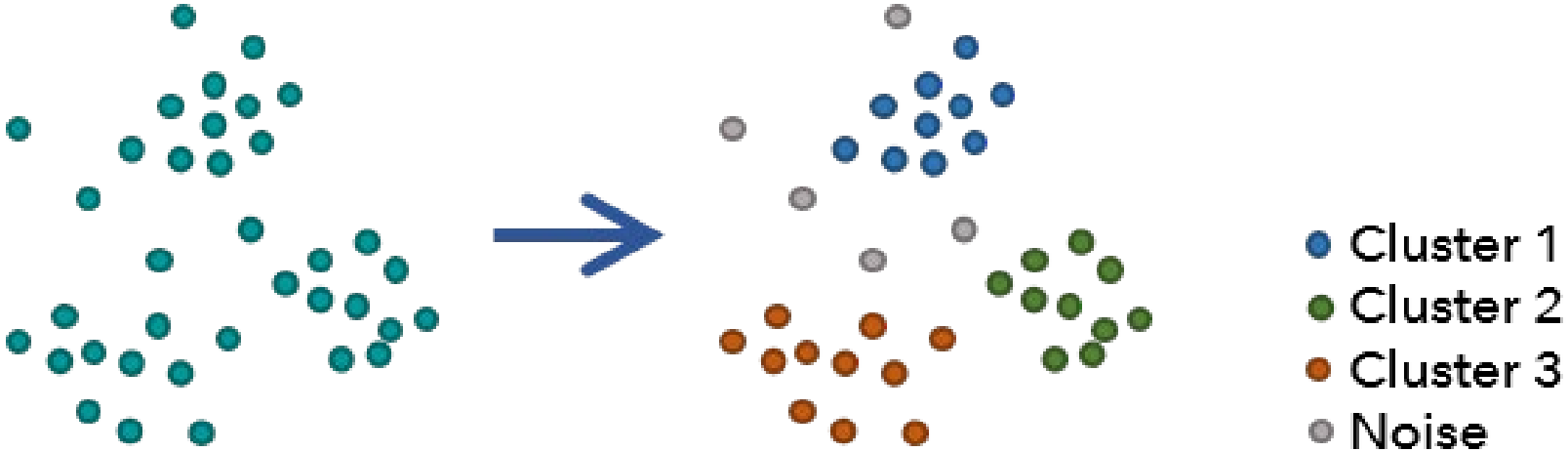
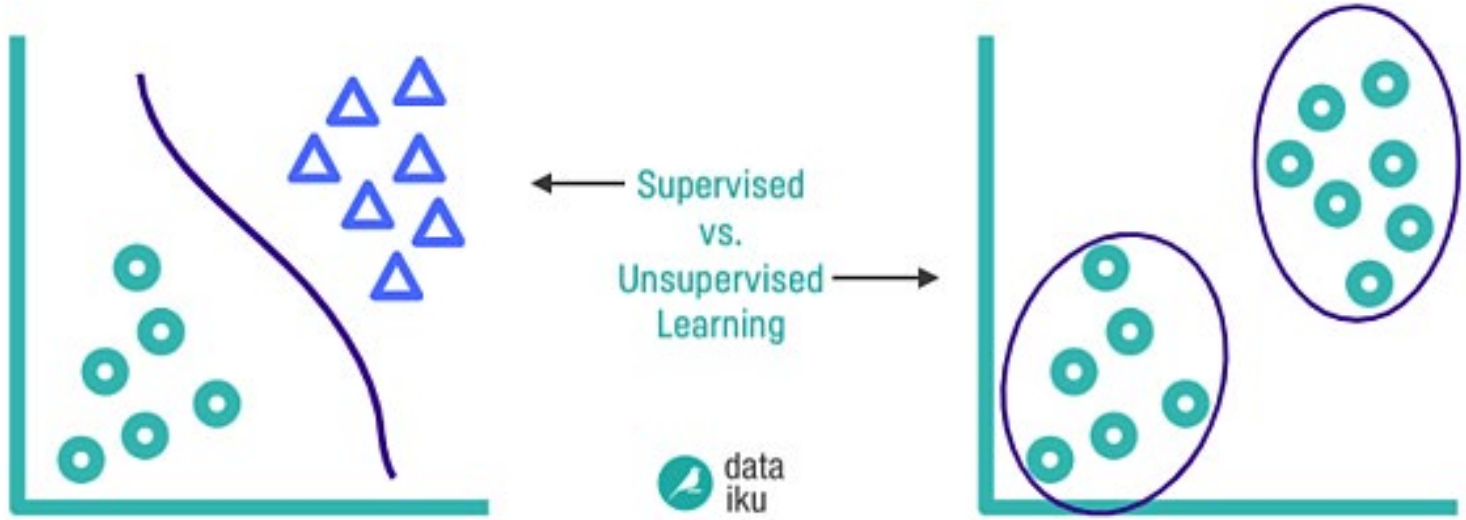
Лекция 07
Обучение без учителя (неподконтрольное)

А. Кластеризация: К-средних, DBSCAN, Иерархическая
Б. Карта Кохонена и ее варианты
В. Уменьшение размерности и визуализация: PCA, t-SNE

Кластеризация vs классификация

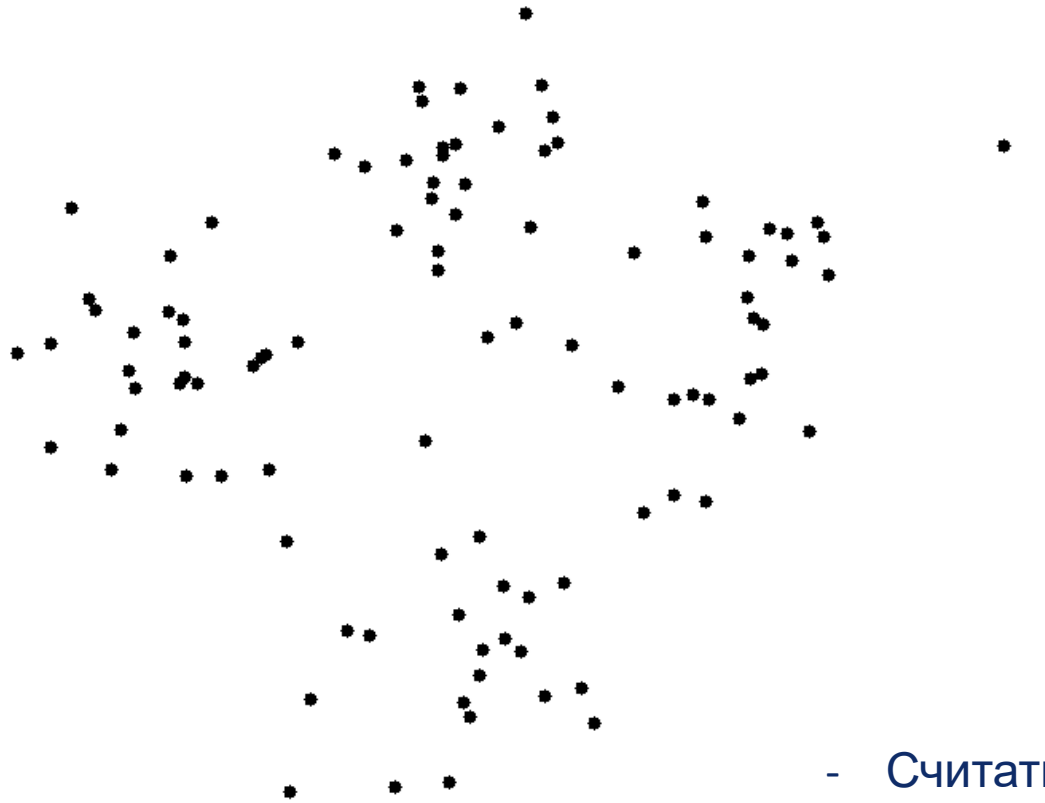




https://pro.arcgis.com/ru/pro-app/2.9/tool-reference/spatial-statistics/densitybasedclustering.htm

https://blog.dataiku.com/clustering-how-it-works-in-plain-english

Метод К-средних



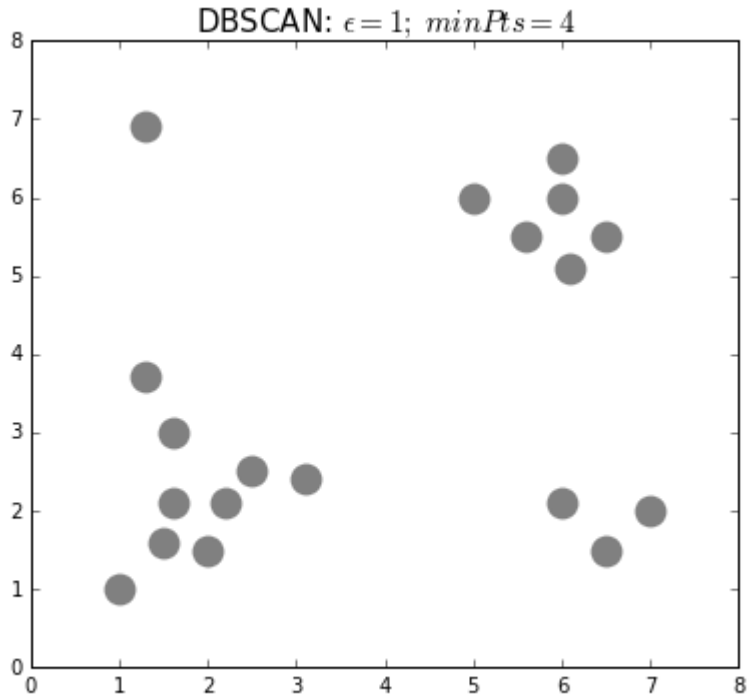
0. Назначить К «центроидов»
1. Посчитать расстояния от «центроидов» до всех точек
2. Распределить точки к ближайшим «центроидам»
3. Посчитать геометрический центр точек, принадлежащих каждому «центроиду»
4. Передвинуть «центроиды» в геометрический центр его точек
5. Повторить с п.1.

- Считать медианы
- Другая метрика расстояния, ускорение расчетов расстояния
- Ввести взаимодействие между центроидами
- Инициализация центроидов

http://shabal.in/visuals/kmeans/6.html

Метод DBSCAN


```
DBSCAN(DB, distFunc, eps, minPts) {
   C=0                                                  /* Счётчик кластеров */
   for each point P in database DB {
      if label(P) ≠ undefined then continue               /* Точка была просмотрена во внутреннем цикле */
      Neighbors N=RangeQuery(DB, distFunc, P, eps)      /* Находим соседей */
      if|N|< minPts then {                              /* Проверка плотности */
         label(P)=Noise                                 /* Помечаем как шум */
         continue
      }
      C=C + 1                                           /* следующая метка кластера */
      label(P)=C                                        /* Помечаем начальную точку */
      Seed set S=N \ {P}                                /* Соседи для расширения */
      for each point Q in S {                             /* Обрабатываем каждую зачаточную точку */
         if label(Q)=Noise then label(Q)=C             /* Заменяем метку Шум на Край */
         if label(Q) ≠ undefined then continue            /* Была просмотрена */
         label(Q)=C                                     /* Помечаем соседа */
         Neighbors N=RangeQuery(DB, distFunc, Q, eps)   /* Находим соседей */
         if|N|≥ minPts then {                           /* Проверяем плотность */
            S=S ∪ N                                     /* Добавляем соседей в набор зачаточных точек */
         }
      }
   }
}
```

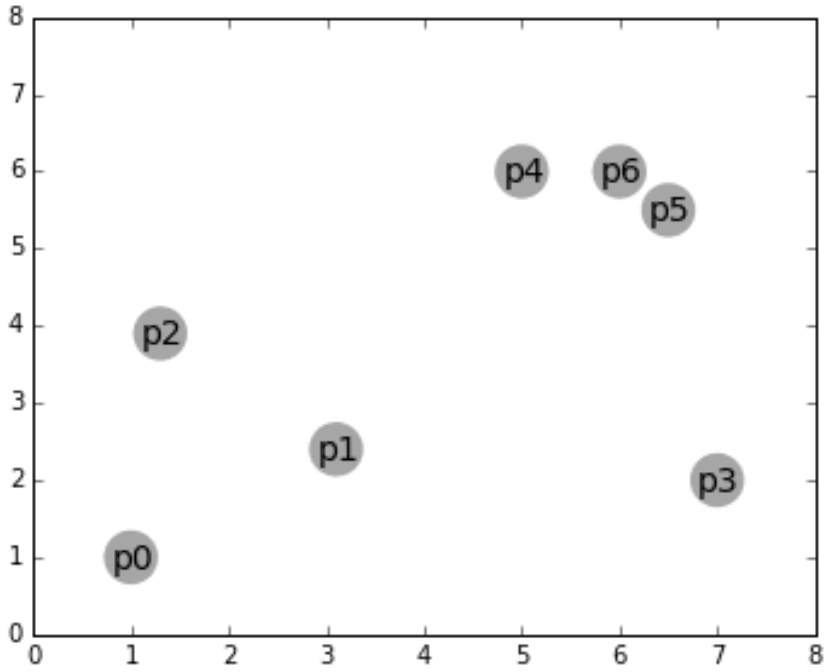


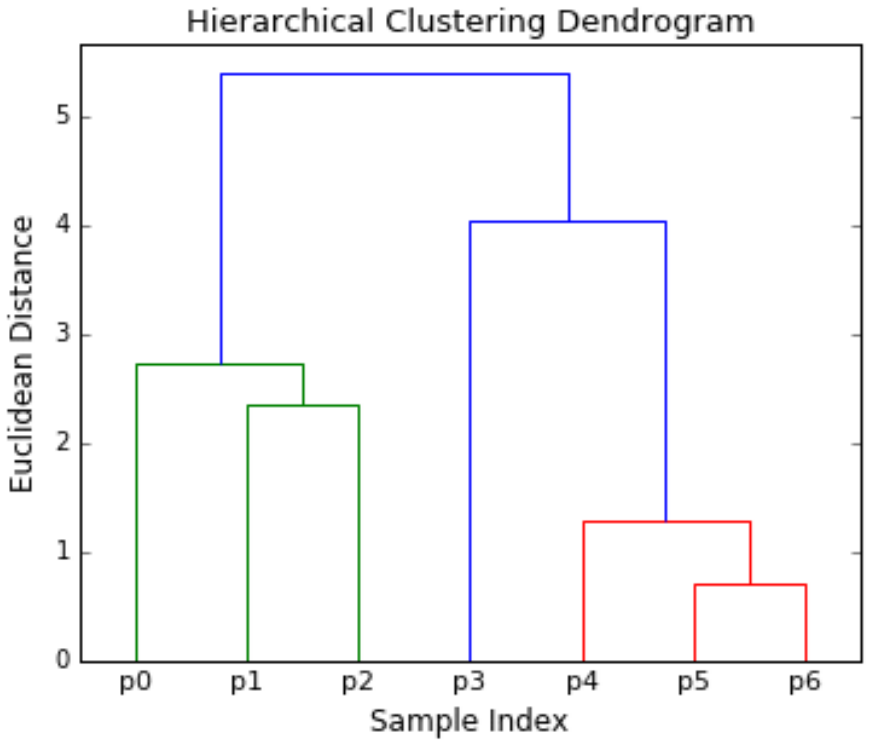
- Другая метрика расстояния, ускорение расчетов расстояния
- Иерархия кластеров по плотности
- Оценка гиперпараметров

https://ru.wikipedia.org/wiki/DBSCAN

https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Иерархическая кластеризация





0. Каждый пример – свой кластер. Считаем расстояние между всеми кластерами.
1. Выбрать ближайшие кластеры.
2. Слить их в один кластер.
3. Удалить слитые из списка кластеров.
4. Добавить новый кластер в список, его координаты. Отметить расстояние, на котором произошло слияние.
5. Пересчитать расстояние между этим кластером и другими.
6. Повторить с п.1.

- Другая метрика расстояния, ускорение расчетов расстояния

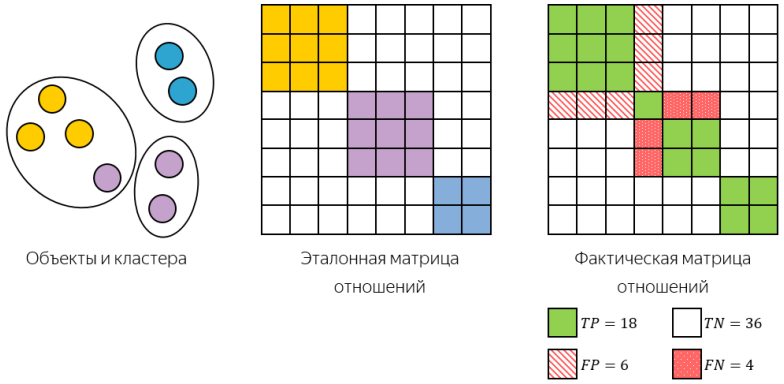
https://neerc.ifmo.ru/wiki/index.php?title=Иерархическая_кластеризация

https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Метрики кластеризации

ВНЕШНИЕ

Есть эталонные кластеры (классы) (редкость!)



$X \backslash Y$	Y_1	Y_2	...	Y_s	Sums
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{22}	...	n_{2s}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
X_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2	...	b_s	n

$$Rand = \frac{TP + TN}{TP + TN + FP + FN} \qquad Jaccard = \frac{TP}{TP + FN + FP}$$

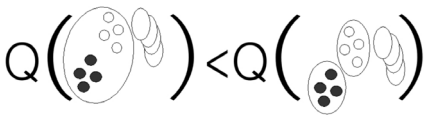
Пусть $p_{ij} = \frac{n_{ij}}{n}, p_i = \frac{a_i}{n}, p_j = \frac{b_j}{n}$.

- Элементы принадлежат одному кластеру и одному классу — TP
- Элементы принадлежат одному кластеру, но разным классам — FP
- Элементы принадлежат разным кластерам, но одному классу — FN
- Элементы принадлежат разным кластерам и разным классам — TN

$$\overbrace{ARI}^{\text{Adjusted Index}} = \frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}} - \overbrace{[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected Index}}}{\underbrace{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}]}_{\text{Max Index}} - \underbrace{[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected Index}}}$$

$$E = -\sum_i p_i \left(\sum_j \frac{p_{ij}}{p_i} log\left(\frac{p_{ij}}{p_i}\right)\right)$$

однородность



Пары примеров. Число перестановок: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

https://habr.com/ru/company/yandex/blog/500742/

https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче_кластеризации

Метрики кластеризации

ВНУТРЕННИЕ

Нет эталонных кластеров.
Сравниваем кластеры между собой и их центроиды

Компактность

$$WSS = \sum_{j=1}^{M} \sum_{i=1}^{|C_j|} (x_{ij} - \overline{x_j})^2$$

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \|x_i - x_j\|$$

Отделимость

$$BSS = n \cdot \sum_{j=1}^{M} (\overline{x_j} - \overline{x})^2$$

$$b(x_i, c_k) = min_{c_l \in C \setminus c_k} \{ \frac{1}{|c_l|} \sum_{x_j \in c_l} \|x_i - x_j\| \}$$

Силуэт

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}},$$

Индекс Дэвиса-Болдуина

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{\|\overline{c_k} - \overline{c_l}\|} \right\},$$

где:

$$S(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \|x_i - \overline{c_k}\|$$

https://habr.com/ru/company/yandex/blog/500742/

https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче_кластеризации

Карта Кохонена

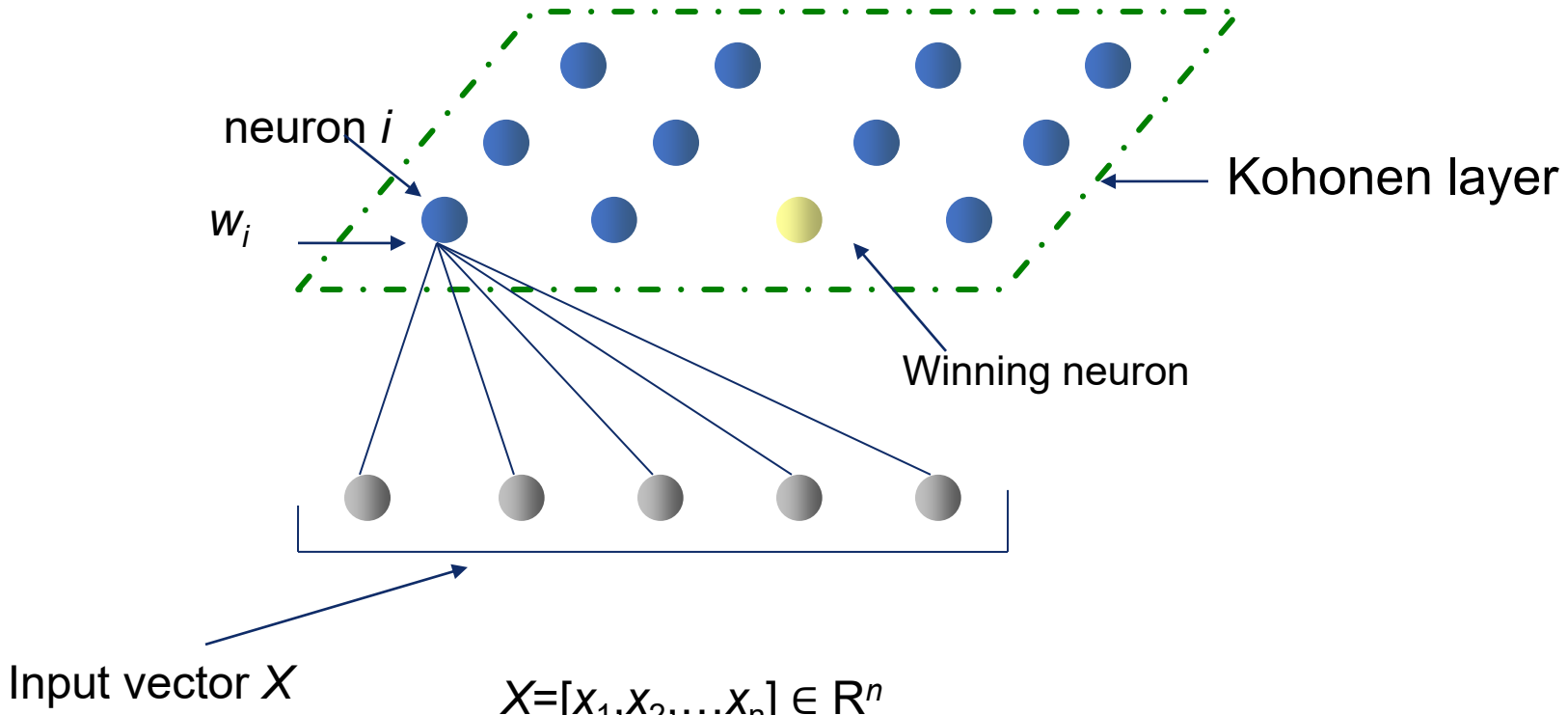
0. Назначить К «центроидов»
1. Посчитать расстояния от «центроидов» до всех точек
2. Распределить точки к ближайшим «центроидам»
3. Посчитать геометрический центр точек, принадлежащих каждому «центроиду»
4. Передвинуть «центроиды» в геометрический центр его точек
5. Повторить с п.1.

НЕЗАВИСИМЫ

- Считать медианы
- Другая метрика расстояния, ускорение расчетов расстояния
- Ввести взаимодействие между центроидами
- Инициализация центроидов

http://shabal.in/visuals/kmeans/6.html

Карта Кохонена. Структура



$X=[x_1,x_2,\ldots x_n] \in R^n$

$w_i=[w_{i1},w_{i2},\ldots,w_{in}] \in R^n$

Карта Кохонена. Обучение

1) Инициализация центроидов нейронов *Wi*

2) Входной *m*-мерный вектор *Xs* подается в карту

3) Вычисляем расстояния до всех нейронов d*i*(*Wi, Xs*):

$$d_i(W_i, X_s) = \sum_{j=1}^{m} \left(w_j - x_j\right)^2$$

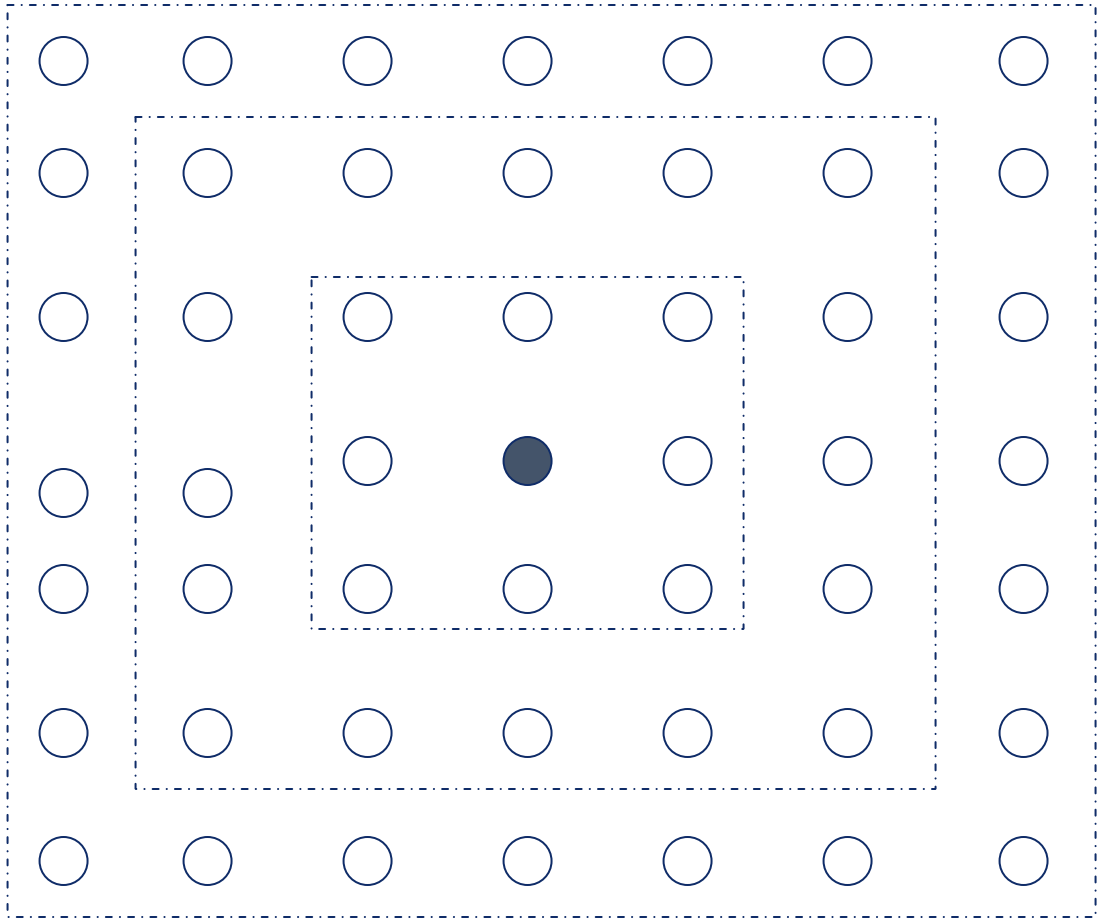
4) Ищем ближайший нейрон “победитель”

5) Сдвигаем нейрон-победитель и его соседей в направлении входа:

$$W_{jnew} = W_{jold} + \alpha(X_i - W_{jold})$$

α – скорость движения, у победителя больше, у соседей меньше

6) Повторяем с новым входом



Карта Кохонена. Применения и варианты

- Уменьшение размерности и Визуализация
- Кластеризация (требует доп. проверок)
- Классификация (аналогично К-средним)

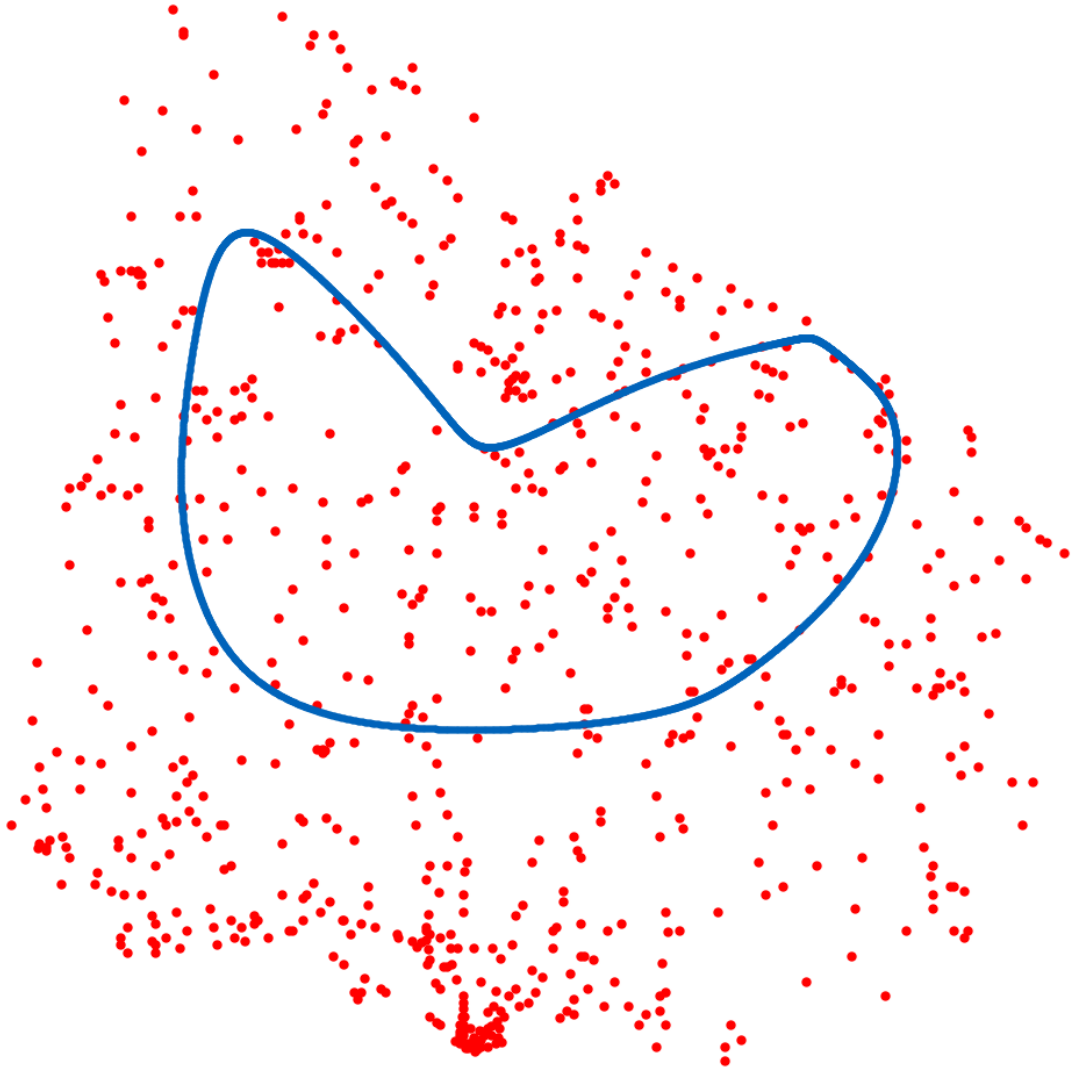
Варианты и модификации:

- Одномерные кольца – задача коммивояжера
- Растущий нейронный газ – добавлены механизмы изменения связей и нейронов – изменение сетки и построение метакластеров

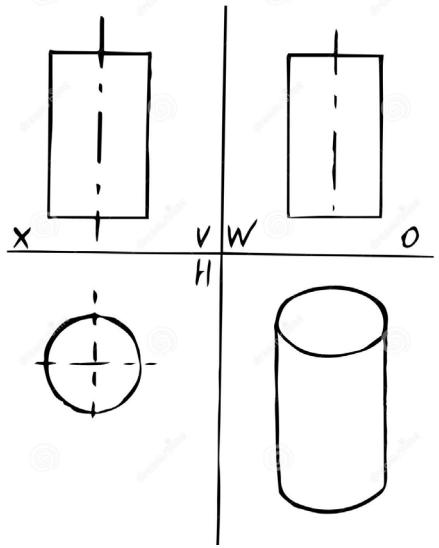
https://www.demogng.de/

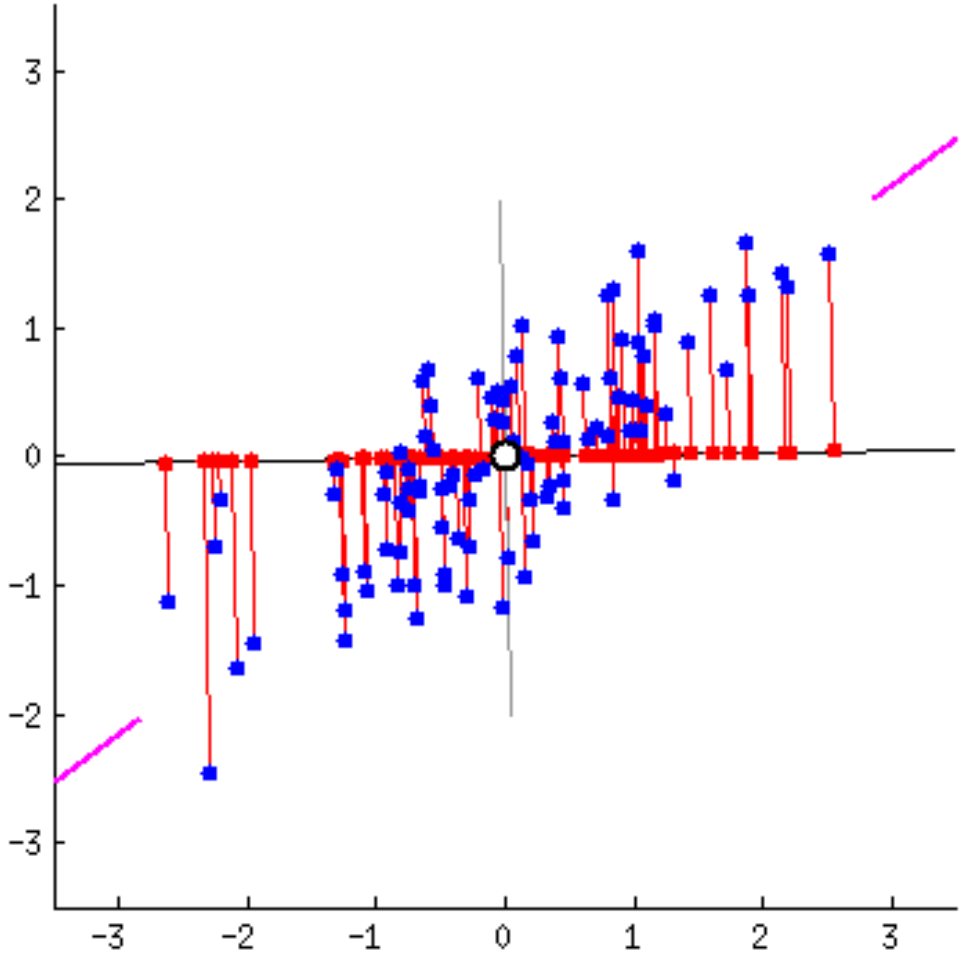
Маршрутизация и коммивояжер

- города это входные данные (точки)
- нейроны замкнуты в **кольцо**
- механизмы удаления нейронс (небыли победителями)
- механизмы дублирования нейронс (победители более 1 раза)



Уменьшение размерности и визуализация. Метод главных компонент





$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

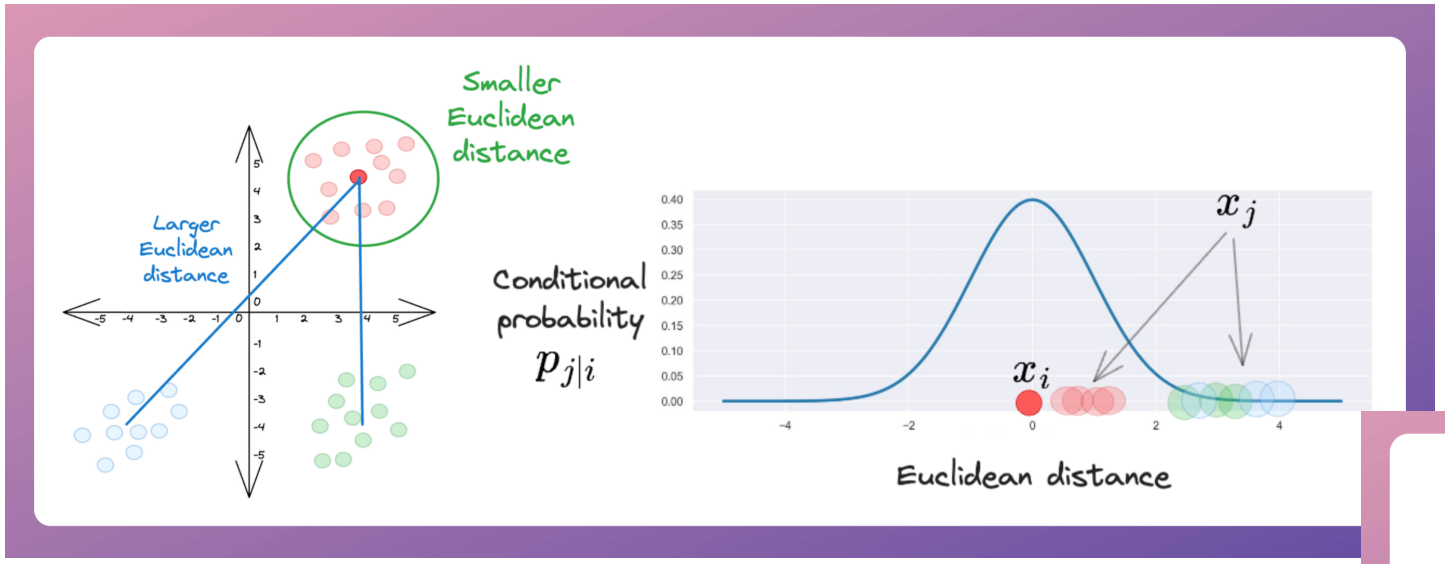
$$C = \frac{1}{m-1} X_{\text{scaled}}^T X_{\text{scaled}}$$

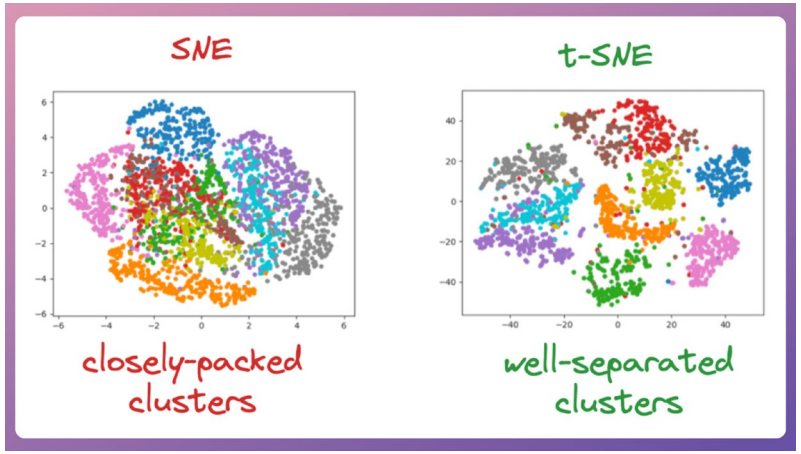
Cv=λv

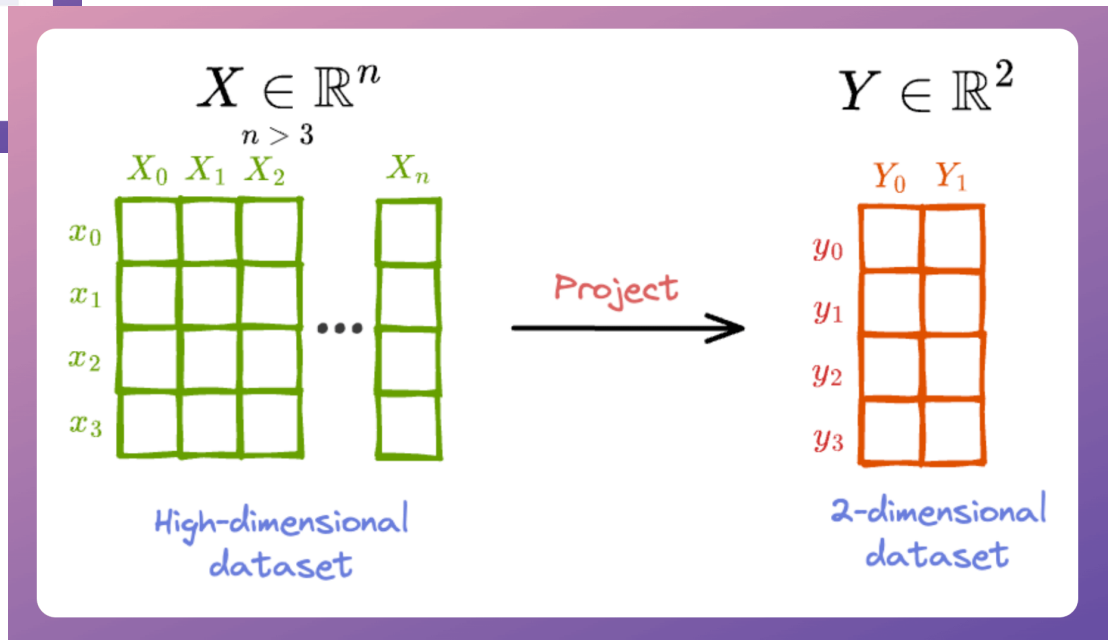
W=[v1,v2,…,vk]

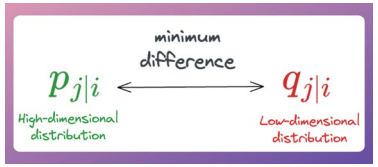
$$X_{\text{PCA}} = X_{\text{scaled}} W$$

Уменьшение размерности и визуализация. SNE, t-SNE









$$D_{KL}(P \parallel Q) = \sum_x P(x) \cdot \log\left(\frac{P(x)}{Q(x)}\right)$$

https://blog.dailydoseofds.com/p/t-sne-vs-sne-whats-the-difference

Q&A и контакты

Группа по дисциплине:

https://t.me/+8dShF1tFSDg0ZmJi