Методы понижения размерности

Машинное обучение

Наиль Баширов

Московский физико-технический институт

28 октября 2025

ФФПМИ

Сингулярное разложение

Определение

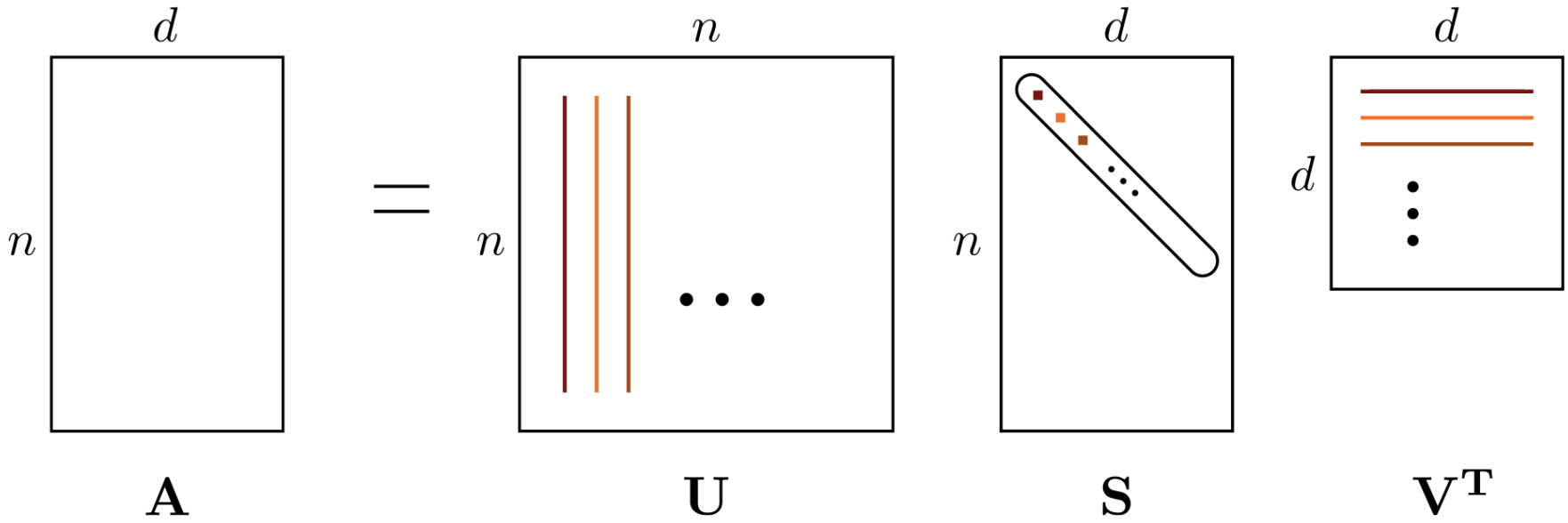
Сингулярным разложением матрицы A размера $n \times d$ является произведение трёх матриц:

$$A = U\Sigma V^\top,$$

где

- матрица U размера $n \times n$ ортогональна,
- матрица V размера $d \times d$ ортогональна,
- матрица Σ размера $n \times d$ такая, что $S_{ii} = \sigma_i = \sqrt{\lambda_i} \geq 0$, и $S_{ij} = 0$, если $i \neq j$ где $\{\lambda_i\}_{i=1}^k$ – собственные числа матрицы $A^\top A$ (и ненулевые собственные значения матрицы $AA^\top$).

Сингулярное разложение



Подсчёт сингулярного разложения

Пусть дана матрица A размера $n \times m$.

1. Посчитаем AA^T (размер - $m \times m$) и A^TA (размер - $n \times n$).
2. Находим собственные значения λ_i и собственные вектора v_i матрицы A^TA.
3. Найденные собственные вектора - колонки матрицы V, единственное, что остаётся сделать - запустить на них процесс Грама-Шмидта, чтобы ортонормировать их.
4. Корни из собственных значения λ_i и есть значения на диагонали матрицы S.
5. Ортонормированные собственные значения матрицы AA^T и есть столбцы матрицы U.

Подсчёт сингулярного разложения

Вопрос: Как можно ускорить выбранные пункты?

1. Посчитаем AA^T (размер - $m \times m$) и A^TA (размер - $n \times n$).
2. **Находим собственные значения λ_i и собственные вектора v_i матрицы A^TA.**
3. **Найденные собственные вектора - колонки матрицы V, единственное, что остаётся сделать - запустить на них процесс Грама-Шмидта, чтобы ортонормировать их.**
4. Корни из собственных значения λ_i и есть значения на диагонали матрицы S.
5. **Ортонормированные собственные значения матрицы AA^T и есть столбцы матрицы U.**

Подсчёт сингулярного разложения: шаги 2-3

Обычно, вместо процесса нахождения собственных значений и векторов (которое крайне затратно), используется алгоритм с образом матрицы, а именно:

- Засэмплировать матрицу Ω, состоящую из нормальных гауссовых векторов;
- Найти образ этих векторов $Y = A\Omega$;
- Применить QR-разложение: $Y = QR$, где Q – ортогональная, а R – верхнетреугольная;
- Вычислить SVD для матрицы $R = U\Sigma V^T$;
- Матрица QU - искомая ортогональная.

Подсчёт сингулярного разложения: шаг 5

Замечание: собственные значения матриц $A^\top A$ и $AA^\top$ связаны. Если мы уже посчитали первое, то можем посчитать второе исходя из первого.

$u_i = \frac{1}{\sqrt{\lambda_i}} A v_i$, где u_i - столбец U, v_i - столбец матрицы V, λ_i - соответствующее собственное значение.

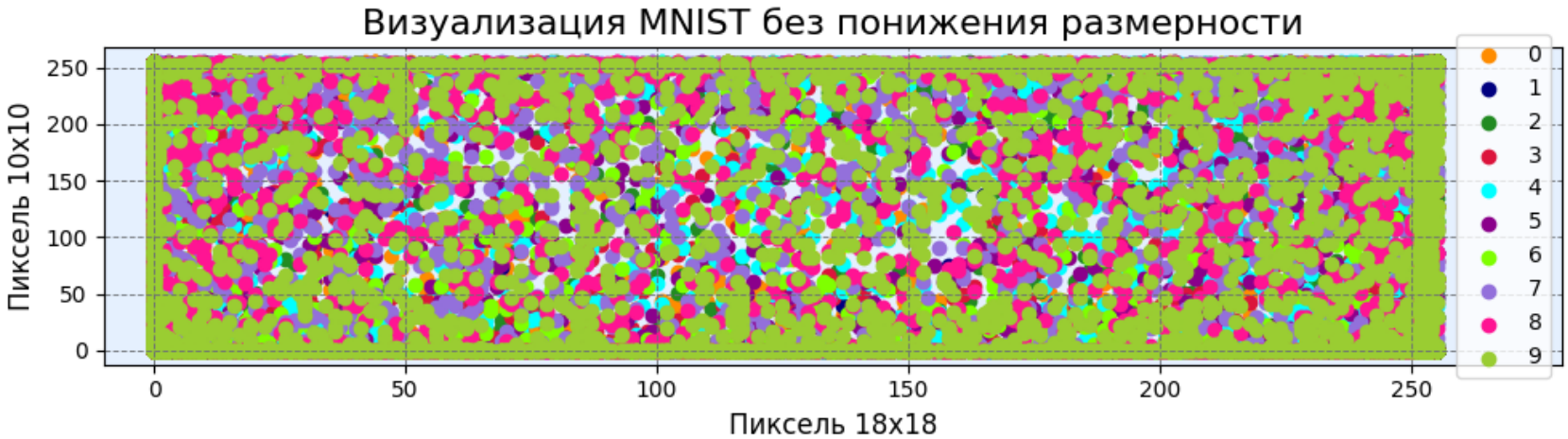
Приложения сингулярного разложения

- Решение обратных задач и вычисление псевдообратной матрицы;
- Понижение размерности (PCA - метод главных компонент);
- Латентно-семантический анализ (LSA) в обработке естественного языка (NLP);
- Сжатие изображений;
- Подавление шума в обработке сигналов;
- Рекомендательные системы;

Мотивация

Задача: распознать цифры на картинке.

Датасет состоит из картинок размера 28×28 пикселей. Все картинки в черно-белом, так что каждый пиксель принимает значение от 0 до 255. Это можно трактовать как датасет на $28 * 28 = 784$ параметра.



Датасет: MNIST dataset

Постановка задачи снижения размерности данных

Пусть $x^1, x^2, \ldots, x^n$ - выборка в пространстве $\mathbb{R}^d$.

Мы хотим выбрать пространство меньшей размерности $k < d$ так, чтобы если расстояние между объектами выборки по некоторой метрике было маленьким, осталось сравнительно маленьким, а если было большим - осталось большим.

Причины понижения размерности

- уменьшение вычислительных затрат;
- сжатие данных для более эффективного хранения информации;
- борьба с зависимыми признаками;
- борьба с переобучением;
- визуализация и интерпретация данных

Метод главных компонент (PCA)

Пусть $x^1, x^2, \ldots, x^n$ - выборка в пространстве $\mathbb{R}^d$. Соберём из объектов выборки дата матрицу X размера $n \times d$.
Хотим перейти от матрицы X к матрице $Z \in \mathbb{R}^{n \times k}$, $k < d$.

Метод главных компонент (PCA)

Пусть $x^1, x^2, \ldots, x^n$ - выборка в пространстве $\mathbb{R}^d$. Соберём из объектов выборки дата матрицу X размера $n \times d$.
Хотим перейти от матрицы X к матрице $Z \in \mathbb{R}^{n \times k}$, $k < d$.

Помимо этого, мы хотим, чтобы старые признаки восстанавливались по новым с приемлемой точностью, т. е. существует матрица $V = (v_j^i)$ размера $d \times k$ такая, что

$$\hat{x}^i = \sum_{l=1}^{k} v_l^i z_l \quad \approx \quad x^i.$$

Метод главных компонент (PCA)

Пусть $x^1, x^2, \ldots, x^n$ - выборка в пространстве $\mathbb{R}^d$. Соберём из объектов выборки дата матрицу X размера $n \times d$.
Хотим перейти от матрицы X к матрице $Z \in \mathbb{R}^{n \times k}$, $k < d$.

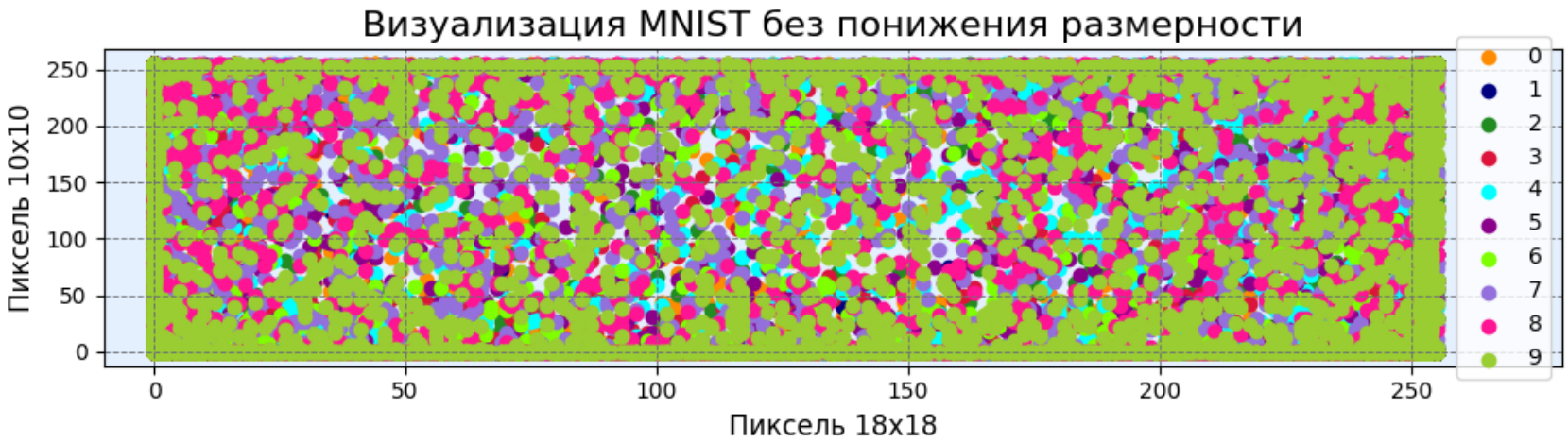
Помимо этого, мы хотим, чтобы старые признаки восстанавливались по новым с приемлемой точностью, т. е. существует матрица $V = (v_j^i)$ размера $d \times k$ такая, что

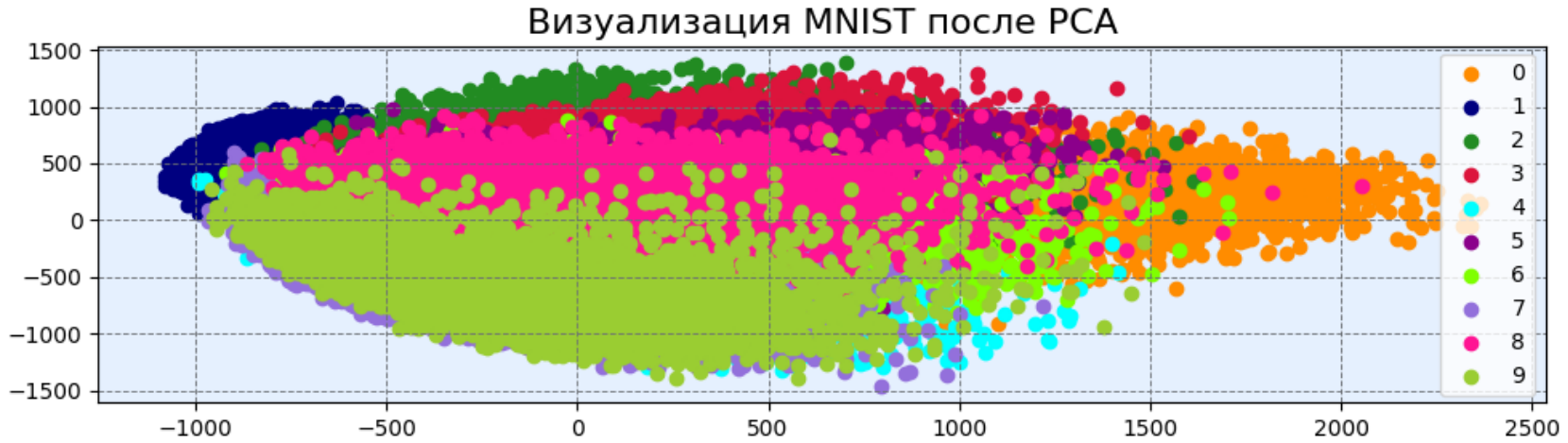
$$\hat{x}^i = \sum_{l=1}^{k} v_l^i z_l \quad \approx \quad x^i.$$

Из условий выше легко записывается задача оптимизации:

$$\Delta^2(Z, V) = \sum_{j=1}^{k} \|\hat{x^j} - x^j\|^2 = \|X - ZV^\top\|_F^2 \to \min_{Z,V}$$

PCA на MNIST


Визуализация MNIST без понижения размерности
Пиксель 10x10
Пиксель 18x18
0
50
100
150
200
250
0
1
2
3
4
5
6
7
8
9


Визуализация MNIST после PCA
1500
1000
500
0
−500
−1000
−1500
−1000
−500
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6
7
8
9

Свойства метода

Теорема

Если $k < rank(X)$, то минимум $\Delta^2(Z, V)$ достигается, когда столбцы матрицы V есть собственные вектора матрицы $X^\top X$, соответствующие k максимальным собственным значениям этой матрицы. При этом $Z = XV$, а матрицы V и Z – ортогональны.

Свойства метода

Теорема

Если $k < rank(X)$, то минимум $\Delta^2(Z, V)$ достигается, когда столбцы матрицы V есть собственные вектора матрицы $X^\top X$, соответствующие k максимальным собственным значениям этой матрицы. При этом $Z = XV$, а матрицы V и Z – ортогональны.

Доказательство:

- Из условий минимума получаем:

$$Z = XV(V^T V)^{-1}, \qquad V = X^T Z(Z^T Z)^{-1}.$$

- Решение задачи определено с точностью невырожденного преобразования R: $ZV^T = (ZR)(R^{-1}V^T)$.
- Выберем R так, чтобы матрица $Z^T Z$ была диагональной, а $V^T V$ – ортогональной (такое преобразование существует).

Свойства метода

Доказательство (продолжение):

- Тогда уравнение из условия минимума сводятся к

$$Z = XV, \qquad V\Lambda = X^T Z.$$

- Подставляя первое во второе получаем: $V\Lambda = X^T XV$, то есть V составлено из собственных векторов матрицы $X^T X$.
- Аналогично $Z\Lambda = XX^T Z$, то есть Z составлено из собственных векторов матрицы XX^T.
- Подставляя в задачу минимизации:

$$\Delta^2(Z, V) = \mathrm{Tr}\left[(X - ZV^T)(X - ZV^T)\right] = \|X\|_F^2 - \mathrm{Tr}\left[V\Lambda V^T\right].$$

Свойства метода

Теорема

Если $k < rank(X)$, то минимум $\Delta^2(Z, V)$ достигается, когда столбцы матрицы V есть собственные вектора матрицы $X^\top X$, соответствующие k максимальным собственным значениям этой матрицы. При этом $Z = XV$, а матрицы V и Z – ортогональны.

Свойства метода

Теорема

Если $k < rank(X)$, то минимум $\Delta^2(Z, V)$ достигается, когда столбцы матрицы V есть собственные вектора матрицы $X^\top X$, соответствующие k максимальным собственным значениям этой матрицы. При этом $Z = XV$, а матрицы V и Z – ортогональны.

Свойства матриц V и Z:

1. Матрица V ортонормирована, т.е. $V^\top V = I$;
2. $Z^\top Z = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$,где $\lambda_1 \geq \cdots \geq \lambda_k$ –k максимальных собственных значений матрицы $X^\top X$.
3. $X^\top XV = V\Lambda,\ X^\top Z = Z\Lambda$;
4. $\|ZV^\top - X\|_F^2 = \|X\| - \text{tr}(\Lambda) = \sum\limits_{j=k+1}^{d} \lambda_j.$

Выбор количества признаков

Наблюдение

Пусть

$$E_k = \frac{\sum_{j=k+1}^{d} \lambda_j}{\sum_{j=1}^{d} \lambda_j}.$$

Чем меньше E_k, тем лучше новые признаки приближают старые.

Выбор количества признаков

Наблюдение

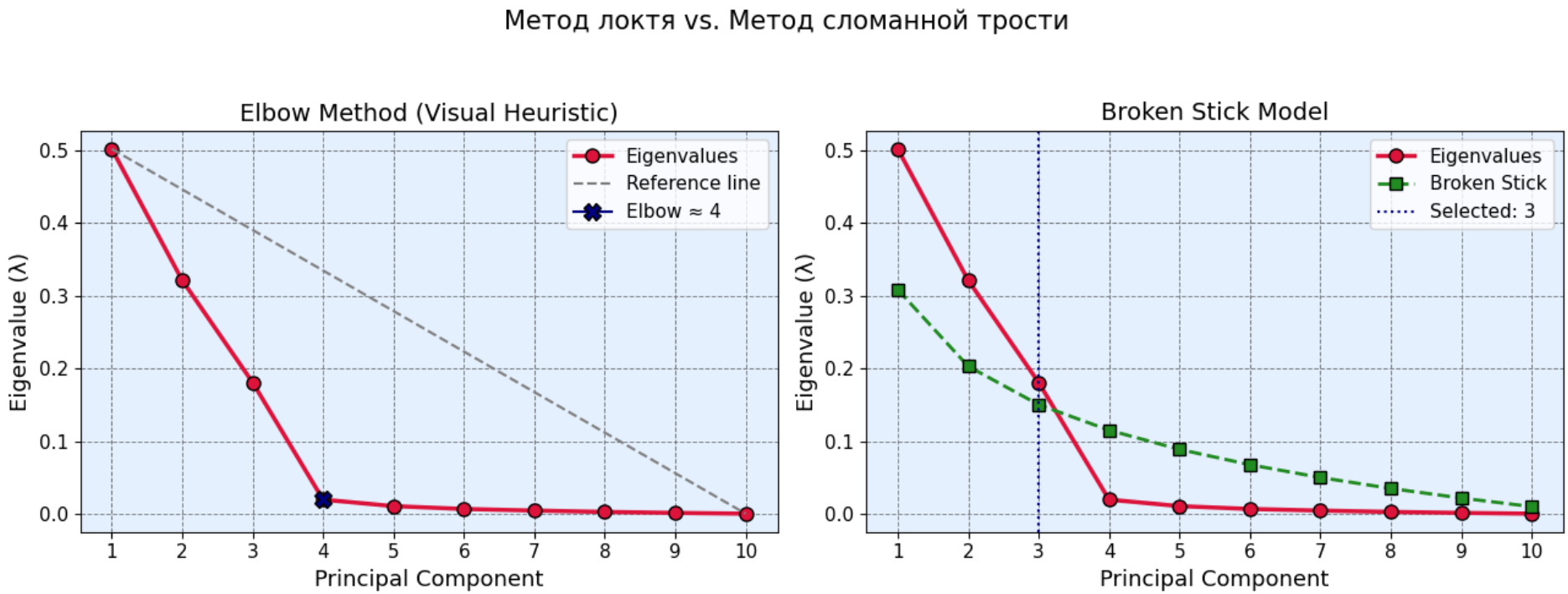
Пусть

$$E_k = \frac{\sum_{j=k+1}^{d} \lambda_j}{\sum_{j=1}^{d} \lambda_j}.$$

Чем меньше E_k, тем лучше новые признаки приближают старые.

- $\widetilde{k} = \min_k E_k < \varepsilon$ - эффективная размерность пространства признаков X.
- **Метода локтя**: Упорядочим собственные значения по убыванию. Если E_{k+1} достаточно мало и $E_k >> E_{k+1}$, то в качестве эффективной размерности берем k.
- **Метод сломанной трости**: $\bar{k} = \inf\{k : \frac{\lambda_k}{\sum_{i=1}^{k} \lambda_i} < \frac{1}{d} \sum_{j=k}^{d} \frac{1}{j}\}$

Метод локтя/метод крутого склона



Главные компоненты

Главные компоненты – это собственные векторы матрицы $X^\top X$, соответствующие k максимальным собственным значениям этой матрицы.
В качестве новых признаков $\{z_j\}_{j=1}^k$ модели мы выбираем проекции старых объектов на эти собственные векторы.

Вероятностная интерпретация: Проекции объектов на первую главную компоненту c_1 имеют наибольшую выборочную дисперсию среди проекций на всевозможные направления $d \in \mathbb{R}^k$.
Далее, $\forall j \geq 2$ проекции объектов на c_j – j-тую главную компоненту – имеют наибольшую выборочную дисперсию среди проекций на всевозможные направления $d \in \mathbb{R}^k$, перпендикулярные $c_1, \ldots, c_{j-1}$.

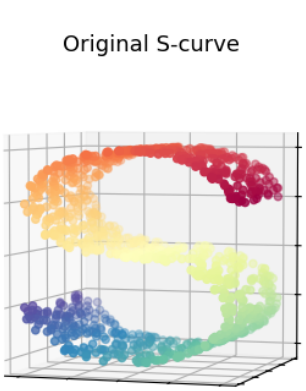
Особенности метода главных компонент

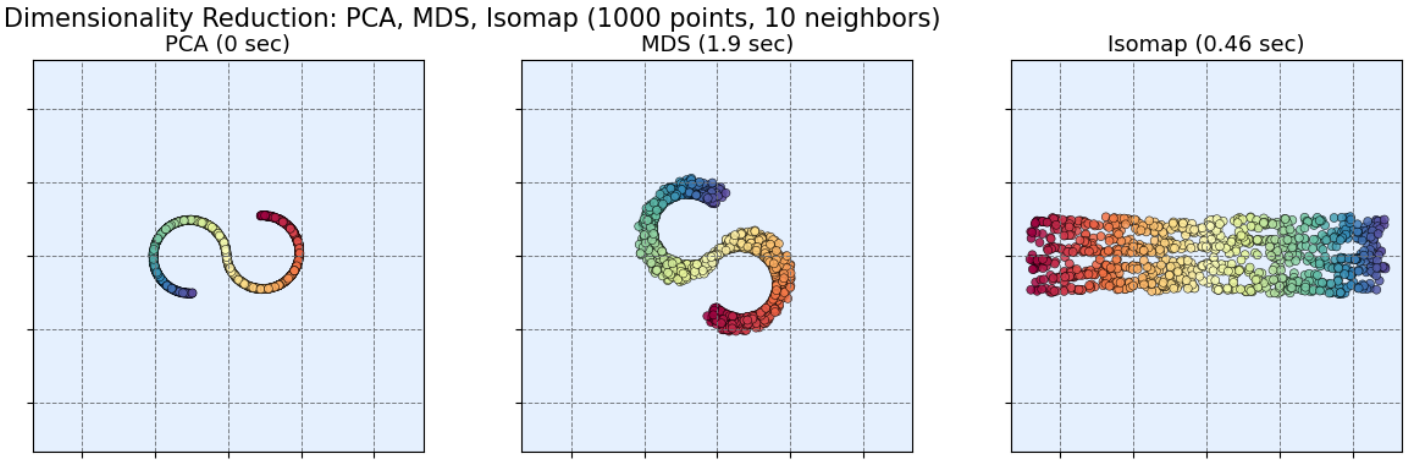
- Выбросы могут сильно помешать работе алгоритма, поэтому стоит их удалить.
- Если 2 признака имеют очень большую корреляцию, то один из признаков тоже стоит удалить, иначе матрицы будут плохо обращаться.
- Если признаки в различных шкалах, то **стандартизация данных обязательна**.
- PCA является инвариантным относительно поворота координат в пространстве переменных.
- PCA завязан на собственных значениях у матриц. Собственные значения могут быть кратными. Если некоторые собственные значения матрицы $X^\top X$, совпали, то новое пространство признаков может определяться неоднозначно. При этом, если все собственные значения различны, то мы будем выделять линейные подпространства исходного пространства.

Нелинейные методы понижения размерности

- **Kernel PCA**: вместо стандартного скалярного произведения рассмотрим скалярное произведение $K(x, y) = \langle \varphi(x), \varphi(y) \rangle$ - тут всё по аналогии с SVM и ядрами там.
- **Многомерное шкалирование (MDS)**: не использует значения x^i, а лишь разницу между ними, чтобы минимизировать разницу между расстояниями $\left|\rho(x^i, x^j) - \rho(z^i, z^j)\right|$ при переходе между пространствами; удобен для задач, в которых тяжело отбирать признаки (например, для картинок).
- **Isomap**: тот же самый PCA, только вместо евклидового расстояния берется геодезическое, которое дает более обширную характеристику в пространстве признаков.

Нелинейные методы понижения размерности





t-SNE

t-SNE (t-distributed Stochastic Neighbor Embedding) – один из алгоритмов понижения размерности, в частности, хорошо подходящий для визуализации данных.

Постановка задачи снижения размерности данных

Задача Снижение размерности данных
Пусть $x^1, x^2, \ldots, x^n$ - выборка в пространстве $\mathbb{R}^d$. Хотим перейти от признаков X к новым признакам $Z = (z_j^i)$, где Z -матрица размера $n \times k,\ k < d$.
Наша цель - отобразить кластерную структуру, сохранив кластеры без сохранения пространственных взаимоотношений кластеров.

SNE

Определим вероятность того, что x^j - сосед x^i, пропорционально плотности $\mathcal{N}(x^i, \sigma_i^2)$ в точке x^j:

$$p_{j|i} = \frac{\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x^j - x^k\|^2/2\sigma_i^2\right)}.$$

SNE

Определим вероятность того, что x^j - сосед x^i, пропорционально плотности $\mathcal{N}(x^i, \sigma_i^2)$ в точке x^j:

$$p_{j|i} = \frac{\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|x^j - x^k\|^2/2\sigma_i^2\right)}.$$

Определим вероятность того, что z^j - сосед z^i, пропорционально плотности $\mathcal{N}(z^i, 1/2)$ в точке z^j:

$$q_{j|i} = \frac{\exp\left(-\|z^j - z^i\|^2\right)}{\sum_{k\neq i}\exp\left(-\|z^j - z^k\|^2\right)}.$$

Будем считать, что $p_{i|i} = 0$ и $q_{i|i} = 0$.

Как выбирать σ_i^2

Тогда **энтропия** имеет вид: $H(\sigma_i) = -\sum_{j=1}^{n} p_{j|i} \log p_{j|i}$.

Перплексия $\text{Perp}(\sigma_i) = 2^{H(\sigma_i)}$ имеет смысл сглаженного показателя эффективного числа соседей точки X_i.

Значение перплексии — гиперпараметр метода. Задается одинаковым для всех i. Числа σ_i подбираются на основе перплексии с помощью бинарного поиска.

Вопрос: Зачем делать разные σ_i?

Как выбирать σ_i^2

Тогда **энтропия** имеет вид: $H(\sigma_i) = -\sum_{j=1}^{n} p_{j|i} \log p_{j|i}$.
Перплексия $\text{Perp}(\sigma_i) = 2^{H(\sigma_i)}$ имеет смысл сглаженного показателя эффективного числа соседей точки X_i.
Значение перплексии — гиперпараметр метода. Задается одинаковым для всех i. Числа σ_i подбираются на основе перплексии с помощью бинарного поиска.
Вопрос: Зачем делать разные σ_i?
Разная σ_i необходима из-за возможного наличия кластеров разных плотностей.

Оптимизационная задача

Хотим выбрать $z^1, z^2, \ldots, z^n$ так, чтобы вероятности $q_{j|i}$ как можно точнее описывали $p_{j|i}$.

Функция потерь: дивергенция Кульбака-Лейблера между $p_{j|i}$ и $q_{j|i}$

$$\mathcal{L}(P, Q) = \sum_{i=1}^{n} \mathrm{KL}(P_i \parallel Q_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \to \min_Z$$

Вопрос: одинаковые ли штрафы, если близкие точки расположены далеко и далекие точки расположены близко?

Оптимизационная задача

Хотим выбрать $z^1, z^2, \ldots, z^n$ так, чтобы вероятности $q_{j|i}$ как можно точнее описывали $p_{j|i}$.

Функция потерь: дивергенция Кульбака-Лейблера между $p_{j|i}$ и $q_{j|i}$

$$\mathcal{L}(P, Q) = \sum_{i=1}^{n} \mathrm{KL}(P_i \parallel Q_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \to \min_{Z}$$

Вопрос: одинаковые ли штрафы, если близкие точки расположены далеко и далекие точки расположены близко?

- большой штраф, если близкие точки будут расположены далеко.
- малый штраф, если далекие точки будут расположены близко.

Симметричный SNE

Рассмотрим «симметричные» вероятности:

$$p_{ji} = \frac{\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}{\sum_{j\neq i}\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}, \quad q_{ji} = \frac{\exp\left(-\|z^j - z^i\|^2\right)}{\sum_{j\neq i}\exp\left(-\|z^j - z^i\|^2\right)}$$

причем $p_{ii} = q_{ii} = 0$
Функция потерь

$$\mathcal{L}(P, Q) = \mathrm{KL}(P \parallel Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Если x_i - выброс, то $\|x^i - x^j\| \gg 0$ и $p_{ij} \approx 0$ для любого j. Значит расположение z^i почти не влияет на $\mathcal{L}$.

Симметричный SNE

Рассмотрим «симметричные» вероятности:

$$p_{ji} = \frac{\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}{\sum_{j\neq i}\exp\left(-\|x^j - x^i\|^2/2\sigma_i^2\right)}, \quad q_{ji} = \frac{\exp\left(-\|z^j - z^i\|^2\right)}{\sum_{j\neq i}\exp\left(-\|z^j - z^i\|^2\right)}$$

причем $p_{ii} = q_{ii} = 0$
Функция потерь

$$\mathcal{L}(P, Q) = \mathrm{KL}(P \parallel Q) = \sum_{i\neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Если x_i - выброс, то $\|x^i - x^j\| \gg 0$ и $p_{ij} \approx 0$ для любого j. Значит расположение z^i почти не влияет на $\mathcal{L}$.
Вместо этого определим p_{ij} как $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}$. Тогда $\sum_j p_{ij} > 1/2n$.

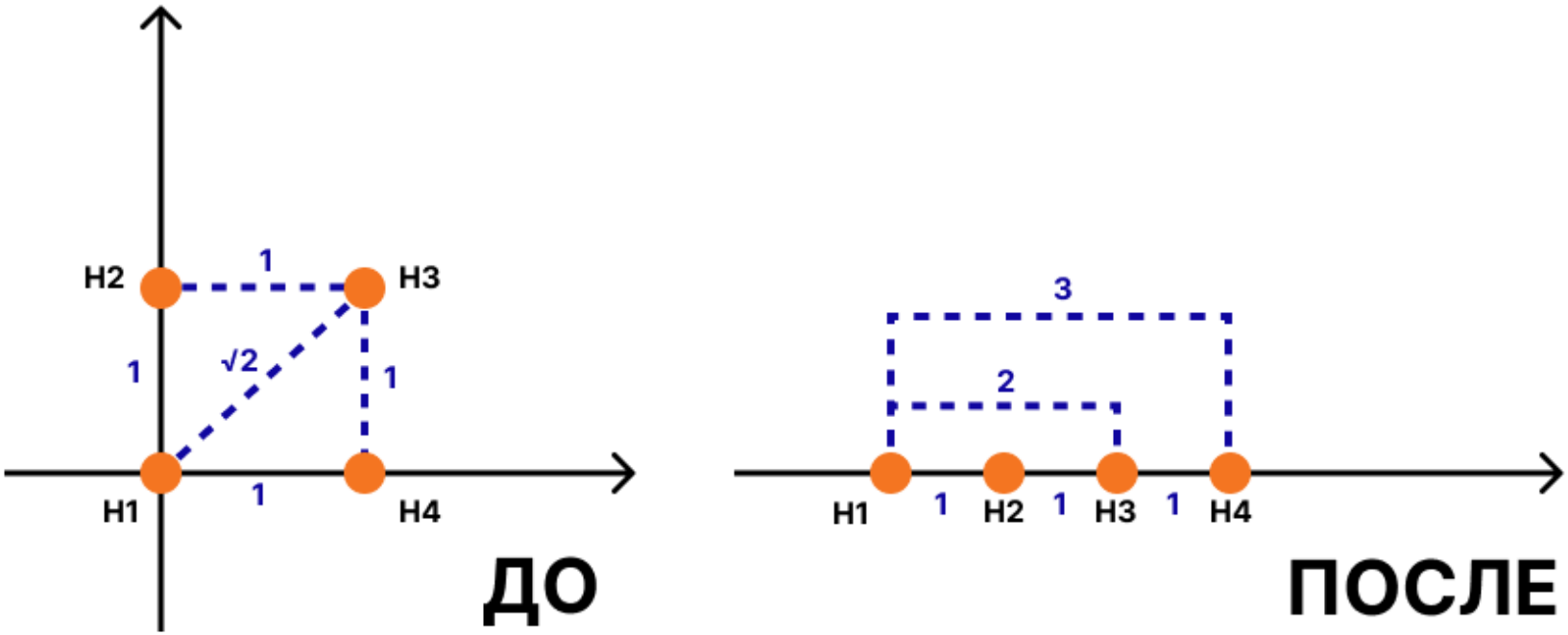
Crowding problem

При вложении в пространство малой размерности при использовании нормального распределения действуют силы сжатия, из-за чего точки сильно сжимаются в кучу.

Пример

Данные: облака точек в вершинах правильного тетраэдра в $\mathbb{R}^3$. В силу симметричности при сжатии в $\mathbb{R}^2$ на точки будут действовать "сжимающие силы" по диагоналям.

Визуализация Crowding problem



Мы не можем поставить в новом пространстве точку H_4 так, чтобы она был на расстоянии 1 и от H_1, и от H_3.

t-SNE

Для снижения частоты возникновения Crowding problem определим q_{ij} на основе **распределения Стьюдента**:

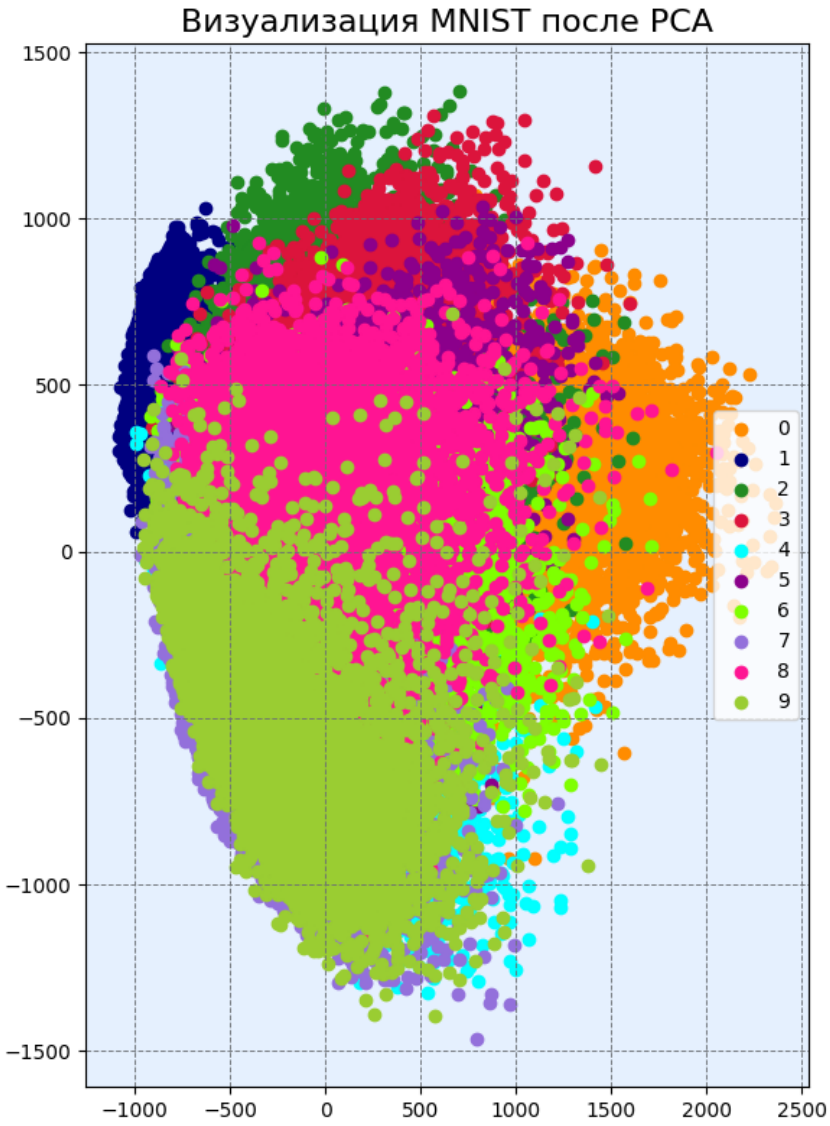
$$q_{ji} = \frac{\left(1 + \|z^i - z^j\|^2\right)^{-1}}{\sum_{i \neq j}(1 + \|z^i - z^j\|^2)^{-1}}$$

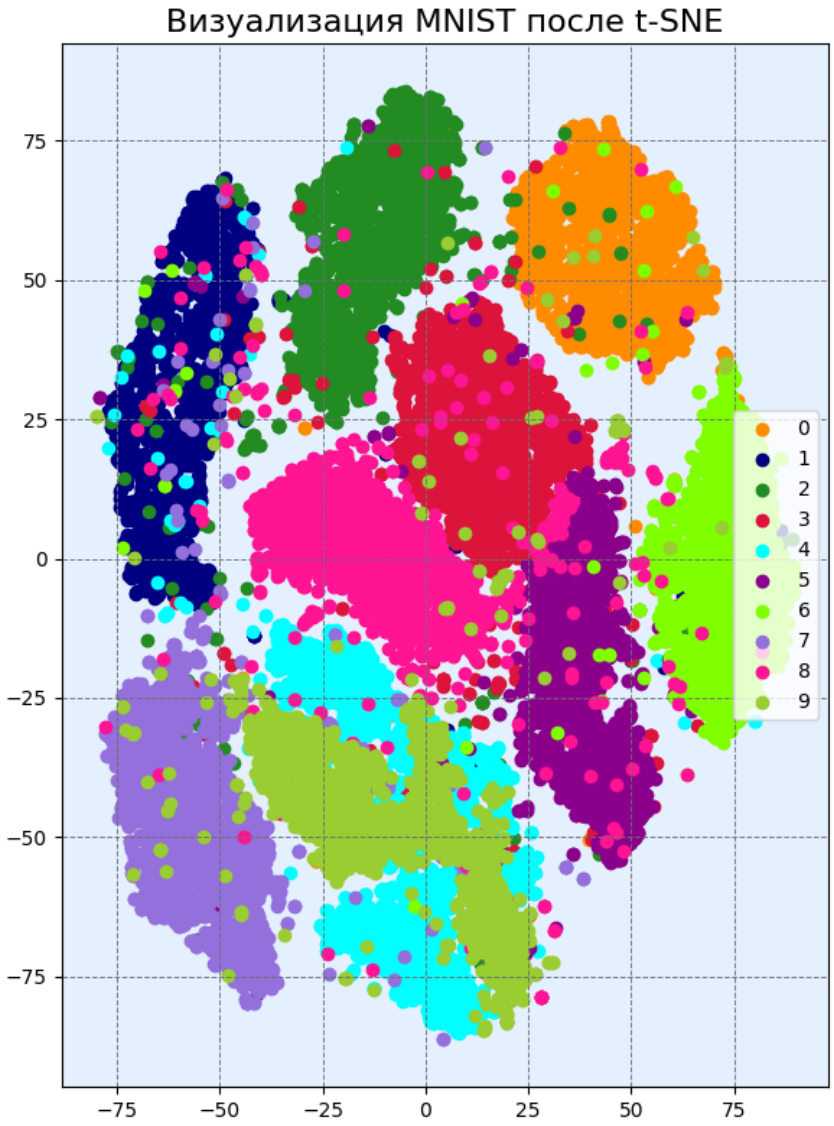
При этом $p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$. Градиент в таком случае

$$\frac{\partial \mathcal{L}}{\partial z^i} = 4 \sum_j^n \frac{(p_{ij} - q_{ij})(z^i - z^j)}{1 + \|z^i - z^j\|^2}$$

- Градиент делится на квадрат расстояния плюс 1.
- Если точки близки, то $\|z^i - z^j\|^2 \approx 0$, и сила остается прежней.
- Если точки далеки, то $\|z^i - z^j\|^2 \gg 0$, сила сжатия становится существенно меньше, и сильного сжатия не происходит.

Сравнение PCA и t-SNE на MNIST





UMAP

Uniform Manifold Approximation and Projection — метод, выполняющий нелинейное снижение размерности. Алгоритм предложен в 2018 году с целью получить аналог t-SNE, но с более сильным математическим обоснованием.

Ориентированный граф

Пусть $X = (x^1, x^2, \ldots, x^n)$ - выборка в пространстве $\mathcal{X}$.
$\rho(x, y)$ - метрика в $\mathcal{X}$.

Определим случайный ориентированный граф на множестве вершин X. Считаем, что каждое ребро появляется независимо от других.

Пусть $x^{i_1}, x^{i_2}, \ldots, x^{i_k}$ - k ближайших соседей объекта x^i в выборке X .
$r_i = \min_s \rho(x^i, x^{i_s})$ — **расстояние до ближайшего соседа.**

Вероятность ребра из x^i в x^{i_s} (для остальных ноль):

$$\mathbb{P}\left\{x^i \to x^{i_s}\right\} = \exp\left(-\frac{\rho(x^i, x^{i_s}) - r_i}{\sigma_i}\right),$$

где σ_i определяется как решение уравнения $\sum_{s=1}^{k} \mathbb{P}\left\{x^i \to x^{i_s}\right\} = \log_2 k$

Неориентированный граф

На основе ориентированного графа построим неориентированный. X — множество вершин. Вероятность ребра между x^i и x^{i_s} :

$$\begin{aligned}\mathbb{P}\left\{x^i \leftrightarrow x^{i_s}\right\} &= \mathbb{P}\left\{\{x^i \to x^{i_s}\} \bigcup \{x^{i_s} \to x^i\}\right\} \\ &= \mathbb{P}\left\{x^i \to x^{i_s}\right\} + \mathbb{P}\left\{x^{i_s} \to x^i\right\} - \mathbb{P}\left\{x^i \to x^{i_s}\right\}\mathbb{P}\left\{x^{i_s} \to x^i\right\}\end{aligned}$$

Новые признаки

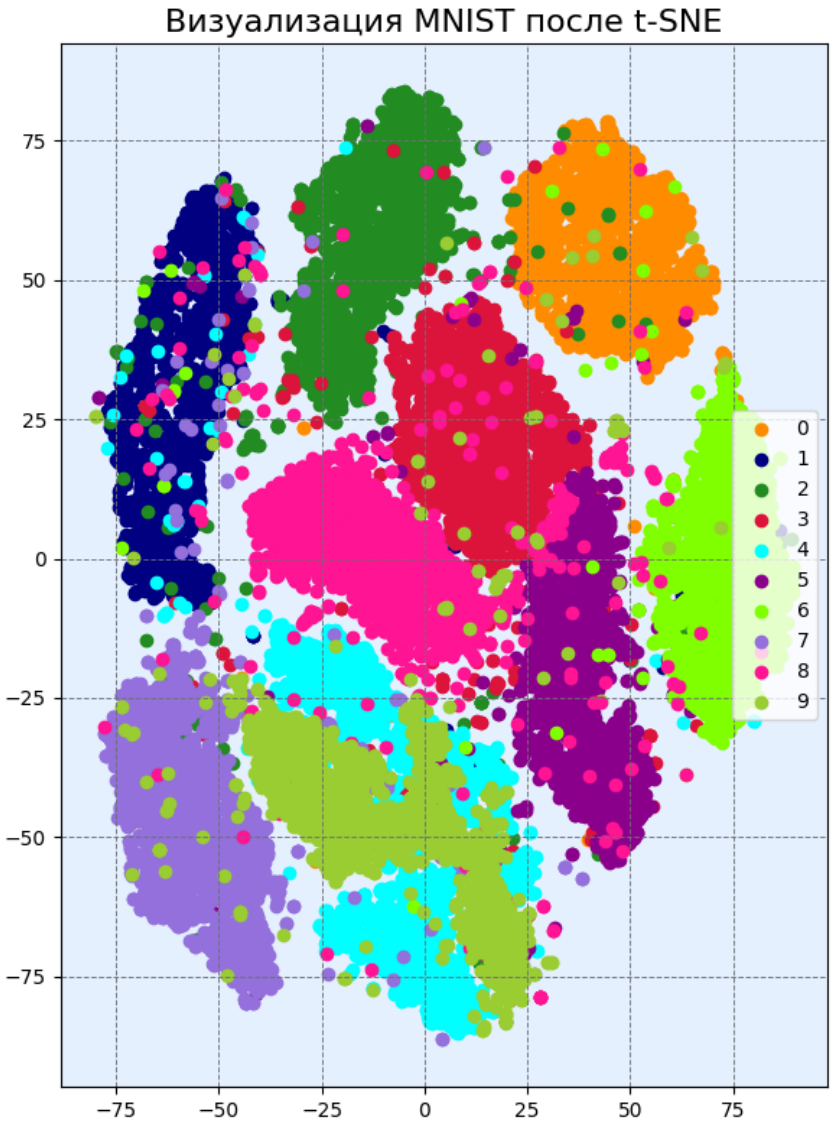
Пусть $z^1, z^2, \ldots, z^n$ - новые признаки, которые хотим получить.
На них определяем случаем неориентированный граф с вероятностями $\mathbb{P}\left\{z^i \leftrightarrow z^j\right\} = (1 + a\|z^i - z^j\|_2^{2b})$, где a и b - гиперпараметры.
Минимизируем дивергенцию Кульбака-Лейблера:

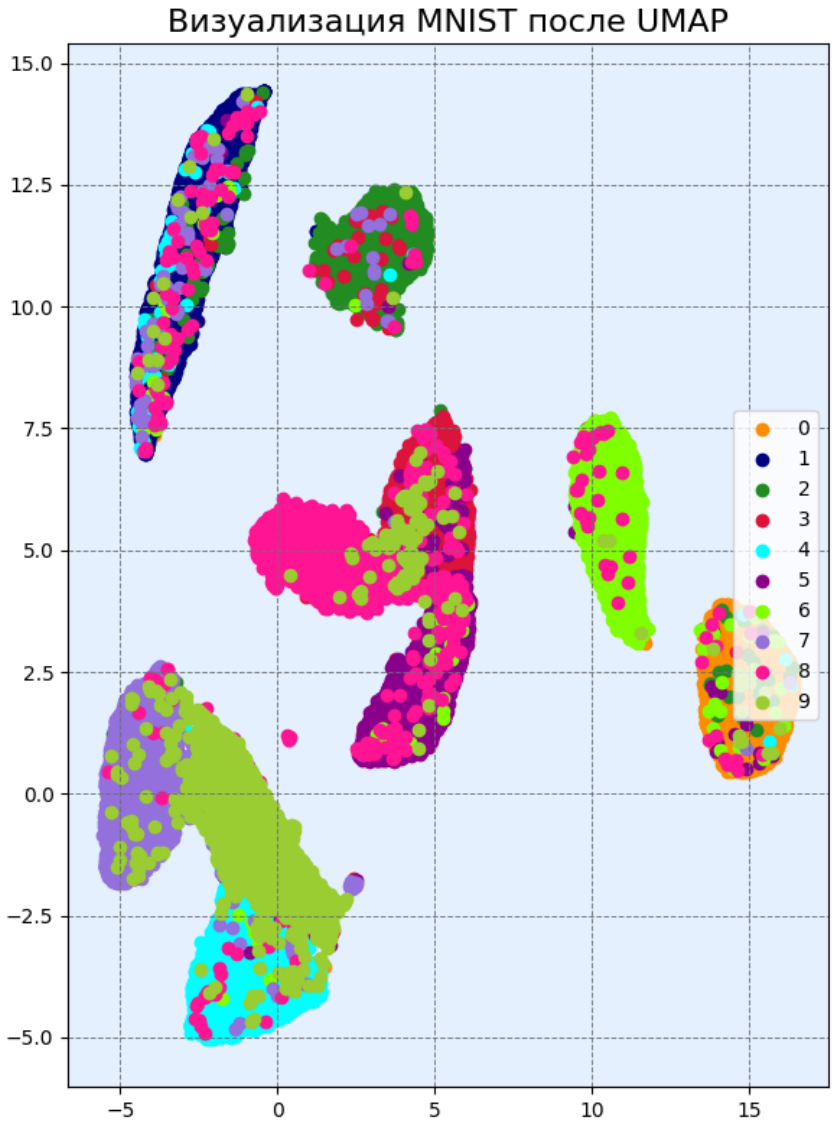
$$\begin{aligned}\mathrm{KL}(\mathbb{P}_x \parallel \mathbb{P}_z) &= \sum_{i,j}\Bigg[\mathbb{P}\left\{x^i \leftrightarrow x^j\right\}\log\frac{\mathbb{P}\left\{x^i \leftrightarrow x^j\right\}}{\mathbb{P}\left\{z^i \leftrightarrow z^j\right\}} \\ &\quad + \left(1 - \mathbb{P}\left\{x^i \leftrightarrow x^j\right\}\right)\log\frac{1 - \mathbb{P}\left\{x^i \leftrightarrow x^j\right\}}{1 - \mathbb{P}\left\{z^i \leftrightarrow z^j\right\}}\Bigg]\end{aligned}$$

Свойства UMAP

- Введение графов и вероятностей на них эквивалентно использованию метрики локальной связности, которая устойчива к проклятию размерности. UMAP можно применять на данных огромных размерностей.
- Возможно сохранение пространственных взаимоотношений между кластерами.
- Можно применять на новых данных и выполнять обратное преобразование.
- Работает в несколько раз быстрее t-SNE.
- Поддерживает различные метрики.

Сравнение t-SNE и UMAP на MNIST





Пример

4 миллиона транскриптомов отдельных клеток взрослого мозга мыши, помеченных по исходному региону мозга и представленных с помощью UMAP (Yao Z. et al., 2023, bioRxiv).