Воронежский Государственный Университет
Факультет прикладной математики, информатики и механики

Кафедра вычислительной математики и прикладных информационных технологий

Применение персистентных гомологий для задачи классификации изображений

Бакалаврская работа
Направление 01.03.02 Прикладная математика и информатика
Профиль Математическое моделирование и вычислительная математика

Обучающийся		Снопов П.М.
Руководитель	д.т.н., проф.	Леденева Т.М.

Содержание

Цель и задачи работы

Цель: Исследование подхода, основанного на топологическом анализе данных, для классификации изображений

Задачи:

- Изучение теоретических и практических основ топологического анализа данных
- Анализ подходов классификации изображений
- Формирование алгоритма на основе персистентных гомологий
- Проведение вычислительного эксперимента, выявление области применимости, плюсов и минусов данного подхода

Актуальность

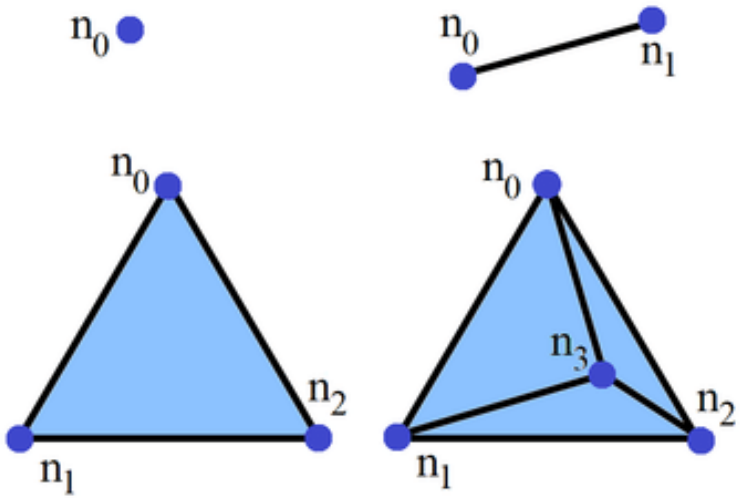
- Классификация изображений – фундаментальная задача анализа данных, одно из основных направлений ее развития, которое в последнее время имеет очень интенсивное развитие, связанное с достижениями нейронных сетей.
- Основным инструментом ТДА являются персистентные гомологии, о которых можно думать как об адаптации понятия гомологии к облаку точек. С помощью такого инструмента можно выявлять топологические характеристики исследуемого объекта.
- Задача классификации по сути является задачей определения характеристических свойств, которым удовлетворяют объекты одного класса. Поэтому персистентные гомологии могут быть полезны в данной задаче, так как зачастую такие свойства имеют геометрическую природу.

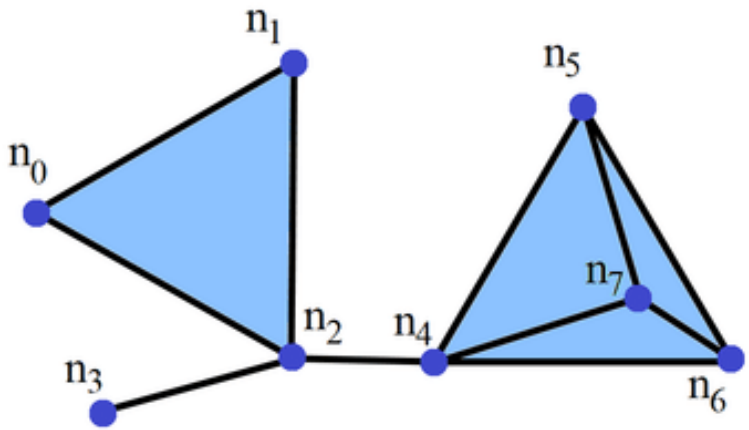
Симплициальные комплексы

Definition

Симплициальный комплекс K – это множество симплексов, т.е. выпуклых оболочек набора $n+1$ точек $\in \mathbb{R}^p$, таких, что векторы $x_1 - x_0, ..., x_n - x_0$ линейно независимы, при этом

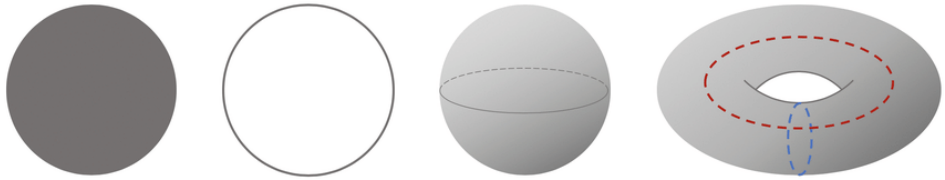
- Для каждого симплекса из K его грани тоже лежат в K,
- Пересечение любых двух симплексов $\sigma, \tau \in K$ либо пусто, либо является гранью и σ, и τ.





Симплициальные гомологии

Для симплициального комплекса можно посчитать его группы симплициальных гомологий H_n. Ранк n-ой группы – это n-ое число Бетти β_n. Оно отражает количество n-мерных особенностей комплекса. Для $n = 0$, β_0 отражает количество компонент связности данного пространства. При $n = 1$ – количество циклов. При $n = 2$ число Бетти β_2 описывает количество "полостей".



$\beta_0 = 1$ $\beta_1 = 0$ $\beta_2 = 0$

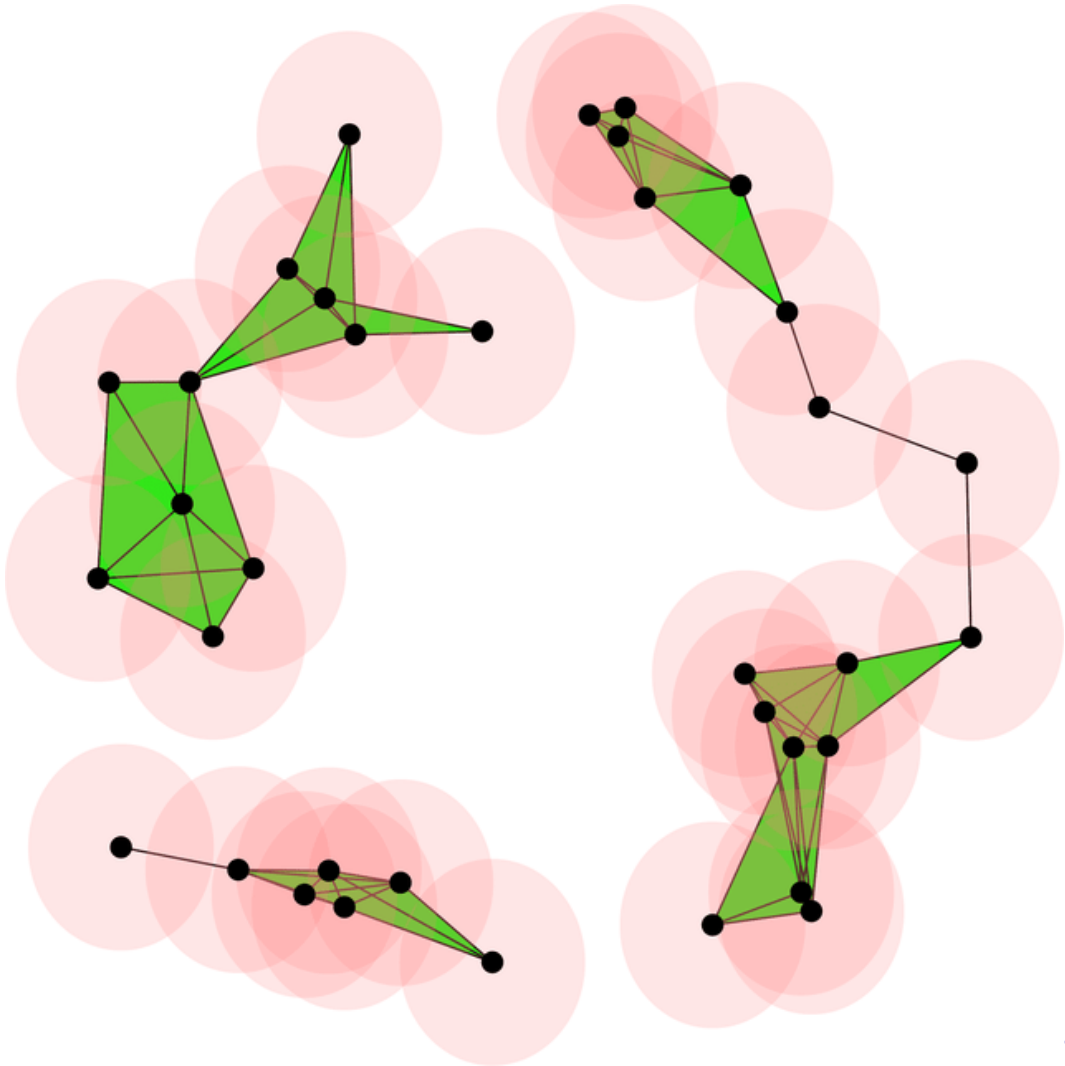
$\beta_0 = 1$ $\beta_1 = 1$ $\beta_2 = 0$

$\beta_0 = 1$ $\beta_1 = 0$ $\beta_2 = 1$

$\beta_0 = 1$ $\beta_1 = 2$ $\beta_2 = 1$

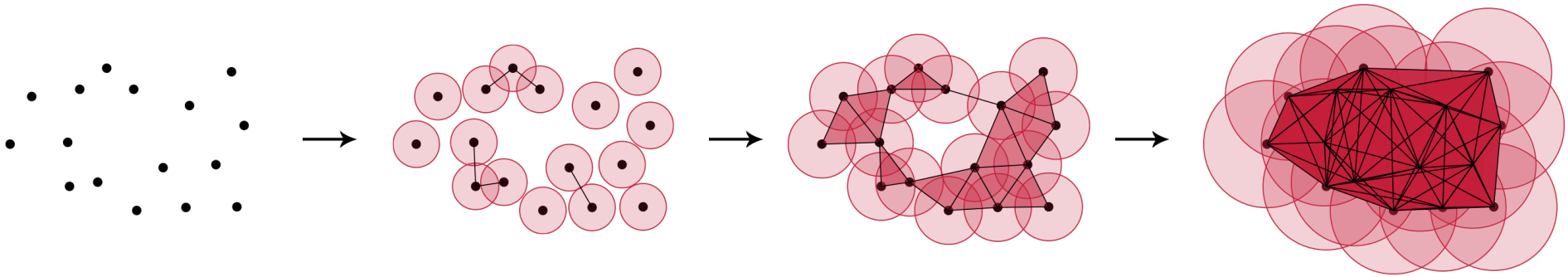
Как построить симплициальный комплекс?

Имея облако точек $X \subset \mathbb{R}^n$, можно построить симплициальный комплекс по нему, например комплекс *Вьеториса—Рипса*.

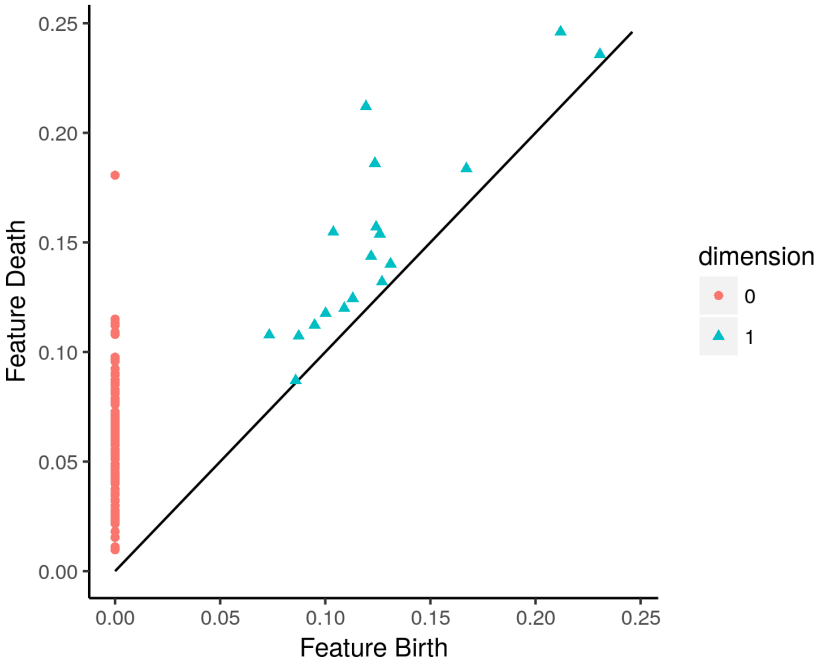


Фильтрации и устойчивые гомологии

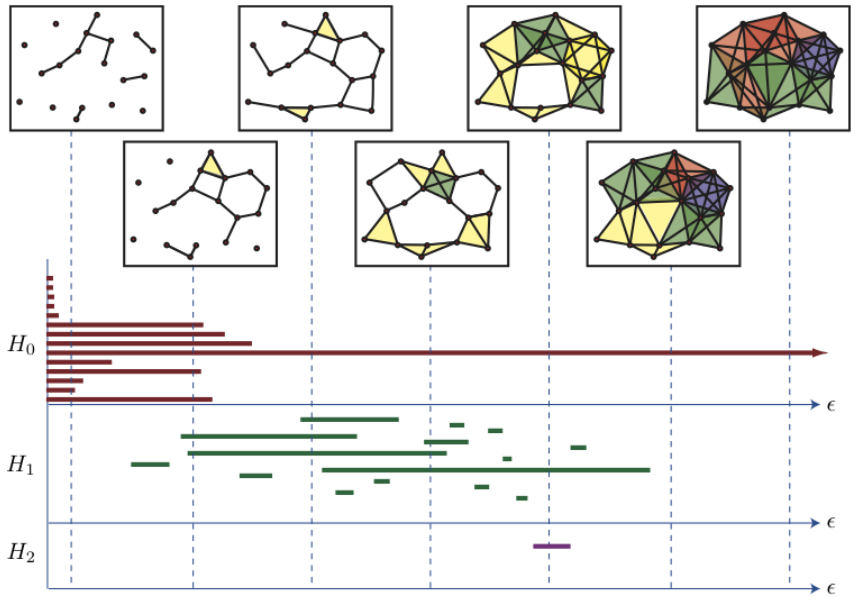
Фильтрацией симплициального комплекса K называют вложенное семейство подкомплексов $(K_\tau)_{\tau \in T}$, такое, что если $\tau < \tau^{'}$, то $K_\tau \subseteq K_{\tau'}$. n-ыми устойчивыми гомологиями будем называть семейство n-гомологий для каждого члена фильтрации. Естественный порядок на фильтрации будет порождать естественные гомоморфизмы между группами гомологий. Таким образом, устойчивые гомологии отслеживают появление и исчезновение топологических особенностей в фильтрации.



Фильтрации и устойчивые гомологии



(a) Диаграмма персистентности

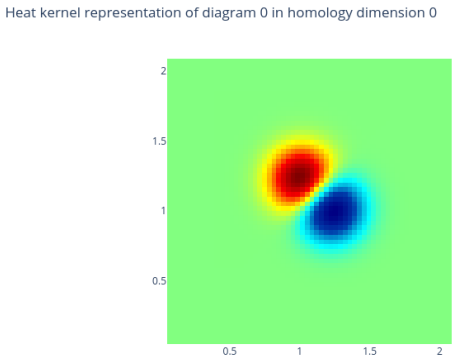


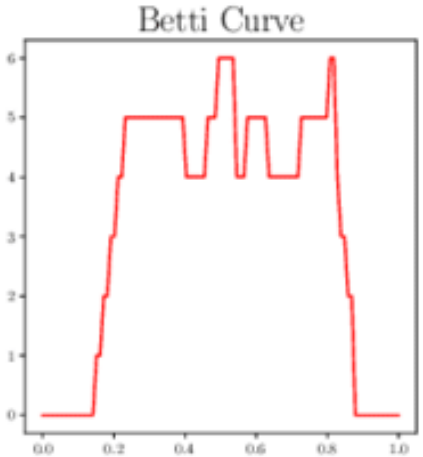
(b) Баркод

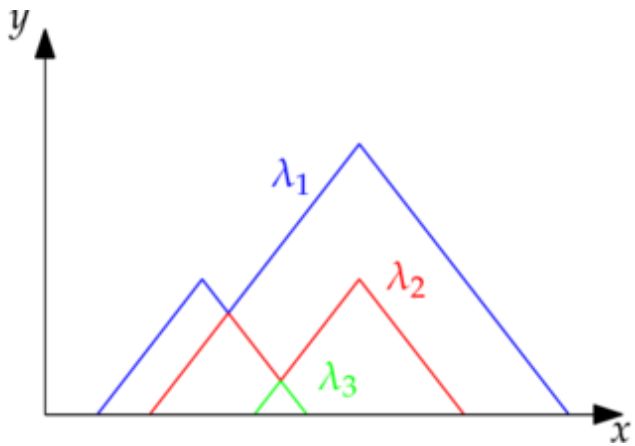
Рис.: Способы кодирования информации об персистентных гомологиях

Метрическое пространство всех диаграмм персистентности

На множестве $\mathcal{D}$ диаграмм персистентности можно ввести структуру метрического пространства, однако для большинства алгоритмов машинного обучения и анализа данных требуется более сложная структура.
Можно векторизовать диаграмму, т.е. построить отображение $\varphi : \mathcal{D} \to V$, где V – нормированное векторное пространство. Тогда диаграмме B можно сопоставить число – $\|\varphi(B)\|$.





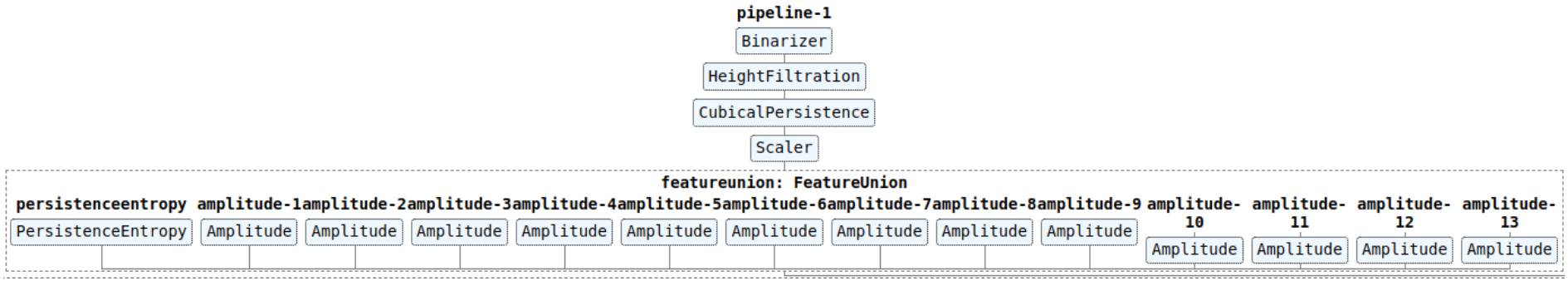


Общая идея алгоритма

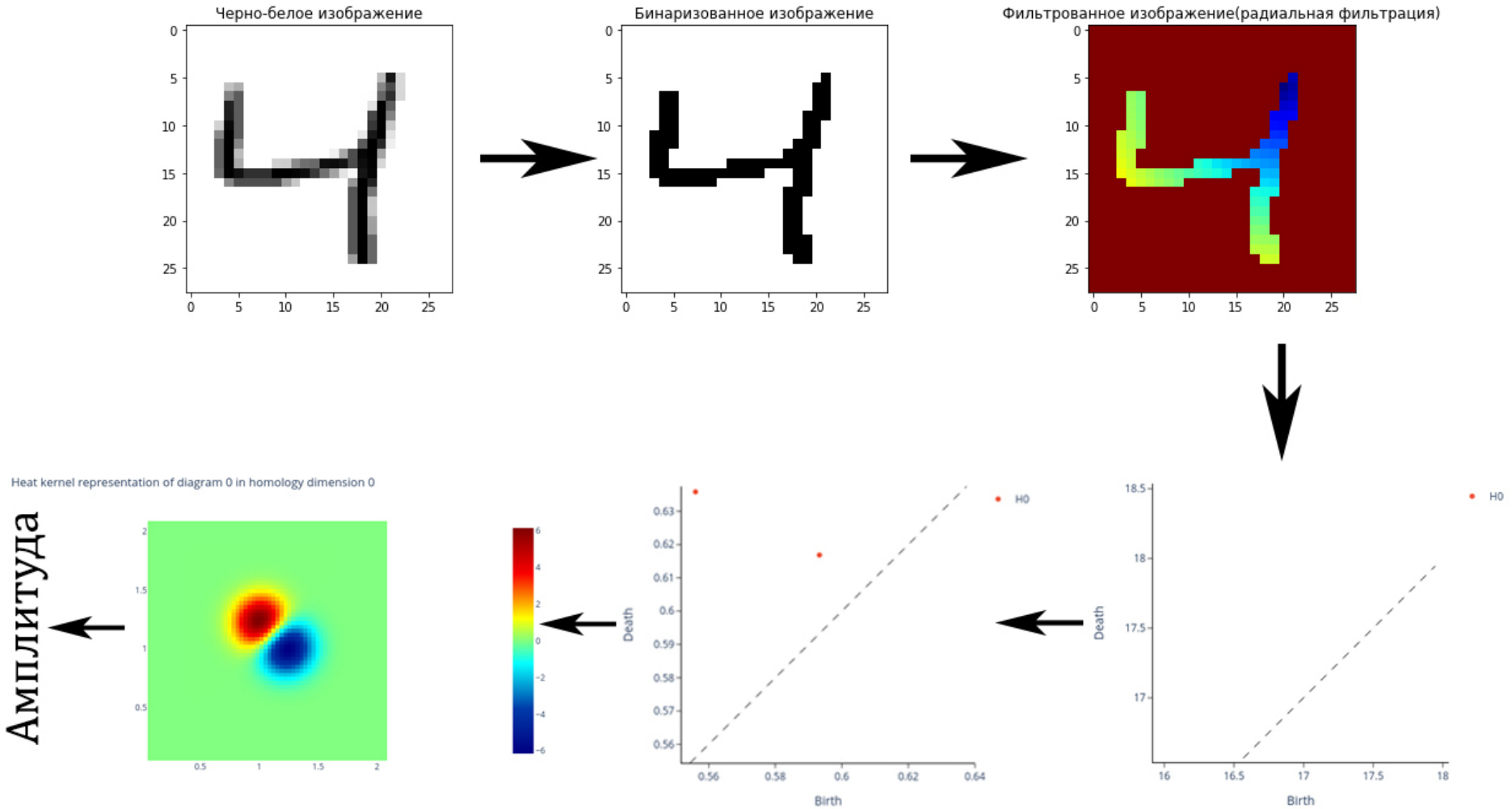
- По изображению строим фильтрацию;
- По построенной фильтрации находим кубический комплекс, персистентные гомологии и строим диаграмму устойчивости;
- Векторизуем диаграмму устойчивости, получаем векторное представление, которое можно использовать в моделях машинного обучения.

Пайплайн используемого метода построения векторного представления

Построить фильтрацию для изображения можно различными методами. В настоящей работе было построено 20 разнообразных фильтрацией, для каждой из которых были получены 0 и 1 персистентные гомологии, из диаграмм которых было получено 14 признаков. Таким образом, для одной картинки было получено $20 \times 2 \times 14 = 560$ признаков.



Процесс получения векторного представления



Сравнение результатов различных методов машинного обучения

Название модели	Значение на тренировочной выборке	Значение на тестовой выборке
Логистическая регрессия	0.989	0.903
Метод опорных векторов	1.0	0.893
Случайный лес	1.0	0.89
CatBoost	1.0	0.893

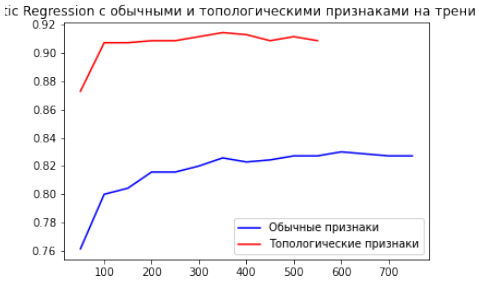
Таблица: Значения базовых моделей на тренировочной и тестовой выборках

Сравнение результатов различных методов машинного обучения

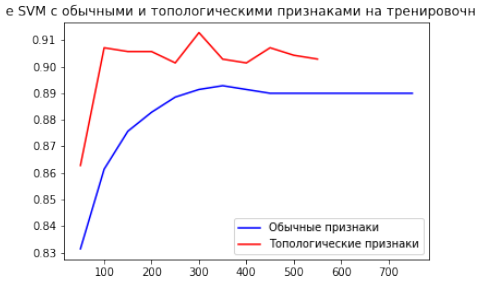
Название модели	Значение на тренировочной выборке	Значение на тестовой выборке
Логистическая регрессия	0.911	0.917
Метод опорных векторов	0.903	0.893
Случайный лес	0.89	0.89
CatBoost	0.905	0.9

Таблица: Значения наилучших моделей, полученных в результате подбора параметров поиском по сетке, на тренировочной и тестовой выборках

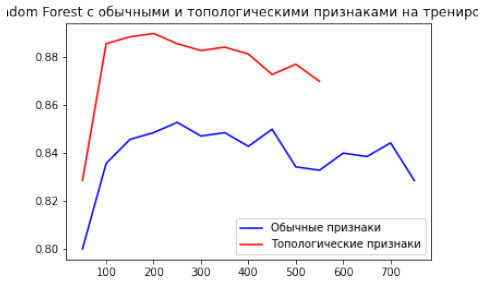
Подбор гиперпараметров и отбор признаков



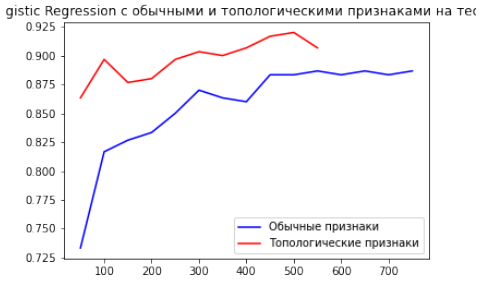
(a) Логистическая регрессия на тренировочной выборке



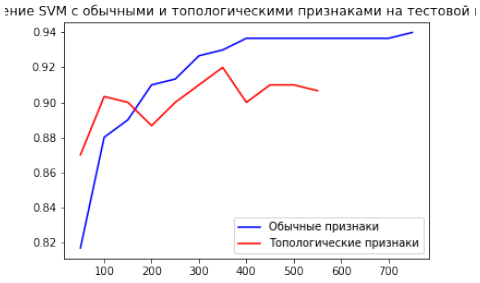
(b) Метод опорных векторов на тренировочной выборке



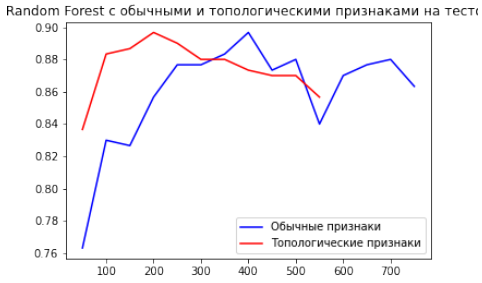
(c) Случайный лес на тренировочной выборке



(d) Логистическая регрессия на тестовой выборке

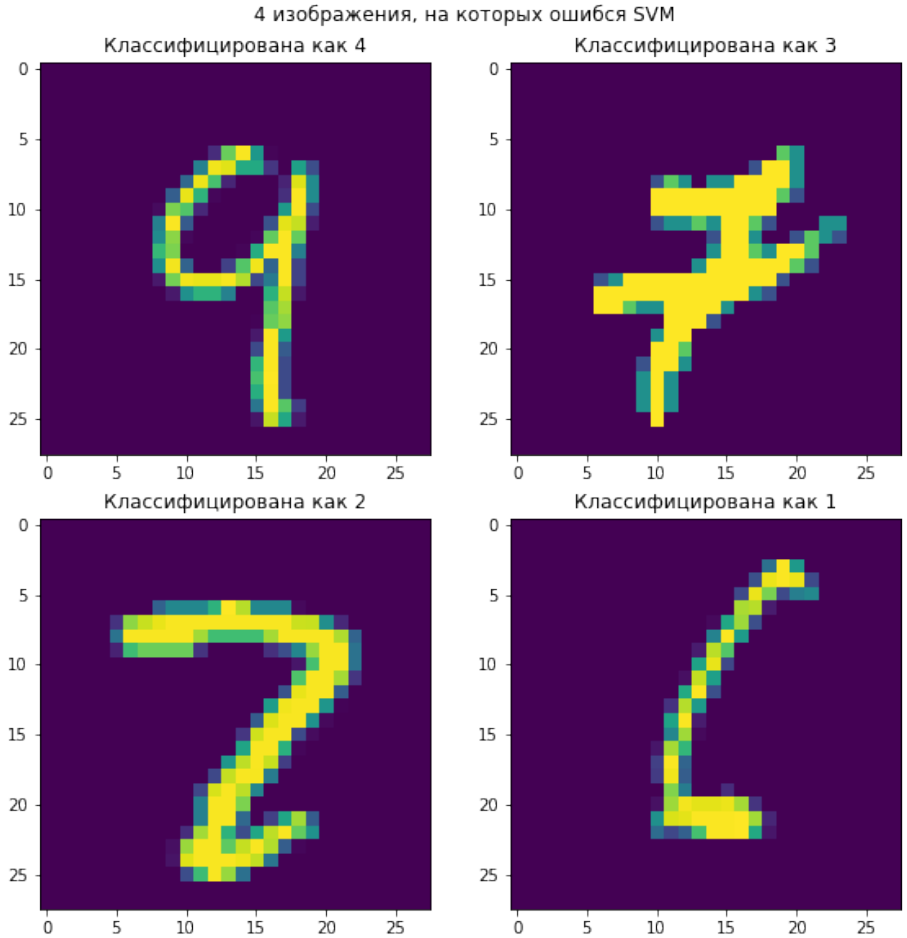


(e) Метод опорных векторов на тестовой выборке



(f) Случайный лес на тестовой выборке

Первые несколько изображений, на которых ошибся классификатор



Заключение

- В ходе данной работы был изучен теоретический материал по алгебраической и прикладной топологии и машинному обучению.
- Был реализован алгоритм классификации датасета MNIST, используя только топологические свойства изображенных рукописных цифр, который выдавал высокий уровень точности.
- В ходе данной работы было выявлено, что реализованный алгоритм является эффективным алгоритмом понижения размерности пространства признаков.
- Было проведено сравнение с другими моделями машинного обучения, в результате которого было обнаружено, что реализованный алгоритм показывает более высокую точность классификации при меньшем числе признаков.

Воронежский Государственный Университет
Факультет прикладной математики, информатики и механики

Кафедра вычислительной математики и прикладных информационных технологий

Применение персистентных гомологий для задачи классификации изображений

Бакалаврская работа
Направление 01.03.02 Прикладная математика и информатика
Профиль Математическое моделирование и вычислительная математика

Обучающийся Снопов П.М.
Руководитель д.т.н., проф. Леденева Т.М.