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.

Актуальность

- Классификация изображений – фундаментальная задача анализа данных, одно из основных направлений приложения машинного обучения к компьютерному зрению.
- Машинное обучение – главный тренд последних 10 лет($\sim$ 150 публ./день).
- Устойчивые гомологии и топологический анализ данных – недавно возникшая и быстро развивающаяся область современной математики и анализа данных.

Постановка задачи

- Задано конечное множество объектов вместе с метками классов.
- Метки классов остальных объектов не известны.
- Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

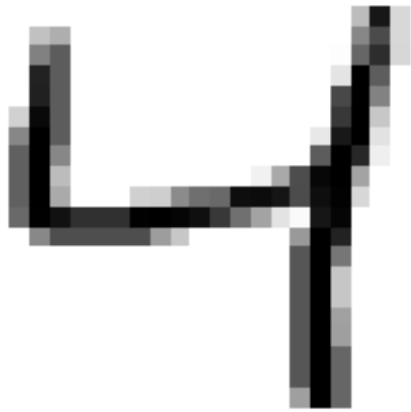


Рис. Пример изображения из датасета MNIST

Цель и задачи работы

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

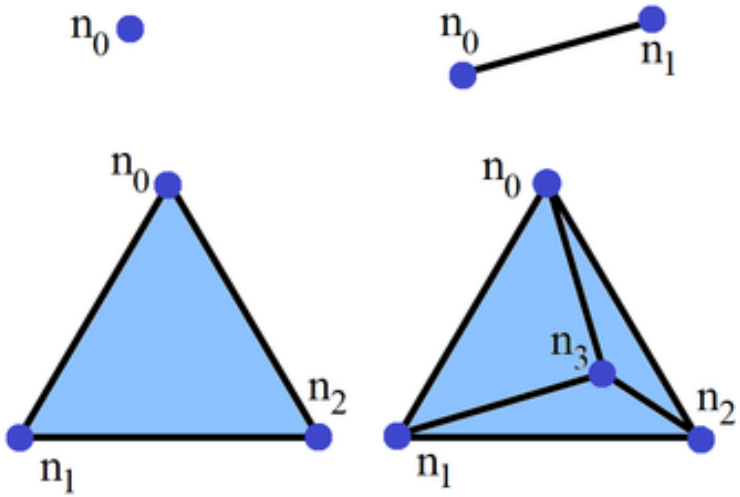
Задачи:

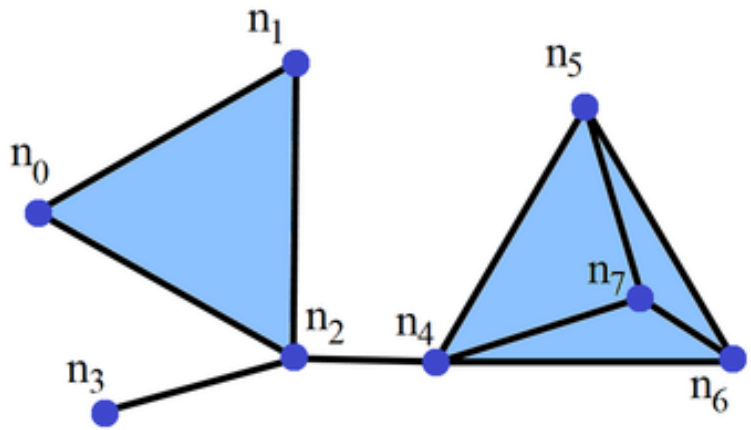
- Изучение теоретических и практических основ топологического анализа данных
- Анализ подходов классификации изображений
- Формирование алгоритма на основе персистентных гомологий
- Проведение вычислительного эксперимента, выявление области применимости, плюсов и минусов данного подхода

Симплициальные комплексы

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.



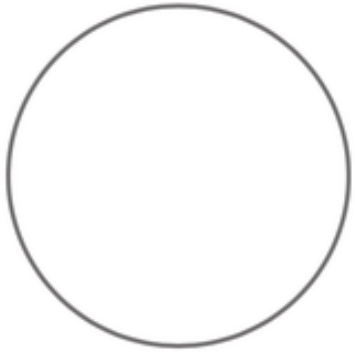


Симплициальные гомологии

Группа симплициальных гомологий $H_n(K)$ размерности n отражает количество n-циклов. Ранг данной группы – *n-ое число Бетти* β_n – количество n-циклов.

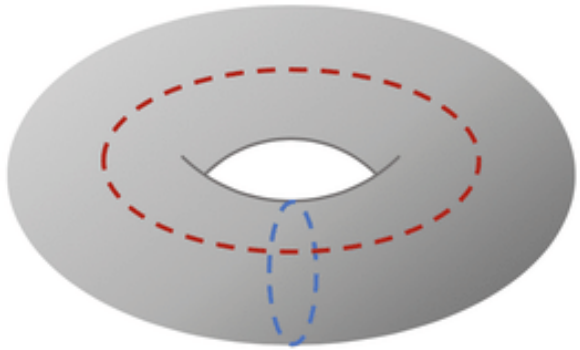
$$\beta_n = \dim(H_n(K))$$

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



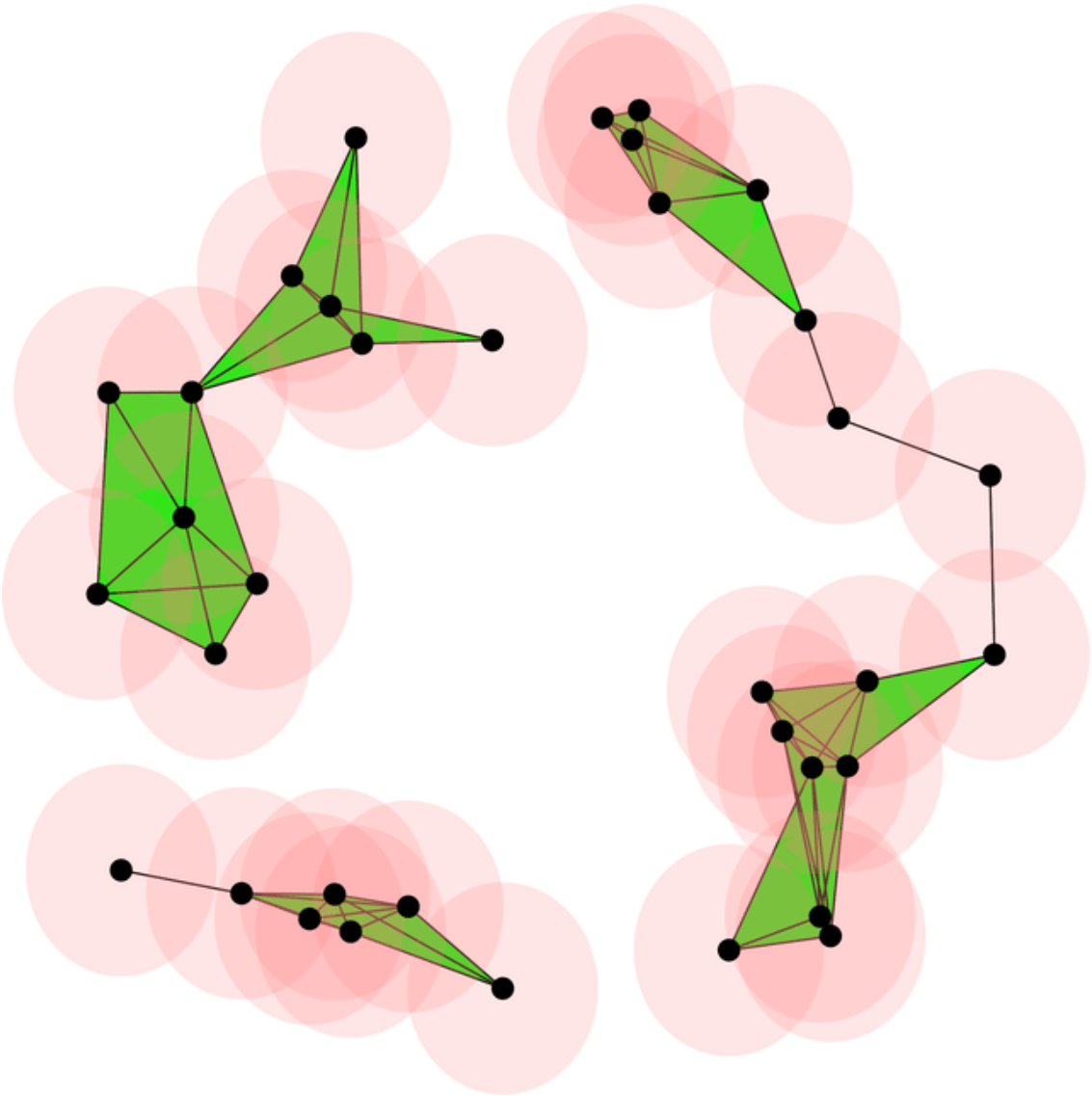
$\beta_0 = 1$
$\beta_1 = 1$
$\beta_2 = 0$

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



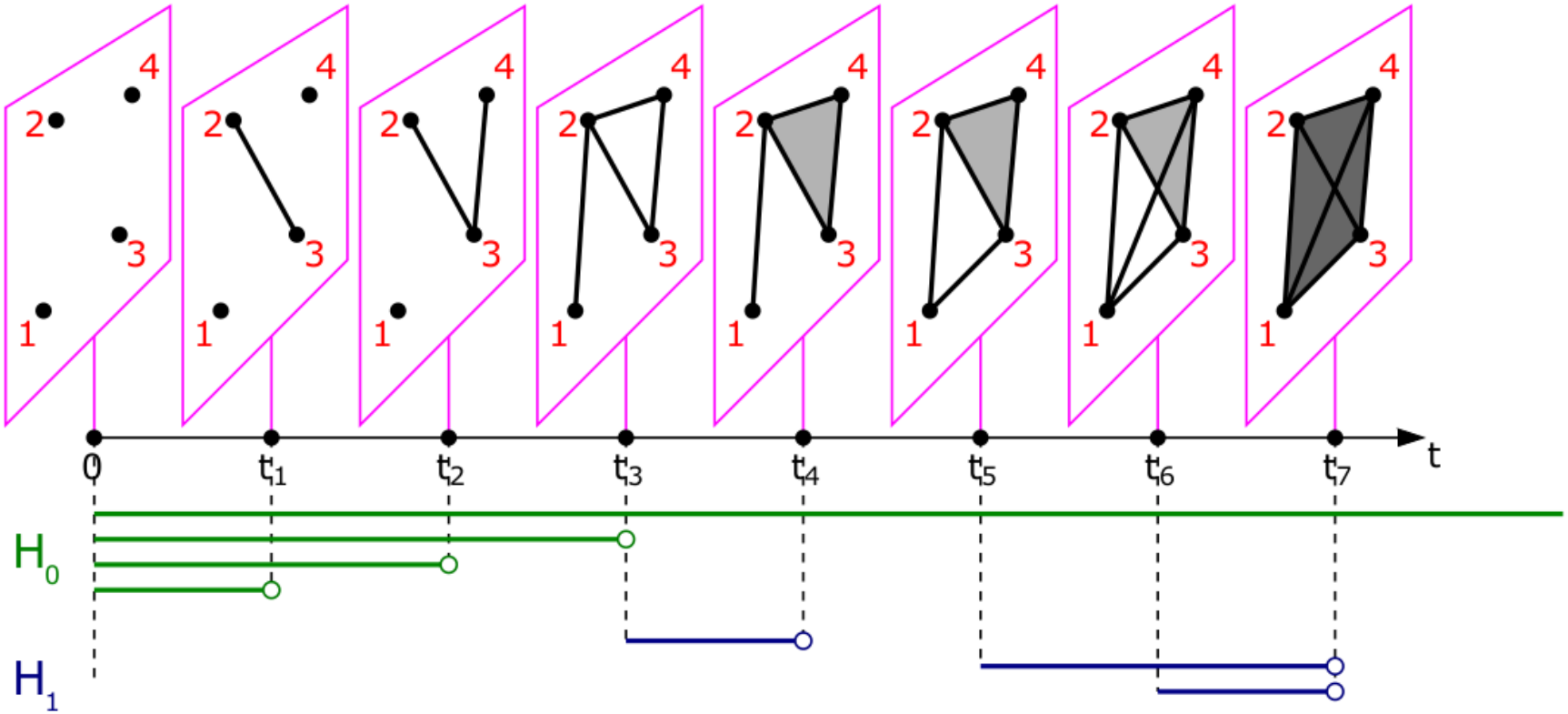
$\beta_0 = 1$
$\beta_1 = 2$
$\beta_2 = 1$

Как построить симплициальный комплекс?



Фильтрации и устойчивые гомологии

Фильтрация – коллекция комплексов.
Устойчивые гомологии – коллекция групп гомологий комплексов фильтрации.



Фильтрации и устойчивые гомологии

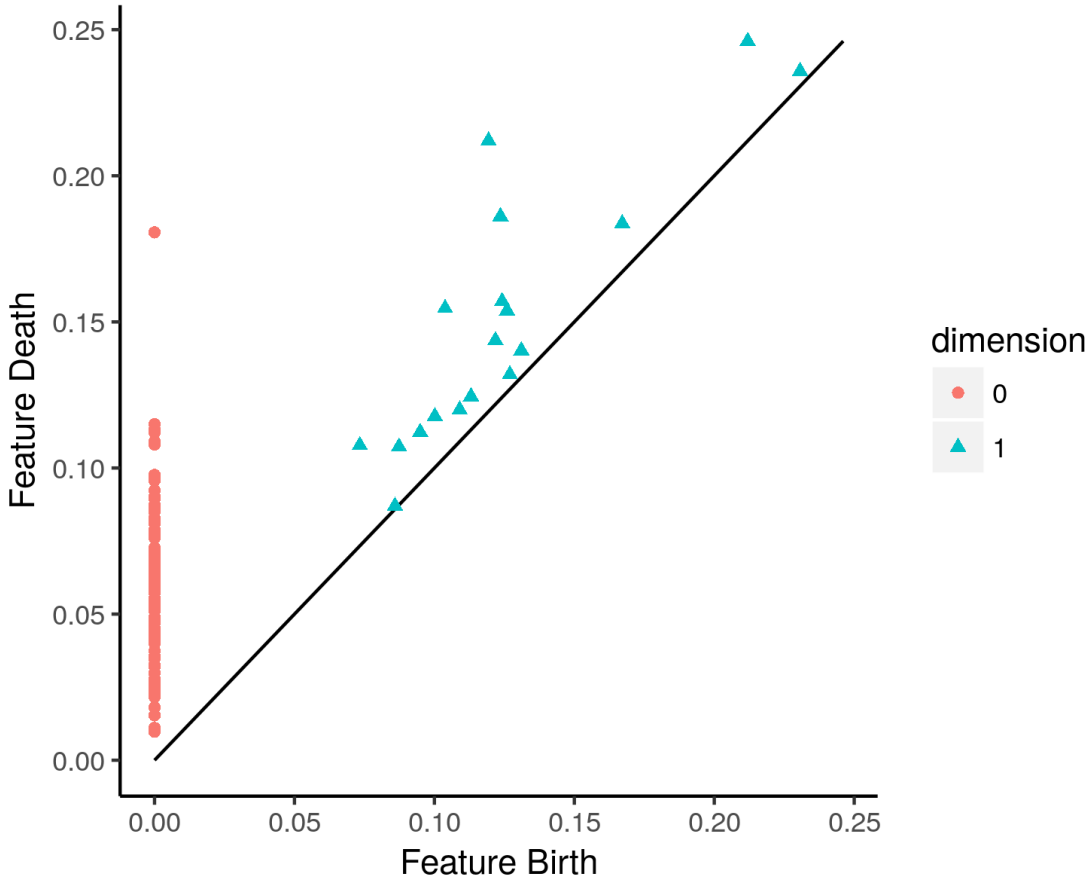


Рис. Диаграмма персистентности

Векторизация диаграмм персистентности

Диаграммы устойчивости с данной метрикой образуют метрическое пространство $\mathcal{D}$:

$$W_p(B, B') = \inf_{\gamma: B \to B'} \left(\sum_{u \in B} \|u - \gamma(u)\|_\infty^p \right)^{\frac{1}{p}}.$$

Векторизация диаграмм – это отображение $\varphi : \mathcal{D} \to V$, где V – нормированное векторное пространство.

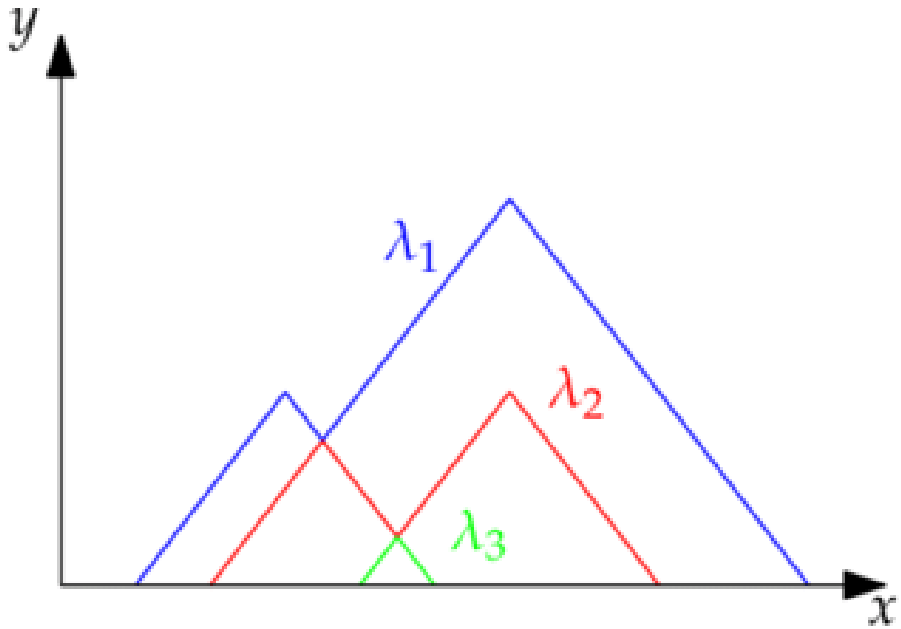


Рис. График некоторой векторизации

Алгоритм построения векторного представления

Исходные данные: Черное-белое изображение

Результат: Вектор размерности d

По изображению построить n разных кубических фильтраций

Для каждой *фильтрации* **выполнять**

- найти ее 0 и 1 кубические гомологии
- построить диаграмму устойчивости
- посчитать k разных векторных представлений

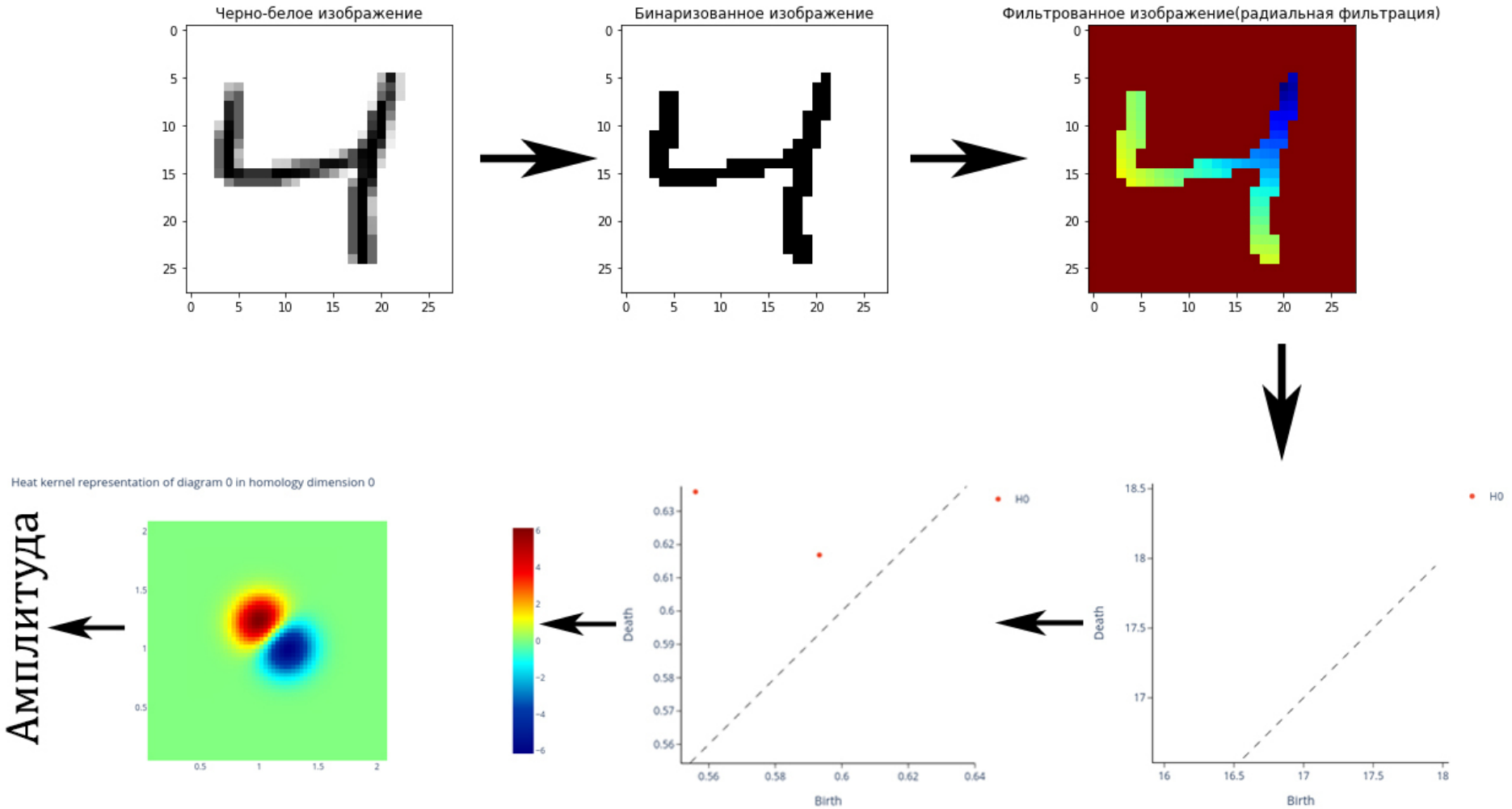
Конец

Собрать $d = 2nk$ представлений в один вектор

Использованное ПО

- Python
- Scikit-Learn
- Giotto-TDA
- NumPy
- Matplotlib

Процесс получения векторного представления



Сравнение результатов различных методов машинного обучения

Таблица: Значения базовых моделей на тренировочной и тестовой выборках

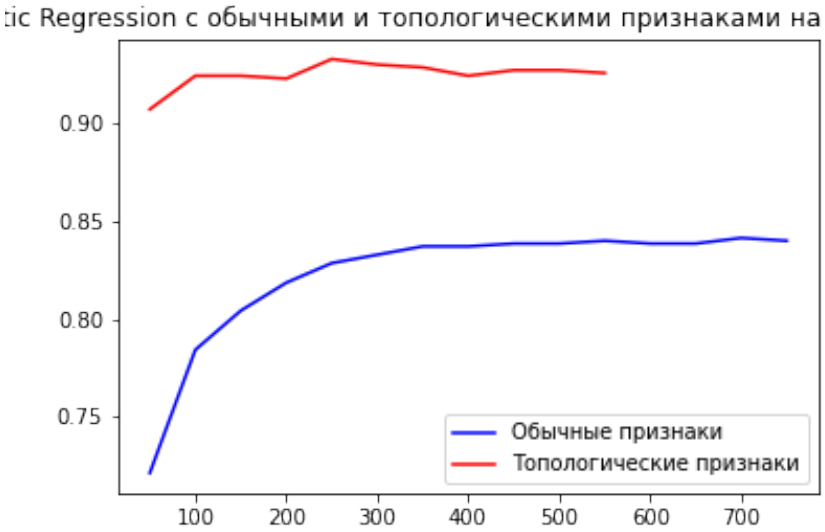
Название модели	Значение на тренировочной выборке	Значение на тестовой выборке
Логистическая регрессия	0.989	0.903
Метод опорных векторов	1.0	0.893
Случайный лес	1.0	0.89
LightGBM	1.0	0.9
XGBoost	1.0	0.883
CatBoost	1.0	0.893

Сравнение результатов различных методов машинного обучения

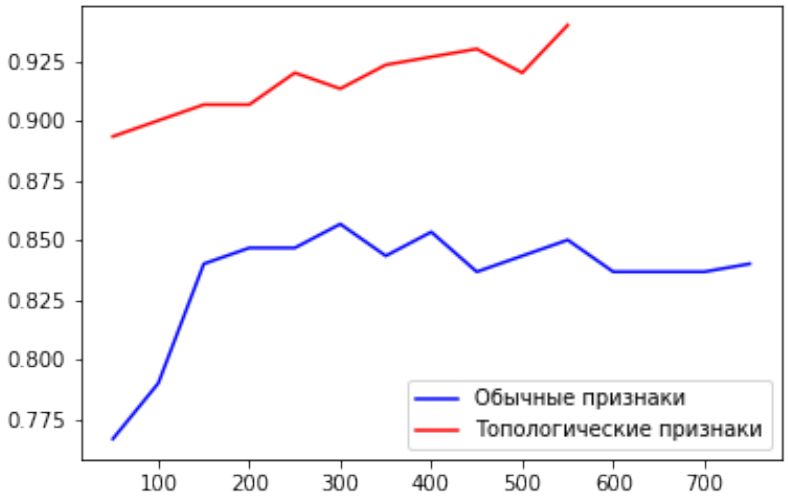
Таблица: Значения наилучших моделей, полученных в результате подбора параметров поиском по сетке, на тренировочной и тестовой выборках

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	0.88	0.87
LightGBM	0.908	0.917
XGBoost	0.907	0.897
CatBoost	1.0	0.927

Подбор гиперпараметров и отбор признаков



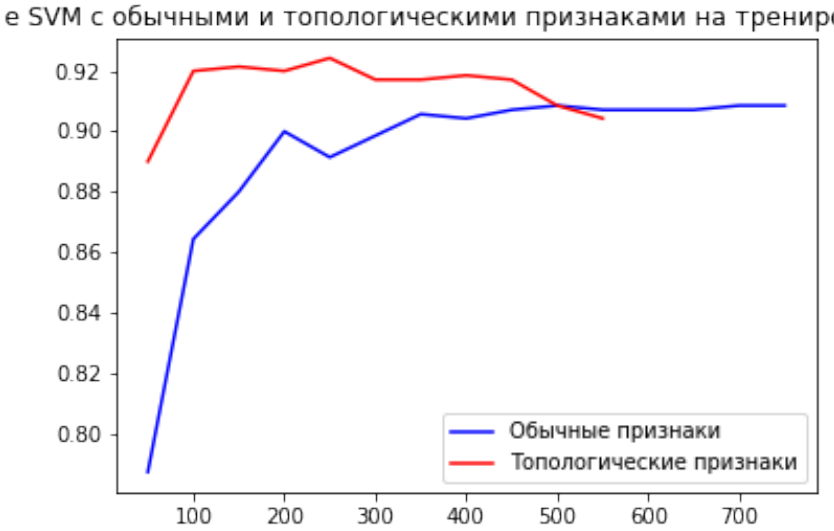
(a) Тренировочная выборка



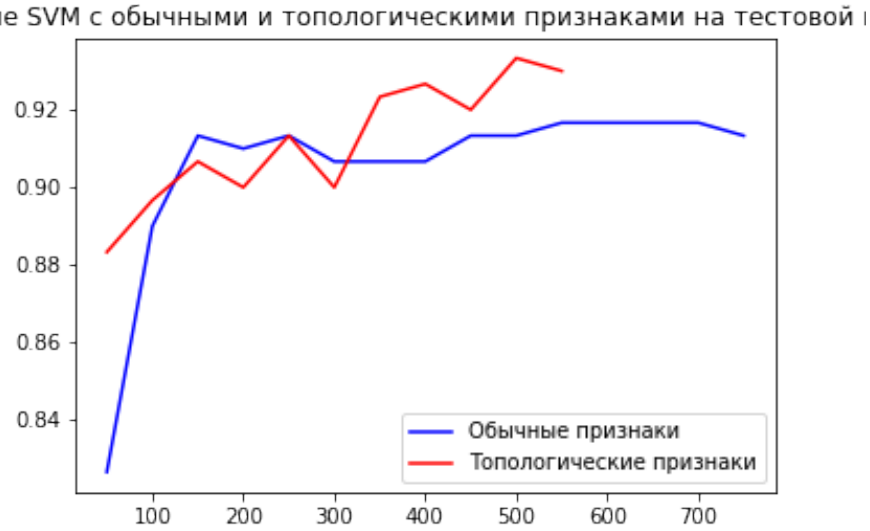
(b) Тестовая выборка

Рис. Логистическая регрессия

Подбор гиперпараметров и отбор признаков



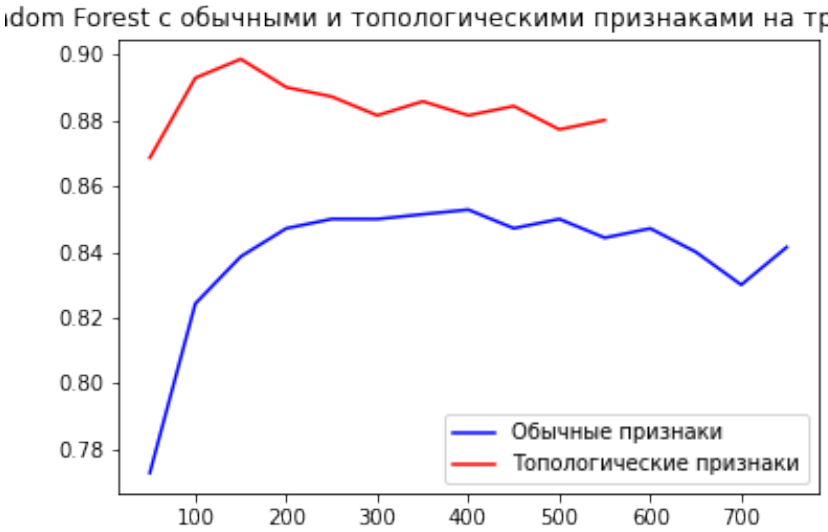
(a) Тренировочная выборка



(b) Тестовая выборка

Рис. Метод опорных векторов

Подбор гиперпараметров и отбор признаков

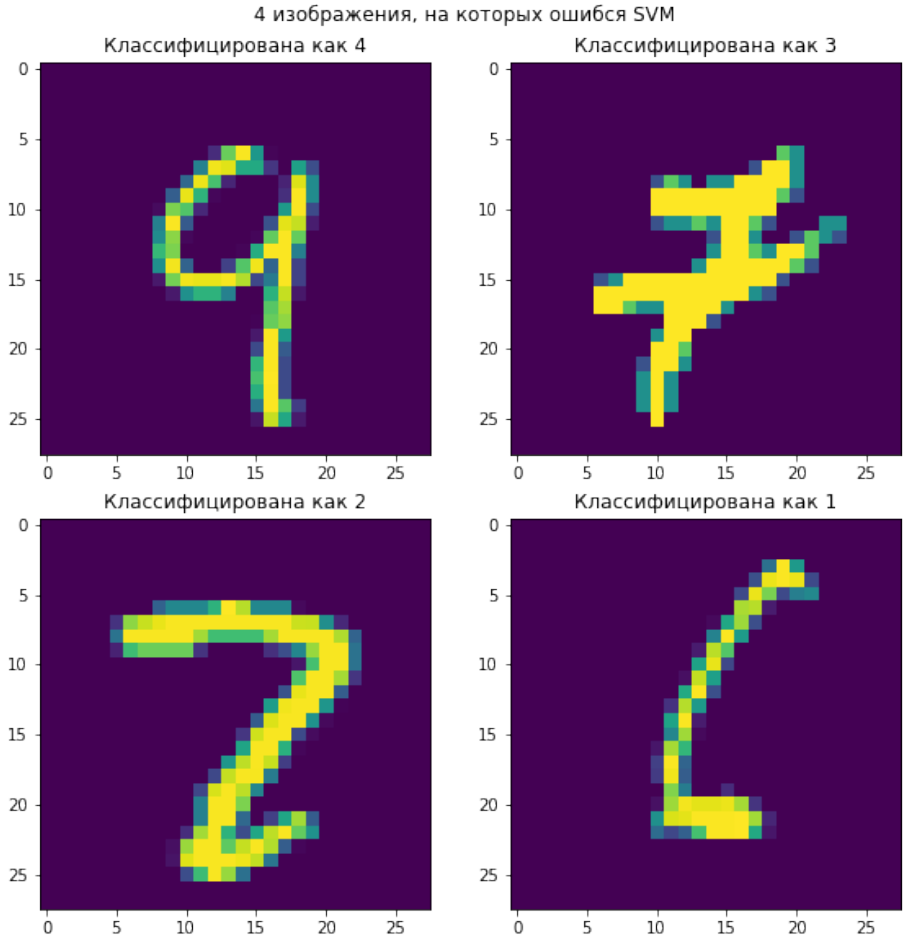


(a) Тренировочная выборка

(b) Тестовая выборка

Рис. Случайный лес

Первые несколько изображений, на которых ошибся классификатор



Результаты

- В ходе данной работы был изучен теоретический материал по алгебраической и прикладной топологии и машинному обучению.
- Было проведено сравнение пакетов топологического анализа данных.
- Был реализован алгоритм классификации датасета MNIST, который использовал только топологические характеристики рукописных цифр в качестве признаков.
- Было проведено сравнение моделей машинного обучения с разными признаками.

Выводы

- Подход, основанный на топологических характеристиках данных, показал свою работоспособность на примере изображений.
- Реализованный алгоритм является эффективным алгоритмом понижения размерности пространства признаков: он показывает более высокую точность классификации при меньшем числе признаков.

Спасибо за внимание!

Мои публикации на данную тему:

 Снопов П. М. — Применение алгебраической топологии в задачах анализа данных. — // Сборник трудов Международной научной конференции «Актуальные проблемы прикладной математики, информатики и механики». — Воронеж, 2020. — с. 1085—1093.

 Снопов П. М. — Сравнительный анализ пакетов для вычисления устойчивых гомологий. — // Межвузовская научная конференция молодых ученых и студентов «Математика, информационные технологии, приложения». — Воронеж, 2021. — с. 234—239.