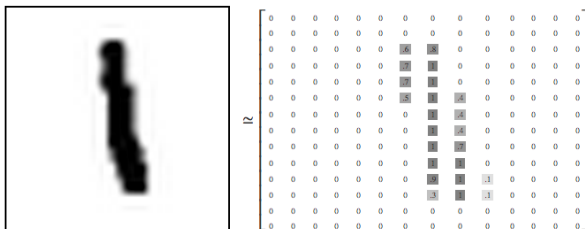


- Рассмотрим для примера набор изображений рукописных цифр MNIST.

A selection from the 64-dimensional digits dataset



- Объект из набора — изображение 28×28 пикселей, в каждом пикселе известна интенсивность цвета — вещественное число из $[0, 1]$.
- Матрицу интенсивностей можно развернуть в вектор признаков длины $28 \times 28 = 784$.




- Что можно сказать про расположение объектов (изображений рукописных цифр) в признаковом пространстве?

- Давайте возьмем случайный вектор из пространства признаков и посмотрим, какие изображения 28×28 будут ему соответствовать.



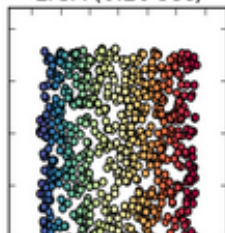
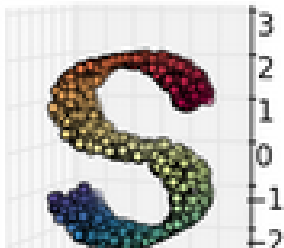
- Случайно взятый вектор признаков с вероятностью 1 не будет соответствовать рукописной цифре. А вектора, соответствующие цифрам, занимают незначительную часть всего пространства признаков.
- Множество векторов признаков, соответствующих рукописным цифрам образует подпространство меньшей размерности в нашем исходном признаковом пространстве.

- Линейная комбинация признаков рукописных цифр может не соответствовать рукописной цифре.

$$0.5 \cdot \begin{array}{|c|} \hline 0 \\ \hline \end{array} + 0.5 \cdot \begin{array}{|c|} \hline 3 \\ \hline \end{array} = \begin{array}{|c|} \hline 3 \\ \hline \end{array}$$


- Подпространство в котором лежат рукописные цифры — не является линейным.
- Было бы удобно работать в признаковом пространстве, содержащем только рукописные цифры.

- Для того чтобы было проще понять, что значит работать в пространстве, содержащем только рукописные цифры, рассмотрим более простой модельный пример.
- В трехмерном признаковом пространстве имеется S-образная поверхность, на которой лежат объекты выборки, которые можно отобразить на двухмерную координатную плоскость, сохранив при этом всю информацию об объектах, но при этом вся плоскость равномерно заполнена объектами выборки.



Дано $\{x_1, \dots, x_\ell\}$ — *выборка* объектов.

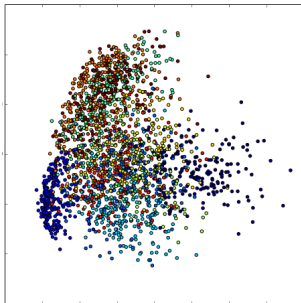
Задача: Отобразить все объекты выборки в пространство малой размерности $x_i \mapsto \tilde{x}_i \in \mathbb{R}^d$.

Требования к маломерному представлению:

- Должно хорошо отражать структуру данных в исходном пространстве;
- Сохраняло интересующие нас закономерности в данных.

- Наиболее популярное приложение методов нелинейного понижения размерности — визуализация наборов данных.
- Человек живет в 3-х мерном мире и «посмотреть» глазами на данные с более чем 3 вещественными признаками — достаточно трудно.
- Визуализировать объекты, представленные векторами большой размерности — расположить их на прямой, плоскости или в виде 3-х мерного облака точек, таким образом, чтобы их расположение хорошо отражало структуру данных в исходном пространстве.
- Возьмем для примера рукописные цифры от 0 до 5 отобразим их в двумерное пространство (на плоскость) при помощи метода главных компонент.

- На этом изображении, координатные оси соответствуют компонентам проекции. Различные цифры выделены цветом.



- Классы перемешаны между собой, хотя человек по изображению отличает достаточно просто.
- Это следствие нелинейности пространства рукописных цифр, заданных числовыми векторами.

Мы рассмотрим несколько простых методов, используемых на практике.

Гипотеза: хорошее малоразмерное представление сохраняет попарные расстояния между объектами выборки.

d_{ij} — расстояние между x_i и x_j .

- признаковые описания не нужны — достаточно расстояний

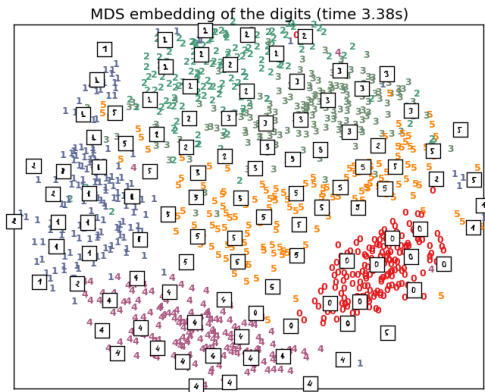
$\tilde{d}_{ij} = \|\tilde{x}_i - \tilde{x}_j\|$ — евклидово расстояние между маломерными представлениями.

Ищем представления, аппроксимирующие d_{ij} :

$$\sum_{i < j}^{\ell} (\|\tilde{x}_i - \tilde{x}_j\| - d_{ij})^2 \rightarrow \min_{(\tilde{x}_i)_{i=1}^{\ell} \subset \mathbb{R}^d}$$

Оптимизация: алгоритм SMACOF.

- Применим многомерное шкалирование (MDS) к выборке рукописных цифр, уменьшим размерность признакового описания до 2х и расположим объекты на плоскости.



Stochastic Neighbor Embedding:

- В точности воспроизвести расстояния — слишком сложно
- Достаточно сохранения пропорций:

$$\rho(x_1, x_2) = c\rho(x_1, x_3) \Rightarrow \rho(\tilde{x}_1, \tilde{x}_2) = c\rho(\tilde{x}_1, \tilde{x}_3).$$

- Опишем объекты нормированными расстояниями до остальных объектов:

$$p(x_j | x_i) = \frac{\exp(\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(\|x_i - x_k\|^2 / 2\sigma^2)}$$

$$q(\tilde{x}_j | \tilde{x}_i) = \frac{\exp(\|\tilde{x}_i - \tilde{x}_j\|^2)}{\sum_{k \neq i} \exp(\|\tilde{x}_i - \tilde{x}_k\|^2)}$$

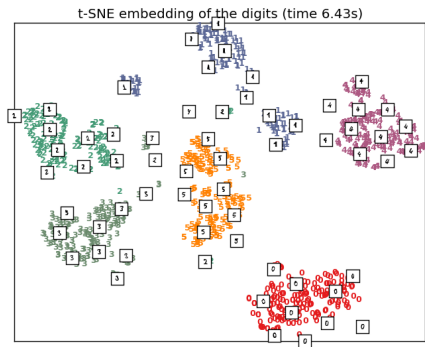
- Минимизируем разницу между распределениями расстояний (мера — дивергенция Кульбака-Лейблера):

$$\sum_{i=1}^{\ell} \sum_{j \neq i} p(x_j | x_i) \log \frac{p(x_j | x_i)}{q(\tilde{x}_j | \tilde{x}_i)} \rightarrow \min_{(\tilde{x}_i)_{i=1}^{\ell} \subset \mathbb{R}^d}$$

t-Distributed Stochastic Neighbor Embedding — развитие SNE:

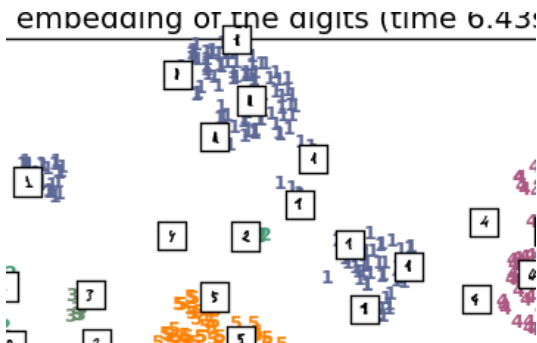
- Чем выше размерность пространства, тем меньше расстояния между парами точек отличаются друг от друга (проклятие размерности).
- Невозможно воспроизвести это свойство в двух- или трех-мерном пространстве.
- Значит, нужно меньше штрафовать за увеличение пропорций в маломерном пространстве.
- Изменим распределение:

$$q(\tilde{x}_j | \tilde{x}_i) = \frac{(1 + \|\tilde{x}_i - \tilde{x}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\tilde{x}_i - \tilde{x}_k\|^2)^{-1}}$$



- Объекты расположились на плоскости в виде явно выраженных кластеров.
- К такому двух-мерному представлению объектов можно применять методы классификации, не проигрывая в точности.

- Такая визуализация позволяет посмотреть глазами на внутреннюю структуру данных
- Рукописные единицы бывают характерно трех видов:
 - С подставкой
 - Без верхней засечки, похожие на "палочку"
 - С выраженной верхней засечкой, сильнее всего похожие на цифру 4, расположенные ближе всего к облаку из четверок.



- Разобрали основную идею MDS, t-SNE
- Применения:
 - эффективное понижение размерности данных;
 - визуализация данных.
- Убедились на примере набора данных MNIST, что методы могут понизить размерность пространства с нескольких сот признаков до двух.