EleWise проекты

Лекция 7

Кластеризация и визуализация

7.1. Кластеризация

7.1. Кластеризация

Дано:

X - пространство объектов; $X^\ell = \{x_i\}_{i=1}^\ell$ - обучающая выборка

$\rho: X \times X \to [0, \infty)$ - ф-ия расстояния между объектами

Найти:

$y_i \in Y$ - метки кластеров объектов:

- каждый кластер состоит из близких объектов
- объекты разных кластеров сильно различны.

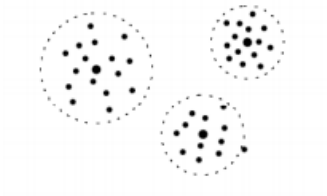
***Кластеризация** - это обучение без учителя*

Решение задачи кластеризации принципиально неоднозначно:

- различные критерии качества кластеризации
- различные эвристические методы кластеризации
- различные варианты ф-ии расстояния ρ

7.1. Кластеризация

- вертикальные расстояния, как правило, меньше межкластерных



- кластеры с центром

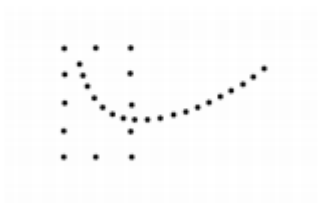
- кластеры могут соединяться перемычками

- кластеры могут накладываться на разреженный фон из редко расположенных объектов

7.1. Кластеризация

- ленточные кластеры

- перекрывающие кластеры



- кластеры могут образовываться не по сходству
  а по другим типам регулярностей

- кластеры могут вообще отсутствовать

7.1. Кластеризация

Метод k-средних

Объекты x_i задаются векторами признаков $(f_1(x_i), \ldots, f_n(x_i))$

Вход: X^ℓ — обучающая выборка, параметр k;

Выход: центры кластеров μ_y, $y \in Y$;

1: начальное приближение центров μ_y, $y \in Y$;

2: **повторять**

3: отнести каждый x_i к ближайшему центру:

$$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \ldots, \ell;$$

4: вычислить новые положения центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \ldots, n;$$

5: **пока** y_i не перестанут изменяться;

7.1. Кластеризация

Мягкий вариант метода k-средних (EM-алгоритм)

Вход: X^ℓ — обучающая выборка, параметр k;

Выход: центры кластеров μ_y, $y \in Y$;

1: начальное приближение центров μ_y, $w_y := \frac{1}{|Y|}$, $y \in Y$;

2: **повторять**

3: оценить близость каждого x_i ко всем центрам:

$g_{iy} := w_y \exp\left(-\frac{1}{2}\rho^2(x_i, \mu_y)\right)$, $i = 1, \ldots, \ell$, $y \in Y$;

$g_{iy} := \dfrac{g_{iy}}{\sum\limits_{z \in Y} g_{iz}}$ — нормированные близости;

4: отнести каждый x_i к ближайшему центру:

$y_i := \arg\max\limits_{y \in Y} g_{iy}$, $i = 1, \ldots, \ell$;

5: новые положения центров и мощности кластеров:

$\mu_{yj} := \frac{1}{\ell w_y} \sum\limits_{i=1}^{\ell} g_{iy} f_j(x_i)$, $w_y := \frac{1}{\ell} \sum\limits_{i=1}^{\ell} g_{iy}$, $y \in Y$, $j = 1, \ldots, n$;

6: **пока** g_{iy} не перестанут изменяться;

7.1. Кластеризация

Формирование начального приближения для k-средних

Вход: X^ℓ — обучающая выборка, параметры q, δ, k

Выход: $U \subset X^\ell$ — начальные приближения центров μ_y, $y \in Y$;

1: среднее расстояние до q ближайших соседей:

$R_i := \frac{1}{q} \sum_{j=1}^{q} \rho\left(x_i, x_i^{(j)}\right)$, для всех $i = 1, \ldots, \ell$,

где $x_i^{(j)}$ — j-й ближайший сосед объекта x_i;

2: отбросить шумовые объекты:

$X' := \left\{x_i \in X^\ell \mid R_i \leqslant \Delta\right\}$ при Δ: $|X'| = (1 - \delta)\ell$;

3: выбрать пару самых удалённых объектов:

$U := \arg \max_{x, x' \in X'} \rho(x, x')$;

далее последовательно присоединять к U по одному объекту, самому удалённому от уже выбранных:

4: **повторять** $k - 2$ **раз**

$U := U \cup \arg \max_{x \in X'} \min_{u \in U} \rho(x, u)$;

7.1. Кластеризация

ElеWise проекты

Недостатки k-средних

- Чувствительность к выбору начального приближения
- Необходимость задать k

Способы устранения

- Эвристика для выбора начального приближения
- Мягкая кластеризация
- Мультистарт: несколько случайных инициализаций; выбор лучшей кластеризации по функционалу качества
- Быстрые алгоритмы (k-means++, сэмплирование)
- Варьирование числа кластеров k в ходе итераций

7.2. Иерархическая кластеризация

7.2. Иерархическая кластеризация

Дано:

X - пространство объектов; $X^\ell = \{x_i\}_{i=1}^\ell$ - обучающая выборка

$\rho: X \times X \to [0, \infty)$ - ф-ия расстояния между объектами

Найти:

$y_i \in Y$ - метки кластеров объектов:

- каждый кластер состоит из близких объектов
- объекты разных кластеров сильно различны.

***Кластеризация** - это обучение без учителя*

Как определить число кластеров?

Вместо этого можно строить иерархическую кластеризацию

7.2. Иерархическая кластеризация

Алгоритм Ланса-Уильямса [1967] основан на оценивании расстояния $R(U, V)$ между парами кластеров U, V

1: сначала все кластеры одноэлементные:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\};$
$R(\{x_i\}, \{x_j\}) := \rho(x_i, x_j);$
2: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
3: найти в C_{t-1} два ближайших кластера:
$(U, V) := \arg\min_{U \neq V} R(U, V);$
$R_t := R(U, V);$
4: слить их в один кластер:
$W := U \cup V;$
$C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$
5: **для всех** $S \in C_t$
6: вычислить $R(W, S)$ по формуле Ланса-Уильямса;

7.2. Иерархическая кластеризация

Как определить расстояние *R(W, S)* между кластерами *W=U ∪ V* и *S*, зная расстояния $R(U,S)$, $R(V,S)$, $R(U,V)$?

Формула обобщающая большинство разумных способов определить это расстояние [Ланс, Уильямс, 1967]

$$\begin{aligned} R(U \cup V, S) = \alpha_U \cdot R(U,S) + \\ + \alpha_V \cdot R(V,S) + \\ + \beta \cdot R(U,V) + \\ + \gamma \cdot \left| R(U,S) - R(V,S) \right|, \end{aligned}$$

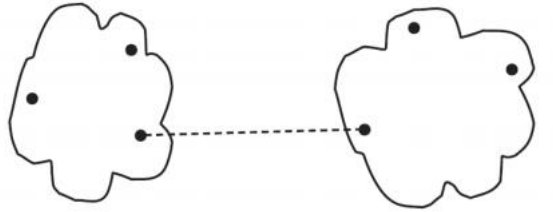
где α_U, α_V, β, γ - числовые параметры

7.2. Иерархическая кластеризация

1. Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

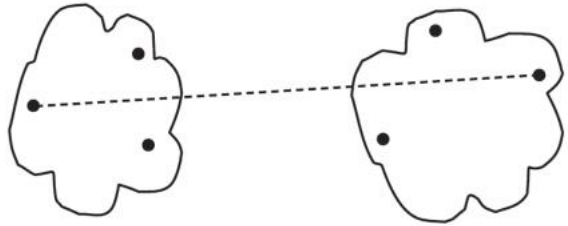
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

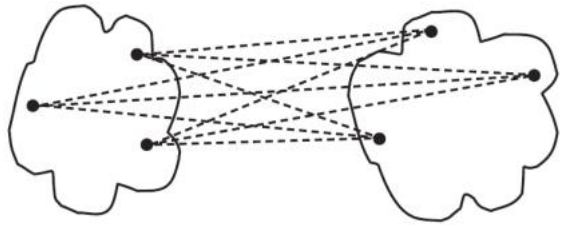
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{г}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},\ \ \beta = \gamma = 0.$$



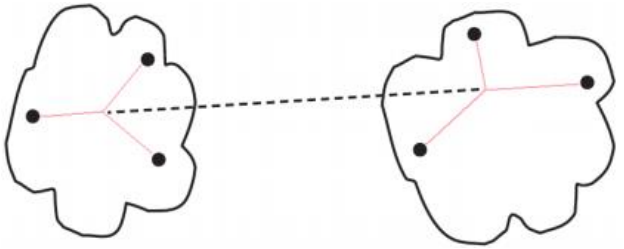
7.2. Иерархическая кластеризация

4. Расстояние между центрами:

$$R^{ч}(W,S) = \rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|U|}{|W|},\ \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V,\ \gamma = 0.$$



5. Расстояние Уорда:

$$R^{у}(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|}, \sum_{s\in S}\frac{s}{|S|}\Big);$$

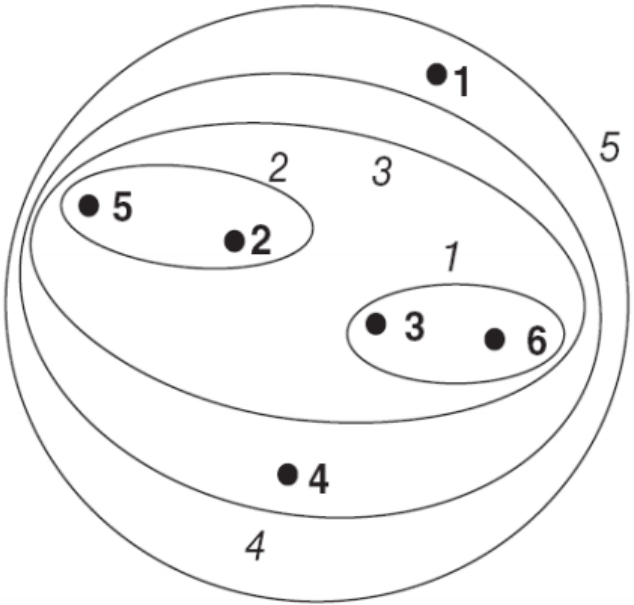
$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

Проблема выбора: **какая ф-ия расстояний лучше?**

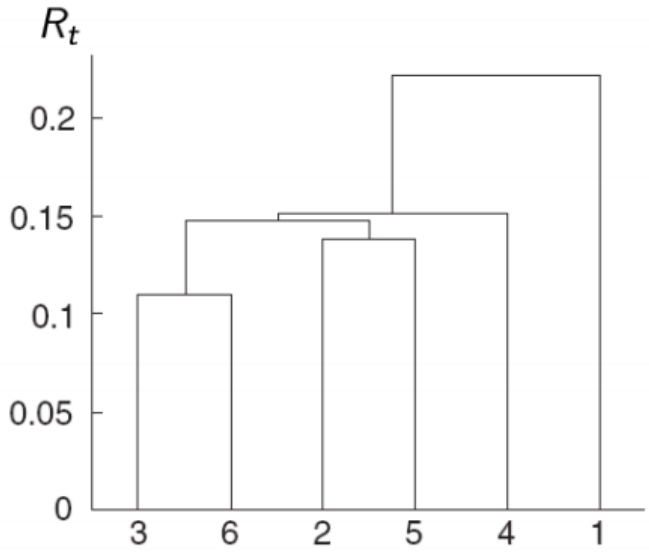
7.2. Иерархическая кластеризация

1. Расстояние ближайшего соседа:

Диаграмма вложения



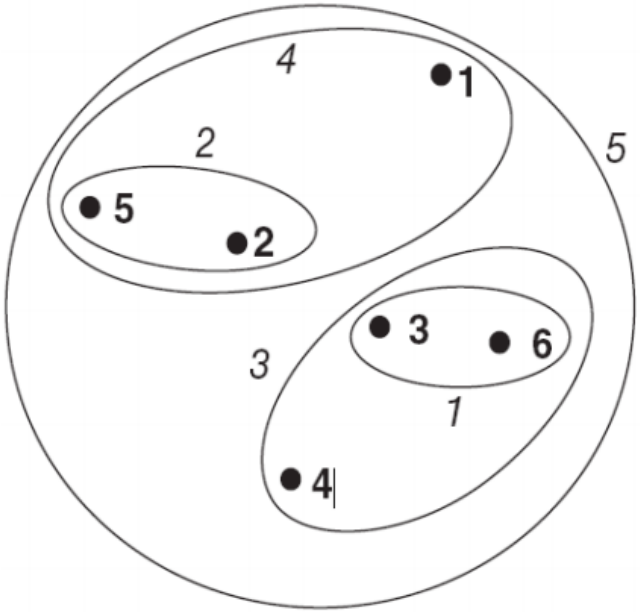
Дендрограмма



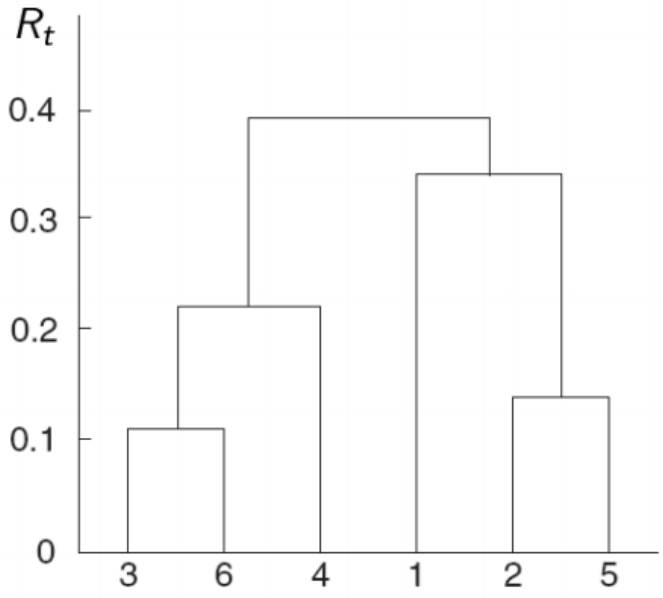
7.2. Иерархическая кластеризация

2. Расстояние дальнего соседа:

Диаграмма вложения



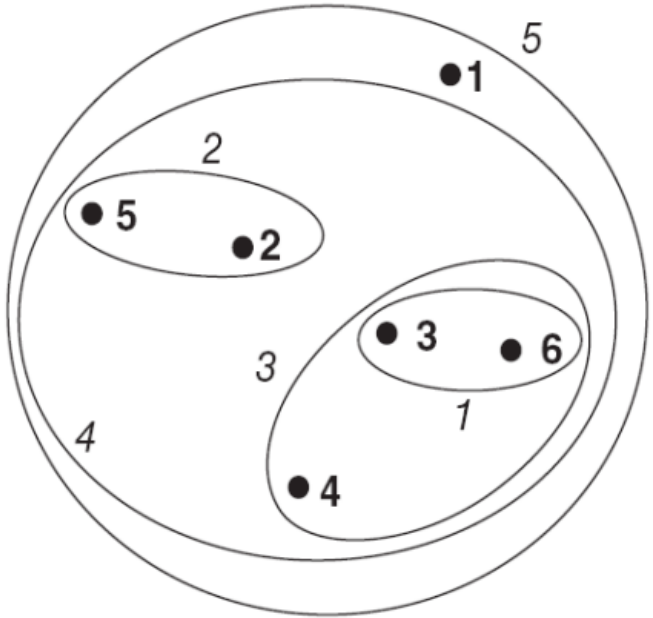
Дендрограмма



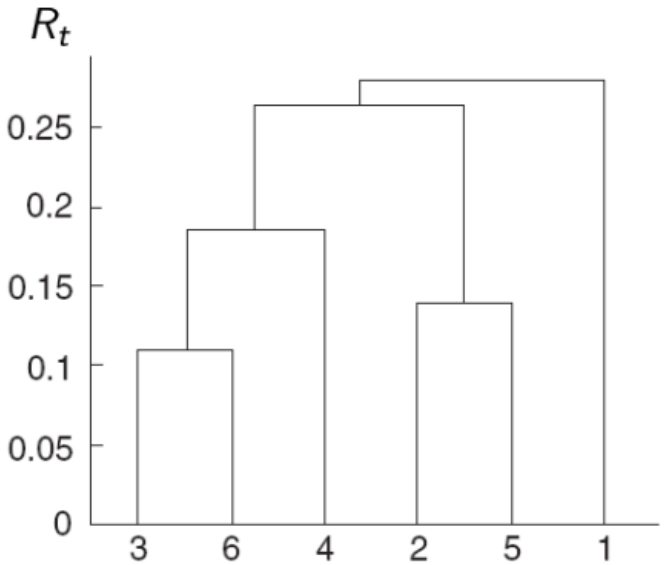
7.2. Иерархическая кластеризация

3. Групповое среднее расстояние:

Диаграмма вложения



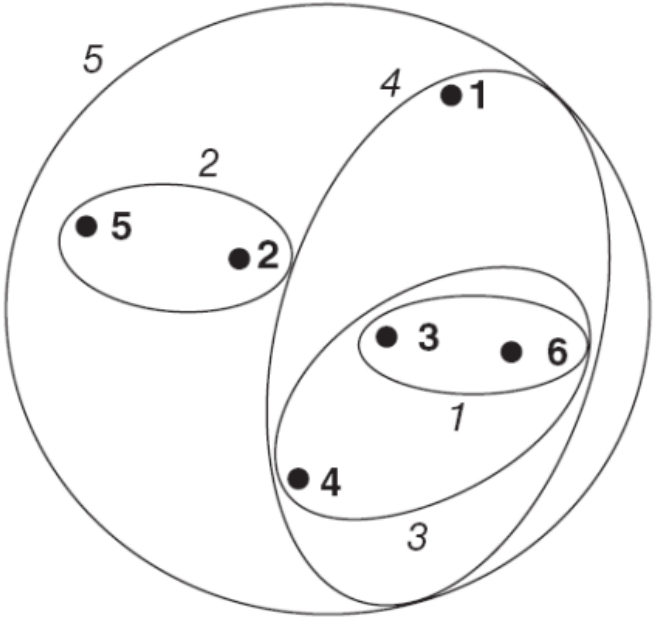
Дендрограмма



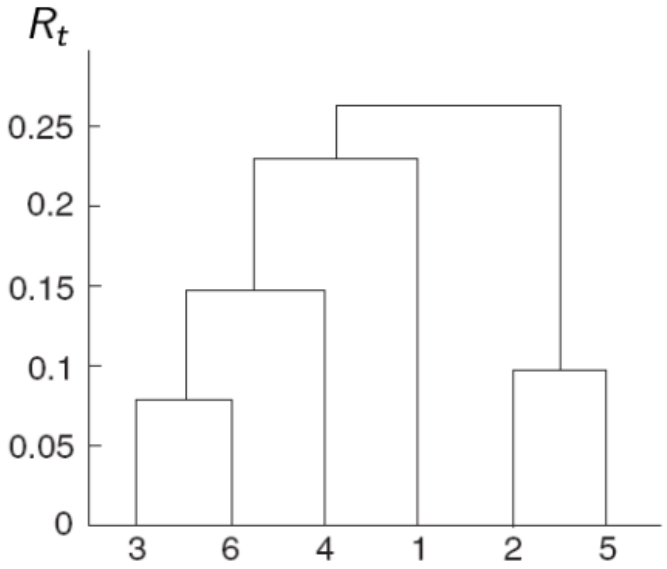
7.2. Иерархическая кластеризация

5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



7.2. Иерархическая кластеризация

Основные свойства иерархической классификации:

- Монотонность: дендрограмма не имеет самопересечений, при каждом слиянии расстояний между объединяемыми кластерами только увеличивается: $R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell$.
  Достаточное условие монотонности:
  $$\alpha_U \geqslant 0, \quad \alpha_V \geqslant 0, \quad \alpha_U + \alpha_V + \beta \geqslant 1, \quad \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$$
- Сжимающее расстояние: $R_t \leqslant \rho(\mu_U, \mu_V), \ \forall t.$
- Растягивающее расстояние: $R_t \geqslant \rho(\mu_U, \mu_V), \ \forall t$

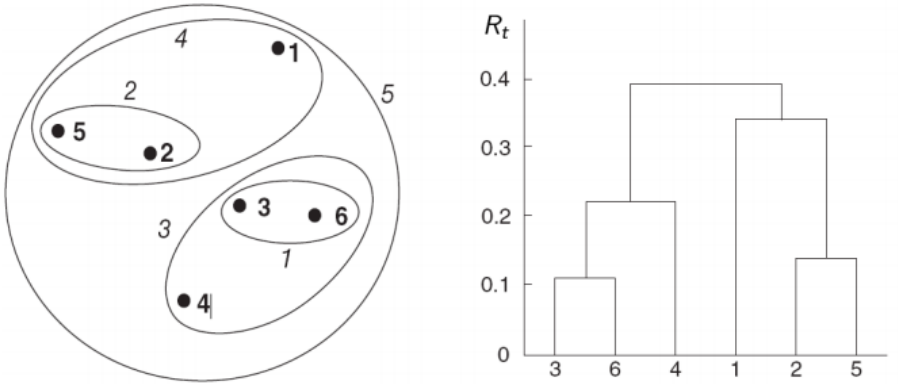
$R^{ц}$ не монотонно; $R^{б}$, $R^{д}$, $R^{г}$, $R^{у}$ монотонны

$R^{б}$ сжимающее; $R^{д}$, $R^{у}$ растягивающие

7.2. Иерархическая кластеризация

Рекомендации и выводы:

- рекомендуется пользоваться расстоянием Уорда R^y
- обычно строят несколько вариантов и выбирают лучший визуально по дендрограмме
- определение числа кластеров – по максимуму $|R_{t+1} - R_t|$, тогда результирующее множество кластеров := C_t



7.3. Нелинейные методы понижения размерности

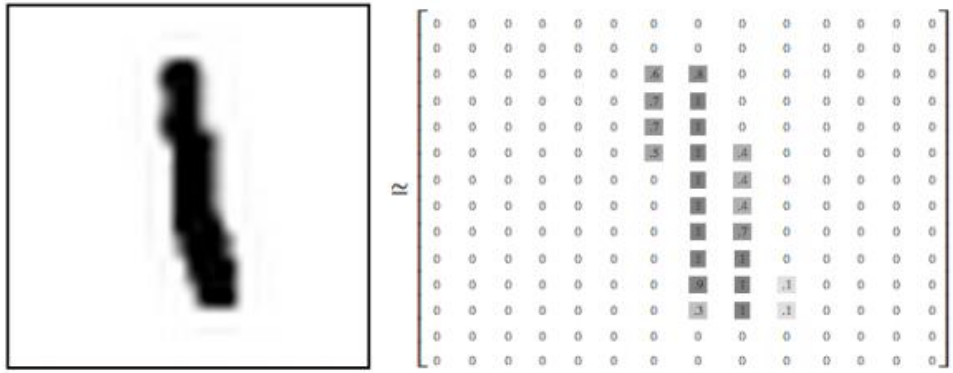
7.3. Нелинейные методы понижения размерности

Пример: набор изображений рукописных цифр MNIST

A selection from the 64-dimensional digits dataset

7.3. Нелинейные методы понижения размерности

- Объект из набора - изображение 28x28 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?

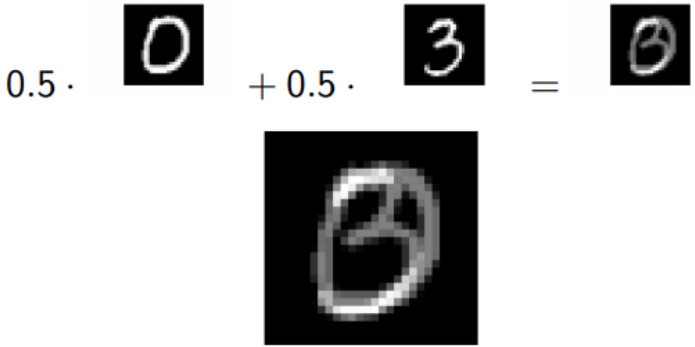
7.3. Нелинейные методы понижения размерности

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

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

7.3. Нелинейные методы понижения размерности

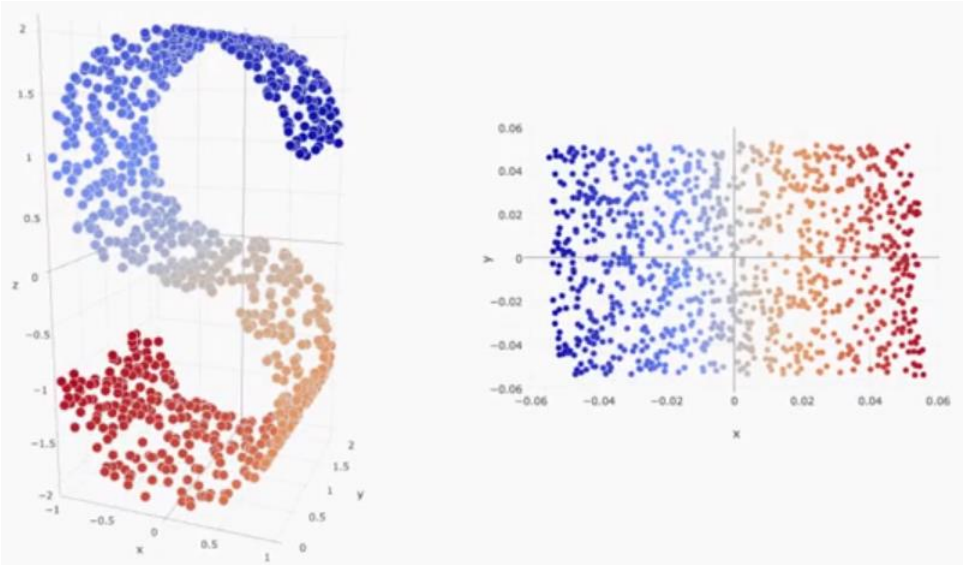
- Линейная комбинация признаков изображений рукописных цифр может не соответствовать рукописной цифре:



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

7.3. Нелинейные методы понижения размерности

В трехмерном признаковом пространстве имеется S-образная поверхность, на которой лежать объекты выборки, которые можно отобразить на двухмерную координатную плоскость с сохранением информации об объектах.



7.3. Нелинейные методы понижения размерности

Дано: $\{x_1, \ldots, x_\ell\}$ - выборка объектов

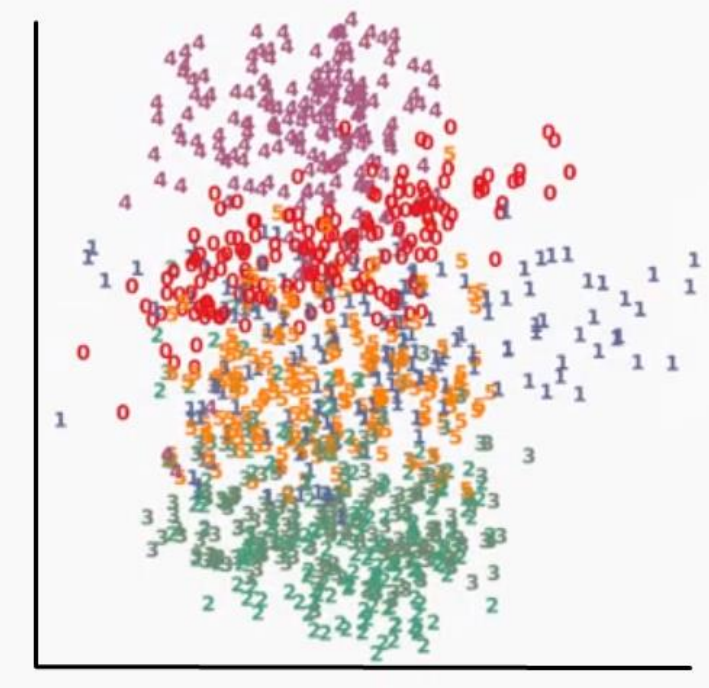
Задача: отобразить все объекты выборки в пространство малой размерности $x_i \mapsto \tilde{x}_i \in \mathbb{R}^d$.

Требование к маломерному представлению:

- Должно хорошо сохранять структуру данных в исходном пространстве.
- Сохранять интересующие нас закономерности в данных.

7.3. Нелинейные методы понижения размерности

Координаты оси соответствуют компонентам проекции. Различные цифры выделены цветом.



7.3. Нелинейные методы понижения размерности

Многомерное шкалирование (Multidimensional Scaling, MDS)

Гипотеза: хорошее малоразмерное представление сохраняет попарные расстояния между объектами

d_{ij} - расстояние между x_i и x_j

Признаковые описания не нужны, достаточно расстояний

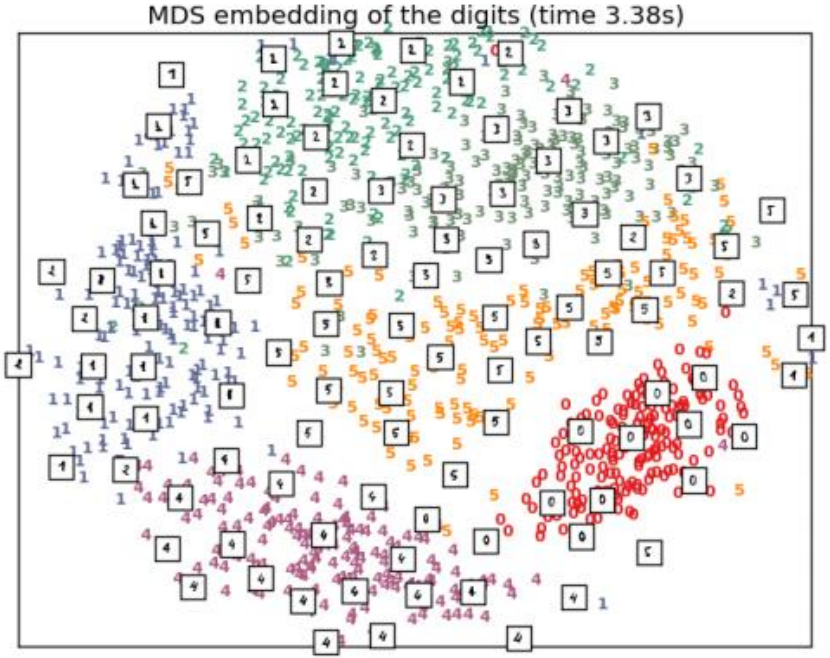
$\tilde{d}_{ij} = \|\tilde{x}_i - \tilde{x}_j\|$ - евклидово расстояние между маломерными представлениями

Ищем представление, аппроксимирующее d_{ij} (оптимизируем методом SMACOF)

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

7.3. Нелинейные методы понижения размерности

Уменьшение признакового пространства до 2х при помощи метода MDS



7.3. Нелинейные методы понижения размерности

Stochastic Neighbor Embedding (SNE):

- В точности воспроизвести расстояния - слишком сложно, достаточно сохранения пропорции $\rho(x_1, x_2) = c\rho(x_1, x_3) \Rightarrow \rho(\tilde{x}_1, \tilde{x}_2) = c\rho(\tilde{x}_1, \tilde{x}_3)$.
- Опишем объекты нормированными расстояниями до других объектов

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

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

- Минимизируем разницу между распределениями расстояний (мера - дивергенция Кльбака-Лейблера)

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

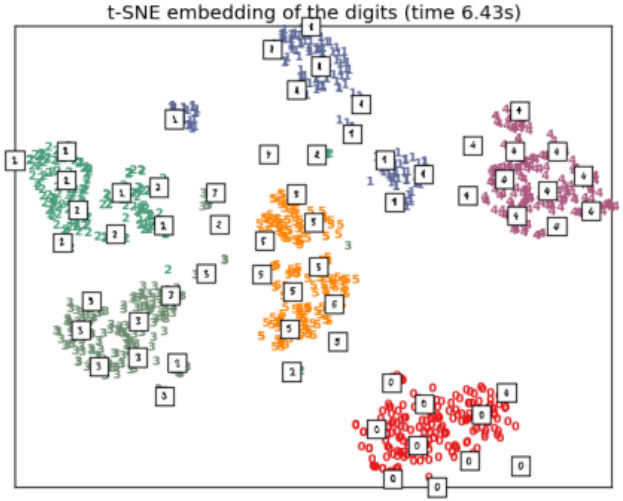
7.3. Нелинейные методы понижения размерности

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:

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

7.3. Нелинейные методы понижения размерности



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

7.3. Нелинейные методы понижения размерности

Рукописные единички 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.