\epsilon(\mathbf{x})$ denote a ball of radius ϵ around $\mathbf{x}$. We refer the reader to the appendix or [3] for the details of definition 1.

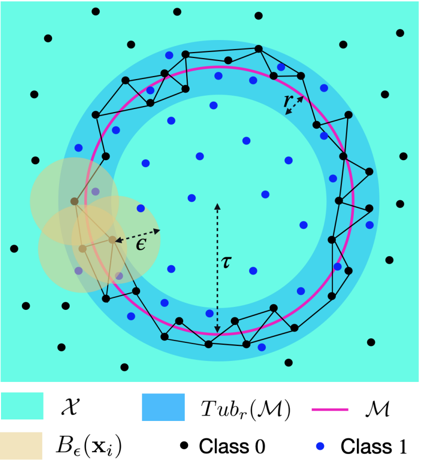


Figure 1: An example of labeled Čech complex constructed in a tubular neighborhood $\mathrm{Tub}_r(\mathcal{M})$ of radius r, for a manifold $\mathcal{M}$ of condition number $1/\tau$. The complex is constructed on samples in class 0, by placing balls of radius ϵ ($B_\epsilon(\mathbf{x}_i)$), and is "witnessed" by samples in class 1. $\mathcal{X}$ is the compact probability space for the data. Each triangle is assumed to be a $2-$simplex in the simplicial complex.

Definition 1. *Given $\epsilon, \gamma > 0$, an (ϵ, γ)-labeled Čech complex is a simplicial complex constructed from a collection of simplices such that each simplex σ is formed on the points in the set $\mathcal{D}^0$ witnessed by the reference set $\mathcal{D}^1$ satisfying the following conditions: **(a)** $\bigcap_{\mathbf{x}_i \in \sigma} B_\epsilon(\mathbf{x}_i) \neq \emptyset$, where $\mathbf{x}_i \in \mathcal{D}^0$ are the vertices of σ. **(b)** $\forall \mathbf{x}_i \in \sigma$, $\exists \mathbf{x}_j \in \mathcal{D}^1$ such that, $\|\mathbf{x}_i - \mathbf{x}_j\| \leq \gamma$.*

The set $\mathcal{D}^0$ is used to construct the LČ complex witnessed by the reference set $\mathcal{D}^1$. This allows us to infer that each vertex of the simplices σ are within distance γ to some point in $\mathcal{D}^1$. The authors in [3] show that, under certain assumption on the manifold and the distribution, provided sufficiently many random samples (and their labels) drawn according p_{XY}, the set $U = \bigcup_{\mathbf{x}_i \in \sigma} B_\epsilon(\mathbf{x}_i)$ forms a cover of $\mathcal{M}$ and deformation retracts to $\mathcal{M}$. Moreoever, the nerve of the covering is homotopy equivalent to $\mathcal{M}$. The assumptions under which the above result holds also turns out to be critical to achieve the results of this paper, and hence we will devote the rest of this section to elaborating on these.

Before stating our assumptions, we need a few more definitions. For the distribution p_{XY}, we will let

$$\mathfrak{D} = \{\mathbf{x} \in \mathcal{X} : p_{X|Y}(\mathbf{x} \mid 1) p_{X|Y}(\mathbf{x} \mid 0) > 0\}.$$

In other words, $\mathfrak{D}$ denotes the region of the feature space where both classes overlap, i.e., both class conditional distributions $p_{X|Y}(\cdot \mid 1)$ and $p_{X|Y}(\cdot \mid 0)$ have non-zero mass. For any $r > 0$, we let $\mathrm{Tub}_r(\mathcal{M})$ denote a "tubular" neighborhood of radius r around $\mathcal{M}$. Figure 1 illustrates this pictorially; we refer the reader to the appendix for a formal definition. The main assumption underlying the results of this paper (similar to those in [3, 11]) is as follows.

Assumption 1. *There is an $r > 0$ such that* **(a)** $\mathfrak{D} \in \mathrm{Tub}_r(\mathcal{M})$ *and* ***(b)*** $r < (\sqrt{9} - \sqrt{8})\tau$. *Further, we assume that **(c)** $\epsilon \in \left(\frac{(r+\tau)-\sqrt{r^2+\tau^2-6\tau r}}{2}, \frac{(r+\tau)+\sqrt{r^2+\tau^2-6\tau r}}{2}\right)$.*

Assumption 1 (a) implies that if one were to construct the LČ complex from points in the overlapping region $\mathfrak{D}$, this will be fully contained in $\mathrm{Tub}_r(\mathcal{M})$ provided γ is chosen appropriately. This also implies that $\mathcal{D}^0$ is fully contained in $\mathrm{Tub}_r(\mathcal{M})$, and the reference set $\mathcal{D}^1$ is contained in $\mathrm{Tub}_{r+\gamma}(\mathcal{M})$. Also the upper bound on r is a constant fraction of τ, and at this upper bound, ϵ will be $\frac{r+\tau}{2}$. In the stylized example in Figure 1 where $\mathcal{M}$ is a circle, $\mathrm{Tub}_r(\mathcal{M})$ is an annulus and the radius ϵ of the covering ball $B_\epsilon(\mathbf{x})$ is constrained by τ.

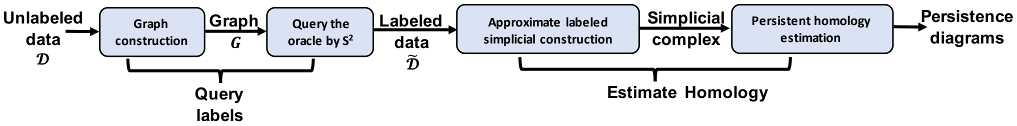

Unlabeled data $\mathcal{D}$
Graph construction
Graph G
Query the oracle by S^2
Labeled data $\tilde{\mathcal{D}}$
Approximate labeled simplicial construction
Simplicial complex
Persistent homology estimation
Persistence diagrams
Query labels
Estimate Homology

Figure 2: The proposed active learning framework for finding the homology of decision boundaries.

3 Active Learning for Finding the Homology of Decision Boundaries

As the definitions above and results from [3] make clear, constructing a useful LČ complex requires sampling both class-conditional distributions in the region around the decision boundary to a sufficient resolution. The key insight of our paper is to devise a framework based on active learning that sequentially and adaptively decides where to obtain data and therefore query-efficiently samples points near the decision boundary. In what follows, we will provide a brief description of our algorithm, and then we will establish rigorous theoretical guarantees on the query complexity of the proposed algorithm.

3.1 The Active Learning Algorithm

A schematic diagram of the proposed active learning framework is presented in Figure 2. As illustrated, the framework takes as input an unlabeled dataset $\mathcal{D}$, and this dataset is used to generate an appropriate graph on the data. This graph is then used to iteratively query labels near the decision boundary. The subset of labeled samples are used to estimate the homology, resulting in the persistence diagram of the LČ complex. We briefly outline the label query and homology estimation phases below, and refer the reader to the supplementary material for the full details.

Label query phase: The label query phase starts with constructing a graph $G = (\mathcal{D}, E)$ from the unlabeled dataset $\mathcal{D}$. While other choices are possible, we will suppose that the graph we construct is either a k-radius near neighbor or a k-nearest neighbors graph[2]. After graph construction, a graph-based active learning algorithm (S^2) path [13] accepts the graph $G = (\mathcal{D}, E)$ and selects the data points whose labels it would like to see. This selection is based on the structure of the graph and all previous gathered labels. Specifically, S^2 continually queries for the label of the vertex that bisects the shortest shortest path between any pair of oppositely labeled vertices. The authors in [13] show that S^2 provably query efficiently locates the cut-set in this graph (i.e., the edges of the graph that have oppositely labeled vertices). As a result, the query phase outputs a set $\tilde{\mathcal{D}}$ associated with the labels that is near the decision boundary.

Homology estimation phase: Entering the homology estimation stage, we construct an approximation of the LČ complex from the query set $\tilde{\mathcal{D}}$. Specifically, we construct the locally scaled labeled Vietoris-Rips (LS-LVR) complex introduced in [3]. Sticking to the pre-defined dataset $\mathcal{D}$ as an example, there are two steps to construct the LS-LVR complex: (1) Induce edges from $\mathcal{D}$ to generate an edge set $E \subseteq \{\{\mathbf{x}_i, \mathbf{x}_j\} | (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D}^2 \wedge y_i \neq y_j \wedge \|\mathbf{x}_i - \mathbf{x}_j\| \leq \kappa\sqrt{\rho_i \rho_j}\}$. Here, κ is a scale parameter, ρ_i is the smallest radius of a sphere centered at $\mathbf{x}_i$ to enclose k-nearest opposite class neighbors and ρ_j has a similar definition. Apparently, we have generated a bipartite graph where every edges connect points in the opposing classes; (2) Connect all 2-hop neighbors to build a simplicial complex. Varying scale parameter κ produces a filtration of the simplicial complex and generates the persistent homology such as the persistent diagrams.

3.2 Theoretical results

Let $G = (\mathcal{D}, E)$ denote a k-radius neighbor/$k-$nearest neighbors graph constructed from the dataset $\mathcal{D}$. This allows us to define the cut-set $C = \{(\mathbf{x}_i, \mathbf{x}_j) | y_i \neq y_j \wedge (\mathbf{x}_i, \mathbf{x}_j) \in E\}$ and cut-boundary $\partial C = \{\mathbf{x} \in V : \exists e \in C \text{ with } \mathbf{x} \in e\}$. Recalling the Assumption 1(a), we immediately have $\partial C \subseteq \mathrm{Tub}_r(\mathcal{M})$ for an appropriately chosen r. The key intuition behind our approach is that this implies that S^2 is naturally turned to focusing the labels acquired within $\mathrm{Tub}_r(\mathcal{M})$. As we show below, this is done in a remarkably query efficient manner, and furthermore, when labeled data obtains thus is used to construct an LČ complex, this allows us to find the homology of the manifold $\mathcal{M}$. We

[2] The k-radius near neighbor graph connects all pairs of vertices that are a distance of at most k away, and the k-nearest neighbor graph connects a vertex to its k nearest neighbors

begin by sketching a structural lemma about the graph G and refer the reader to the supplementary materials for a full statement and proof.

Lemma 1 (sketch)**.** *Suppose $\mathcal{D}^0$ and $\mathcal{D}^1$ are r-dense in $\mathcal{M}$, then the graph $G = (\mathcal{D}, E)$ constructed from $\mathcal{D}$ is such that $\mathcal{D}^0 \bigcap \partial C$ and $\mathcal{D}^1 \bigcap \partial C$ are both r-dense in $\mathcal{M}$ for an appropriate choice of k.*

Lemma 1 tells us that an appropriate constructed graph induces a boundary ∂C that contains sufficient examples covering $\mathcal{M}$. We next need some structural assumptions about the manifold $\mathcal{M}$.

Assumption 2. *For the r from Assumption 1, we have: **(a)** $\inf_{\mathbf{x}\in\mathcal{M}} \mu_{\mathcal{X}|y}(B_{r/2}(\mathbf{x})) > k^y_{r/2}, y \in \{0, 1\}$. **(b)** $\sup_{\mathbf{x}\in\mathcal{M}} \mu_{\mathcal{X}}(B_{r/2}(\mathbf{x})) < h_{r/2}$ **(c)** $\mu_{\mathcal{X}}(\mathrm{Tub}_r(\mathcal{M})) \leq N_{r/2}h_{r/2}$.*

The assumption 2(a) ensures sufficient mass in both class. The assumption 2(b)(c) upper-bounds the measure of $\mathrm{Tub}_r(\mathcal{M})$. Recall that $G = (\mathcal{D}, E)$ in assumption 1 is a labeled graph, we further write β to denote the proportion of the smallest connected component with all the examples identically labeled.

Theorem 1. *Let $N_{r/2}$ be the covering number of the manifold $\mathcal{M}$. Under Assumptions 1 and 2, for any $\delta > 0$, we have that the $(\epsilon, 2r)$-LČ complex estimated by our framework is homotopy equivalent to $\mathcal{M}$ with probability at least $1 - \delta$ provided*

$$|\tilde{\mathcal{D}}| > \frac{\log\left\{1/\left[\beta\left(1-\sqrt{1-\delta}\right)\right]\right\}}{\log\left[1/(1-\beta)\right]} + |\mathcal{D}|N_{r/2}h_{r/2}(\lceil log_2|\mathcal{D}|\rceil + 1) \tag{2}$$

where

$$\begin{aligned}|\mathcal{D}| = \max\Bigg\{&\frac{1}{P(y=0)k^0_{r/2}}\left[\log\left(2N_{r/2}\right) + \log\left(\frac{1}{(1-\sqrt{1-\delta})}\right)\right],\\ &\frac{1}{P(y=1)k^1_{r/2}}\left[\log\left(2N_{r/2}\right) + \log\left(\frac{1}{(1-\sqrt{1-\delta})}\right)\right]\Bigg\}\end{aligned}. \tag{3}$$

Remark 1. Theorem 1 demonstrates that our active learning framework has a query complexity of $\mathcal{O}(NN_{r/2}h_{r/2}log_2N)$. That is, after $\mathcal{O}(NN_{r/2}h_{r/2}log_2N)$ queries at most, a $(\epsilon, 2r) - L\check{C}$ complex constructed from the queried examples will be homotopy equivalent to $\mathcal{M}$ with high probability. Notice that the intrinsic complexity of the manifold naturally plays a significant role, and the less complex the manifold the more significant gains the active learning framework has over its passive counterpart (cf. eq. 3). In the supplementary material, we also provide a simple and concrete example that numerically shows the superiority of the sample complexity of our proposed framework with respect to its passive counterpart.

Remark 2. The results of Theorem 1 can be improved by carrying out a more intricate analysis of the active learning algorithm as in [13]. Indeed, one may also replace the S^2 algorithm in our framework with a different graph-based active learning algorithm seamlessly to leverage the properties of that algorithm for active homology estimation of decision boundaries. These, and the relaxation of Assumption 1 (a), are promising avenues for future work.

We provide a complete proof of Theorem 1 in the supplementary material. However, we will provide some intuition about the operation of our algorithm, and hence to the proof of the theorem here. The S^2 algorithm is split into two phases: uniform sampling and path bisection. The uniform sampling serves to finding a path connecting vertices of opposite labels. The path bisection phase queries at the mid-point of the shortest path that connects oppositely labeled vertices in the underlying graph. As the authors in [13] show, this endows S^2 with the ability to quickly narrow in on the cut-boundary ∂C. The uniform sampling phase accounts for the first term in eq. 2, which guarantees that there are sufficient paths to identify ∂C completely. In the path bisection phase, we take $(\lceil \log_2 |\mathcal{D}|\rceil + 1)$ (this may be tightened using the techniques in [13]) queries at most to find the end point of the cut-edge inside a path; this needs to be done at most $|\partial C|$ to complete the querying phase. Next, Assumption 1(a) guarantees that $\partial C \subseteq \mathrm{Tub}_r(\mathcal{M})$. Therefore, we may use the measure $N_{r/2}h_{r/2}$ from Assumption 2(b)(c) to upper-bound $|\partial C|$ which results in the second term of eq. 2. This naturally ties in the query complexity to the manifold complexity via $N_{r/2}$ and $\mathrm{Tub}_r(\mathcal{M})$. Eq. 3 comes from the necessary condition for the LČ complex being homotopy equivalent to $\mathcal{M}$, following along the lines of [3].

4 Experimental Results

We compare the homological properties estimated from our active learning algorithm to a passive learning approach on both synthetic data and real data. In the experiments we use the characteristics of homology group of dimension 1 (β_1, PD_1). We chose to use dimension 1 since [3] shows that this provides the best topological summaries for applications related to model selection. Using the synthetic data, we study the sample complexity of active learning by examining the homological summaries β_1 and PD_1. For real data, we estimate PD_1 of the Banknote, MNIST and CIFAR10 and then utilize PD_1 to do model selection from several families of classifiers.

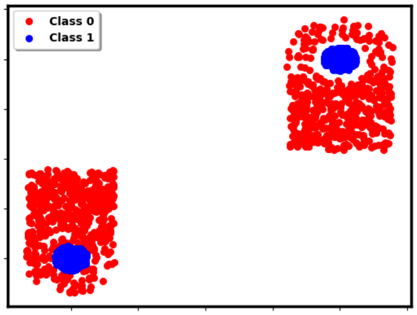


Figure 3: Visualization of the synthetic data.

4.1 Experiments on Synthetic Data

The synthetic data in Figure 3 has decision boundaries that are homeomorphic to two disjoint circles. This dataset has 2000 examples. Please see the appendix for more details. Clearly from Figure 3, β_1 of the decision boundary is two.

Per the first step of our active learning algorithm, we construct a k-radius NN graph with $k = 0.65$. The scale parameter is set assuming we have full knowledge of the decision boundary manifold. Subsequently, we use S^2 to query the labels of examples on the created graph. After the label query phase, we construct the LS-LVR complex with the queried samples and compute β_1 and PD_1 using the Ripser package [14] and its interface [15]. For the passive learning baseline, we uniformly query the examples with all other aspects of the experiment remaining identical to the active case. We also compute β_1 and PD_1 from the complete dataset and consider them as the “ground-truth" homology summaries. We evaluate the similarities between the estimated homology summaries and the ground-truth homology summaries to show the effectiveness of our active learning framework. We compare the bottleneck distance [16, 17] between the ground-truth and estimated values of PD_1 for different percent of data labeled. These results are shown on Figure 4. As is clear from the figure, the bottleneck distance for our active learning framework decreases faster than the passive learning approach and perfectly recovers the homology with only 50% of data. A visualization of the query process is shown on Figure 5. As expected, the active learning framework selects more examples to query near the decision. Please refer to the appendix to evaluate the performance of the active learning framework for different k-radius NN graphs and β_1 recovery.

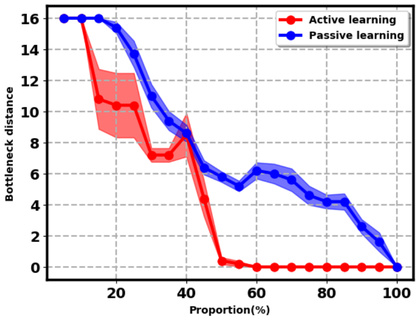


Figure 4: Bottleneck distance from ground-truth PD_1 by passive learning and active learning.

4.2 Experiments on Real Data

To demonstrate the effectiveness of our active learning framework on real data, we consider the classifier selection problem discussed in [3]. A bank of pretrained classifiers is accessible in the marketplace and customers will select a proper one without changing the hyperparameters of the classifiers. We consider two selection strategies: First, a classifier with the smallest bottleneck distance from the PD_1 of queried data is selected; second, we

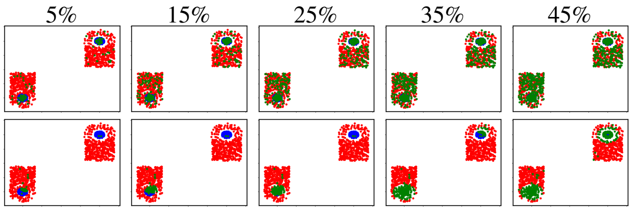


Figure 5: Visualization of the query process by passive learning (top row) and our active learning framework (bottom row). More examples (highlighted by green) near the decision boundaries are selected to query in the proposed framework.

Banknote	KNN	SVM	Neural network	Decision tree
Passive	0.1072±0.0000	0.3753±0.0005	0.4316±0.0000	0.1997±0.0000
Active^1	0.0783±0.0014	0.3231±0.0012	0.4316±0.0000	0.1901±0.0004
Active^2	0.1017±0.0001	0.3431±0.0012	0.3730±0.0138	0.1744±0.0026
Active^3	**0.0346±0.0013**	**0.0836±0.0133**	**0.1058±0.0265**	**0.1613±0.0004**
Passive + Validation	0.0176±0.0000	**0.0259±0.0000**	0.0068±0.0000	0.0741±0.0000
Active^1 + Validation	0.0173±0.0000	**0.0259±0.0000**	**0.0039±0.0000**	**0.0731±0.0000**
Active^2 + Validation	**0.0149±0.0000**	**0.0259±0.0000**	0.0134±0.0001	**0.0731±0.0000**
Active^3 + Validation	**0.0149±0.0000**	**0.0259±0.0000**	0.0072±0.0000	0.0770±0.0000

MNIST	KNN	SVM	Neural network	Decision tree
Passive	0.0129±0.0000	**0.0141±0.0000**	0.0202±0.0000	**0.0332±0.0000**
Active^1	0.0128±0.0000	0.0161±0.0001	**0.0150±0.0000**	0.0388±0.0001
Active^2	0.0122±0.0000	0.0162±0.0001	0.0177±0.0000	**0.0332±0.0000**
Active^3	**0.0104±0.0000**	0.0156±0.0001	0.0388±0.0020	**0.0332±0.0000**
Passive + Validation	0.0119 ±0.0000	0.0124±0.0000	**0.0104±0.0000**	0.0290±0.0000
Active^1 + Validation	0.0123±0.0000	**0.0119±0.0000**	**0.0104±0.0000**	0.0284±0.0000
Active^2 + Validation	0.0108±0.0000	**0.0119±0.0000**	0.0125±0.0000	0.0284±0.0000
Active^3 + Validation	**0.0104±0.0000**	**0.0119±0.0000**	0.0127±0.0000	**0.0274±0.0000**

CIFAR10	KNN	SVM	Neural network	Decision tree
Passive	**0.3065±0.0002**	0.4683±0.0000	0.3185±0.0000	**0.3625±0.0000**
Active^1	0.3201±0.0000	0.4591±0.0005	**0.3058±0.0006**	**0.3625±0.0000**
Active^2	0.3095±0.0001	**0.4007±0.0038**	**0.3058±0.0006**	**0.3625±0.0000**
Active^3	0.3109±0.0001	0.4464 ±0.0005	0.3185±0.0000	**0.3625 ±0.0000**
Passive + Validation	0.2987 ±0.0001	0.2698±0.0000	0.2651±0.0001	**0.3137±0.0002**
Active^1 + Validation	0.2911 ±0.0001	0.2797±0.0000	**0.2558±0.0000**	0.3146±0.0000
Active^2 + Validation	0.2987±0.0001	0.2864±0.0003	0.2649±0.0001	0.3214±0.0005
Active^3 + Validation	**0.2935±0.0000**	**0.2665±0.0000**	0.2615±0.0001	0.3221±0.0004

Table 1: Average validation error (five trials) on banknote, MNIST and CIFAR10 for the model selected with 15% pool data. Passive/Active stands for the classifiers picked by the estimated homology similarities. Passive/Active + Validation stands for the classifiers picked by the ensemble of estimated homology similarities and validation error. The subscript 1, 2 and 3 of the active learning indicates the used 3NN, 5NN and 7NN graphs. Best performance in the non-ensemble classifier selection and ensemble classifier selection are boldfaced.

further select an additional classifier based on the validation error computed from the queried data. We select between the two classifiers based on which results in the lowest error on the validation data - this is similar to an ensemble classifier selection problem similar to that defined in [18].

We split the real data to a validation set, a test set and lastly a pool of unlabeled data. The training set is used to generate four different banks of classifiers: KNN with nearest neighbors number ranging from 1 to 29, SVM with polynomial kernel function degree ranging from 1 to 14, decision tree with maximum depth ranging from 1 to 27, and neural networks with layers number ranging from 1 to 6. The test set is used to evaluate the test error of each classifier. The data pool is prepared for examples query.

We use the proposed active learning framework to estimate the homological properties of the queried data: constructing a k-nearest neighbors graph, query examples by S^2 and computing the PD_1 with the queried examples. We set k =3, 5, and 7. For passive learning, we keep all the operations same as the active learning framework except the queried examples are collected by uniform random sampling. To compute the PD of the decision boundary of the the classifier, we simply use the validation set input and the classifier output. Having estimated the homological summaries from the queried data and the classifiers, we compute the bottleneck distance between the PD_1 of the queried data and the classifiers. For the non-ensemble method, we simply select the classifier with the smallest bottleneck distance. For the ensemble method, we further include an additional classifier selected based on the validation error computed from the queried data. This results in two candidate classifiers and we eventually output the one with the smallest validation error.

We implement the above procedure in the datasets Banknote [19], MNIST [20] and CIFAR10 [21]. Banknote contains 1372 instances in two classes with four input features for a binary classification task. We randomly sample 100 examples as the training set and use the rest data as both the validation set and data pool. For the MNIST and the CIFAR10 datasets, we create a digit 1 vs. 8 classification task from the MNIST and an automobile vs. ship classification task from the CIFAR10. We randomly sample the data to create a training set with the size of 200, a validation set with the size of 2000, and a data pool with the size of 2000.

Table 1 indicates the validation error on banknote, MNIST and CIFAR10 for the classifier selected using 15% pool data. As we observe, the classifiers selected by our proposed active learning framework generally has a lower validation error rate than the passive learning, especially in an ensemble classifier selection framework. The validation error and the homological properties are both computed from the queried data. So, the performance difference between passive ensemble and active ensemble learning is related only to the efficient query. In addition, ensembling homology matching with validation-based model selection works better than selection with just matching homologies.

Figure 6 indicates the performance of classifiers selected by homology similarities at the cost of different proportions of the data pool. As expected, the proposed active learning framework achieves the best model selection faster than the passive learning for all the classifiers families. Note that the selection performance may be unstable with the increasing number of the queries, since active learning exhausts the informative examples rapidly and begin to query noisy examples. In summary, Table 1 and Figure 6 implies the advantage of active learning in finding good homology summaries is transmitted to model selection generating a better performance than the passive learning.

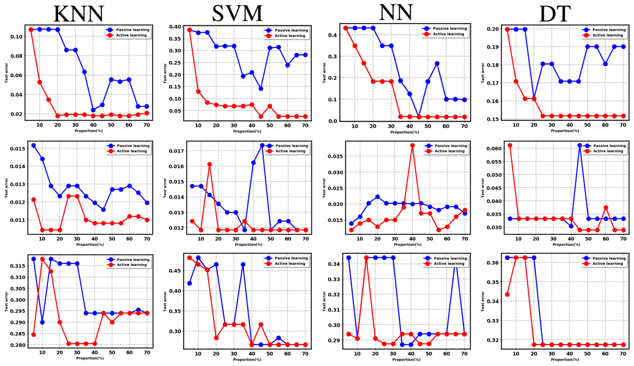
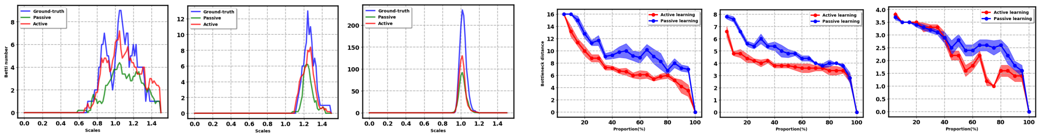


Figure 6: Validation errors as a function of proportions of queried data on banknote (top), MNIST (middle) and CIFAR10 (bottom) for model selection from different families of classifiers. Nonensemble selection by homological similarities.

4.3 Homological Properties Analysis on Real Data

We present the homological properties estimated by the passive learning and the proposed active learning framework. As we observe in the Figure 7(a), β_1 estimated by our active learning algorithm has a more similar trend to ground-truth in all three real datasets. Furthermore, CIFAR10 has a significantly higher β_1 than MNIST and Banknote datasets indicating more complex decision boundaries. This is consistent with the Table 1 which shows CIFAR10 binary classification task is more difficult than the other two tasks of Banknote and MNIST. The bottom row of Figure 7 shows the bottleneck distance between the estimated PD_1 and the ground-truth PD_1 at the cost of different proportion of data pool. Similar to the experiments in the synthetic dataset, we consider the PD constructed from the whole data pool as the ground-truth. We observe that the proposed active learning algorithm maintains a smaller bottleneck distance on early stages of query. Such benefits gradually diminish along with the increasing data proportion used for query.



(a) β_1 estimation comparisons at the cost of 50% data proportion

(b) Bottleneck distance at the cost of different proportions of data pool.

Figure 7: Homological properties for the banknote (left), MNIST (middle) and CIFAR10 (right).

5 Conclusions

We propose an active learning algorithm to find the homology of decision boundaries. We theoretically analyze the query complexity of the proposed algorithm and prove conditions necessary to recover the homology of decision boundaries. The extensive experiments on synthetic and real datasets with the application on model selection corroborate our theoretical results.

Broader Impact

The proposed approach, although has strong algorithmic and theoretical merits, has potential real-world application as we demonstrated.

One of the key uses of this approach is to create efficient summaries of decision boundaries of datasets [22] and models. Such summaries can be quite useful in applications like AI model marketplaces [23], where data and models can be securely matched without revealing too much information about each other. This is helpful in scenarios where the data is private and models are proprietary or sensitive.

A downside of being able to compute homology of decision boundaries with few examples is that malicious users may be able to learn about the key geometric / topological properties of the models with fewer examples than they would use otherwise. While this in itself may be benign, combined with other methods, they may be able to design better adversarial attacks on this model for instance. Ways of mitigating it in sensitive scenarios include ensuring that users do not issue too many queries of examples close to the boundary successively, since this may be revealing of malicious intent.

References

[1] G. Kusano, Y. Hiraoka, and K. Fukumizu, “Persistence weighted gaussian kernel for topological data analysis,” in *International Conference on Machine Learning*, 2016, pp. 2004–2013.

[2] C. Chen, X. Ni, Q. Bai, and Y. Wang, “A topological regularizer for classifiers via persistent homology,” *arXiv preprint arXiv:1806.10714*, 2018.

[3] K. N. Ramamurthy, K. Varshney, and K. Mody, “Topological data analysis of decision boundaries with application to model selection,” vol. 97, pp. 5351–5360, 09–15 Jun 2019. [Online]. Available: http://proceedings.mlr.press/v97/ramamurthy19a.html

[4] B. Rieck, C. Bock, and K. Borgwardt, “A persistent weisfeiler-lehman procedure for graph classification,” in *International Conference on Machine Learning*, 2019, pp. 5448–5458.

[5] W. H. Guss and R. Salakhutdinov, “On characterizing the capacity of neural networks using algebraic topology,” *arXiv preprint arXiv:1802.04443*, 2018.

[6] B. Rieck, M. Togninalli, C. Bock, M. Moor, M. Horn, T. Gumbsch, and K. Borgwardt, “Neural persistence: A complexity measure for deep neural networks using algebraic topology,” *arXiv preprint arXiv:1812.09764*, 2018.

[7] K. R. Varshney and K. N. Ramamurthy, “Persistent topology of decision boundaries,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 3931–3935.

[8] T. K. Ho, M. Basu, and M. H. C. Law, “Measures of geometrical complexity in classification problems,” in *Data complexity in pattern recognition*. Springer, 2006, pp. 1–23.

[9] C. Hofer, R. Kwitt, M. Niethammer, and A. Uhl, “Deep learning with topological signatures,” in *Advances in Neural Information Processing Systems*, 2017, pp. 1634–1644.

[10] J. Kim, J. Shin, F. Chazal, A. Rinaldo, and L. Wasserman, “Homotopy reconstruction via the cech complex and the vietoris-rips complex,” in *The 36th International Symposium on Computational Geometry (SoCG 2020)*, 2020.

[11] P. Niyogi, S. Smale, and S. Weinberger, “Finding the homology of submanifolds with high confidence from random samples,” *Discrete & Computational Geometry*, vol. 39, no. 1-3, pp. 419–441, 2008.

[12] H. Edelsbrunner and J. Harer, “Persistent homology - a survey,” *Contemporary mathematics*, vol. 453, pp. 257–282, 2008.

[13] G. Dasarathy, R. Nowak, and X. Zhu, “S2: An efficient graph based active learning algorithm with application to nonparametric classification,” in *Conference on Learning Theory*, 2015, pp. 503–522.

[14] U. Bauer, “Ripser: efficient computation of vietoris-rips persistence barcodes,” Aug. 2019, preprint.

[15] N. Saul and C. Tralie, “Scikit-tda: Topological data analysis for python,” 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2533369

[16] A. Efrat, A. Itai, and M. J. Katz, “Geometry helps in bottleneck matching and related problems,” *Algorithmica*, vol. 31, no. 1, pp. 1–28, 2001.

[17] M. Kerber, D. Morozov, and A. Nigmetov, “Geometry helps to compare persistence diagrams,” *Journal of Experimental Algorithmics (JEA)*, vol. 22, pp. 1–20, 2017.

[18] R. M. Cruz, R. Sabourin, G. D. Cavalcanti, and T. I. Ren, “Meta-des: A dynamic ensemble selection framework using meta-learning,” *Pattern recognition*, vol. 48, no. 5, pp. 1925–1935, 2015.

[19] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.

[22] H. A. Edwards and A. J. Storkey, “Towards a neural statistician,” *ArXiv*, vol. abs/1606.02185, 2016.

[23] A. Bridgwater, “Enough Training, Let’s Get Down To The AI Supermarket,” *Forbes*, Sep 2018. [Online]. Available: https://www.forbes.com/sites/adrianbridgwater/2018/09/18/enough-training-lets-get-down-to-the-ai-supermarket/#5f13cdbc10c3