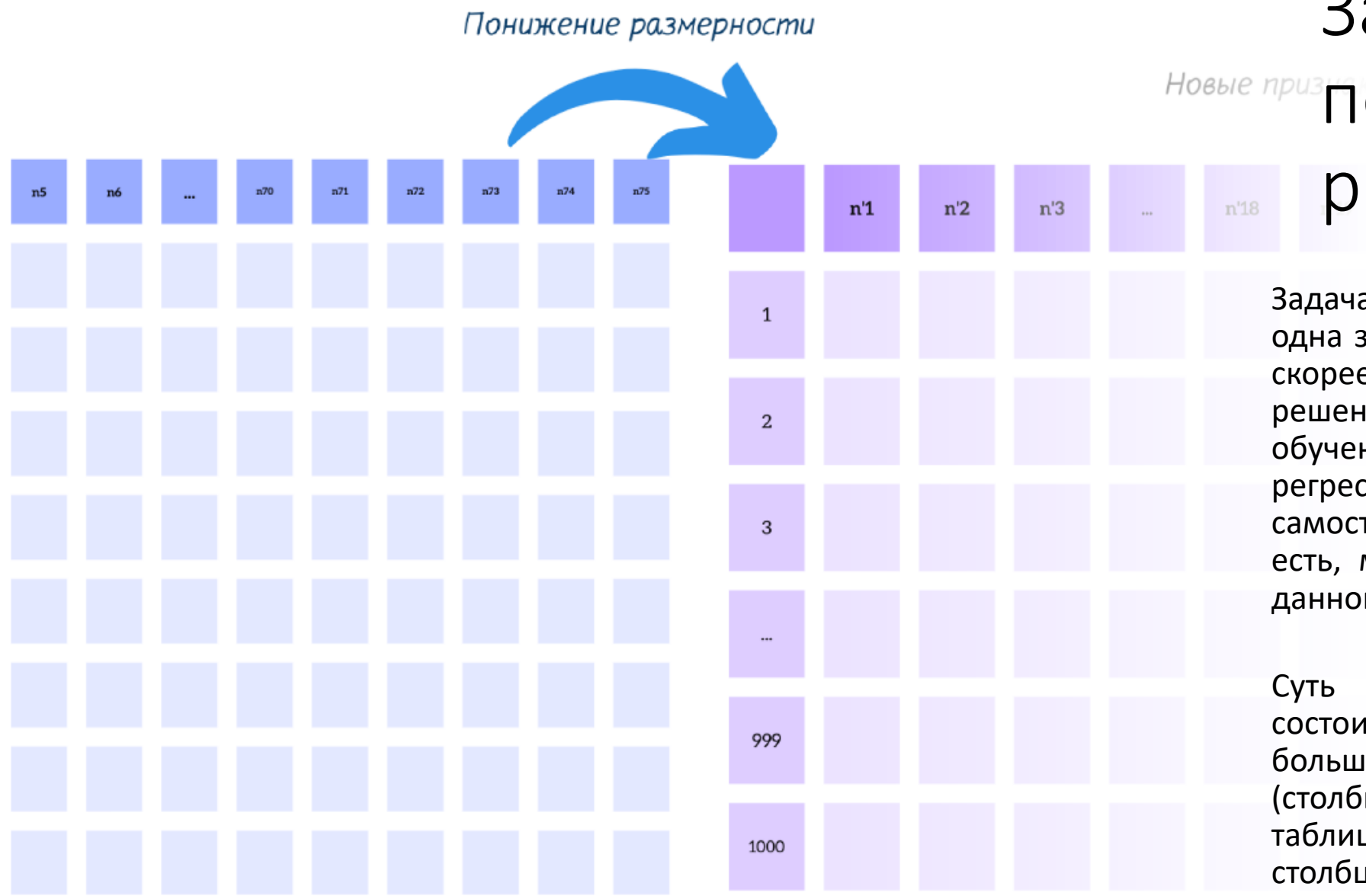
Лекция 10. Задача понижения размерности

2025/2026 учебный год

Доцент кафедры ИВЭ, Махно В.В.

ИНСТИТУТ
МАТЕМАТИКИ
МЕХАНИКИ
КОМПЬЮТЕРНЫХ
НАУК
имени И.И. Воровича



Задача понижения размерности

Задача понижения размерности (еще одна задача обучения без учителя)— это скорее вспомогательная задача для решения других задач машинного обучения, например классификации, регрессии или кластеризации, однако и самостоятельные применения у нее тоже есть, мы рассмотрим их ближе к концу данного блока.

Суть задачи понижения размерности состоит в том, чтобы имея данные с большим количеством признаков (столбцов), преобразовать их в новую таблицу с меньшим количеством столбцов. Количество строк (объектов) при этом останется неизменным.

Для чего же нужно понижение размерности?

- Сжатие данных

Большая таблица данных может занимать много места на жестком диске, и уменьшив количество признаков в N раз, мы уменьшим размер файла с данными в те же самые N раз.

- Ускорение предсказаний

Если алгоритму предсказания (например, алгоритму предсказания ухода клиента или алгоритму кластеризации) нужно обработать тысячи признаков, это будет занимать гораздо больше времени, чем если он будет обрабатывать десятки признаков. При этом во многих задачах, особенно связанных с онлайн-сервисами, существуют ограничения на время выполнения предсказаний, например поисковой запрос должен выполняться за доли секунды.

- Визуализация данных

Если алгоритм понижения размерности составит новую таблицу с двумя столбцами, такие данные будет легко визуализировать, отложив по осям два новых признака.

- Более компактное и «правильное» описание объектов

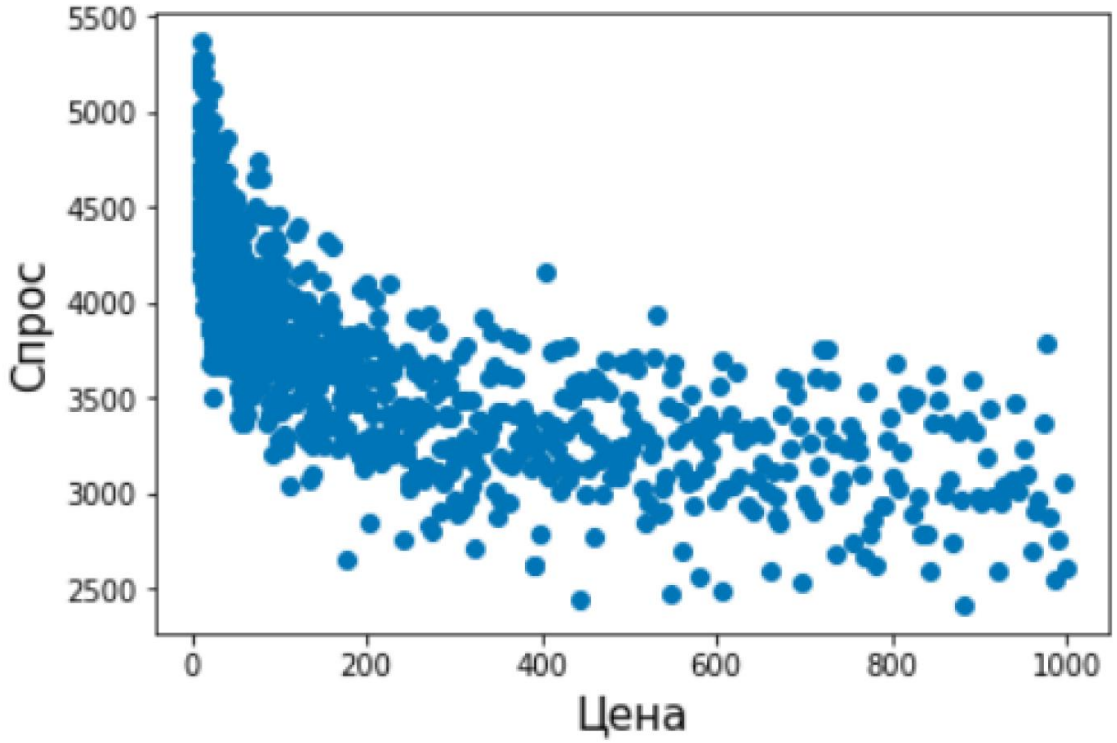
- Новые признаки могут описывать объекты более емко, чем исходные, что упростит работу другим алгоритмам. Более того, среди исходных признаков могут быть такие, которые ухудшают предсказания, и при понижении размерности эти признаки будут удалены.

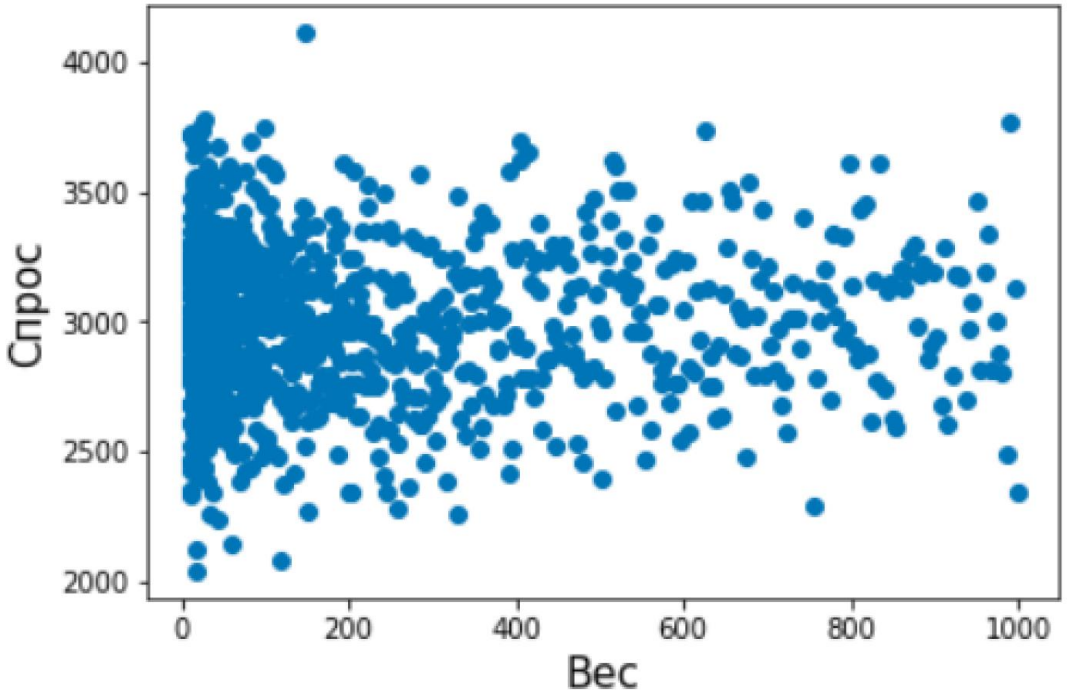
Отбор признаков

Обычно выделяют два типа алгоритмов понижения размерности: алгоритмы отбора признаков и алгоритмы выделения новых признаков на основе исходных. Первые просто удаляют столбцы из таблицы, а вторые вычисляют новые столбцы по формулам, включающим все исходные столбцы. Сначала сосредоточимся на первом типе.

Метод фильтрации признаков

Самые простые методы отбора признаков подразумевают анализ каждого признака отдельно и называются методами фильтрации признаков. Например, если у нас есть признаки «стоимость товара» и «вес товара», то мы можем построить графики стоимость-спрос и вес-спрос и по ним постараться увидеть, зависит ли спрос от какой-либо величины.





Оберточные методы отбора признаков

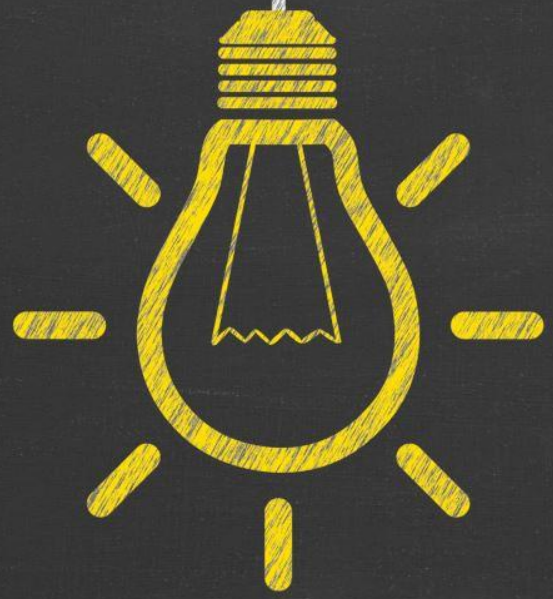
Эти методы — более сложный отбор признаков, включающий обучение алгоритмов и измерение их качества на тестовой выборке. Например, чтобы оценить, важен ли признак «вес товара», оберточный метод сначала обучит алгоритм на данных, включающих этот признак, затем на данных без этого признака (столбец «вес товара» удален) и сравнит, ухудшилось ли качество: если спрос на товары стал предсказываться менее точно, значит, признак «вес товара» важный, иначе он удаляется.

Встроенные методы отбора признаков

Это методы, подразумевающие, что алгоритм обучения сам может определить, какие признаки важные, а какие нет. В задаче регрессии мы обсуждали линейные модели: те, которые умножают значения признаков на веса и складывают. Если у какого-то признака после обучения получится вес, равный нулю, это значит, что признак не влияет на предсказание и его можно удалить — это и есть пример встроенного метода. А точнее, встроенным методом является регуляризация — специальный механизм, который помогает настроить в линейных моделях такие веса, что среди них будет много нулевых.

Выделение признаков

Методы отбора признаков выполняют понятное действие — удаляют столбцы. Методы же выделения новых признаков делают более сложную операцию: они создают новые столбцы, которые вычисляются по формулам, зависящим от имеющихся в данных столбцов.



Метод главных компонент

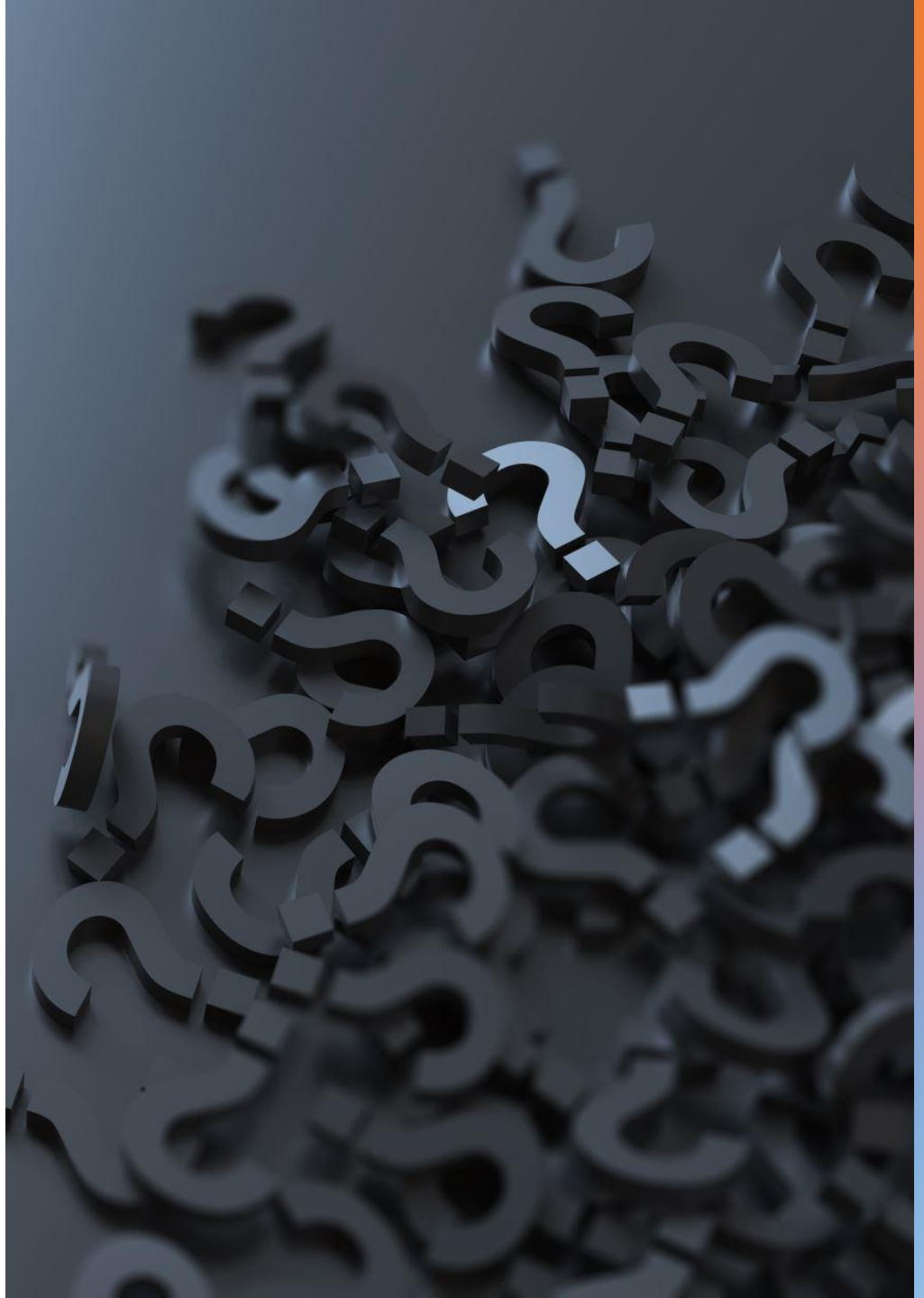
Один из наиболее популярных методов выделения признаков называется метод главных компонент (англ. Principal Component Analysis, PCA): он задает новые признаки как линейные формулы от исходных. Иными словами, каждый новый признак будет равен сумме исходных признаков, умноженных на веса, и веса настраиваются в процессе обучения. Очень похоже на линейные модели регрессии, только в регрессии целевая переменная известна, а здесь алгоритм сам «придумывает», что будет означать полученная сумма. Достоинства и недостатки такие же, как у других линейных методов: метод работает быстро, но выделяет слишком простые зависимости. Похожим образом на PCA работает другой алгоритм — Singular Value Decomposition, SVD.

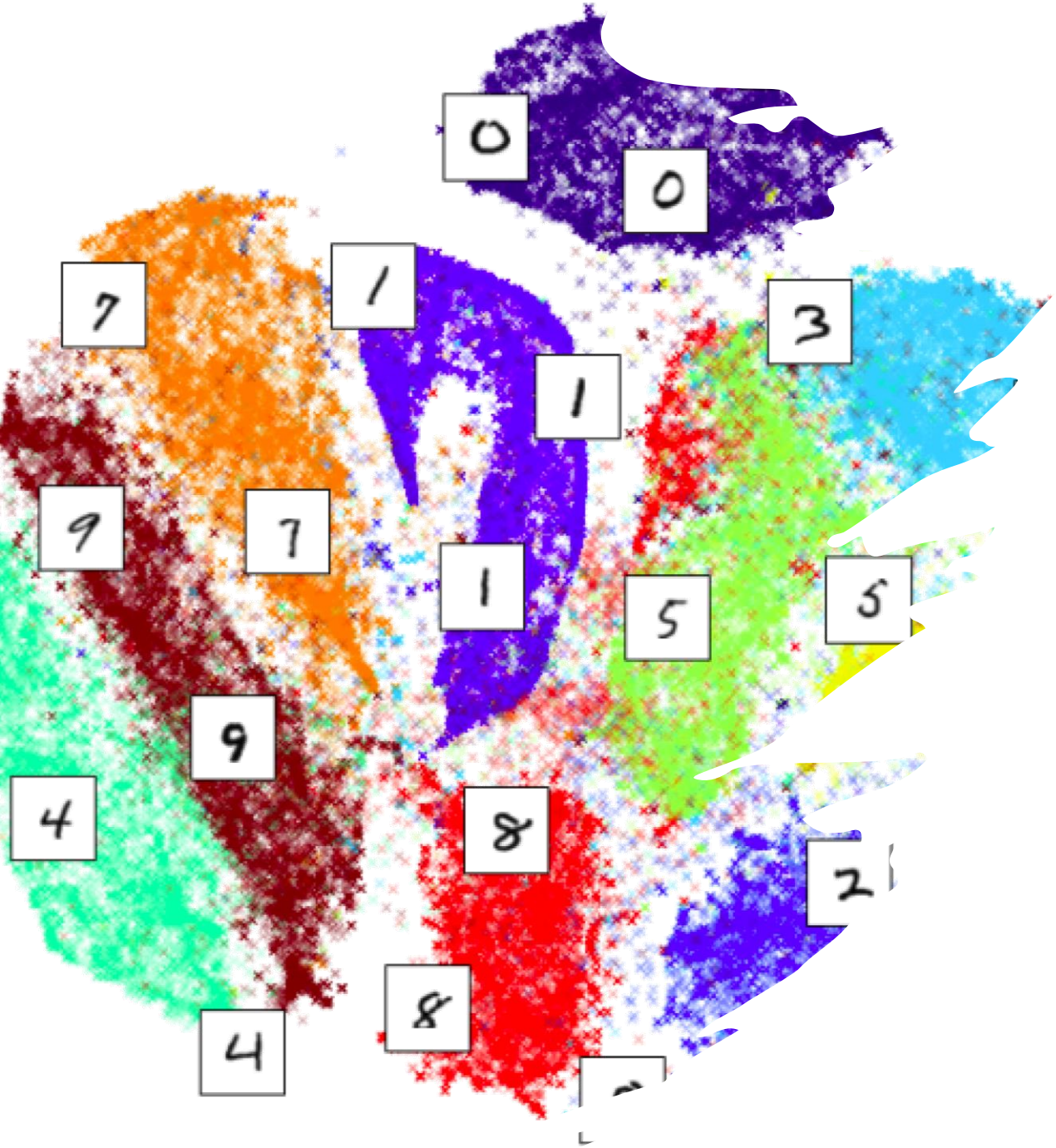
Автокодировщик

Автокодировщик (autoencoder) - это специальная архитектура нейронных сетей, которые используют вместо линейной формулы. Такая архитектура также позволяет осуществить обучение без учителя с использованием алгоритма обратного распространения ошибки. Автокодировщики часто применяются для видео или изображений, чтобы, например, убирать лишний шум, находить похожий контент или искать материалы по текстовому запросу.

Понижение размерности в работе с текстами

Для выделения признаков из текстов часто используется тематическое моделирование, например алгоритмы LSA (Latent Semantic Analysis) и LDA (Latent Dirichlet Allocation). В этом случае сам текст представляется как набор слов (неупорядоченный), иными словами, вычисляются частоты слов. Новые признаки задают темы: один новый признак — одна тема, при этом каждый объект (текст) может относиться к нескольким темам. Например, текст может быть одновременно про политику, экономику и немножко литературу.





Визуализация данных

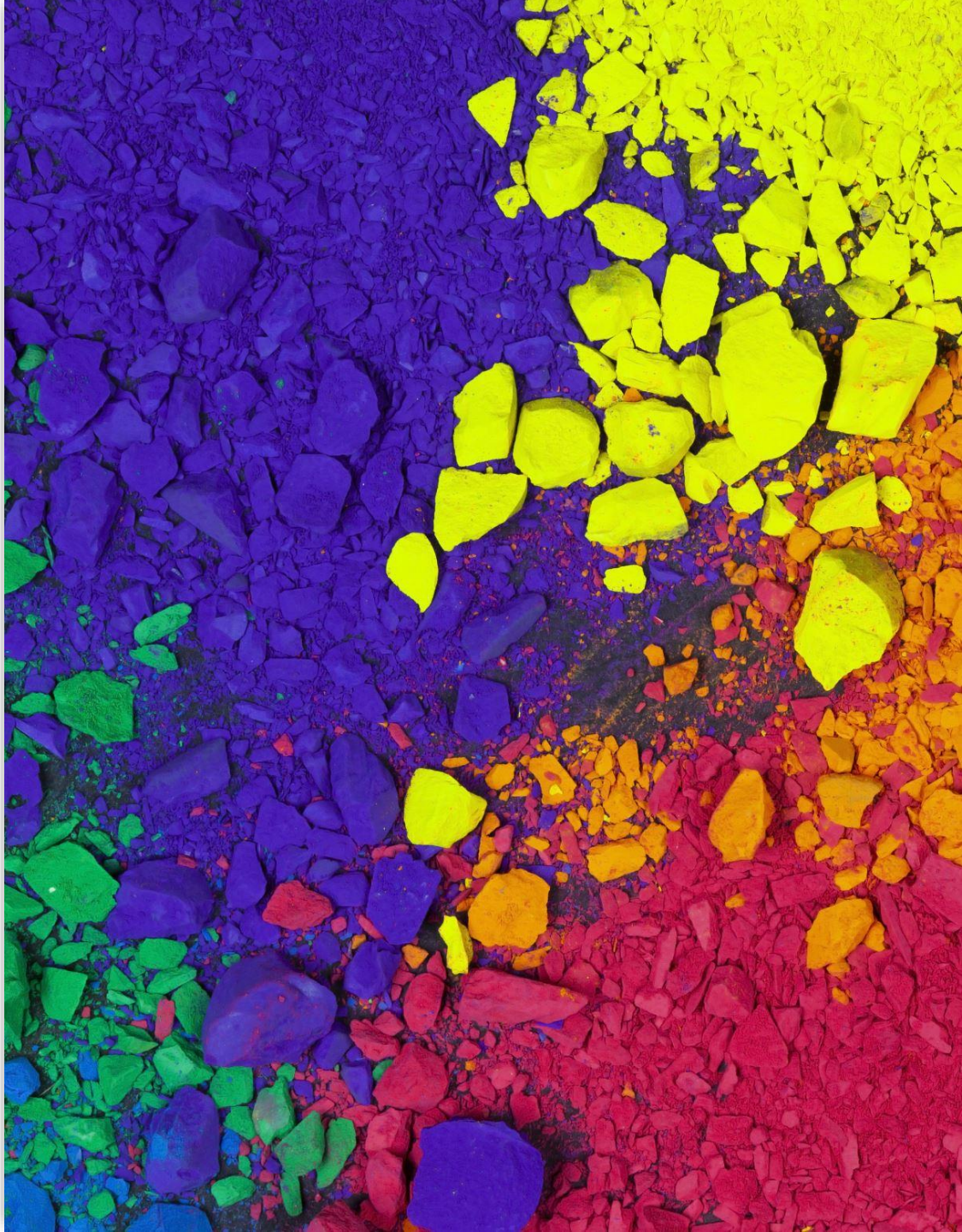
- Многомерное шкалирование

Метод под названием многомерное шкалирование (Multidimensional Scaling, MDS) старается найти такие новые признаки так, чтобы схожести между объектами, измеренные по исходным признакам), были примерно такими же, как схожести между объектами, измеренные по новым признакам.

Применение визуализации

Если выполнялась кластеризация объектов, на такой визуализации можно отобразить ее цветами и оценить, разделились ли объекты по цветам, как на изображении выше (качественная кластеризация), или цвета перемешаны (плохая кластеризация). Однако это будет лишь приблизительным способом оценки качества кластеризации.

Визуализация с помощью t-SNE позволяет посмотреть на данные «свысока», но окончательных выводов по ней лучше не делать, потому что значения осей интерпретировать невозможно.



Бесплатный курс от Сбера по генеративному искусству

https://courses.sberuniversity.ru/generative_art?utm_source=tg&utm_medium=organic&utm_campaign=courses&utm_content=gen_i&utm_term=01_09_2023