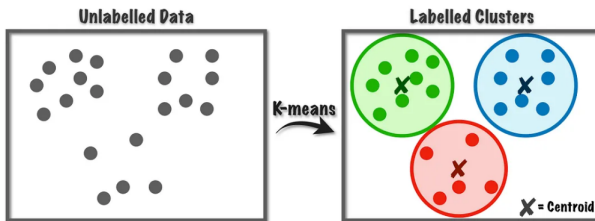


## Продвинутая кластеризация

Виктор Китов

[v.v.kitov@yandex.ru](mailto:v.v.kitov@yandex.ru)



# Содержание

- 1 Кластеризация, основанная на плотности объектов
  - Алгоритм DBScan
- 2 Иерархическая кластеризация
- 3 Оценка качества кластеризации

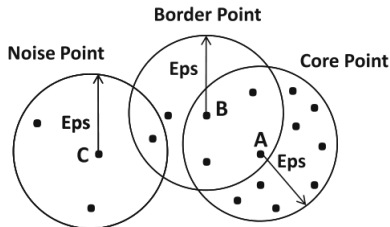
- 1 Кластеризация, основанная на плотности объектов
  - Алгоритм DBScan

# DBScan

$k, \varepsilon$  - параметры метода.

Разделим множество объектов на 3 категории:

- основные точки: имеющие  $\geq k$  точек внутри  $\varepsilon$ -окрестности
- пограничные точки: не основные, но содержащие хотя бы одну основную внутри  $\varepsilon$ -окрестности
- шумовые точки: не основные и не пограничные



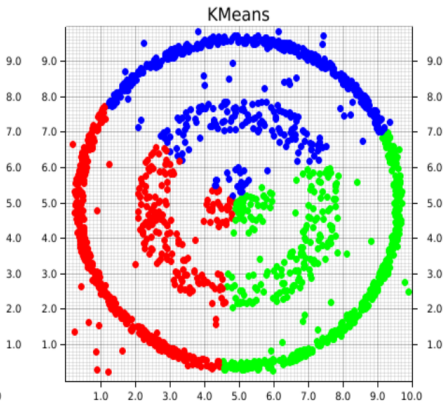
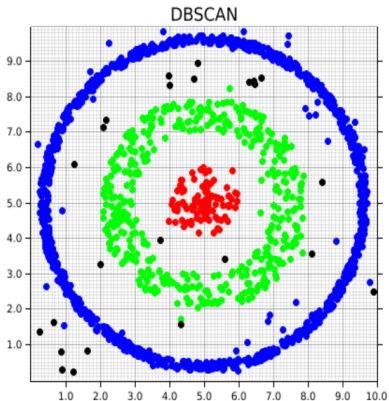
# Алгоритм

**ВХОД:** выборка, параметры  $\varepsilon, k$ .

- 1) Определить основные/пограничные/шумовые точки, используя  $\varepsilon, k$ .
- 2) Создать граф: узлы-основные точки, связи - если точки на расстоянии  $\leq \varepsilon$  друг от друга.
- 3) Определить компоненты связности в графе =кластеры (методом распространения).
- 4) Соотнести основные точки кластерам=компонентам связности, а пограничные-по основным в их  $\varepsilon$  окрестности.

**ВЫХОД:** разбиение на кластеры  
(основных и пограничных точек)

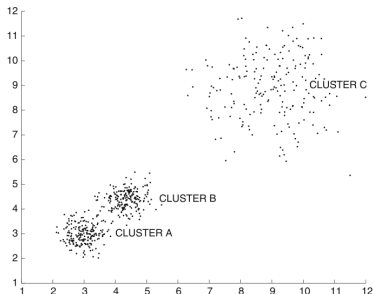
## Пример работы DBScan<sup>1</sup>



<sup>1</sup>Источник иллюстрации.

## Комментарии

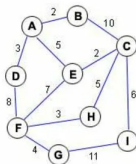
- Соединение основных точек - метод одиночной связи в аггломеративной кластеризации с остановкой  $\rho > \varepsilon$ .
- Преимущества: автоматически определяется  $\#$  кластеров, устойчиво к выбросам.
- Недостаток: не работает с кластерами разной плотности
  - высокое  $k$ -пропуским C; низкое  $k$ -A и B объединяться:



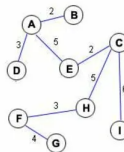
# Альтернативный графовый метод кластеризации

- Альтернативный графовый метод кластеризации:
  - 1 построить граф (узлы-объекты, вес связи-расстояние)
  - 2 построить минимальное остовное дерево<sup>2</sup>
  - 3 удалить  $K - 1$  ребро с макс. связью
- Получим  $K$  кластеров.
  - альтернатива: #кластеров определять автоматически, удаляя ребра с  $dist \geq threshold$ .

Исходный граф



Минимальное остовное дерево

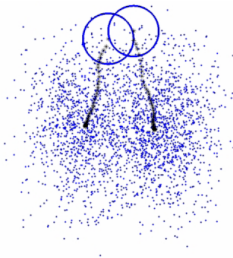


<sup>2</sup> Минимальное остовное дерево.



## Кластеризация сдвигом среднего значения

Кластеризация сдвигом среднего значения (mean shift): точки итеративно сдвигаются в направлении локального увеличения плотности по правилу



Пример сходимости для top-hat ядра  $K = \mathbb{I} \left[ \frac{\rho(z, x)}{h} \leq 1 \right]$

Кластер - итоговый локальный максимум плотности (отбрасываем максимумы с  $\rho(x) < \tau$ ).

## Комментарии

- Правило сдвига:

$$z_0 = x_n, \quad z = \frac{\sum_{k=1}^N K(\rho(z_i, x_k)/h) x_k}{\sum_{k=1}^N K(\rho(z, x_k)/h)}$$

- Ядро  $K(\cdot)$  - некоторая  $\downarrow$  ф-ция (ядро).
- Пример: Гауссово ядро

$$K(\rho(x, x')/h) = e^{-\rho(x, x')^2/h^2}$$

- Преимущества:
  - автоматически определяется #кластеров, кластеры могут быть произвольной формы
- Недостаток: вычислительная сложность, нет фильтрации выбросов

# Кластеризация mean shift

ВХОД: выборка  $x_1, \dots, x_N$ , ядро  $K(\cdot)$ , ширина окна  $h$ .

ДЛЯ  $n = 1, \dots, N$ :

$$z_n := x_n$$

ПОВТОРЯТЬ до сходимости:

$$z_n := \frac{\sum_{k=1}^N K(\rho(z_n, x_k)/h) x_k}{\sum_{k=1}^N K(\rho(z, x_k)/h)}$$

ассоциировать  $x_n$  пику  $z_n$

Объединить почти одинаковые расположения пиков  $z_1, \dots, z_N$ .

ВЕРНУТЬ кластеры точек, отнесенных одинаковым пикам плотности.

# Содержание

- 1 Кластеризация, основанная на плотности объектов
- 2 Иерархическая кластеризация
  - Иерархическая кластеризация сверху вниз
  - Иерархическая кластеризация снизу вверх
- 3 Оценка качества кластеризации

## Мотивация иерархической кластеризации

- #кластеров  $K$  заранее неизвестно.
- Кластеризация обычно не плоская, а иерархическая с разными уровнями детализации:
  - сайты в интернете
  - книги в библиотеке
  - животные в природе
- Подходы к иерархической кластеризации:
  - сверху вниз
    - более естественное для людей
  - снизу вверх (агломеративная кластеризация)

## 2 Иерархическая кластеризация

- Иерархическая кластеризация сверху вниз
- Иерархическая кластеризация снизу вверх

## Алгоритм

ВХОД:

выборка объектов, алгоритм плоской кластеризации  $A$ ,  
правила выбора листа и остановки

инициализировать дерево корнем, содержащим все объекты

ПОВТОРЯТЬ

выбрать лист  $L$  по правилу выбора листа  
используя  $A$  разбить  $L$  на кластеры  $L_1, \dots, L_K$   
добавить листы к  $T$ , соответствующие  $L_1, \dots, L_K$

ПОКА выполнено условие остановки

## Комментарии

- Алгоритм выбора листа:
  - ближайший к корню  
=> сбалансированное дерево по высоте
  - с максимальным числом элементов  
=> сбалансированное дерево по #объектов в листах

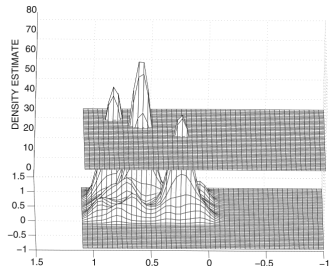
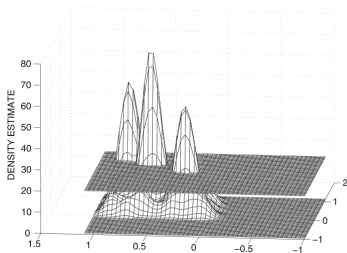


## 2 Иерархическая кластеризация

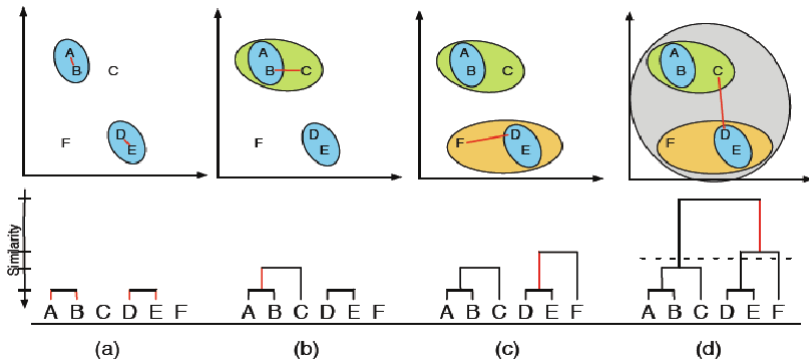
- Иерархическая кластеризация сверху вниз
- Иерархическая кластеризация снизу вверх

# DENCLUE - иерархическое обобщение mean shift

- 1 Производим кластеризацию методом mean shift.
- 2 Объединяем кластеры с пиками, соединяемые цепочкой высоко вероятных значений плотности  $p(x_{i(k)}) \geq h$ .
  - варьируя  $h$  получаем иерархическую кластеризацию



## Аггломеративная кластеризация - идея



## Аггломеративная кластеризация - алгоритм

инициализировать матрицу попарных расстояний  $M \in \mathbb{R}^{N \times N}$  между кластерами из отдельных объектов  $\{x_1\}, \dots, \{x_N\}$

ПОВТОРЯТЬ:

- 1) выбрать ближайшие кластеры  $i$  и  $j$
- 2) объединить  $i, j \rightarrow \{i + j\}$
- 3) удалить строки/столбцы  $i, j$  из  $M$
- 4) добавить строку/столбец для нового  $\{i + j\}$

ПОКА не выполнено условие остановки

ВЕРНУТЬ иерархическую кластеризацию

- Условие остановки:
  - Остался 1 кластер либо осталось  $\leq K$  кластеров
  - расстояние между ближайшими кластерами  $\geq$  порога.
- Частичное обучение: если часть классов известна - объединяем  $i$  и  $j$ , только если там представители одного класса.

## Расстояние между кластерами

- Расстояние между объектами  $\Rightarrow$  расстояние между кластерами:

- Метод одиночной связи (single linkage)

$$\rho(A, B) = \min_{a \in A, b \in B} \rho(a, b)$$

- Метод полной связи (complete linkage)

$$\rho(A, B) = \max_{a \in A, b \in B} \rho(a, b)$$

- Метод средней связи (group average link)

$$\rho(A, B) = \text{mean}_{a \in A, b \in B} \rho(a, b)$$

- Центроидный метод (pair-group method using the centroid average)

$$\rho(A, B) = \rho(\mu_A, \mu_B)$$

$$\text{где } \mu_U = \frac{1}{|U|} \sum_{x \in U} x \text{ или } m_U = \text{median}_{x \in U} \{x\}$$

## Свойства межкластерных расстояний<sup>4</sup>

- Метод одиночной связи
  - извлекает кластеры произвольной формы
  - может случайно объединить разные кластеры цепочкой выбросов
  - $M_{(i \cup j)k} = \min\{M_{ik}, M_{jk}\}$
- Метод полной связи
  - создает компактные кластеры
  - $M_{(i \cup j)k} = \max\{M_{ik}, M_{jk}\}$
- Метод средней связи<sup>3</sup> и центроидный метод-компромисс между одиночной и полной связью.

---

<sup>3</sup>Как  $M_{(i \cup j)k}$  будет пересчитываться для него?

<sup>4</sup>Пусть мы модифицируем  $\rho(x, x')$  монотонным преобразованием  $F$ :  
 $\rho'(x, x') = F(\rho(x, x'))$ . Which of the cluster distances will not be affected by this change?

## Свойства межкластерных расстояний

Метод средней связи предпочтительнее центроидного, поскольку

- центроидный метод может приводить к немонотонной последовательности расстояний дендрограммы.
  - методы одиночной, полной и средней связи дают монотонную последовательность
- представление кластера его центром не учитывает структуру кластера
- центроидный метод предпочитает более крупные кластера, для которых центроиды получаются в среднем ближе

## Комбинация K-средних и аггломеративной

- Сложность аггломеративной кластеризации  $K$  объектов:  
 $O(K^2 \ln K)$ 
  - через алгоритм кучи
- Для снижения вычислений:
  - 1 применим K средних к  $N$  объектам (сложность  $O(N)$ )
  - 2 применим аггломеративную кластеризацию к найденным  $K$  кластерам
  - она позволяет выделять невыпуклые кластера



# Содержание

- 1 Кластеризация, основанная на плотности объектов
- 2 Иерархическая кластеризация
- 3 Оценка качества кластеризации
  - Оценки не использующие разметку
  - Оценки, использующие разметку

# Оценка качества кластеризации

## Оценка качества кластеризации:

- если кластеризация-промежуточный этап: по качеству итоговой задачи
- если нет разметки:
  - используют идею, что кластеризация хороша, если:
    - объекты одного кластера похожи
    - объекты разных кластеров непохожи
- если есть разметка:
  - учитывать инвариантность к переименованию
  - имеет смысл для малого #размеченных объектов
    - иначе - классификация

- 3 Оценка качества кластеризации
  - Оценки не использующие разметку
  - Оценки, использующие разметку

## Метрики качества<sup>5</sup>

- Пусть  $z_n$  - номер кластера для  $x_n$ .
- Среднее внутрикластерное расстояние:

$$F_0 = \frac{\sum_{i < j} \mathbb{I} \|z_i = z_j\| \rho(x_i, x_j)}{\sum_{i < j} \mathbb{I} \|z_i = z_j\|}$$

- Среднее межкластерное расстояние:

$$F_1 = \frac{\sum_{i < j} \mathbb{I} \|z_i \neq z_j\| \rho(x_i, x_j)}{\sum_{i < j} \mathbb{I} \|z_i \neq z_j\|}$$

- Композитные метрики:

$$F_0/F_1, F_1 - F_0$$

---

<sup>5</sup>Какие метрики нужно максимизировать, а какие - минимизировать?

## Индекс Дэвиса-Болдуина

- $s_i = \frac{1}{|C_i|} \sum_{n \in C_i} \rho(\mu_i, x_n)$  - диаметр кластера  $i$ .
- $d_{ij} = \rho(\mu_i, \mu_j)$  - расстояние между центроидами  $i$  и  $j$ .
- Качество разделения кластеров  $i$  и  $j$ :

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

- Индекс Дэвиса-Болдуина:

$$DB = \frac{1}{K} \sum_{k=1}^K \max_{i \neq j} R_{ij}$$

- ⊕ : Быстро вычисляется.
- ⊖ : Поощряет выпуклые кластера
- ⊖ : Из-за центроидов - завязано на Евклидово расстояние

## Коэффициент силуэта<sup>6</sup>

Качество кластеризации каждого объекта  $x_i$  определим по формуле:

$$Silhouette_i = \frac{d_i - s_i}{\max\{d_i, s_i\}}$$

где среднее расстояние от  $x_i$  до объектов

- $s_i$  - того же кластера
- $d_i$  - ближайшего чужого кластера

Общее качество классификации (коэффициент силуэта):

$$Silhouette = \frac{1}{N} \sum_{i=1}^N \frac{d_i - s_i}{\max\{d_i, s_i\}}$$

---

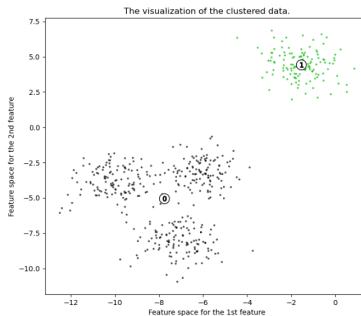
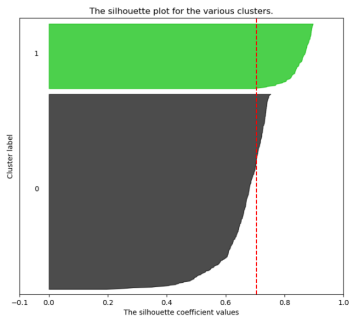
<sup>6</sup>Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65.

# Обсуждение

- Преимущества
  - Интерпретируемость:  $Silhouette \in [-1, 1]$ ,
    - 1: идеальная кластеризация
    - 0: случайная кластеризация
    - -1: полностью некорректная (инвертированная) кластеризация
- Недостатки
  - сложность  $O(N^2 D)$ 
    - можно рассчитывать по случайной подвыборке
  - поощряет выпуклые кластеры

## Подбор #кластеров по силуэту<sup>7</sup>

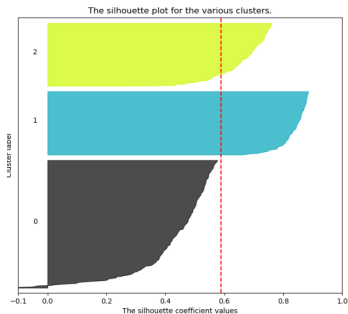
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.





## Подбор #кластеров по силуэту<sup>7</sup>

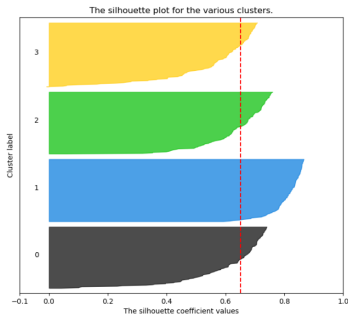
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



<sup>7</sup> Эксперимент в sklearn.

## Подбор #кластеров по силуэту<sup>7</sup>

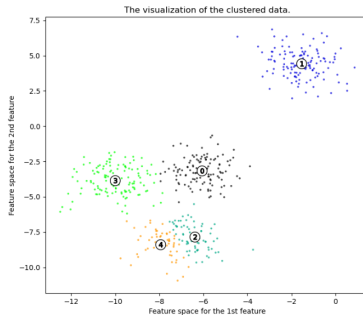
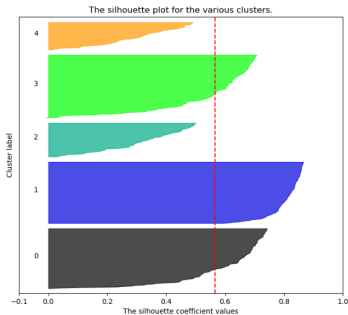
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



<sup>7</sup>Эксперимент в sklearn.

## Подбор #кластеров по силуэту<sup>7</sup>

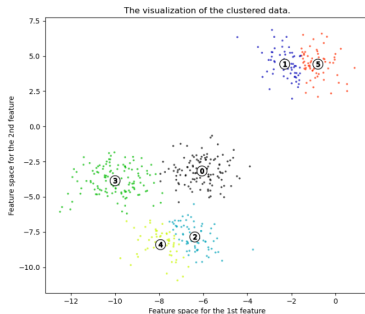
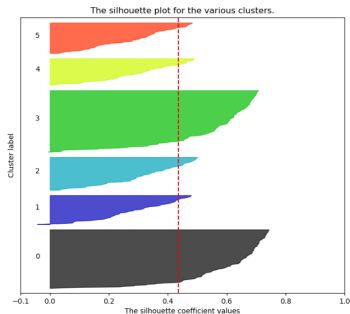
- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



<sup>7</sup>Эксперимент в sklearn.

## Подбор #кластеров по силуэту<sup>7</sup>

- Отсортируем объекты в каждом кластере по коэффициенту силуэта.
- Качество кластеризации - среднее значение коэффициента и отсутствие отрицательных значений.



<sup>7</sup>Эксперимент в sklearn.

## Индекс Калинского<sup>8</sup>

- Внутрикластерная (within cluster) ковариационная матрица

$$W = \frac{1}{N - K} \sum_{k=1}^K \sum_{x \in C_k} (x - \mu_k)(x - \mu_k)^T$$

- Межкластерная (between cluster) ковариационная матрица

$$B = \frac{1}{K - 1} \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

- Индекс Калинского:

$$I = \frac{\text{tr } B}{\text{tr } W} = \frac{N - K}{K - 1} \frac{\text{tr} \left\{ \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T \right\}}{\text{tr} \left\{ \sum_{k=1}^K \sum_{x \in C_k} (x - \mu_k)(x - \mu_k)^T \right\}}$$

---

<sup>8</sup>[https://www.researchgate.net/publication/233096619\\_A\\_Dendrite\\_Method\\_for\\_](https://www.researchgate.net/publication/233096619_A_Dendrite_Method_for_)

# Индекс Калинского

- Используем свойства

$$\sum_i \operatorname{tr} \{ \alpha_i A_i \} = \sum_i \alpha_i \operatorname{tr} A_i$$

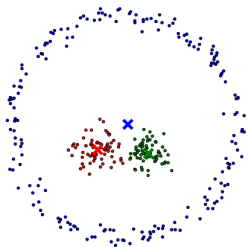
$$\operatorname{tr} \{ AB \} = \operatorname{tr} \{ BA \}, \operatorname{tr} a = a \quad \forall a \in \mathbb{R}$$

$$\begin{aligned} I &= \frac{\operatorname{tr} B}{\operatorname{tr} W} = \frac{N - K}{K - 1} \frac{\operatorname{tr} \left\{ \sum_{k=1}^K N_k (\mu_k - \mu) (\mu_k - \mu)^T \right\}}{\operatorname{tr} \left\{ \sum_{k=1}^K \sum_{x \in C_k} (x - \mu_k) (x - \mu_k)^T \right\}} \\ &= \frac{N - K}{K - 1} \frac{\sum_{k=1}^K N_k \operatorname{tr} \left\{ (\mu_k - \mu)^T (\mu_k - \mu) \right\}}{\sum_{k=1}^K \sum_{x \in C_k} \operatorname{tr} \left\{ (x - \mu_k) (x - \mu_k)^T \right\}} = \frac{N - K}{K - 1} \frac{\sum_{k=1}^K N_k \|\mu_k - \mu\|^2}{\sum_{k=1}^K \sum_{x \in C_k} \|x - \mu_k\|^2} \end{aligned}$$

- Измеряем отношение межкластерного к внутрикластерному разбросу.

## Ограничение для невыпуклого кластера

- Сложность  $O(ND)$ , но поощряет выпуклые кластеры.



- Из-за невыпуклости синего кластера коэффициент силуэта и индекс Калинского будут занижать хорошее качество кластеризации, т.к.
  - $s_i$  велико, а  $d_i$  - мало
  - $\sum_{k=1}^K N_k \|\mu_k - \mu\|^2$  мало, а  $\sum_{k=1}^K \sum_{x \in C_k} \|x - \mu_k\|^2$  велико

- 3 Оценка качества кластеризации
  - Оценки не использующие разметку
  - Оценки, использующие разметку



## Перекрестная таблица

- Пример перекрестной таблицы (contingency matrix):

	кластер 1	кластер 2	кластер 3
класс 1	5	2	0
класс 2	0	3	4

- ⊕ : Определяем разброс каждого класса по кластерам и разброс кластера по классам.
- ⊖ : Сложно анализировать для большого числа кластеров/классов. Не числовая метрика качества.

## Перекрестная таблица

- Пример перекрестной таблицы (contingency matrix):

	кластер 1	кластер 2	кластер 3
класс 1	5	2	0
класс 2	0	3	4

- $\oplus$  : Определяем разброс каждого класса по кластерам и разброс кластера по классам.
- $\ominus$  : Сложно анализировать для большого числа кластеров/классов. Не числовая метрика качества.
- Числовая мера качества - Unsupervised Clustering Accuracy:
  - $\Pi$  - все возможные перенумеровки результатов кластеризации

$$ACC(\mathbf{c}, \mathbf{z}) = \max_{\pi \in \Pi} \frac{1}{N} \sum_{n=1}^N \mathbb{I}[c_n = \pi(z_n)]$$

## Матрица сочетаемости

- Матрица сочетаемости  $\in \mathbb{R}^{2 \times 2}$  вычисляет счётчики  $\# \text{пар}$   $(x_i, x_j)$ .

	$z_i = z_j$	$z_i \neq z_j$
$y_i = y_j$	$n_{11}$	$n_{12}$
$y_i \neq y_j$	$n_{21}$	$n_{22}$

- Как понять по матрице качество кластеризации?

## Матрица сочетаемости

- Матрица сочетаемости  $\in \mathbb{R}^{2 \times 2}$  вычисляет счётчики  $\# \text{пар}$   $(x_i, x_j)$ .

	$z_i = z_j$	$z_i \neq z_j$
$y_i = y_j$	$n_{11}$	$n_{12}$
$y_i \neq y_j$	$n_{21}$	$n_{22}$

- Как понять по матрице качество кластеризации?
- $\oplus$  : Определяем сочетаемость разбиения по классам-кластерам.
- $\ominus$  : Не даёт единой метрики качества.

## Rand index

- Rand index - единая метрика по матрице сочетаемости.
- Пусть  $y_1, \dots, y_N$  - истинная разметка. Обозначим<sup>9</sup>

$$n_{11} = |\{(x_i, x_j) : z_i = z_j \ \& \ y_i = y_j\}|$$

$$n_{22} = |\{(x_i, x_j) : z_i \neq z_j \ \& \ y_i \neq y_j\}|$$

$$\text{RandIndex} = RI = \frac{n_{11} + n_{22}}{C_2^N} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \in [0, 1]$$

- В чем недостаток?

---

<sup>9</sup>Это loss или score?

## Rand index

- Rand index - единая метрика по матрице сочетаемости.
- Пусть  $y_1, \dots, y_N$  - истинная разметка. Обозначим<sup>9</sup>

$$n_{11} = |\{(x_i, x_j) : z_i = z_j \ \& \ y_i = y_j\}|$$

$$n_{22} = |\{(x_i, x_j) : z_i \neq z_j \ \& \ y_i \neq y_j\}|$$

$$\text{RandIndex} = RI = \frac{n_{11} + n_{22}}{C_2^N} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \in [0, 1]$$

- В чем недостаток? Для случайной кластеризации  $\text{RandIndex} > 0$  ( $\uparrow RI$  с  $\uparrow \#$ кластеров). Чтобы было  $= 0$  нужно использовать

$$\text{AdjustedRandIndex} = ARI = \frac{RI - \mathbb{E}\{RI\}}{\max(RI) - \mathbb{E}\{RI\}}$$

---

<sup>9</sup>Это loss или score?

Гомогенность<sup>10</sup>

- Обозначим  $N = \# \text{объектов}$ ,  $n_k = \# \text{объектов в кластере } k$ ,  $m_c = \# \text{объектов в классе } c$ ,  $n_{ck} = \# \text{объектов класса } c \text{ в кластере } k$ .

$$H_{class} = - \sum_{c=1}^C \frac{m_c}{N} \log \frac{m_c}{n}$$

$$H_{clust} = - \sum_{k=1}^K \frac{n_k}{N} \log \frac{n_k}{n}$$

$$H_{class|clust} = - \sum_{k=1}^K \frac{n_k}{N} \sum_{c=1}^C \frac{n_{ck}}{n_k} \log \frac{n_{ck}}{n}$$

$H_{class|clust} = 0$  при полном объяснении,  $H_{class|clust} = 1$  нет связи

<sup>10</sup><https://aclanthology.org/D07-1043.pdf>

## Гомогенность

$$\text{Homogeneity} = 1 - \frac{H(class|clust)}{H(class)}$$

- Гомогенность показывает долю информации о классах, объясненной кластеризацией.
  - 0: в кластерах представители только 1 класса
  - 1: в кластерах распределение классов=априорному распределению
- Какой недостаток?



## Гомогенность

$$\text{Homogeneity} = 1 - \frac{H(\text{class}|\text{clust})}{H(\text{class})}$$

- Гомогенность показывает долю информации о классах, объясненной кластеризацией.
  - 0: в кластерах представители только 1 класса
  - 1: в кластерах распределение классов=априорному распределению
- Какой недостаток? Гомогенность поощряет  $\uparrow \#$  кластеров
  - $=1$ , когда каждый объект - в своём кластере

Полнота<sup>11</sup>

- Нужна доп. мера полноты (насколько объекты одного класса оказываются в одном кластере)

$$\text{Completeness} = 1 - \frac{H(\text{clust}|\text{class})}{H(\text{clust})}$$

- Полнота = 1, если класс полностью определяет кластер (все объекты класса-в одном кластере)
- Какой недостаток?

---

<sup>11</sup><https://aclanthology.org/D07-1043.pdf>

## Полнота<sup>11</sup>

- Нужна доп. мера полноты (насколько объекты одного класса оказываются в одном кластере)

$$\text{Completeness} = 1 - \frac{H(\text{clust}|\text{class})}{H(\text{clust})}$$

- Полнота =1, если класс полностью определяет кластер (все объекты класса-в одном кластере)
- Какой недостаток? Полнота поощряет ↓#кластеров
  - =1, когда все объекты в одном кластере

---

<sup>11</sup><https://aclanthology.org/D07-1043.pdf>

## V-мера<sup>12</sup>

- V-мера - среднее гармоническое от гомогенности и полноты.

$$V = \frac{1}{\frac{1}{2}\text{Homogeneity} + \frac{1}{2}\text{Completeness}}$$

- Взвешенный учёт гомогенности и полноты:

$$V_{\beta} = \frac{1}{\left(\frac{\beta}{1+\beta}\right)\text{Homogeneity} + \frac{1}{1+\beta}\text{Completeness}}$$

- $V = V_{\beta}$  при  $\beta = 1$ .

---

<sup>12</sup><https://aclanthology.org/D07-1043.pdf>

## Нормализованная взаимная информация

- Взаимная информация - степень связи сл. вел.  $X, Y$ :

$$\begin{aligned}
 MI(X, Y) &= KL(P(X, Y) || P(X)P(Y)) \\
 &= \sum_{x \in \text{dom}(X)} \sum_{y \in \text{dom}(Y)} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}
 \end{aligned}$$

$$MI(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

- Нормализованная взаимная информация ( $NMI \in [0, 1]$ )<sup>13</sup> - др. вариант агрегации полноты и гомогенности:

$$\begin{aligned}
 NMI(\text{clust}, \text{class}) &= \frac{MI(\text{clust}, \text{class})}{\max\{H_{\text{clust}}, H_{\text{class}}\}} \\
 &= \frac{H(\text{clust}) - H(\text{clust}|\text{class})}{\max\{H_{\text{clust}}, H_{\text{class}}\}} = \frac{H(\text{class}) - H(\text{class}|\text{clust})}{\max\{H_{\text{clust}}, H_{\text{class}}\}}
 \end{aligned}$$

<sup>13</sup> Это loss или score?

## Заключение

- Плоская кластеризация:
  - К представителей
    - $\mu_k$  - вычисляемый (среднее: K-means [доступно ядерное обобщение], медиана: K medians)
    - $\mu_k$  - существующий объект
  - Основанная на плотности
    - DB-scan, mean-shift, DENCLUE
- Иерархическая кластеризация
  - сверху-вниз: рекурсивная плоская кластеризация
  - снизу-вверх (агломеративная)
- Оценка качества кластеризации:
  - размеры кластеров vs. межкластерные расстояния
  - сопоставление кластеров с истинными метками
    - важна инвариантность к перенумеровке кластеров