**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**
**ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**
**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук

Образовательная программа «Прикладная математика и информатика»

УДК _____

**Отчет об исследовательском проекте**

на тему «Топологический анализ социальных данных»

(промежуточный, этап 1)

**Выполнил:**

студент группы БПМИ 198 _____ С.В. Пилипенко
Подпись И.О. Фамилия

05.02.2021
Дата

**Принял:**

руководитель проекта Айзенберг Антон Андреевич
Имя, Отчество, Фамилия

кандидат физико-математических наук, доцент
Должность, ученое звание

международная лаборатория алгебраической топологии и ее приложений
Место работы (Компания или подразделение НИУ ВШЭ)

Дата проверки _____2021 _____ _____
Оценка (по 10-тибалльной шкале) Подпись

**Москва 2021 г.**

# Содержание

# 1 Основые термины и определения

**Симплициальный комплекс** *Симплициальным комплексом* на конечном множестве вершин $M$ называется совокупность $K \subset 2^M$ подмножеств множества $M$, удовлетворяющая следующим двум условиям:

1. если $I \in K$ и $J \subset I$, то $J \in K$;
2. $\varnothing \in K$.

**Симплекс** *Симплексом* называются элементы симплециального комплекса $K$.

**Порядковая переменная** Переменная, которая принимает значения из конечного упорядоченного множества.

**Номинальная переменная** Переменная, которая принимает значения из конечного неупорядоченного множества.

# 2 Введение

Современные технологии становятся частью нашей повседневной жизни. Каждый из нас в кармане имеет мгновенный доступ к любому человечеством знанию, описанному в интернете; каждый может связаться с другим на каком бы расстоянии мы не были. И без этого уже не представить наш мир. Человечеству открываются невероятные горизонты бытия, расширяется кругозор и появляются всё больше возможностей для развития. Как личностного развития, так и образовательного.

Конечно, это повлияло на все сферы нашего общества. Образование не исключение, здесь мы видим множество изменений. Учителя имеют возможность проводить интерактивные уроки с помощью красочных презентаций и образовательных фильмов, дети могут достать любую законную литературу, администрация же подстраивается под новые реалии и организует онлайн классы. Почти в каждой школе есть свой Wi-Fi и компьютерный класс. Оценить, насколько изменения в образовании положительно влияют на общий процесс образования я попытаюсь с помощью современных технологий, методами машинного обучения и топологического анализа данных, на основе опросов, проведенных социологами в 2020 году. Основная задача данных опросов собрать данные об удовлетворенности учителей, школьников и администрации школы нововведениями, связанными с усовершенствованием технологий и мира.

В рамках данного проекта я буду искать интересную информацию про то, какое отражение возымело бурное развитие IT технологий на среднее образование с точки зрения школьников, учителей и сотрудников школьной администрации. Для этого я изучу основы топологии, которая дает теоретическое обоснование алгоритмам понижения размерности данных UMAP [0] и Mapper [0], и изображу форму данных, что должно помочь в их кластеризации. Также, я сравню результаты топологических алгоритмов с алгоритмами классического машинного обучения t-SNE [0] и PCA. В результате надеюсь увидеть некоторые не очевидные зависимости между ответами людей.

# 3 Обзор источников

Нашим основным источником информации касательно топологии является [0]. В этой книге рассматриваются темы вроде симплециальных комплексов, сиплексов, нерв-теоремы на уровне, достаточном для понимания работы алгоритмов топологического анализа данных UMAP, Mapper. Для задачи препроцессинга данных мы ориентировались на устоявшиеся практики современной Data Science, подробно рассмотренные в [0].

# 4 Основная часть

Начальным этапом проекта было ознакомление с предоставленными нам данными, полученных в результате опроса большого количества учеников, учителей и членов школьных администраций, проведенного в 2020 году. В анкетах содержалось большое количество вопросов касательно общего влияния цифровых устройств и IT инфрастуктуры школы на организацию учебного процесса с точки зрения различных его участников: школьников, преподавателей, директоров.

На этом этапе перед нами стояла задача предварительной обработки полученных от социологов данных, их нормализация и представление в виде облака точек в пространстве достаточно большой размерности. Из-за своей природы, полученные нами данные являлись весьма не структурированными и нуждались в дополнительной обработке. Так, например, были отброшены все ответы на вопросы, которые были даны в свободной формы: тематика секции школьника, названия онлайн-сервисов, которыми пользуются ученики. Среди оставшихся данных были отсеяны очевидные статистические выбросы, такие как противоречивые ответы на вопросы, ответы, в которых почти на все вопросы был дан ответ с одинаковой нумерацией (например, школьник на все вопросы отвечал «Да» или «1»). Над оставшимся массивом данных уже можно было проводить преобразования для отображения ответов каждого респондента в пространство $\{0,1\}^m$. Для этого мы воспользовались тремя разными алгоритмами: **Ordinal Encoding**, **One-Hot Encoding** и **Dummy Variable Encoding**.

## 4.1 Ordinal Encoding

В случае алгоритма **Ordinal Encoding**, каждой уникальной категории присваивается уникальное целое число. Например, категории «Да», «Нет» и «Затрудняюсь ответить» могут быть закодированы через последовательность 1, 2, 3. Поскольку такое кодирование является весьма естественным и легко обратимым, мы попробовали применить этот алгоритм для кодирования наших данных, однако быстро отказались от этой идеи. Поскольку такое кодирование не является бинарным, его результаты придется дополнительно нормировать, однако мы не сможем избавиться от относительного порядка на ответах (ответ «Нет» будет считаться более важным, чем ответ «Да»). Таким образом, в текущем виде данный алгоритм нам не подходит.

## 4.2 One-Hot Encoding

Как уже было сказано, нам важно сохранить отсутствие относительного порядка между ответами респондентов, чтобы не вводить будущую модель в заблуждение. В таких случаях, обычно, применяют алгоритм **One-Hot Encoding**, который преобразует упорядоченные данные в неупорядоченные посредством удаления каждой целочисленной категории и присваивания ей некоторого бинарного значения за каждое уникальную целочисленную категорию этой переменной [0]. То есть, полученные значения категорий 1, 2, 3 из результата работы предыдущего алгоритма, раскроются в следующую матрицу:

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Такой алгоритм кодирования данных прекрасно подходит для наших нужд.

## 4.3 Dummy Variable Encoding

Можно заметить, что в примере из предыдущего абзаца мы храним лишнюю информацию. В самом деле, нам совершенно не нужен третий столбец матрицы $A$, поскольку из матрицы

$$A' = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

однозначно восстанавливается принадлежность объекта к какой-то категории. То есть, если человек не ответил «Да» и не ответил «Нет», то мы сразу же делаем вывод, что он затрудняется ответить. Помимо этого, в некоторых случаях, кодирование через второй алгоритм может сделать матрицу вырожденной [0], что плохо сказывается на эффективности и точности некоторых алгоритмов классического машинного обучения (таких как линейная регрессия). В итоге, для наших наборов данных мы применили именно этот алгоритм с целью приведения ответов респондентов к булевым $m$-мерным векторам для последующего применения алгоритмов топологического анализа данных.

# Список литературы


0. *Alice Zheng A. C.* Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. — O'Reilly Media, 2018.

0. *Ayzenberg A.* Introduction to Simplicial Complexes and Homologies. — 2020.

0. *Kuhn M.*, *Johnson K.* Feature Engineering and Selection. — Chapman, Hall/CRC, 2019.

0. *Maaten L. van der*, *Hinton G.* Visualizing Data using t-SNE // Journal of Machine Learning Research. — 2008. — T. 9. — C. 2579—2605.

0. *McInnes L.*, *Healy J.*, *Melville J.* UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. — 2018. — URL: https://arxiv.org/pdf/1802.03426.pdf.

0. *Singh G.*, *Memoli F.*, *Carlsson G.* Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. — 2007. — DOI: 10.2312/spbg/spbg07/091-100.

# 5 Приложение

Таблица 1: Календарный план работ по проекту

| Дата | Название задачи |
|---|---|
| 12.03.2021 | Ознакомление с методичкой по симплициальным комплексам и гомологиям |
| 22.03.2021 | Изучение работы классических алгоритмов анализа данных на соц. опросах |
| 30.03.2021 | Изучение работы топологических алгоритмов анализа данных на соц. опросах |
| 14.04.2021 | Сравнительный анализ результатов и их визуализация |
| 26.04.2021 | Отчет по КТ2 |
| 01.05.2021 | Изучение алгоритмов машинного обучения с учителем |
| 07.05.2021 | Разработка модели (нейронной сети), которая будет предсказывать по ответам ученика его школу |
| 15.05.2021 | Итоговый отчет по проекту |