Реферат

Наша работа была мотивированна изучением молодой и активно развивающейся областью топологии — топологического анализа данных (TDA), возникшая посредством множества работ в вычислительной геометрии и прикладной топологии. В основе TDA лежит идея о том, что топология и геометрия обеспечивают крайне эффективный подход к получению надежной качественной и, иногда, количественной информации о структуре данных. TDA стремится предоставить достаточно обоснованные математические, статистические и алгоритмические методы для вывода, анализа и использования топологических структур, лежащих в основе данных, которые часто представлены в виде точек в метрическом пространстве.

Целью исследования нашей работы были социальные опросы школьников, руководителей и учителей с помощью которых мы надеялись выявить паттерны посредством использования методов TDA, в частности, UMAP и PCA. В первом этапе работы, мы занялись очисткой данных от статистических выбросов и изучением методички по топологии, во втором этапе мы сформулировали гипотезы, изучили алгоритмы UMAP и PCA и реализовали их на наших, уже очищенных датасетах, в третьем этапе мы занялись обоснованием корректности нашей работы.

Результаты нашей работы показали, что ответы участников опроса не зависели от пола, класс учеников не влиял на их мнение и распределение ответов сотрудников школьной администрации не зависит от должности.

Содержание

1 Основые термины и определения **4**

2 Введение **5**

3 Теоретическая часть **6**

3.1 Ordinal Encoding 6

3.2 One-Hot Encoding 6

3.3 Dummy Variable Encoding 7

3.4 Principal Component Analysis 7

4 Описание вычислительного эксперимента **9**

4.1 Распределение ответов на опрос не зависит от гендера участников 9

4.2 Распределение ответов учеников на опрос не зависит от класса, в котором ученик обучается 11

4.3 Распределение ответов членов школьной администрации зависит от должности 12

5 Заключение **13**

1 Основые термины и определения

Метрическое пространство *Метрическим пространством* (M, ρ) называют множество M с функцией $\rho : M \times M \to \mathbb{R}_+$, называющейся расстояние, для любых $x, y, z \in M$, что:

1. $\rho(x, y) \geq 0$ и $\rho(x, y) = 0$ iff $x = y$
2. $\rho(x, y) = \rho(y, x)$
3. $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$

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.

Кросс-Энтропия кол-во информации, в среднем, необходимое для идентификации событий из распределения.

2 Введение

Следующим этапом моей работы над проектом были непосредственные манипуляции с данными с целью выявить в них какие-то закономерности. Для этого, пока что, были рассмотрены алгоритмы UMAP [4] и PCA [2]. Для достижения этой цели, нужно было написать код, который берет обработанный массив данных, представленный матрицей $A \in \mathbb{R}^{m \times n}$ со стандартизированными столбцами (из каждого вычли его среднее, и поделили на дисперсию). Чтобы упростить читаемость отчета и работу с исходными данными, я закодировал все «содержательные» вопросы — такие вопросы, по которым очень сложно делать хорошие предсказания (например, ввиду их большого количества), или те вопросы, по которым мы не будем делать предсказания и красить точки в пространстве. Также я выделил несколько гипотез относительно данных, которые хотел проверить:

1. Распределение ответов на опрос не зависит от гендера участников.
2. Распределение ответов учеников на опрос не зависит от класса, в котором ученик обучается.
3. Распределение ответов членов школьной администрации зависит от должности.

3 Теоретическая часть

Для начала нам нужно было очистить данные от статистических выбросов – таким образом, мы избавились от неоднозначных ответов, на подобии тематики секции, названий онлайн-сервисов, которые использовали ученики. Так же мы убрали из данных противоречивые ответы на вопросы 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**.

3.1 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.

3.2 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 [1]. 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.

3.3 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 [3], 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.

3.4 Principal Component Analysis

Большие и массивные наборы данных стали не редкостью и часто включают в себя измерения на многих переменных. Зачастую можно значительно уменьшить количество переменных, при этом сохранив большую часть информации в исходном наборе данных. Метод главных компонент (PCA в дальнейшем), вероятно, является наиболее популярным и широко используемым методом уменьшения размерности для этого. Положим, что у нас есть k измерений на векторе x из p случайных переменных и мы хотим уменьшить его размерность с p к q, где q, обычно, намного меньше чем p. PCA достигает этого посредством нахождения таких линейных комбинаций $a_1'x, a_2'x, ..., a_q'x$, называемых главными компонентами, которые последовательно имеют максимальную вариацию данных, не связанных с предыдущими $a_k'x$. Решая эту максимизационную задачу мы находим, что векторы $a_1, a_2, ...a_q$ являются собственными значениями матрицы ковариций данных S, которая определена как

$$S = \frac{1}{p}\sum_{n=1}^{p}(x_n - \overline{x})(x_n - \overline{x})^T$$

(где $\overline{x} = \frac{1}{p}\sum_{n=1}^{p} x_n$, т.е среднее) Собственные значения дают представляют собой дисперии главных компонент, а отношение суммы первых q собственных значений к сумме дисперсий

всех p изначальных переменных является долей общей дисперсии в исходном наборе данных, приходящиеся на q главных компонент.

Пример

Таким образом, мы можем можем привести пример из [2]

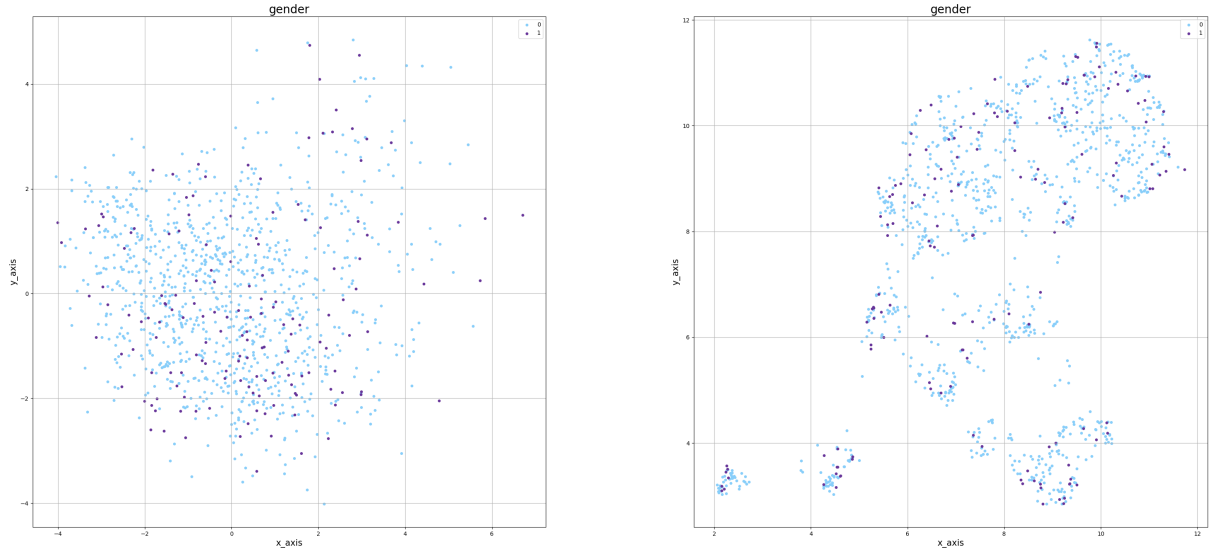
Переменная	a_1	a_2
x_1	0.34	0.39
x_2	0.34	0.37
x_3	0.35	0.10
x_4	0.30	0.24
x_5	0.34	0.32
x_6	0.27	-0.24
x_7	0.32	-0.27
x_8	0.30	-0.51
x_9	0.23	-0.22
x_10	0.36	-0.33

Данные состоят из оценок между 0 и 20 для 150 детей возраста $4\frac{1}{2}$ – 6 лет с острова Уайт по 10 предметам. Пять тестов были вербальными и пять перфомативными. Наша таблица показывает векторы a_1 и a_2, которые являются двумя главными компонентами для этих данных. Первая компонента является линейной комбинацией первых десяти оценок с примерно равным весом (0.36 — максимум, 0.23 минимум) для каждой оценки. Сама по себе, эта компонента оценивает около 48% оригинальной изменчивости. Вторая компонента сравнивает пять вербальных тестов и пять перфомативных. Это учитывает ещё 11% изменчивости. Эта форма говорит нам, что после того как мы учли общие способности детей, следующий наиболее важный для нас (линейный) источник изменчивости это разница между детьми, которые хорошо себя показывают в вербальных тестах, относительно успеваемости детей, показатели которых имеют обратный паттерн.

4 Описание вычислительного эксперимента

Для того, чтобы проверить выдвинутые гипотезы, я построил и проанализировал 12 графиков: на каждом наборе данных (результаты опросов школьников и администрации) я применил алгоритмы UMAP и PCA, а потом сравнил их результаты. Интересно, что распределения у студентов, полученные с помощью PCA и UMAP, практически не отличаются друг от друга, в то время как распределение ответов администрации может меняться в зависимости от того, какую целевую переменную мы предсказываем. Этот эффект объясняется тем, что я использовал две различные техники кластеризации в процессе работы с данными. В случае администрации, я выделял целевую переменную, оставшиеся кодировал, нормализировал, а потом понижал размерность. В случае же ответов учеников, я заранее определил набор целевых переменных, по которым буду различать результаты эксперимента, и понижал размерность оставшихся данных.

4.1 Распределение ответов на опрос не зависит от гендера участников

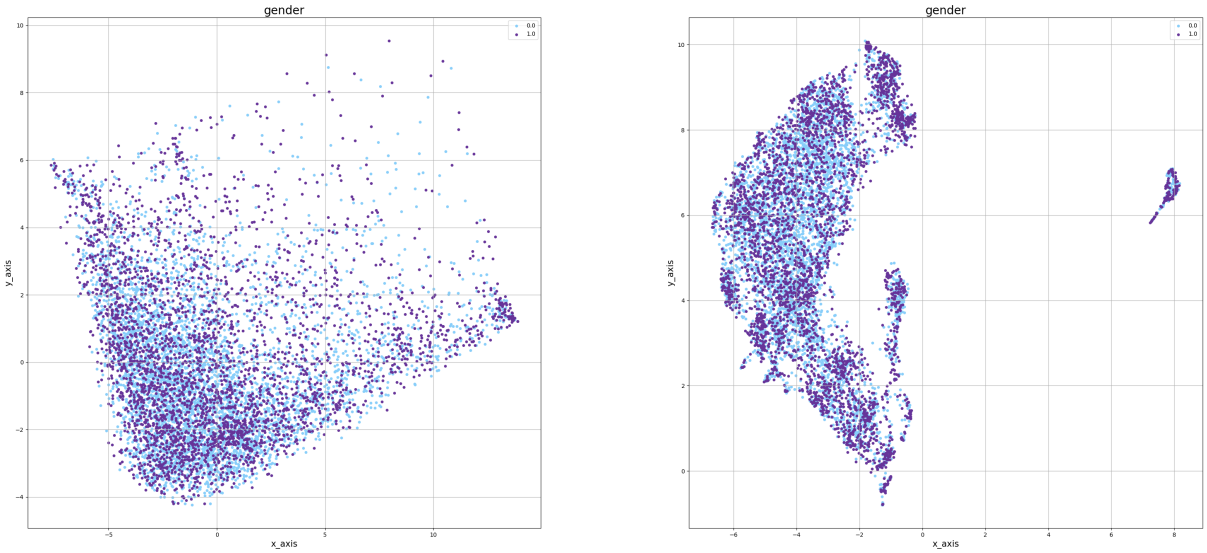


(a) PCA (b) UMAP

Рис. 1: Распределение ответов администрации в зависимости от пола

На построенных распределениях для администрации (рис. 1a и 1b) мы можем видеть, что вне зависимости от гендера респондентов, распределяются по кластерам они примерно

одинаково, с той лишь разницей, что женского персонала в составе администрации школ сильно больше, чем мужского. Аналогичное распределение мы можем наблюдать и на ответах школьников (рис. 2a и 2b), только в этом случае у нас примерно одинаковое количество респондентов разного пола. Таким образом, гипотеза подтверждается, ответы участников действительно не зависят от пола.

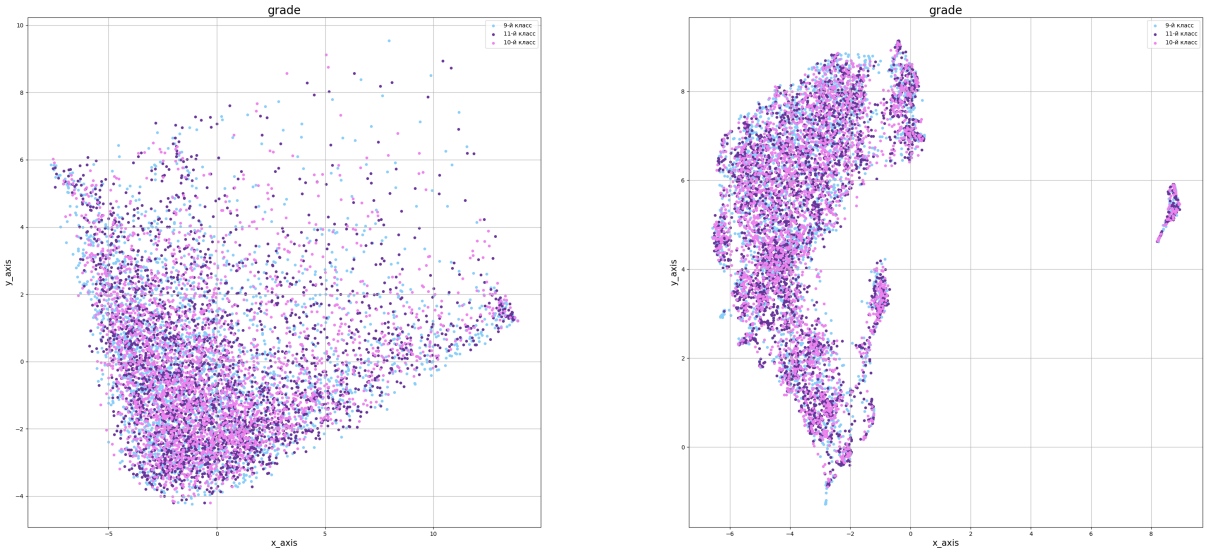


(a) PCA

(b) UMAP

Рис. 2: Распределение ответов школьников в зависимости от пола

4.2 Распределение ответов учеников на опрос не зависит от класса, в котором ученик обучается



(a) PCA

(b) UMAP

Рис. 3: Распределение ответов студентов в зависимости от класса

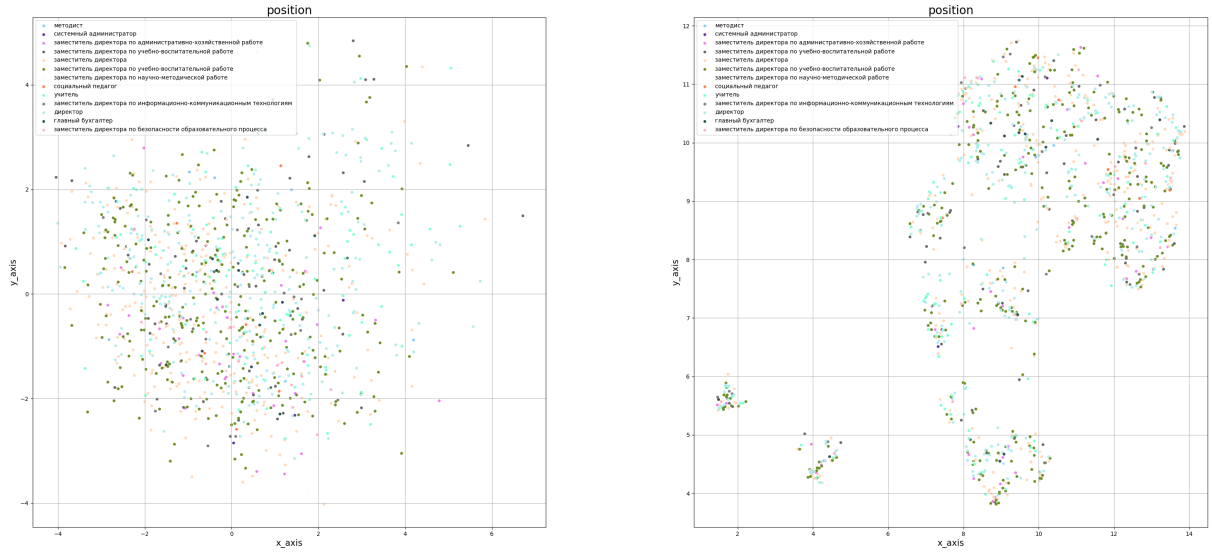
На рис. 3a и 3b мы можем заметить две вещи:

1. Распределение абсолютно не отличается от того, которое было в секции 4.1, что не удивительно, учитывая тот факт, что я просто подменил целевую переменную, по которой раскрашиваю точки в пространстве.

2. Распределение по классам весьма равномерное в двух кластерах в случае рис. 3b.

Таким образом, класс учеников не влияет на их ответы в опросе про техническую оснащенность школы. Это, на самом деле, весьма логично, ведь если школа использует IT технологии в образовательном процессе отдельной параллели, то она будет эти же технологии распространять и на все остальные параллели.

4.3 Распределение ответов членов школьной администрации зависит от должности



(a) PCA (b) UMAP

Рис. 4: Распределение ответов администрации школы в зависимости от должности

Как мы можем наблюдать, моя гипотеза не подтвердилась, поскольку почти все должности присутствуют в каждом из полученных кластеров на рис. 4a и 4b. Это говорит, как я предполагаю, о широкой осведомленности почти каждого сотрудника школьной администрации о технологической оснащенности школы.

5 Заключение

В результате проведенных нами вычислительных экспериментов, мы установили, что ответы респондентов не зависят от пола. Также, мы заметили, что должность респондентов, являющихся сотрудниками школьной администрации, не влияет на их ответы, что является индикатором того, что все сотрудники осведомлены о процессах в своей школе одинаково широко. Тоже самое можно сказать и про школьников, что, в целом, было нами ожидаемо.

В тоже время, разные алгоритмы кластеризации давали очень разный результат на одном и том же наборе данных. Алгоритм UMAP каждый раз находил больше кластеров данных, чем, например, PCA, который мог не найти ни одного кластера во входных данных. Однако, стоит отметить, что ни один из найденных с помощью UMAP кластеров не соответствовал нашим гипотезам, поэтому причина их наличия в данных остается открытым вопросом.

Список литературы

1. *Alice Zheng A. C.* Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. — O'Reilly Media, 2018.
2. *Jolliffe I.* Principal Component Analysis // International Encyclopedia of Statistical Science / под ред. M. Lovric. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2011. — С. 1094—1096. — ISBN 978-3-642-04898-2. — DOI: 10.1007/978-3-642-04898-2_455. — URL: https://doi.org/10.1007/978-3-642-04898-2_455.
3. *Kuhn M., Johnson K.* Feature Engineering and Selection. — Chapman, Hall/CRC, 2019.
4. *McInnes L., Healy J., Melville J.* UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. — 2018. — URL: https://arxiv.org/pdf/1802.03426.pdf.