Semi-Supervised Machine Learning: A Topological Approach

Adrián Inés, César Domínguez, Jónathan Heras, Gadea Mata and Julio Rubio

Departamento de Matemáticas y Computación
Universidad de La Rioja

UNIVERSIDAD DE LA RIOJA

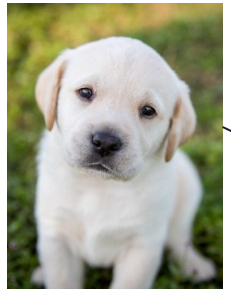
Introduction

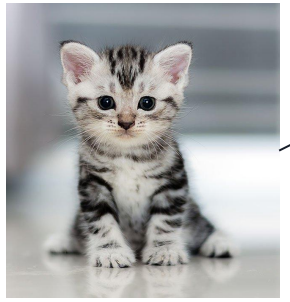
Problems

Amount of data

Amount of resources

Supervised learning

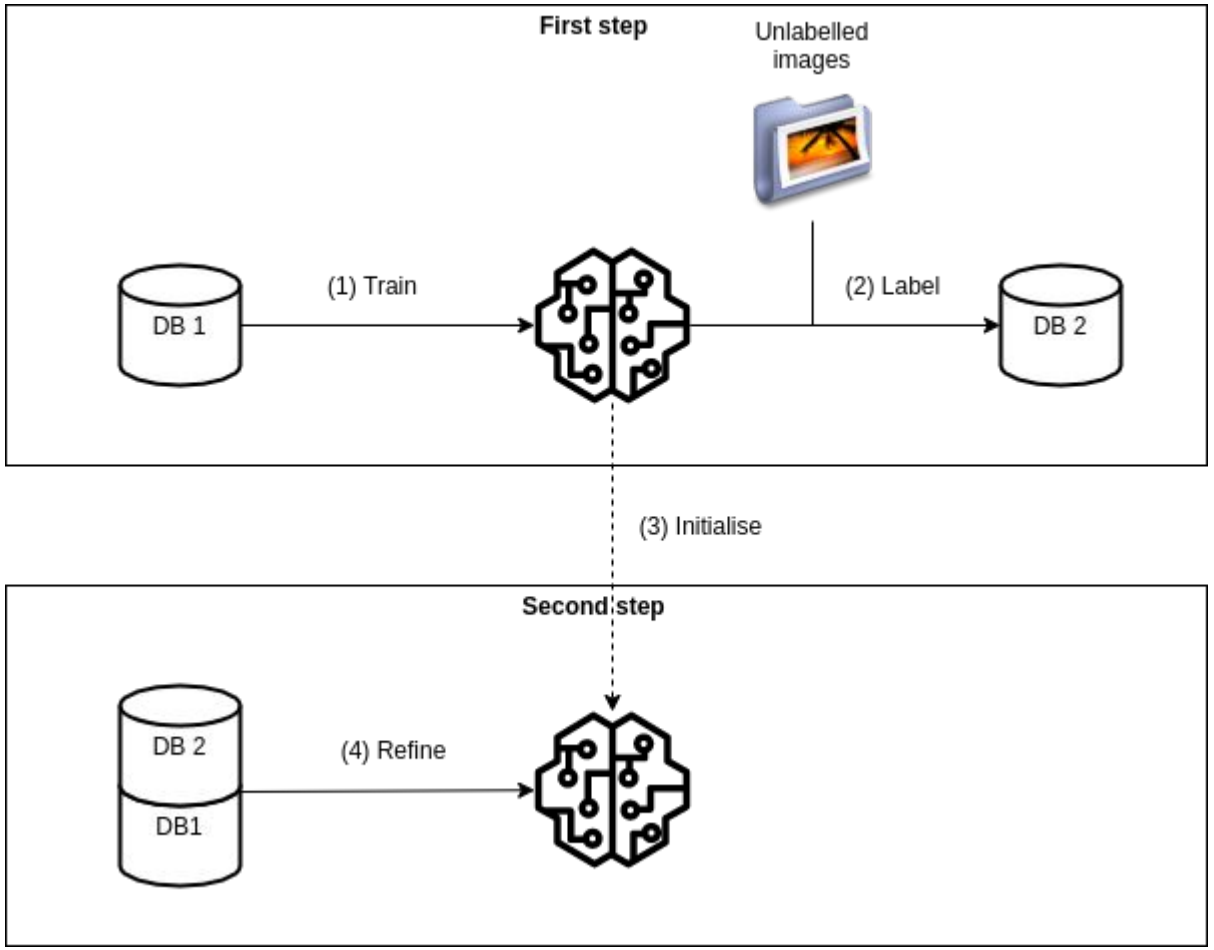




Labels

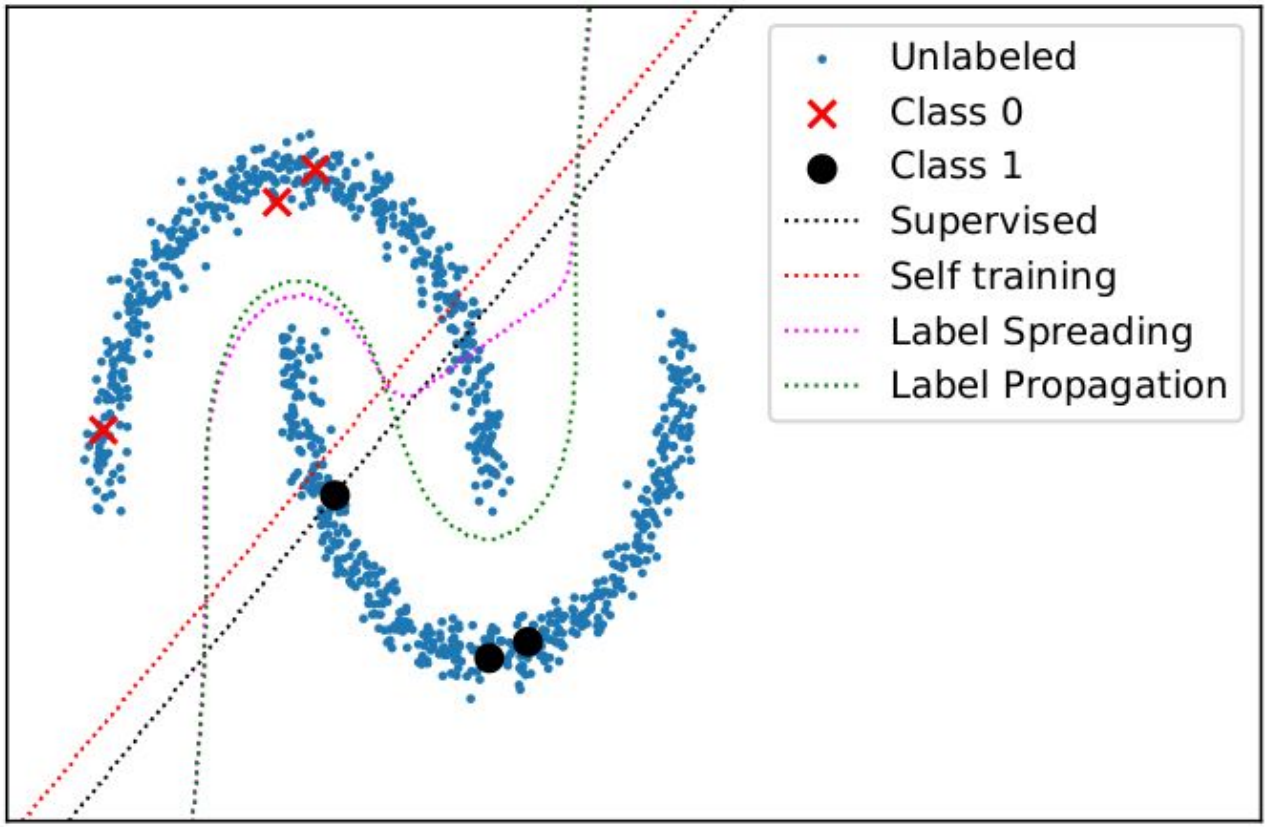
- Lion
- Cat
- Dog
- Bear
- Bird
- Elephant

Semi-Supervised Learning



Introduction

Traditional semi-supervised learning methods are based on the distance among the points.



Introduction

Topological Data Analysis (TDA) is a field devoted to extract topological and geometrical information from data.

Our inspiration comes from the Manifold Hypothesis that explores when high dimensional data could tend to lie in low dimensional manifolds.

Our method works under the hypothesis that each class in the dataset lies on a manifold.

Background

Background

Definition 2 (Vietoris-Rips complex) *Let (M, d) be a finite metric space. For every $\epsilon > 0$, the Vietoris-Rips complex VR_ϵ is defined as follows:*

$$VR_\epsilon(M) = \{\sigma \subseteq M \mid \forall u, v \in \sigma : d(u, v) \leq \epsilon\}$$

We can notice that in the previous definition we do not have a single simplicial complex, but rather we have a set of simplicial complexes that depend on ϵ, a value called the radius. Such a sequence of simplicial complexes is a called a filtration.

Background

Definition 3 (Persistence diagram) *Let (M, d) be a finite metric space, $\{\epsilon_0, \epsilon_1, \ldots \epsilon_n\}$ be real numbers that verify $0 \leq \epsilon_0 < \epsilon_1 < \ldots \epsilon_n < \infty$ and*

$$VR_{\epsilon_0} \subset VR_{\epsilon_1} \subset \cdots \subset VR_{\epsilon_{n-1}} \subset VR_{\epsilon_n}$$

be the Vietoris-Rips filtration of M. Then we define the persistence diagram of M as

$$X = \{a_1, \ldots, a_m\}$$

where $a_i = (\epsilon_r, \epsilon_s)$ is the pair birth and death of a feature.

Background

Example 2 *Let us consider the points* $x_1 = (0, 0)$, $x_2 = (3, 0)$ *and* $x_3 = (2, 2)$ *in the Euclidian space. The Vietoris-Rips complex for three different values of* ϵ ($\epsilon = 0.5$, $\epsilon = 2.5$, *and* $\epsilon = 2.9$) *can be seen in Figure 2*

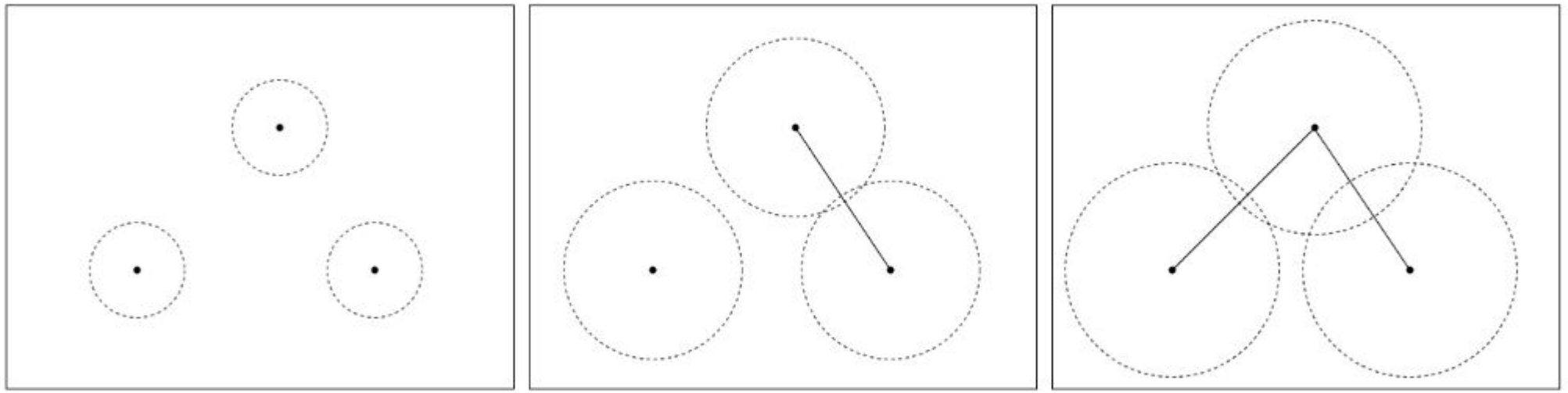


Figure 2: From left to right, Vietoris-Rips complex associated with the three points $x_1 = (0, 0)$, $x_2 = (3, 0)$ and $x_3 = (2, 2)$ for $\epsilon = 0.5$, $\epsilon = 2.5$, and $\epsilon = 2.9$ respectively.

Background

Example 3 *Let us consider the points* $x_1 = (0,0)$, $x_2 = (3,0)$ *and* $x_3 = (2,2)$ *in the Euclidean space, and* VR_ϵ *the Vietoris-Rips filtration of these points. Then, the associated persistence diagram can be seen in Figure 3*

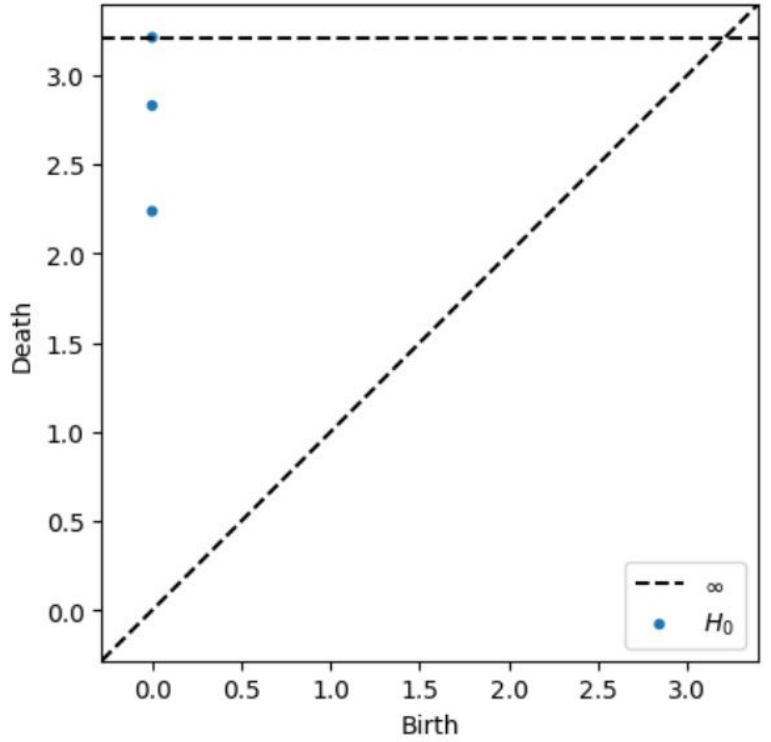


Figure 3: Persistence diagram of the points $x_1 = (0,0)$, $x_2 = (3,0)$ and $x_3 = (2,2)$.

Background

Definition 4 (Bottleneck distance) *Let P and Q be multisets in $\mathbb{R}^2$. The Bottleneck distance between P and Q is defined as*

$$d_B(P,Q) = inf\{c(X) \mid X \text{ is a matching between } P \text{ and } Q\}$$

Definition 5 (Wasserstein distance) *Let P and Q be multisets in $\mathbb{R}^2$. The r-Wasserstein distance between P and Q is defined as*

$$W_r(P,Q) = inf(\sum_{(x,y)\in X} \|q-p\|_\infty^r + \sum_{x\in X^c} \mid p_2 - p_1 \mid^r)^{\frac{1}{r}}$$

where X is a matching and X^c is the set of unmatching points.

Method

Method

Idea: The variation that a manifold suffers when adding a point that belongs to such a manifold is minimal.

Two approaches

We created several semi-supervised learning methods following two different topological data approaches:

- Homological approach
- Connectivity approach

Homological method

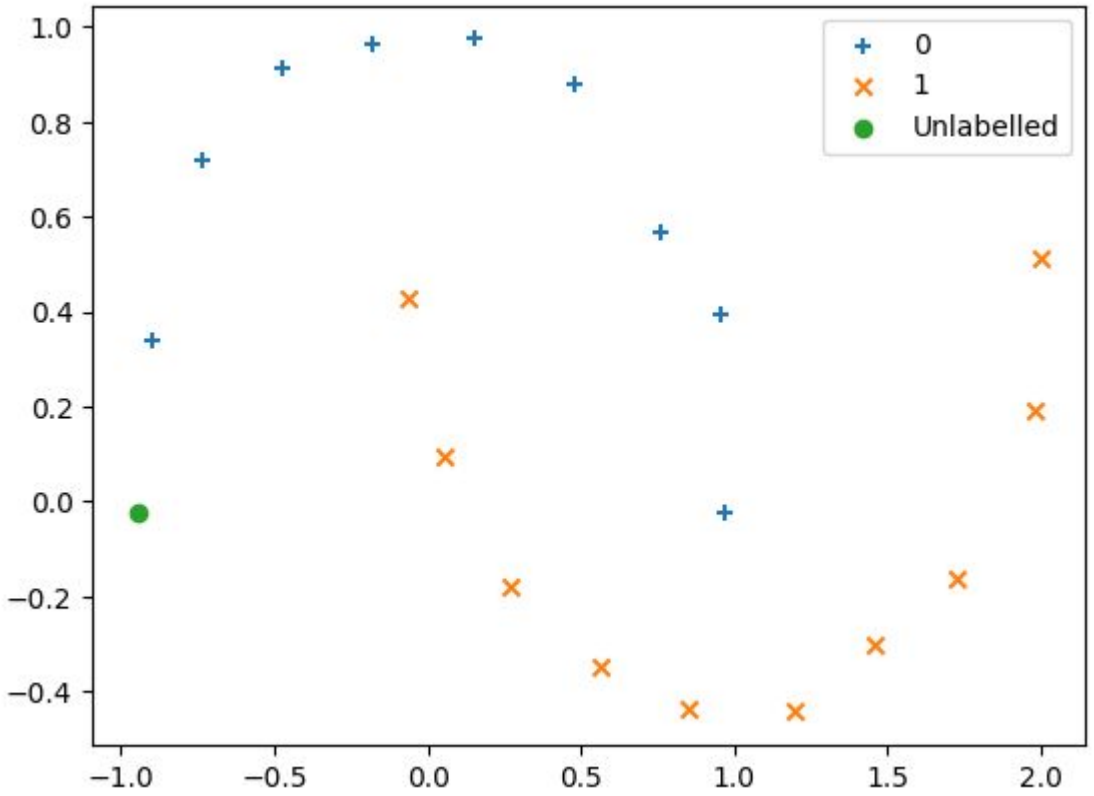
This method is based on the homology of the varieties associated with each data set. We study:

- 0-homology
- Persistence diagrams

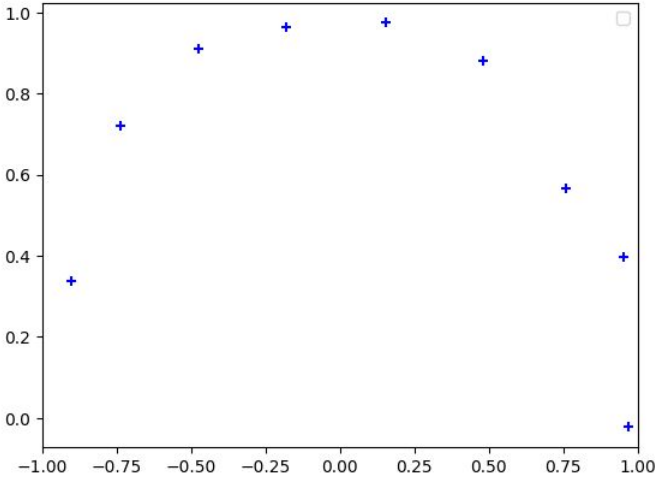
Homological method

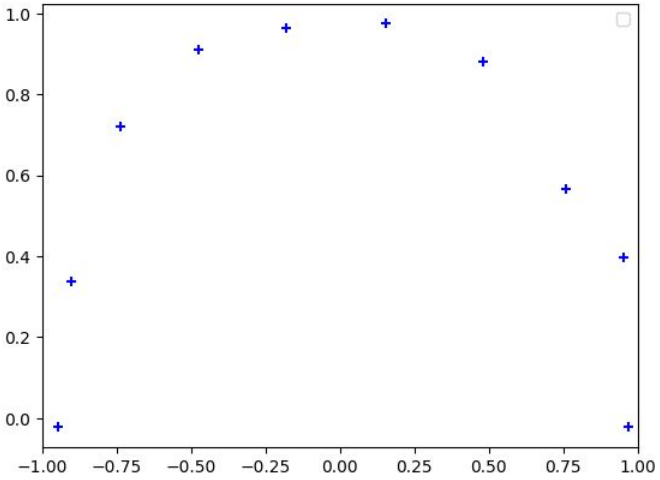
(1) Construct the Vietoris-Rips filtrations V_{X_1}, V_{X_2}, $V_{X_1 \cup \{x\}}$ and $V_{X_2 \cup \{x\}}$;
(2) Construct the persistence diagrams $P(V_{X_1})$, $P(V_{X_2})$, $P(V_{X_1 \cup \{x\}})$ and $P(V_{X_2 \cup \{x\}})$;
(3) Compute the distances $d(P(V_{X_1}), P(V_{X_1 \cup \{x\}}))$ and $d(P(V_{X_2}), P(V_{X_2 \cup \{x\}}))$, from now on d_1 and d_2 respectively;
(4) If both d_1 and d_2 are greater than the threshold t, return none; otherwise, return the set associated with the minimum of the distances d_1 and d_2.

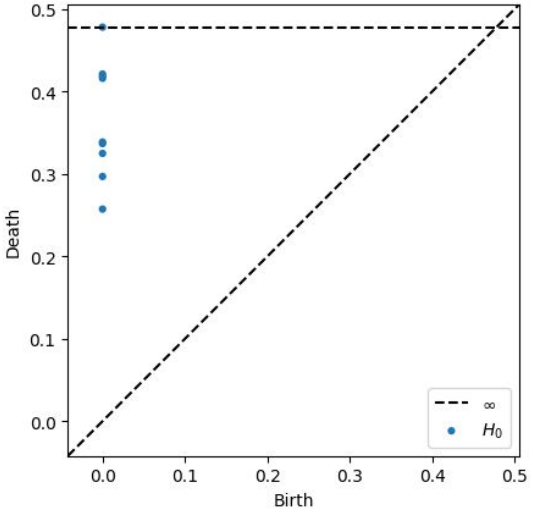
Homological method

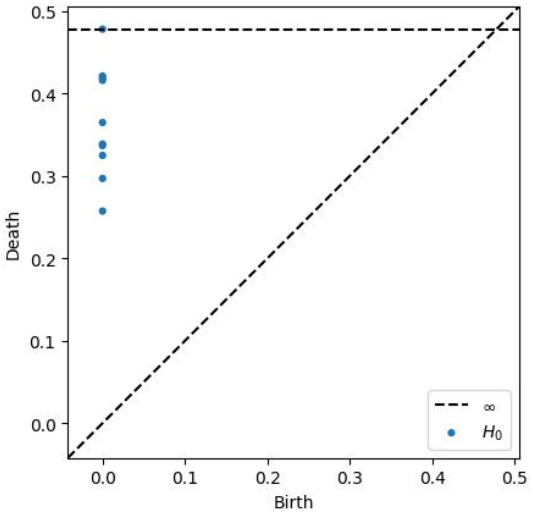


Homological method



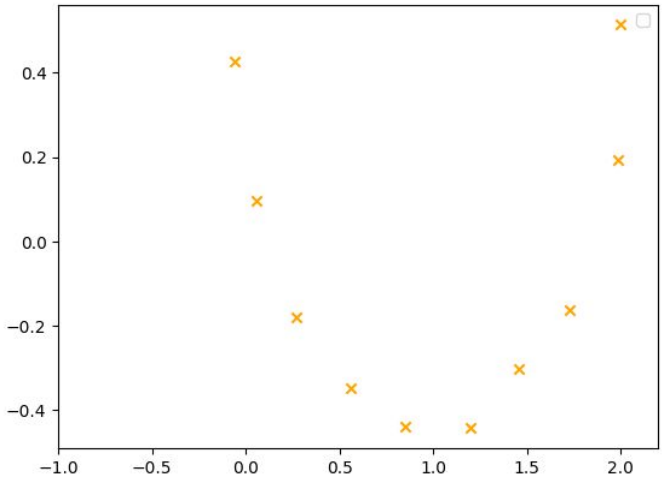


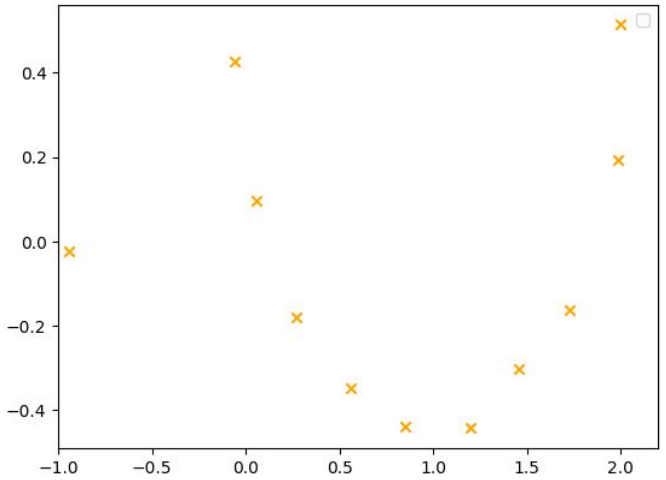


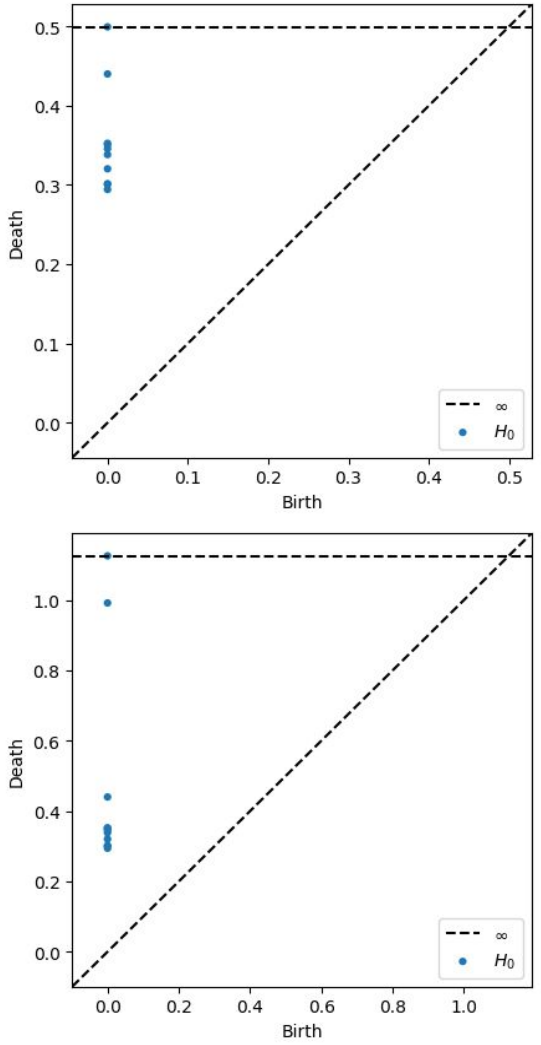


d1 = 0.1285

Homological method

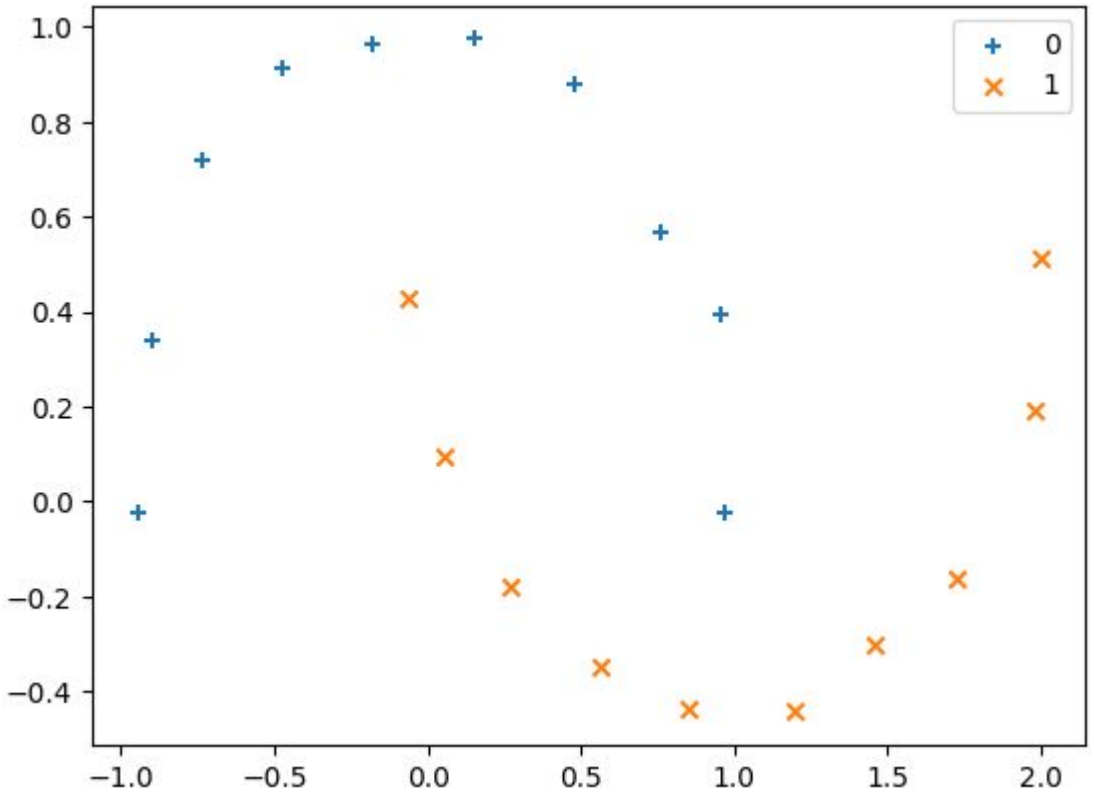






d2 = 0.4958

Homological method



Connectivity method

This method is based on the connectivity of the complex associated with each data set. We study:

- connectivity
- minimum radius

Connectivity method

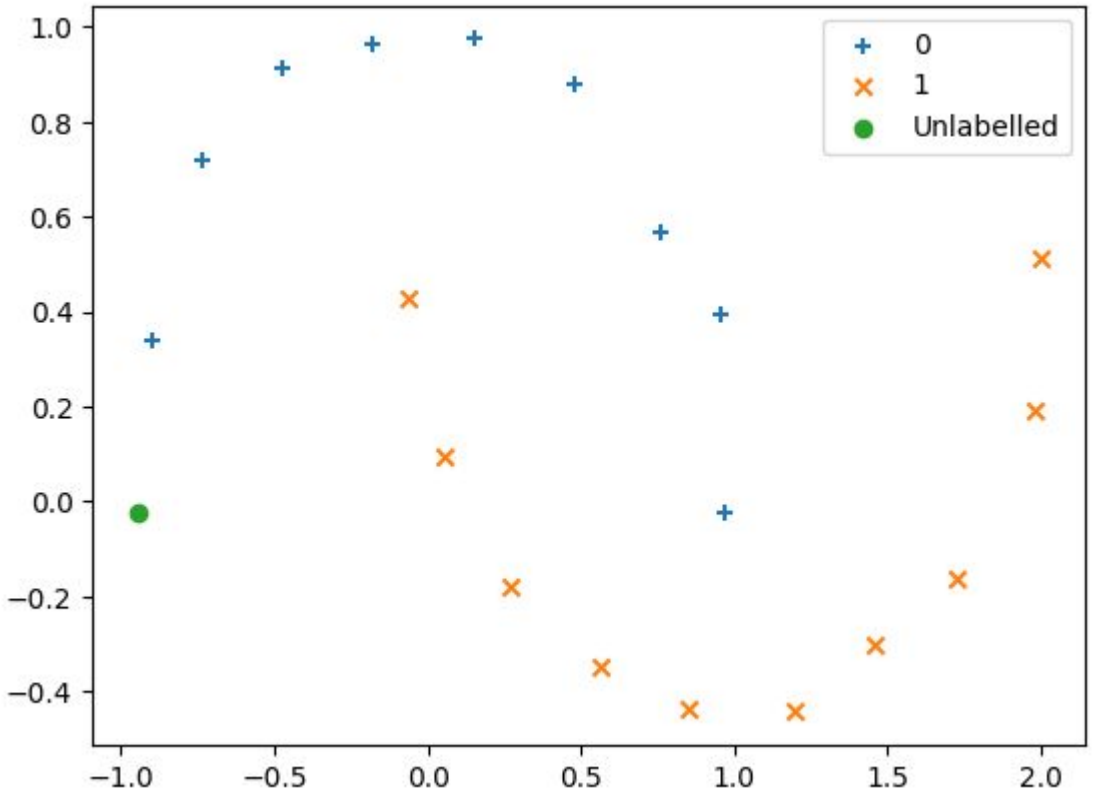
1. Construct the Vietoris-Rips complex V_{X_1}, V_{X_2}, $V_{X_1 \cup \{x\}}$ and $V_{X_2 \cup \{x\}}$;

2. Compute the minimum connectivity radius $r(V_{X_1})$, $r(V_{X_2})$, $r(V_{X_1 \cup \{x\}})$ and $r(V_{X_2 \cup \{x\}})$, from now on r_1, r_2, r_1' and r_2' respectively;

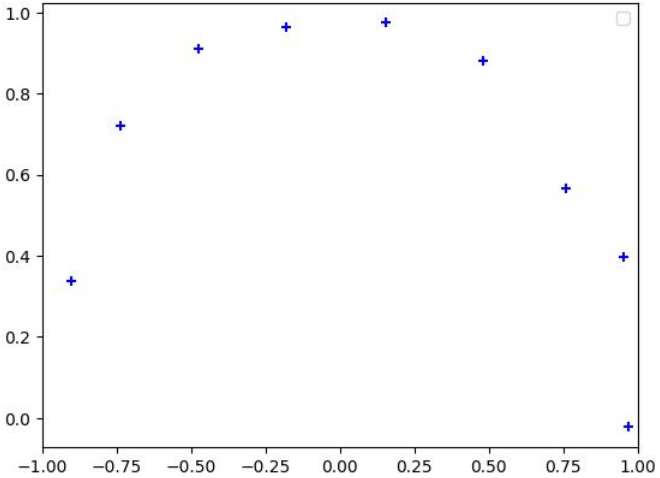
3. Compute the radius variation $|r_1 - r_1'|$ and $|r_2 - r_2'|$ from now on d_1 and d_2 respectively;

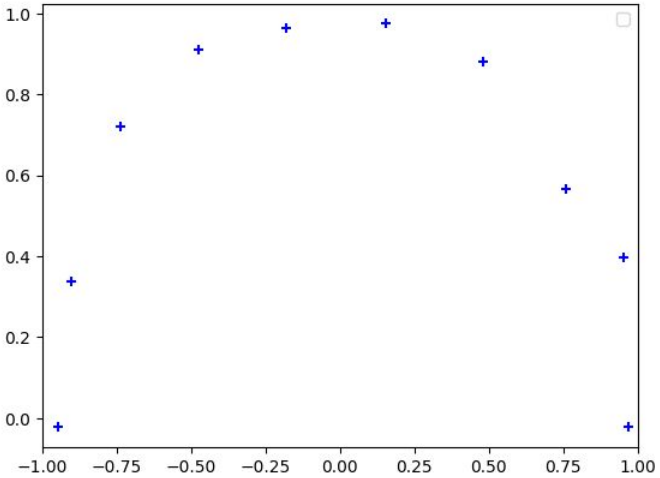
4. If both d_1 and d_2 are zero, return none; otherwise, return the set associated with the minimum of the differences d_1 and d_2.

Connectivity method



Connectivity method



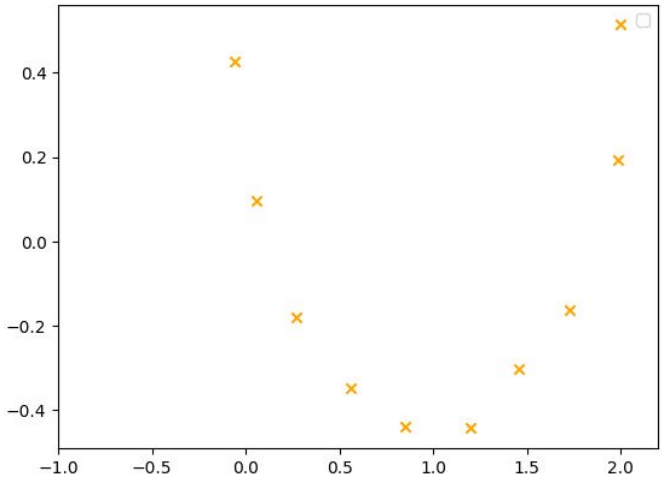


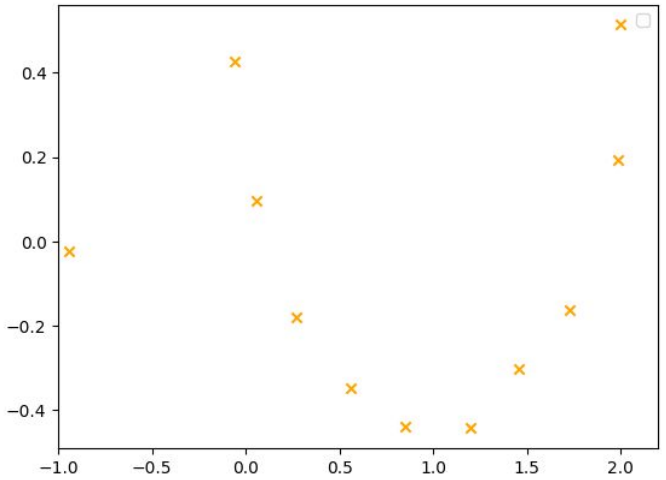
$R_{min} = 1.90$

$R_{min} = 1.94$

d1 = 0.04

Connectivity method





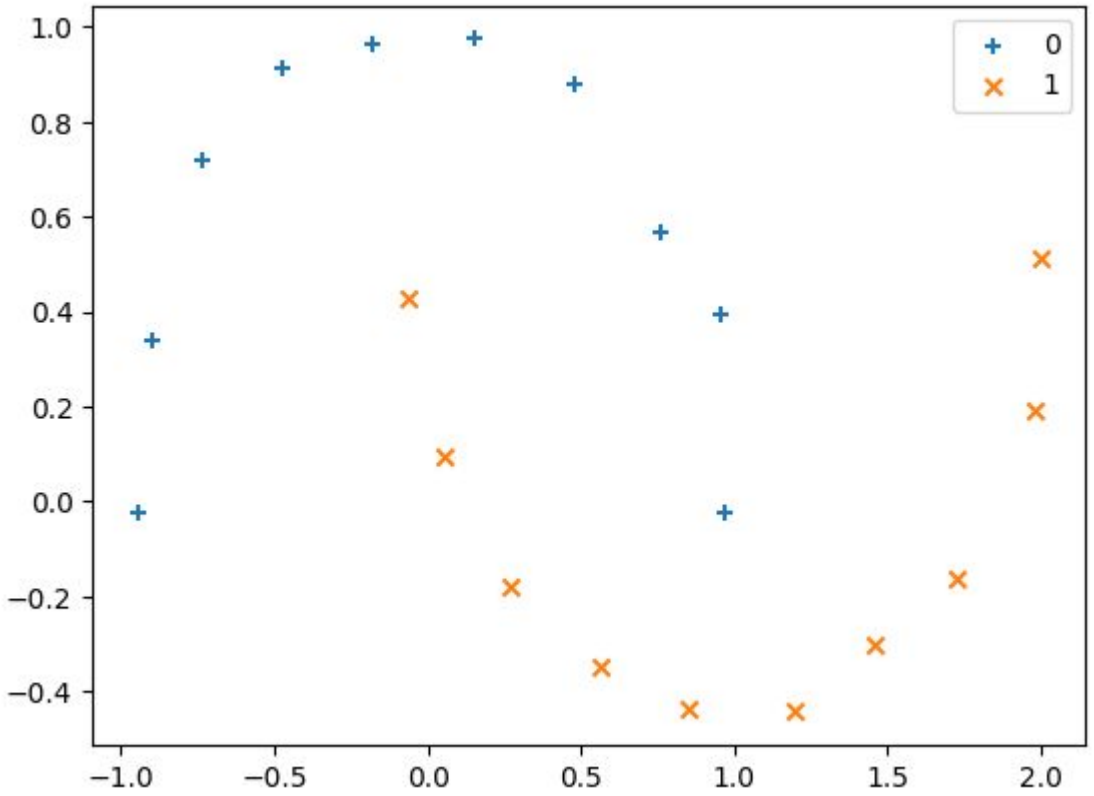
$R_{min} = 2.06$

$R_{min} = 2.99$

d2 = 0.93

Finally we look at which distance is smaller and we write down the point with the class associated with said distance, in this case, class 0.

Connectivity method



Experiments

We have used 5 structured datasets

Dataset	# Examples	# Unlabelled examples	# Features
Banknote	1372	1322	4
Breast Cancer	569	519	30
Ionosphere	351	301	34
Pima Indian Diabetes	768	718	8
Sonar	208	158	60

And 2 different machine learning methods (SVM and Random Forest)

Experiments

We have compared:

- 10 different variants of our homological method
  - 5 methods using bottleneck distance (different th)
  - 5 methods using Wasserstein distance (different th)

- 2 different variants of our connectivity method

- 3 classical methods
  - LabelPropagation
  - LabelSpreading
  - Self-Training classifier

- Base approach (only labeled data)

Results

Method	Banknote		Breast Cancer		Ionosphere		Pima Indian		Sonar		Mean (STD)	
	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Base	97.0	88.6	89.3	**96.1**	80.0	93.3	65.7	60.8	61.3	64.5	78.7(15.2)	80.7(16.7)
Label Propagation	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Label Spreading	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Self Training classifier	95.1	93.6	35.9	35.9	85.0	86.7	66.4	66.4	58.1	67.7	68.1(23.2)	70.1(22.4)
Bottleneck	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	77.1(22.6)	74.1(19.5)
Bottleneck threshold 0.8	**99.2**	92.4	93.2	91.3	78.3	**95.0**	63.6	64.3	61.3	64.5	79.1(17.0)	**81.5(15.6)**
Bottleneck threshold 0.6	**99.2**	91.3	89.3	90.3	75.0	88.3	59.4	63.6	48.4	45.2	74.3(20.9)	75.7(20.6)
Bottleneck threshold 0.4	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Bottleneck threshold 0.2	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Wasserstein	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.8	97.4	89.8	92.2	88.4	80.0	**95.0**	**68.5**	67.8	61.3	64.5	79.9(15.3)	81.1(13.9)
Wasserstein threshold 0.6	**99.2**	93.6	89.3	87.4	70.0	91.7	61.5	61.5	**74.2**	61.3	78.9(15.2)	79.1(16.3)
Wasserstein threshold 0.4	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.2	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Connectivity1	93.6	87.9	89.3	93.2	76.7	88.3	61.5	62.9	64.5	61.3	77.1(14.3)	78.7(15.3)
Connectivity2	93.6	87.5	89.3	93.2	71.7	83.3	60.1	58.7	64.5	64.5	75.8(14.9)	77.5(15.0)

Results

Method	Banknote		Breast Cancer		Ionosphere		Pima Indian		Sonar		Mean (STD)	
	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Base	97.0	88.6	89.3	96.1	80.0	93.3	65.7	60.8	61.3	64.5	78.7(15.2)	80.7(16.7)
Label Propagation	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Label Spreading	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Self Training classifier	95.1	93.6	35.9	35.9	85.0	86.7	66.4	66.4	58.1	67.7	68.1(23.2)	70.1(22.4)
Bottleneck	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	77.1(22.6)	74.1(19.5)
Bottleneck threshold 0.8	99.2	92.4	93.2	91.3	78.3	95.0	63.6	64.3	61.3	64.5	79.1(17.0)	81.5(15.6)
Bottleneck threshold 0.6	99.2	91.3	89.3	90.3	75.0	88.3	59.4	63.6	48.4	45.2	74.3(20.9)	75.7(20.6)
Bottleneck threshold 0.4	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Bottleneck threshold 0.2	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Wasserstein	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.8	97.4	89.8	92.2	88.4	80.0	95.0	68.5	67.8	61.3	64.5	79.9(15.3)	81.1(13.9)
Wasserstein threshold 0.6	99.2	93.6	89.3	87.4	70.0	91.7	61.5	61.5	74.2	61.3	78.9(15.2)	79.1(16.3)
Wasserstein threshold 0.4	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.2	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Connectivity1	93.6	87.9	89.3	93.2	76.7	88.3	61.5	62.9	64.5	61.3	77.1(14.3)	78.7(15.3)
Connectivity2	93.6	87.5	89.3	93.2	71.7	83.3	60.1	58.7	64.5	64.5	75.8(14.9)	77.5(15.0)

Worst → Best

Results

Method	Banknote		Breast Cancer		Ionosphere		Pima Indian		Sonar		Mean (STD)	
	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF	SVM	RF
Base	97.0	88.6	89.3	**96.1**	80.0	93.3	65.7	60.8	61.3	64.5	78.7(15.2)	80.7(16.7)
Label Propagation	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Label Spreading	97.4	93.2	90.3	89.3	86.7	86.7	64.3	68.5	58.1	54.8	79.3(17.1)	78.5(16.3)
Self Training classifier	95.1	93.6	35.9	35.9	85.0	86.7	66.4	66.4	58.1	67.7	68.1(23.2)	70.1(22.4)
Bottleneck	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	77.1(22.6)	74.1(19.5)
Bottleneck threshold 0.8	**99.2**	92.4	93.2	91.3	78.3	**95.0**	63.6	64.3	61.3	64.5	79.1(17.0)	**81.5(15.6)**
Bottleneck threshold 0.6	**99.2**	91.3	89.3	90.3	75.0	88.3	59.4	63.6	48.4	45.2	74.3(20.9)	75.7(20.6)
Bottleneck threshold 0.4	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Bottleneck threshold 0.2	97.4	90.5	87.4	85.4	78.3	86.7	63.6	62.9	45.2	45.2	74.4(20.5)	74.1(19.5)
Wasserstein	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.8	97.4	89.8	92.2	88.4	80.0	**95.0**	**68.5**	67.8	61.3	64.5	79.9(15.3)	81.1(13.9)
Wasserstein threshold 0.6	**99.2**	93.6	89.3	87.4	70.0	91.7	61.5	61.5	**74.2**	61.3	78.9(15.2)	79.1(16.3)
Wasserstein threshold 0.4	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Wasserstein threshold 0.2	97.0	96.2	87.4	87.4	76.7	81.7	60.8	62.9	71.0	71.0	78.6(14.1)	79.8(13.2)
Connectivity1	93.6	87.9	89.3	93.2	76.7	88.3	61.5	62.9	64.5	61.3	77.1(14.3)	78.7(15.3)
Connectivity2	93.6	87.5	89.3	93.2	71.7	83.3	60.1	58.7	64.5	64.5	75.8(14.9)	77.5(15.0)

Time comparison

Method	Time/point (ms)			
	Banknote	BreastCancer	Ionosphere	PrimaIndian
Bottleneck	47,7963	44,3350	64,0693	51,6814
Wasserstein	2,1548	2,9557	3,4632	2,4779
Classic	0,3918	0,9852	1,7316	0,7080

Parallelization

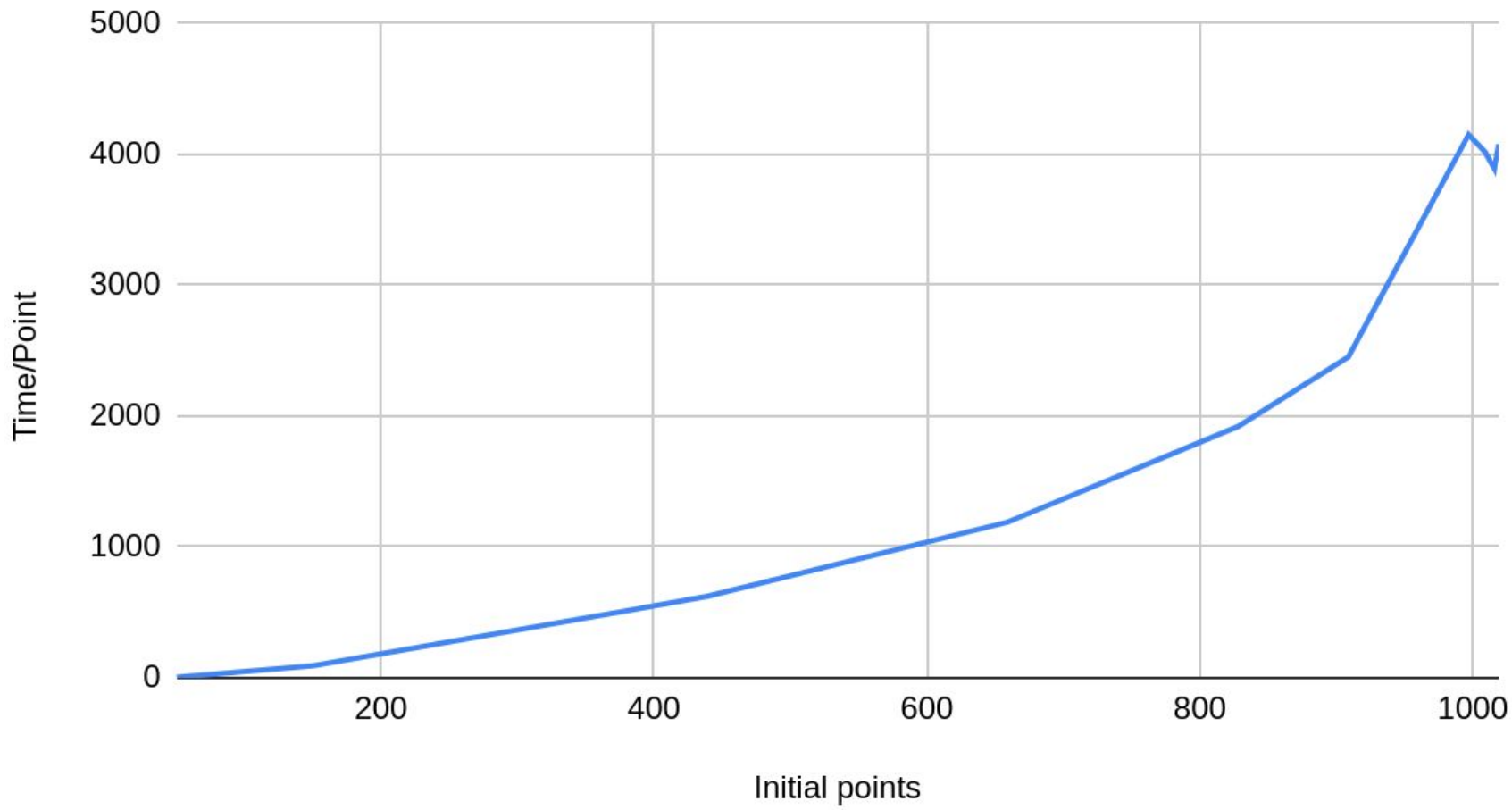
We have parallelized the annotation of each point using:

- Joblib Library
- Parallel method
- Delayed functions

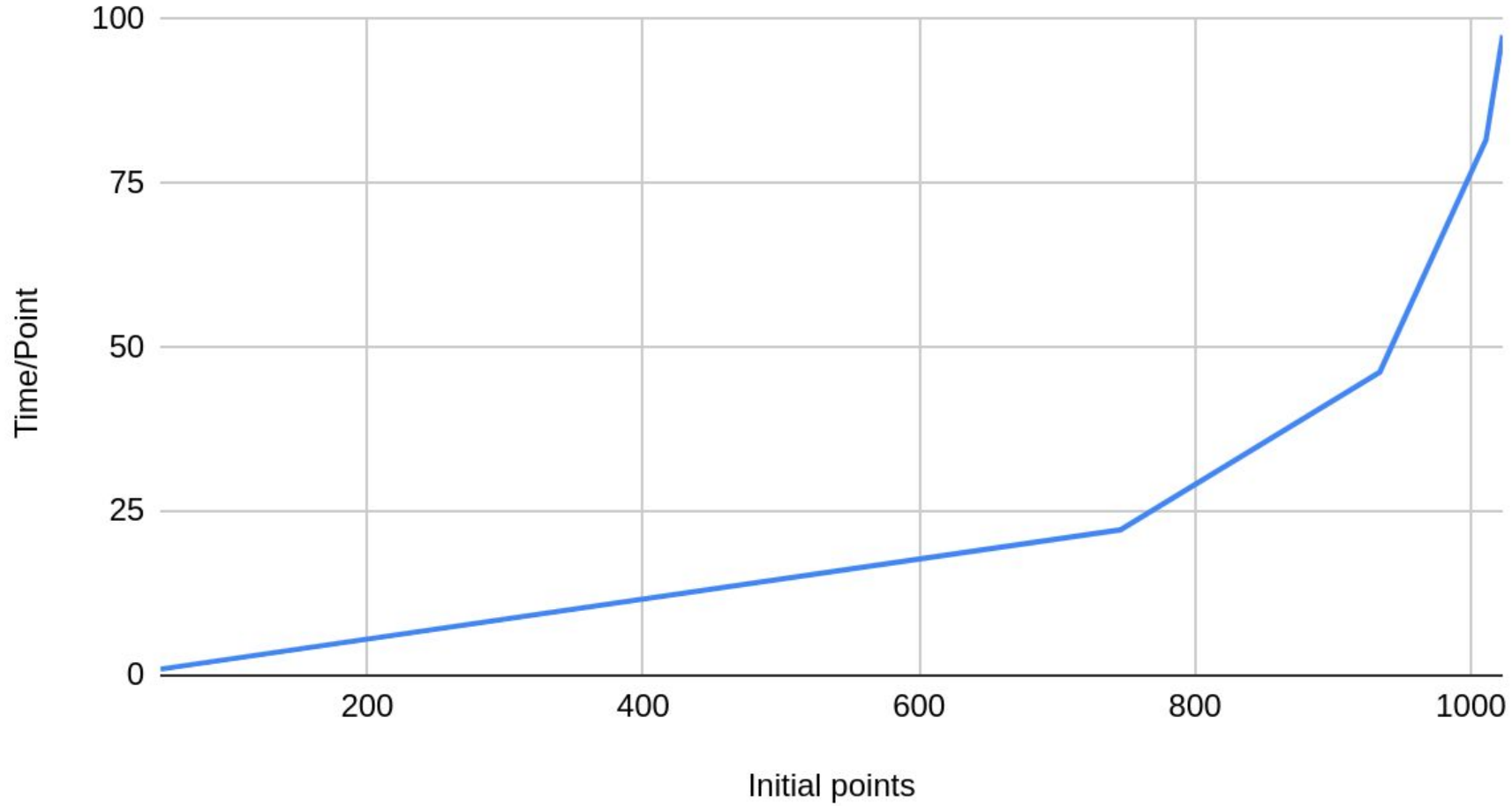
Time comparison - Parallelization

Method	Time/point (ms)			
	Banknote	BreastCancer	Ionosphere	PrimaIndian
Bottleneck	4,7013	6,6502	9,4372	11,1504
Wasserstein	0,9794	1,3547	6,4935	3,5398
Classic	0,3918	0,9852	1,7316	0,7080

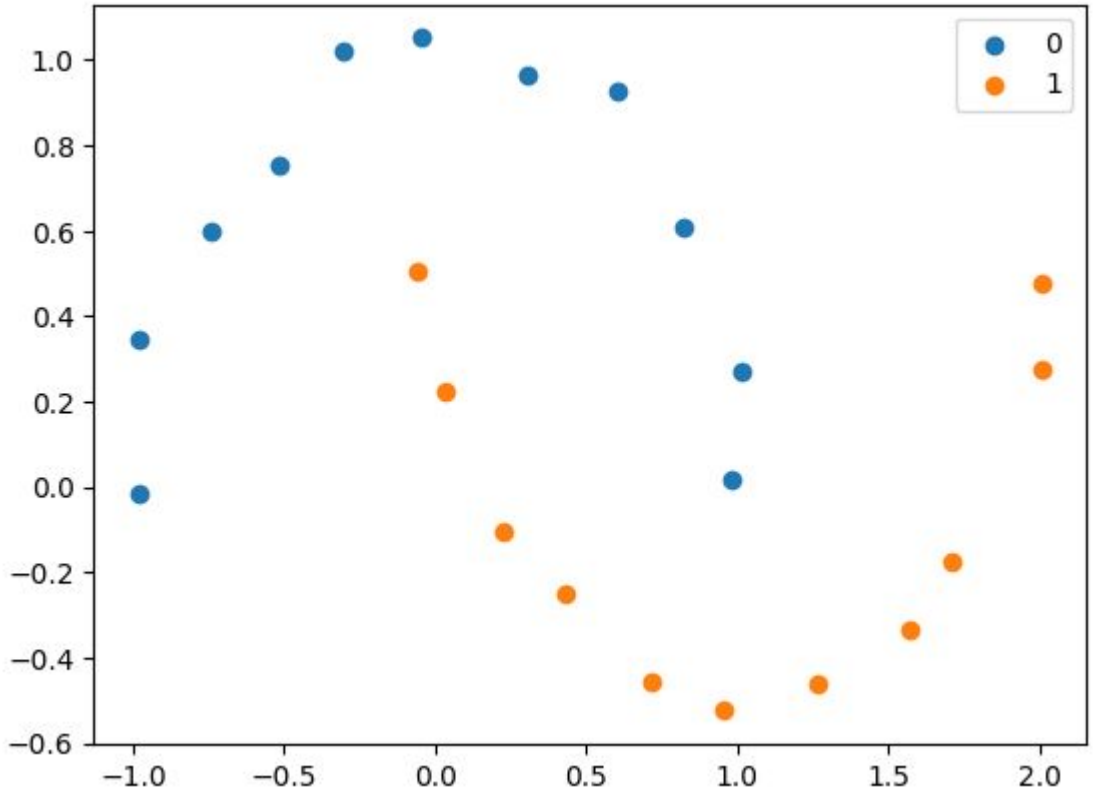
Bottleneck


5000
4000
3000
2000
1000
0
Time/Point
200
400
600
800
1000
Initial points

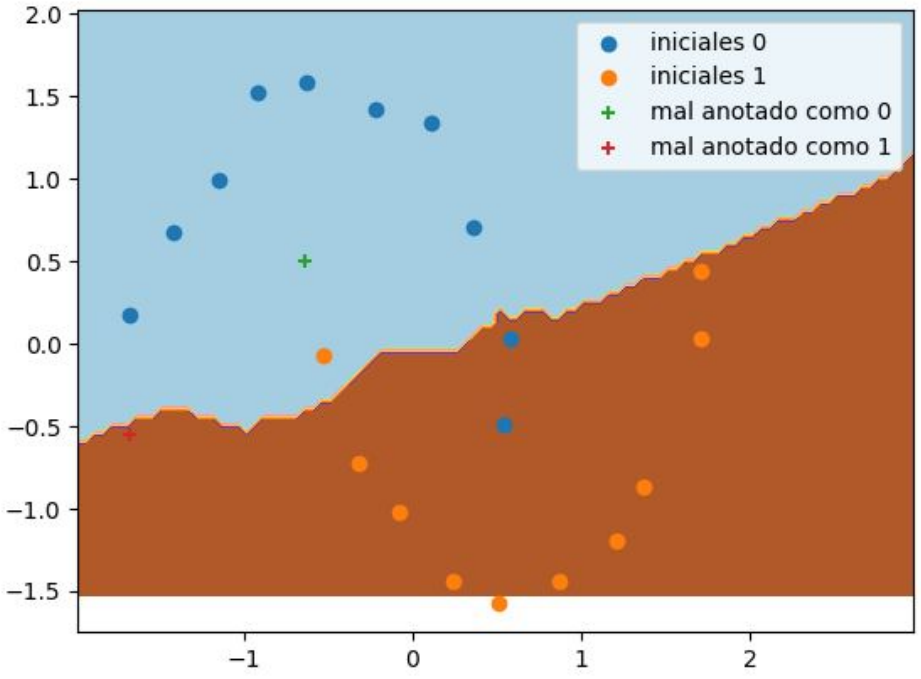
Wasserstein



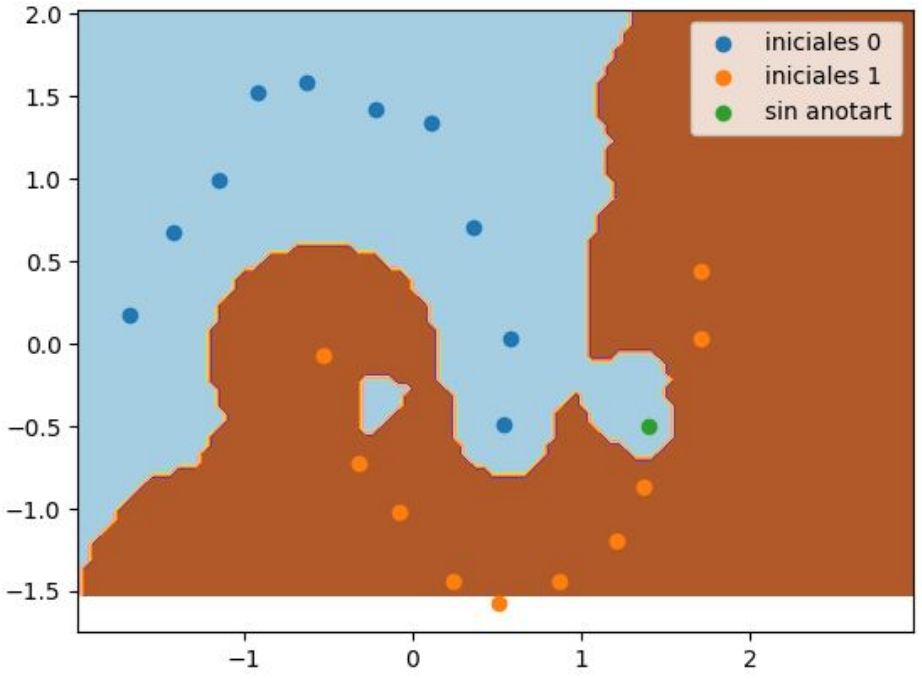
Example Moons



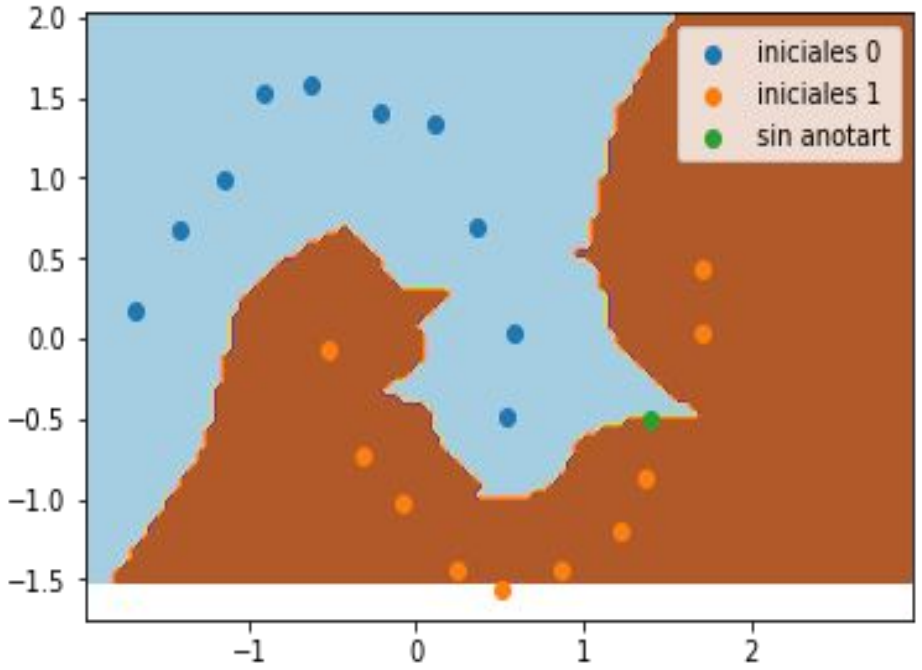
Example Moons - Label Propagation



Example Moons - Bottleneck distance



Example Moons - Wasserstein distance



Conclusions and further work

Conclusions

Our method can create better classification models than the obtained when using classical semi-supervised learning methods.

Further work

The proposed method can be expanded to multi-class classification tasks.

We plan to design new semi-supervised learning algorithms based on other notions from TDA.

Thanks