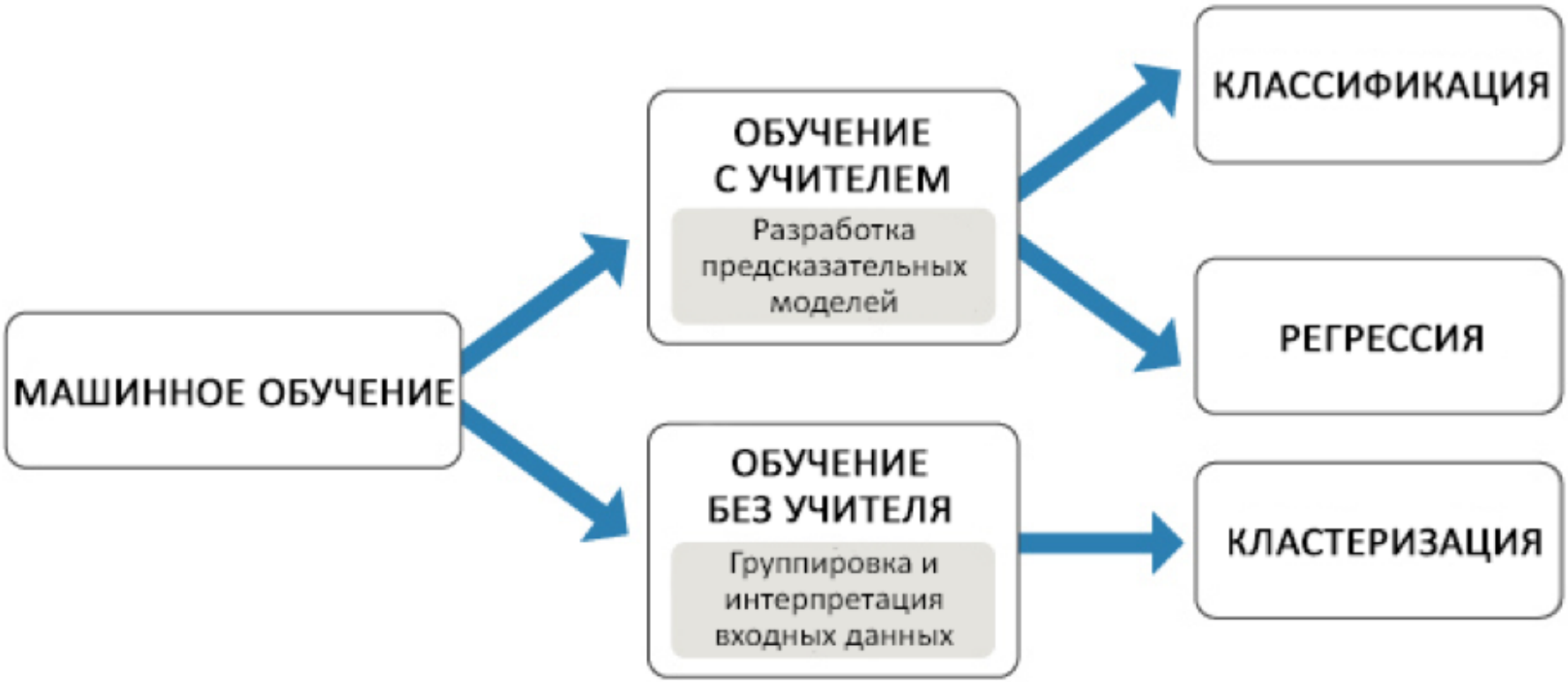
Интеллектуальный анализ данных
(Data Mining)

Шорохов С.Г.

кафедра математического моделирования и искусственного интеллекта

Лекция 3. Кластеризация данных

Основные задачи машинного обучения



Постановка задачи кластеризации

Пусть имеется набор данных из n точек (объектов) в d-мерном пространстве признаков $\mathbf{D} = \{\mathbf{x}_i\}_{i=1}^{n}$, $\mathbf{x}_i \in \mathbb{R}^d$ и известно, что каждая из этих точек относится к одной из k групп, причём признаки точек (объектов) из одного класса не слишком сильно различаются, а признаки объектов из разных классов различаются более существенно.

Задача кластеризации состоит в том, чтобы разделить набор данных $\mathbf{D}$ на k групп (или кластеров) $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ так, чтобы

1. кластеры C_i не пересекались, то есть
$$\forall i, j = \overline{1,k} : \; i \neq j \Rightarrow C_i \cap C_j = \varnothing$$
2. каждый объект из набора данных $\mathbf{D}$ относился к одному из кластеров, то есть
$$\bigcup_{i=1}^{k} C_i = \mathbf{D}$$
3. некоторый показатель ошибки кластеризации $J(\mathcal{C})$ был минимален.

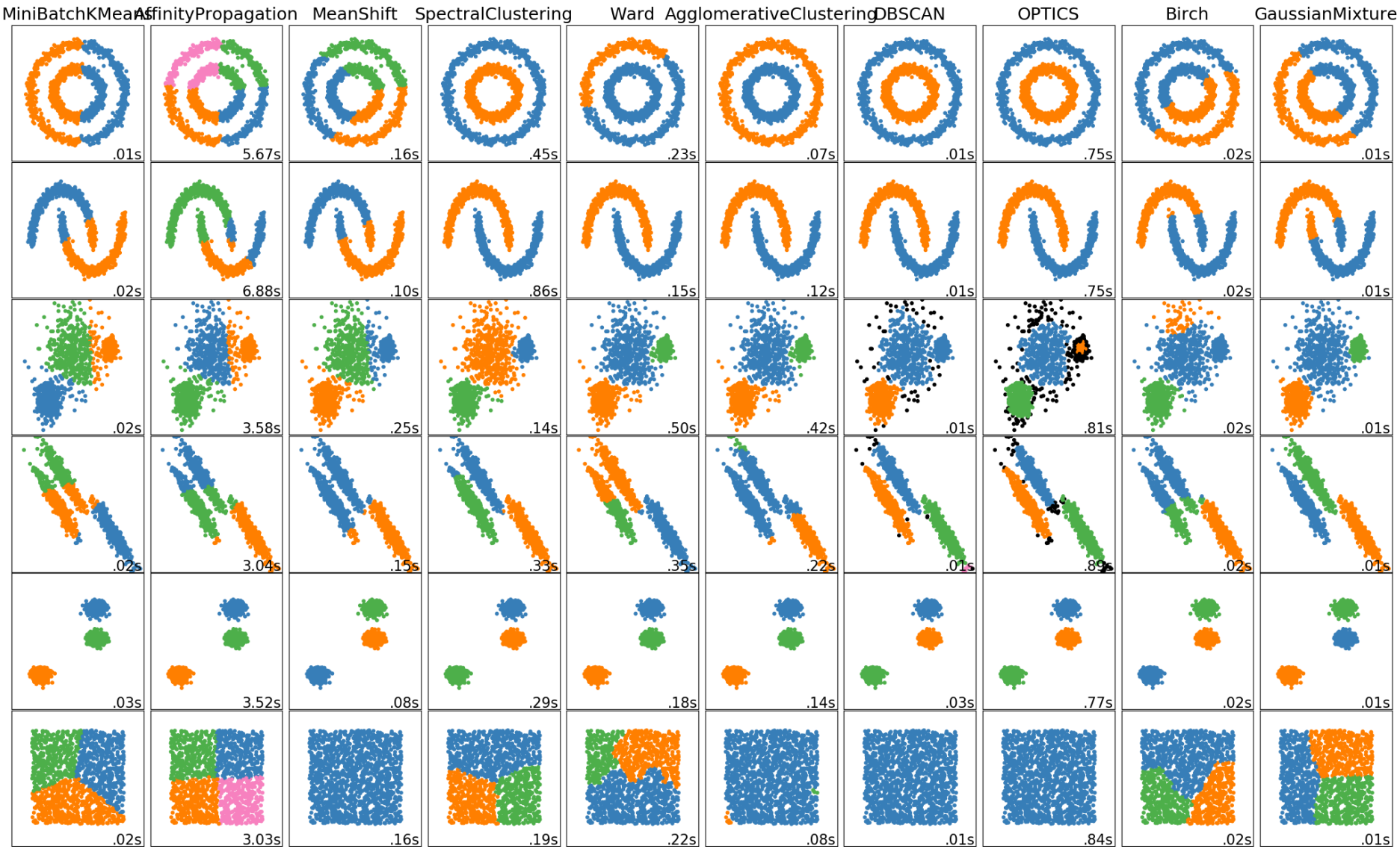
Кластеризация методом полного перебора

Задача кластеризации теоретически может быть решена методом полного перебора (метод «грубой силы», англ. brute force), когда рассматриваются все возможные разбиения набора данных $\mathbf{D}$ из n точек на k кластеров, для каждого разбиения вычисляется показатель ошибки кластеризации и в качестве решения выбирается разбиение с лучшей (минимальной) ошибкой. Точное число способов разбиения n точек на k непустых и непересекающихся множеств задается числом Стирлинга второго рода

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k+j}\binom{k}{j}j^n.$$

Понятно, что перебор всех вариантов и оценка ошибки кластеризации для каждого варианта является является практически нереализуемым для больших наборов данных.

Алгоритмы кластеризации в scikit-learn



Сходство объектов и расстояние между ними

Многие методы решения задач машинного обучения (в т.ч. методы кластеризации) требуют определения сходства (подобия) объектов (или расстояния между объектами), а именно:

Для двух заданных объектов $\mathbf{\bar{O}_1}$ и $\mathbf{\bar{O}_2}$ требуется определить их подобие $\mathbf{Sim}\left(\mathbf{\bar{O}_1}, \mathbf{\bar{O}_2}\right)$ или расстояние между ними $\mathbf{Dist}\left(\mathbf{\bar{O}_1}, \mathbf{\bar{O}_2}\right)$.

Для функций подобия большие значения означают большее сходство, тогда как для функций расстояния меньшие значения подразумевают большее сходство.

Расстояние (подобие) нужно уметь измерять для числовых и категориальных данных, смешанных данных, текстов, множеств, числовых рядов, последовательностей и т.д.

Расстояние между числовыми данными

Наиболее распространенной функцией расстояния для числовых (количественных) данных является расстояние (метрика) L_p. Расстояние L_p между двумя точками данных $\bar{\mathbf{X}} = (\mathbf{x_1}, ..., \mathbf{x_d})$ и $\bar{\mathbf{Y}} = (\mathbf{y_1}, ..., \mathbf{y_d})$ определяется как:

$$\mathbf{Dist}\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \left(\sum_{\mathbf{i=1}}^{\mathbf{d}} |\mathbf{x_i} - \mathbf{y_i}|^{\mathbf{p}}\right)^{\frac{\mathbf{1}}{\mathbf{p}}}$$

Важными частными случаями расстояния L_p являются евклидово $(\mathbf{p = 2})$ и манхэттенское $(\mathbf{p = 1})$ расстояния, а также случай, когда $\mathbf{p} = \infty$, а расстояние задается формулой:

$$\mathbf{Dist}\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \max_{\mathbf{i=\overline{1,d}}} |\mathbf{x_i} - \mathbf{y_i}|$$

Если в конкретной задаче некоторые признаки важнее других, то может применяться обобщенное расстояние L_p (расстояние Минковского):

$$\mathbf{Dist}\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \left(\sum_{\mathbf{i=1}}^{\mathbf{d}} \mathbf{a_i} |\mathbf{x_i} - \mathbf{y_i}|^{\mathbf{p}}\right)^{\frac{\mathbf{1}}{\mathbf{p}}}, \mathbf{a_i} \geqslant \mathbf{0}.$$

Сходство категориальных данных (1)

Рассмотрим две записи $\bar{\mathbf{X}} = (\mathbf{x_1}, ..., \mathbf{x_d})$ и $\bar{\mathbf{Y}} = (\mathbf{y_1}, ..., \mathbf{y_d})$ с категориальными данными. Простейшее возможное сходство между записями $\bar{\mathbf{X}}$ и $\bar{\mathbf{Y}}$ – это сумма сходств отдельных значений атрибутов. Другими словами, если $\mathbf{S}(\mathbf{x_i}, \mathbf{y_i})$ – это сходство между значениями атрибутов $\mathbf{x_i}$ и $\mathbf{y_i}$, то общее сходство определяется следующим образом:

$$\mathbf{Sim}(\bar{\mathbf{X}}, \bar{\mathbf{Y}}) = \sum_{\mathbf{i=1}}^{\mathbf{d}} \mathbf{S}(\mathbf{x_i}, \mathbf{y_i}).$$

Наиболее очевидным выбором является следующая простая мера сопоставления:

$$\mathbf{S}(\mathbf{x_i}, \mathbf{y_i}) = \begin{cases} \mathbf{1}, & \mathbf{x_i} = \mathbf{y_i} \\ \mathbf{0}, & \mathbf{x_i} \neq \mathbf{y_i} \end{cases}$$

Главный недостаток этой меры заключается в том, что она не учитывает относительные частоты среди различных атрибутов.

Сходство категориальных данных (2)

Обратная частота появления (inverse occurrence frequency) является обобщением простой меры сопоставления. Эта мера взвешивает сходство между совпадающими признаками двух записей с помощью функции частоты совпадающего значения $\mathbf{p_k(x)}$, которая равна доле записей, в которых $\mathbf{k}$-й признак принимает значение $\mathbf{x}$ в наборе данных. Другими словами, когда $\mathbf{x_i = y_i}$, сходство по $\mathbf{k}$-му признаку равно $\mathbf{1/p_k(x_i)^2}$:

$$\mathbf{S}\left(\mathbf{x_i}, \mathbf{y_i}\right) = \begin{cases} \frac{\mathbf{1}}{\mathbf{p_k(x_i)^2}}, & \mathbf{x_i = y_i} \\ \mathbf{0}, & \mathbf{x_i \neq y_i} \end{cases}$$

В мере Гудолла также совпадению присваивается более высокое значение сходства, когда значение встречается нечасто. В простом варианте этой меры сходство по $\mathbf{k}$-му признаку определяется как $\mathbf{1 - p_k(x_i)^2}$, когда $\mathbf{x_i = y_i}$, и $\mathbf{0}$ в противном случае:

$$\mathbf{S}\left(\mathbf{x_i}, \mathbf{y_i}\right) = \begin{cases} \mathbf{1 - p_k}\left(\mathbf{x_i}\right)^{\mathbf{2}}, & \mathbf{x_i = y_i} \\ \mathbf{0}, & \mathbf{x_i \neq y_i} \end{cases}$$

Смешанные числовые и категориальные данные

Рассмотрим две записи $\bar{\mathbf{X}} = (\bar{\mathbf{X}}_\mathbf{n}, \bar{\mathbf{X}}_\mathbf{c})$ и $\bar{\mathbf{Y}} = (\bar{\mathbf{Y}}_\mathbf{n}, \bar{\mathbf{Y}}_\mathbf{c})$, где $\bar{\mathbf{X}}_\mathbf{n}, \bar{\mathbf{Y}}_\mathbf{n}$ – подмножества числовых атрибутов, а $\bar{\mathbf{X}}_\mathbf{c}, \bar{\mathbf{Y}}_\mathbf{c}$ – подмножества категориальных атрибутов записей. Тогда общее сходство между записями $\bar{\mathbf{X}}$ и $\bar{\mathbf{Y}}$ определяется следующим образом:

$$\mathbf{Sim}\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \lambda\, \mathbf{NumSim}\left(\bar{\mathbf{X}}_\mathbf{n}, \bar{\mathbf{Y}}_\mathbf{n}\right) + (1-\lambda)\, \mathbf{CatSim}\left(\bar{\mathbf{X}}_\mathbf{c}, \bar{\mathbf{Y}}_\mathbf{c}\right)$$

Параметр λ регулирует относительную важность категориальных и числовых атрибутов. При отсутствии информации об относительной важности атрибутов естественным выбором является использование значения λ, равного доле числовых атрибутов в данных.

Если известны стандартные отклонения значений подобия $\sigma_\mathbf{c}$ и $\sigma_\mathbf{n}$ категориальных и числовых признаков, то формула для общего сходства может быть нормализована следующим образом:

$$\mathbf{Sim}\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \frac{\lambda}{\sigma_\mathbf{n}}\, \mathbf{NumSim}\left(\bar{\mathbf{X}}_\mathbf{n}, \bar{\mathbf{Y}}_\mathbf{n}\right) + \frac{\mathbf{1}-\lambda}{\sigma_\mathbf{c}} \mathbf{CatSim}\left(\bar{\mathbf{X}}_\mathbf{c}, \bar{\mathbf{Y}}_\mathbf{c}\right)$$

Меры сходства текстов (1)

Строго говоря, текст можно рассматривать как количественные многомерные данные, если рассматривать его как мешок слов (bag of words). Частоту каждого слова можно рассматривать как количественный атрибут, а базовую лексику можно рассматривать как полный набор атрибутов.

Такой показатель, как расстояние L_p, плохо адаптируется к разной длине различных текстов. Например, расстояние L_2 между двумя длинными текстами почти всегда будет больше, чем расстояние между двумя короткими текстами, даже если два длинных текста имеют много общих слов, а короткие тексты полностью не пересекаются.

В качестве меры сходства текстов можно использовать косинус-меру, вычисляющую угол между двумя текстами, который нечувствителен к абсолютной длине текста. Пусть $\bar{\mathbf{X}} = (\mathbf{x_1}, ..., \mathbf{x_d})$ и $\bar{\mathbf{Y}} = (\mathbf{y_1}, ..., \mathbf{y_d})$ – текста в лексиконе размера $\mathbf{d}$. Тогда косинус-мера $\cos\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right)$ между текстами $\bar{\mathbf{X}}$ и $\bar{\mathbf{Y}}$ может быть определена следующим образом:

$$\cos\left(\bar{\mathbf{X}}, \bar{\mathbf{Y}}\right) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{d}} \mathbf{x_i y_i}}{\sqrt{\sum_{\mathbf{i=1}}^{\mathbf{d}} \mathbf{x_i^2}} \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{d}} \mathbf{y_i^2}}}$$

Меры сходства текстов (2)

Косинус-мера использует частоты слов. Однако, если два текста содержат необычное слово, это более указывает на сходство, чем случай, когда два текста содержат очень часто встречающееся слово. Для нормализации обычно используется обратная частота id_i, которая является убывающей функцией количества текстов n_i, в которых встречается i-е слово:

$$id_i = \log\left(\frac{n}{n_i}\right)$$

Здесь общее количество текстов в наборе обозначено через n.

Для того, чтобы чрезмерное присутствие одного слова не нарушало меры сходства, к частотам перед вычислением подобия может применяться функция демпфирования $f(\cdot)$, такая как квадратный корень или логарифм:

$$f\left(x_i\right) = \sqrt{x_i},\ f\left(x_i\right) = \log\left(x_i\right)$$

Меры сходства текстов (3)

Если функция демпфирования не используется, то это эквивалентно равенству $f(x_i) = x_i$. Поэтому нормализованная частота $h(x_i)$ для i-го слова может быть определена следующим образом:

$$h(x_i) = f(x_i)\ id_i$$

Тогда косинус-мера может быть переопределена с использованием нормированных частот слов:

$$\cos\left(\bar{X}, \bar{Y}\right) = \frac{\sum_{i=1}^{d} h(x_i)\, h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2}\sqrt{\sum_{i=1}^{d} h(y_i)^2}}$$

Сходство бинарных данных и данных из множеств

Бинарные многомерные данные представляют собой представление данных на основе множества, где значение 1 указывает на присутствие элемента в множестве. Бинарные данные часто встречаются в задачах анализа рыночной корзины, в которых транзакции содержат информацию, соответствующую тому, присутствует ли товар в транзакции. Это можно рассматривать как частный случай текстовых данных, в которых частота слов равна 0 или 1. Если S_X и S_Y – это два набора с бинарными представлениями $\bar{X}$ и $\bar{Y}$, то сходство бинарных представлений $\bar{X}$ и $\bar{Y}$ оценивается при помощи коэффициента Жаккара (Jaccard coefficient):

$$J\left(\bar{X},\bar{Y}\right)=\frac{\sum_{i=1}^{d}x_iy_i}{\sum_{i=1}^{d}x_i^2+\sum_{i=1}^{d}y_i^2-\sum_{i=1}^{d}x_iy_i}=\frac{|S_X\cap S_Y|}{|S_X\cup S_Y|}$$

Данная мера достаточно хорошо характеризует сходство, так как она тщательно учитывает количество общих и непересекающихся элементов в двух наборах.

Кластеризация через представителей

Для каждого кластера C_i может быть выбрана точка-представитель, характеризующая кластер. Наиболее часто в качестве представителя выбирается среднее значение $\overline{\mu}_i$ всех точек кластера, также называемое центром кластера:

$$\overline{\mu}_i = \frac{1}{n_i} \sum_{\overline{\mathbf{x}} \in C_i} \overline{\mathbf{x}},$$

где $n_i = |C_i|$ – количество точек в кластере C_i.

Показатель ошибки кластеризации может выбираться, например, как сумма квадратов расстояний от каждой точки до центра соответствующего кластера:

$$J = \sum_{i=1}^{k} \sum_{\overline{\mathbf{x}} \in C_i} \rho^2 \left(\overline{\mathbf{x}}, \overline{\mu}_i \right),$$

где $\rho(\overline{\mathbf{x}}, \overline{\mathbf{y}})$ – расстояние между точками $\overline{\mathbf{x}}$ и $\overline{\mathbf{y}}$ (метрика).

Алгоритм k средних (k-means)

Алгоритм k средних относится к методам разбиений, которые основаны на поэтапном улучшении некоторого начального разбиения исходного множества до получения оптимального значения некоторой целевой функции.

Для кластеризации $\mathcal{C} = \{C_1, C_2, ..., C_k\}$ в качестве скоринговой функции (функции ошибки), оценивающей качество разбиения, будем выбирать сумму квадратов ошибок (sum of squared errors)

$$SSE\left(\mathcal{C}\right) = \sum_{i=1}^{k} \sum_{\overline{\mathbf{x}}_\mathbf{j} \in C_i} \rho_2^2\left(\overline{\mathbf{x}}_j, \overline{\mu}_i\right) = \sum_{i=1}^{k} \sum_{\overline{\mathbf{x}}_\mathbf{j} \in C_i} \left\|\overline{\mathbf{x}}_j - \overline{\mu}_i\right\|_2^2,$$

где $\overline{\mu}_i$ – центр кластера C_i.

Задача состоит в том, чтобы найти кластеризацию $\mathcal{C}^*$, которая минимизирует функцию SSE:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \left\{SSE\left(\mathcal{C}\right)\right\}.$$

Детали алгоритма k средних

Идея алгоритма k средних заключается в последовательном пересчёте внутригрупповых средних $\overline{\mu}_i$. Пусть вначале для каждого из k кластеров имеется некоторое начальное внутригрупповое среднее $\overline{\mu}_i^{(0)}$. Если внутригрупповые средние заданы, то каждый кластер C_i очевидным образом определяется, как множество точек, которые находятся ближе к центру кластера $\overline{\mu}_i^{(0)}$, чем к центрам других кластеров. Другими словами, каждая точка $\overline{\mathbf{x}}_j$ приписывается к кластеру C_m, где

$$m = \arg\min_{i=\overline{1,k}} \left\{ \rho_2^2 \left(\overline{\mathbf{x}}_j, \overline{\mu}_i \right) \right\} = \arg\min_{i=\overline{1,k}} \left\{ \left\| \overline{\mathbf{x}}_j - \overline{\mu}_i \right\|_2^2 \right\}$$

После этого внутригрупповое среднее можно пересчитать по формуле, после чего снова переопределить кластеры и т.д., пока внутригрупповые средние не перестанут меняться.

Шаги алгоритма k средних

Любую s-ую итерацию алгоритма можно описать тремя шагами.

1. Распределить объекты наблюдения по кластерам:

$$C_l^{(s)} = \left\{ \overline{\mathbf{x}} \in \mathbf{D} \mid \forall j = \overline{1,k},\ j \neq l :\ \rho\left(\overline{\mathbf{x}},\ \bar{\mu}_l^{(s-1)}\right) < \rho\left(\overline{\mathbf{x}},\ \bar{\mu}_j^{(s-1)}\right) \text{ или } \rho\left(\overline{\mathbf{x}},\ \bar{\mu}_l^{(s-1)}\right) = \rho\left(\overline{\mathbf{x}},\ \bar{\mu}_j^{(s-1)}\right),\ l < j \right\}.$$

   При этом объекты просто относятся к тому кластеру, до центра которого расстояние от этого объекта меньше, а в случае равенства наименьших расстояний – в кластер с меньшим номером.

2. Пересчитать центры кластеров:

$$\bar{\mu}_l^{(s)} = \frac{1}{\left|C_l^{(s)}\right|} \sum_{\bar{\mathbf{x}} \in C_l^{(s)}} \bar{\mathbf{x}}.$$

3. Если центры кластеров не изменились с прошлой итерации, то есть

$$\forall j = \overline{1,k}\ \ \bar{\mu}_l^{(s)} = \bar{\mu}_l^{(s-1)},$$

   то закончить выполнение алгоритма, иначе перейти к шагу 1 для следующей итерации $(s+1)$.

Алгоритм k средних

K-means $(\mathbf{D}, k, \varepsilon)$:

$s = 0$

Случайным образом выбрать k центров: $\mu_1^0, \mu_2^0, \ldots, \mu_k^0 \in \mathbb{R}^d$

repeat

  $s \leftarrow s + 1$

  $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, k$

  // Шаг отнесения к кластеру

  foreach $\mathbf{x}_j \in \mathbf{D}$ do

    $j^* \leftarrow \arg\min_i \left\{ \left\| \mathbf{x}_j - \mu_i^{s-1} \right\|^2 \right\}$ // Отнести $\mathbf{x}_j$ к центру j^*

    $C_{j^*} \leftarrow C_{j^*} \cup \{\mathbf{x}_j\}$

  // Шаг обновления центров

  foreach $i = 1$ to k do

    $\mu_i^s \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

until $\sum_{i=1}^k \left\| \mu_i^s - \mu_i^{s-1} \right\|^2 \leq \varepsilon$

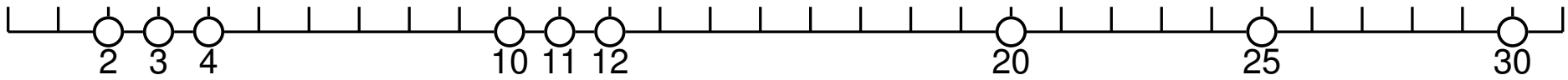
О сходимости алгоритма k средних

Алгоритм k средних всегда сходится за конечное количество итераций, поскольку значение ошибки кластеризации $SSE(\mathcal{C})$ не увеличивается в процессе работы алгоритма, а число возможных разбиений конечного множества на подмножества конечно. Тем не менее, этот алгоритм имеет экспоненциальную сложность в худшем случае, хотя на практике обычно сходится довольно быстро. Кроме того, этот алгоритм может сойтись не к глобальному, а к локальному минимуму функции $SSE(\mathcal{C})$.

Ещё одной особенностью является тот факт, что количество кластеров k должно быть известно заранее. В некоторых прикладных задачах оно действительно известно. Если же это не так, то всегда можно перебирать это количество, оценивая результат в каждом случае. Тем не менее, нужно понимать, что этот алгоритм не предназначен для случая, когда количество кластеров не известно.

Пример кластеризации с алгоритмом k средних

Рассмотрим пример для одномерных данных $\mathbf{D} = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$ $(n = 9)$, указанных на рисунке ниже, причем допустим, что данные должны быть сгруппированы в две группы $(k = 2)$.



Пусть в качестве начальных центров выбраны точки $\mu_1 = 2$ и $\mu_2 = 4$.

На начальном этапе определяем разбиение на кластеры в виде

$$C_1 = \{2, 3\}\,,\; C_2 = \{4, 10, 11, 12, 20, 25, 30\}\,,$$

относя каждую точку набора данных к кластеру с соответствующей центральной точкой (центром).

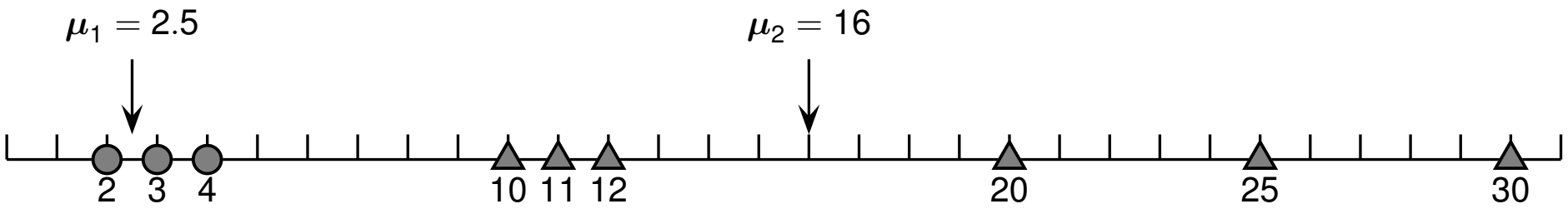
Пример кластеризации с алгоритмом k средних (2)

На первой итерации пересчитываем положение центральных точек кластеров

$$\mu_1 = \frac{1}{2}(2+3) = 2.5,\ \mu_2 = \frac{1}{7}(4+10+11+12+20+25+30) = 16$$

и определяем новое разбиение на кластеры в виде

$$C_1 = \{2,3,4\},\ C_2 = \{10,11,12,20,25,30\}.$$



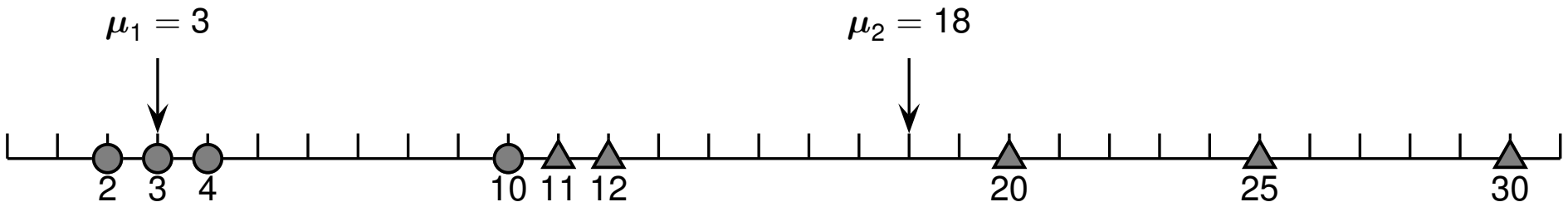
Пример кластеризации с алгоритмом k средних (3)

На второй итерации пересчитываем положение центральных точек кластеров

$$\mu_1 = \frac{1}{3}(2+3+4) = 3,\ \mu_2 = \frac{1}{6}(10+11+12+20+25+30) = 18$$

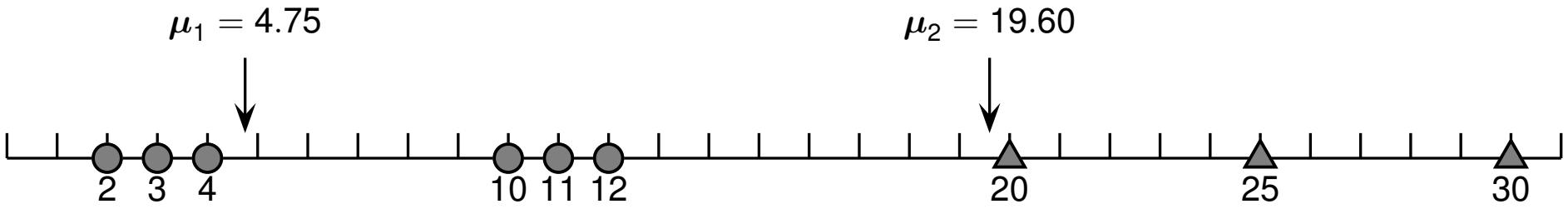
и определяем следующее разбиение на кластеры в виде

$$C_1 = \{2,3,4,10\},\ C_2 = \{11,12,20,25,30\}.$$

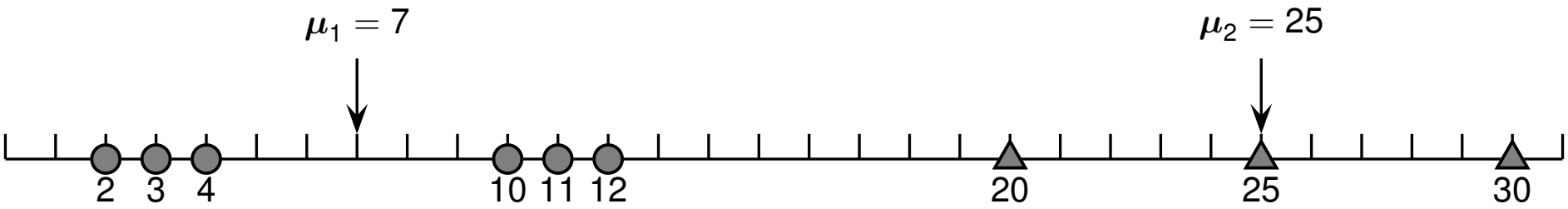


Пример кластеризации с алгоритмом k средних (4)

На третьей итерации получаем



Процесс останавливается на четвертой итерации, когда центральные точки и состав кластеров остаются неизменными.



В результате приходим к разбиению

$$C_1 = \{2, 3, 4, 10, 11, 12\}, C_2 = \{20, 25, 30\}$$

с центральными точками (представителями) $\mu_1 = 7$ и $\mu_2 = 25$.

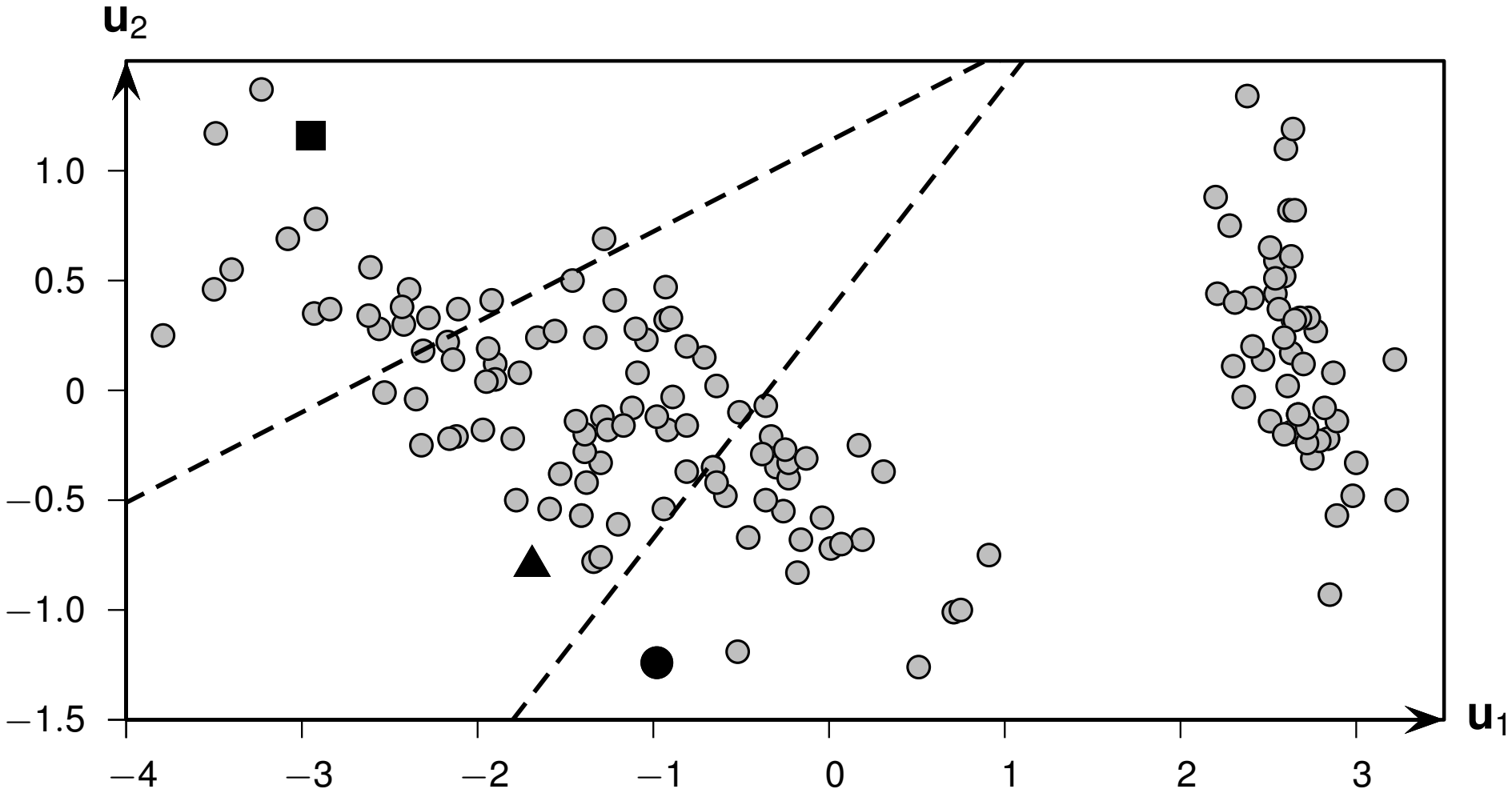
Пример кластеризации набора «Ирисы»

Рассмотрим в качестве следующего примера применение алгоритма k средних к набору данных «Ирисы», используя две главные компоненты набора. В наборе «Ирисы» $n = 150$ и требуется найти разбиение набора на $k = 3$ кластера, соответствующих трем типам ирисов.

Пример кластеризации набора «Ирисы» (2)

Случайный выбор начальных средних значений кластеров дает значения

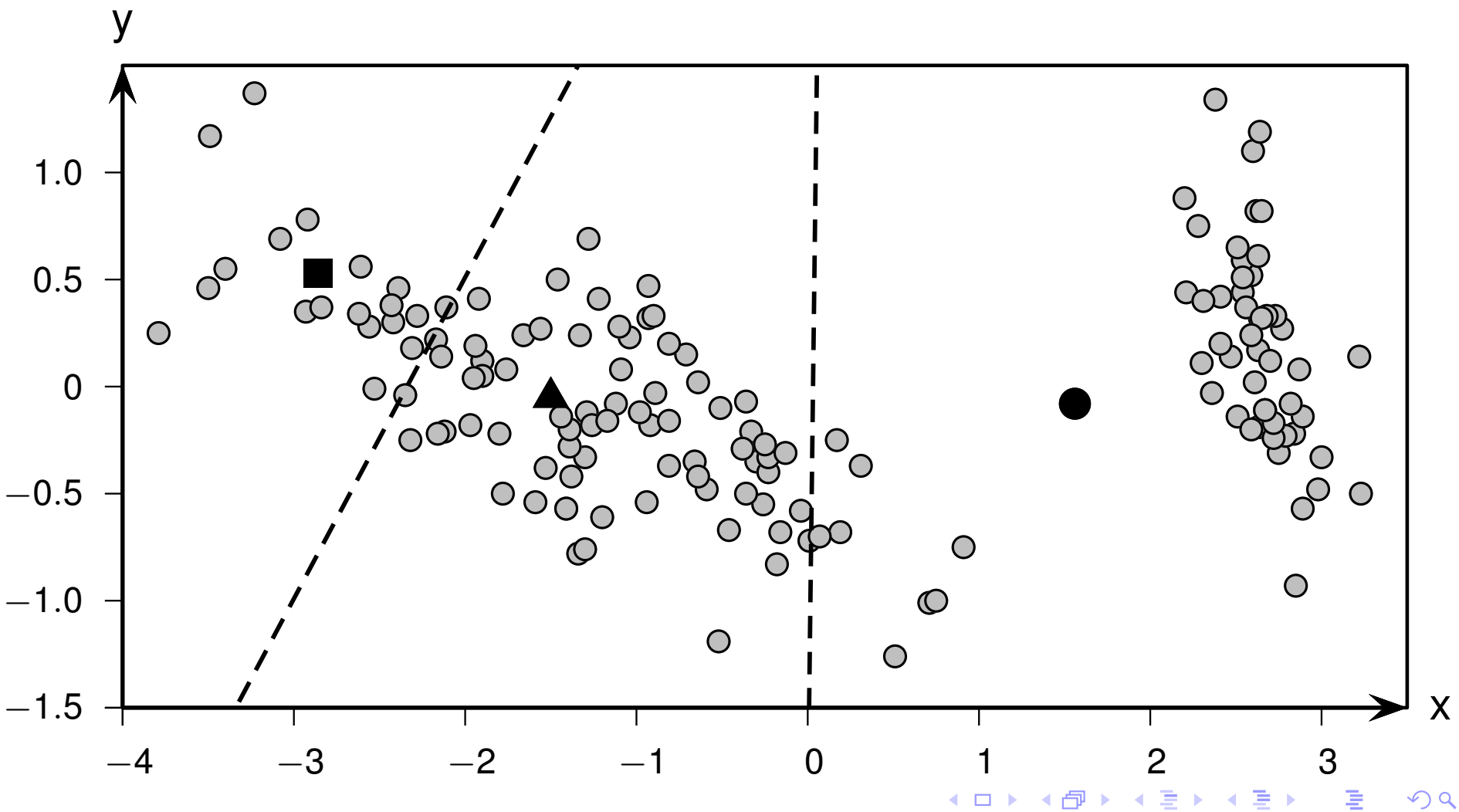
$$\mu_1 = (-0.98, -1.24)^T, \ \mu_2 = (-2.96, 1.16)^T, \ \mu_3 = (-1.69, -0.80)^T.$$



Пример кластеризации набора «Ирисы» (3)

На следующей итерации получаем средние значения кластеров в виде

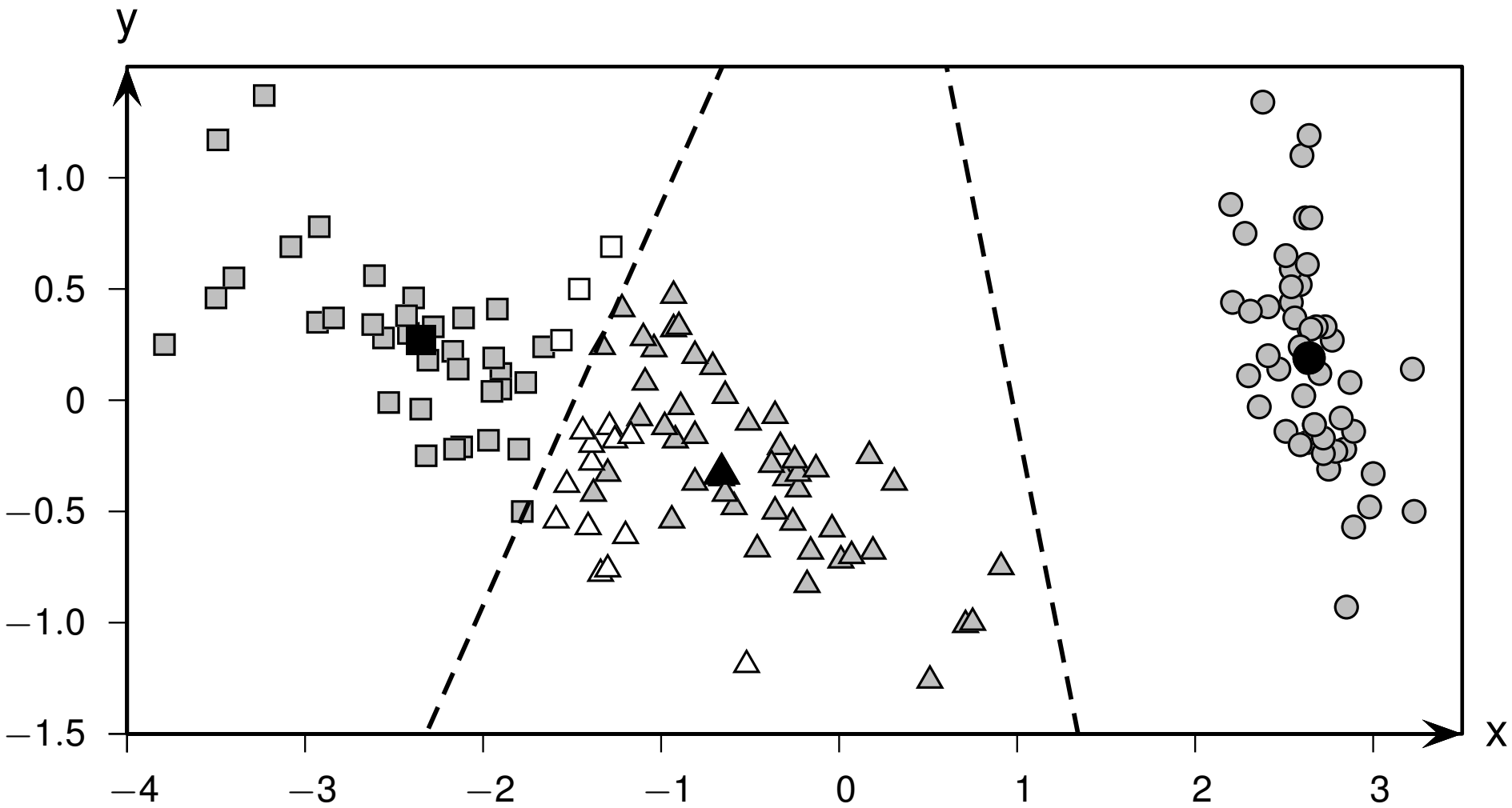
$$\mu_1 = (1.56, -0.08)^T, \ \mu_2 = (-2.86, 0.53)^T, \ \mu_3 = (-1.50, -0.05)^T.$$



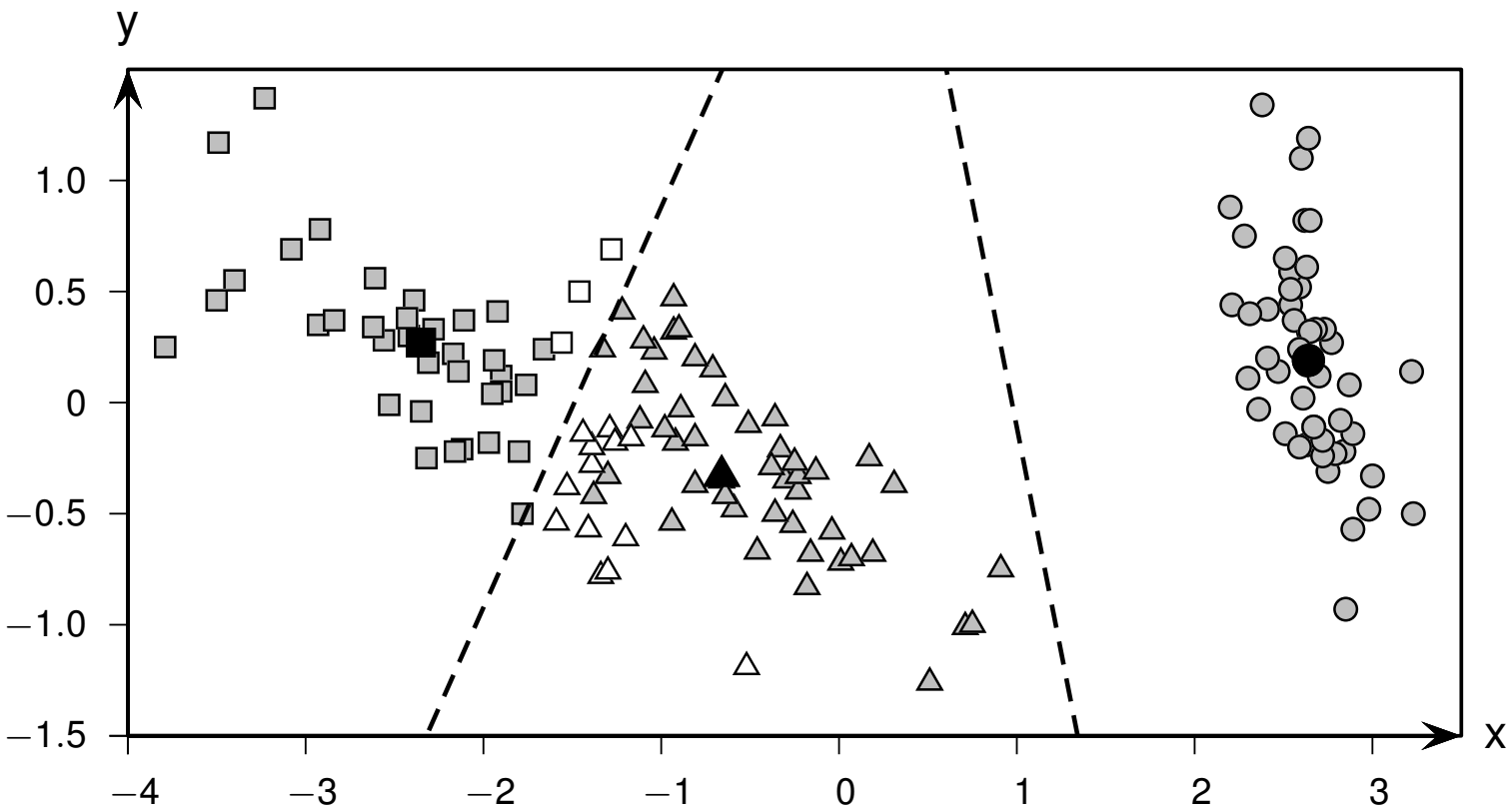
Пример кластеризации набора «Ирисы» (4)

Алгоритм сходится после 8 итераций и мы окончательно получаем

$$\mu_1 = (2.64, 0.19)^T,\ \mu_2 = (-2.35, 0.27)^T,\ \mu_3 = (-0.66, -0.33)^T.$$



Пример кластеризации набора «Ирисы» (5)



Пунктирными линиями показаны прямые, разделяющие точки, принадлежащие различным кластерам. Тремя финальными кластерами являются C_1 (круги), C_2 (квадраты), C_3 (треугольники). Белым цветом выделены объекты, сгруппированные неверно. Кластер C_1 точно соответствует виду iris-setosa, большинство точек кластеров C_2 и C_3 соответствует видам iris-virginica и iris-versicolor.

Выбор начальных центров кластеров

В методах кластеризации через представителей многое зависит от выбора начальных центров $\bar{\mu}_l^{(0)}$. Единственное требование к ним – ни один кластер изначально не должен быть пустым. Один из простых способов выбора начальных центров, который гарантирует выполнение этого условия, – взять в качестве них любые k различных точек из набора $\mathbf{D}$:

$$\bar{\mu}_l^{(0)} \in \mathbf{D},\ \forall i, j = \overline{1,k},\ i \neq j \Rightarrow \bar{\mu}_i^{(0)} \neq \bar{\mu}_j^{(0)}.$$

Тогда, по крайней мере, сами точки, выбранные в качестве центров, изначально попадут в соответствующие кластеры. Кроме того, существуют некоторые более эффективные способы автоматического выбора изначальных центров кластеров.

Две основные проблемы, возникающие при неудачном выборе в качестве центров кластеров нескольких случайных объектов наблюдения, состоят в большом количестве итераций алгоритма k средних и в достижении локального минимума функции J вместо глобального минимума. Использование специфического алгоритма автоматического выбора начальных центров кластеров (k-means++) может решить эти проблемы.

Алгоритм выбора начальных центров кластеров

1. Первый начальный центр $\bar{\mu}_1^{(0)}$ выбирается равновероятно среди имеющихся в наборе $\mathbf{D}$ точек: $\bar{\mu}_1^{(0)} \in \mathbf{D}$.
2. Пусть ранее уже определено s начальных центров кластеров из k. Точки набора распределяются по имеющимся кластерам в соответствии с текущими центрами, как это делается на шаге 1 алгоритма k средних. Каждая точка относится к тому кластеру, к центру которого она ближе. Для каждой точки сохраняется расстояние до центра кластера, к которому она относится.
3. Очередной центр $\bar{\mu}_{s+1}^{(0)}$ выбирается случайно среди всех точек набора. При этом каждая точка x, относящаяся к кластеру C_l, выбирается с вероятностью $p(x) = \dfrac{\rho^2\left(x, \bar{\mu}_l^{(0)}\right)}{\sum_{j=1}^{s}\sum_{y \in C_j} \rho^2\left(y, \bar{\mu}_j^{(0)}\right)}$, то есть вероятность выбора конкретной точки в качестве очередного начального центра кластера прямо пропорциональна квадрату расстояния от этой точки до ближайшего к нему ранее выбранного начального центра кластера.
4. Если все k начальных центров кластеров уже выбраны, закончить выполнение алгоритма, иначе перейти к шагу 2.

Иерархическая кластеризация

Целью иерархической кластеризации является построение последовательности (иерархии) вложенных разбиений исходного набора данных. Для визуализации результатов кластеризации используется дендрограмма в виде дерева или иерархии кластеров.

Кластеры в иерархии ранжируются от нижнего уровня дерева (листьев), на котором каждая точка находится в отдельном кластере, до верхнего уровня (корня), на котором все точки попадают в один кластер.

Методы агломеративной иерархической кластеризации обрабатывают дерево снизу вверх. Начиная с разбиения, в котором каждая точка находится в отдельном кластере, они последовательно объединяют наиболее похожие (близкие) кластеры до тех пор, пока все точки не станут членами одного кластера или не будет найдено необходимое число кластеров. На каждом шаге необходимо вычислять/пересчитывать расстояние между кластерами. Для вычисления расстояния между кластерами могут использоваться различные функции в зависимости от специфики задачи.

Иерархическая кластеризация: вложенные разбиения

Пусть задан набор данных $\mathbf{D} = \{\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, ..., \overline{\mathbf{x}}_n\}$, где $\overline{\mathbf{x}}_i \in \mathbb{R}^d$ и кластеризация $\mathcal{C} = \{C_1, ..., C_k\}$ является разбиением $\mathbf{D}$.

Разбиение $\mathcal{A} = \{A_1, ..., A_r\}$ называется вложенным в другое разбиение $\mathcal{B} = \{B_1, ..., B_s\}$, если и только если $r > s$ и для каждого множества (кластера) $A_i \in \mathcal{A}$ существует множество (кластер) $B_j \in \mathcal{B}$, такое, что $A_i \subseteq B_j$.

Иерархическая кластеризация порождает последовательность n вложенных разбиений $\mathcal{C}_1, ..., \mathcal{C}_n$, при этом разбиение $\mathcal{C}_{t-1}$ вложено в разбиение $\mathcal{C}_t$.

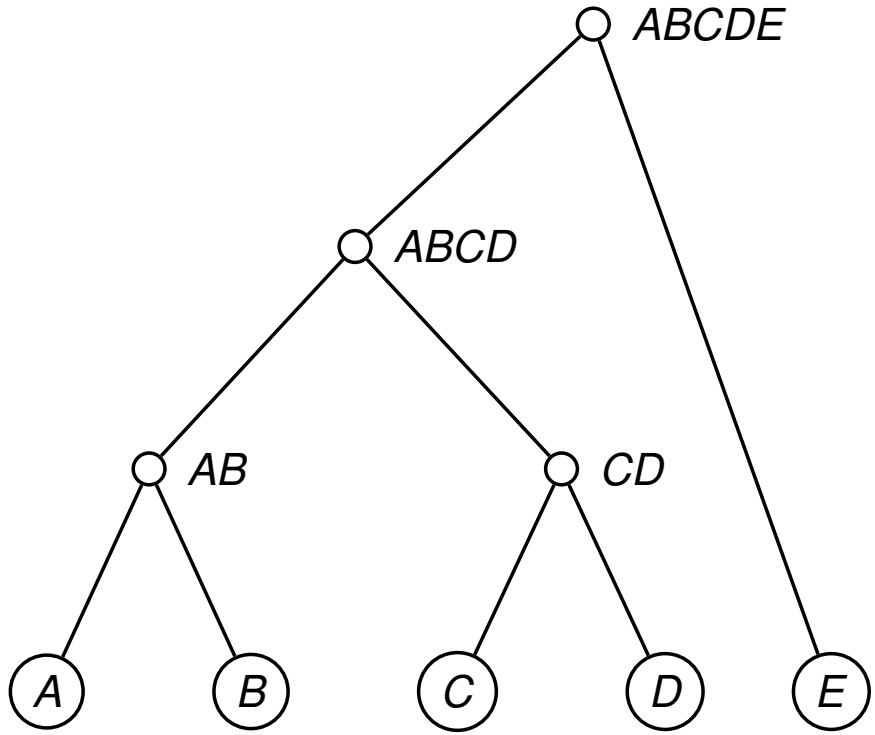
Дендрограмма кластеризации представляет собой бинарное дерево с корнем, отражающее эту структуру вложенных кластеров, с ребрами между кластерами $C_i \in \mathcal{C}_{t-1}$ и $C_j \in \mathcal{C}_t$, если множество C_i вложено в множество C_j, а именно, $C_i \subset C_j$.

Пример: дендрограмма кластеризации

Дендрограмма представляет следующую последовательность вложенных разбиений:

Кластеризация	Кластеры
$\mathcal{C}_1$	$\{A\},\{B\},\{C\},\{D\},\{E\}$
$\mathcal{C}_2$	$\{AB\},\{C\},\{D\},\{E\}$
$\mathcal{C}_3$	$\{AB\},\{CD\},\{E\}$
$\mathcal{C}_4$	$\{ABCD\},\{E\}$
$\mathcal{C}_5$	$\{ABCDE\}$

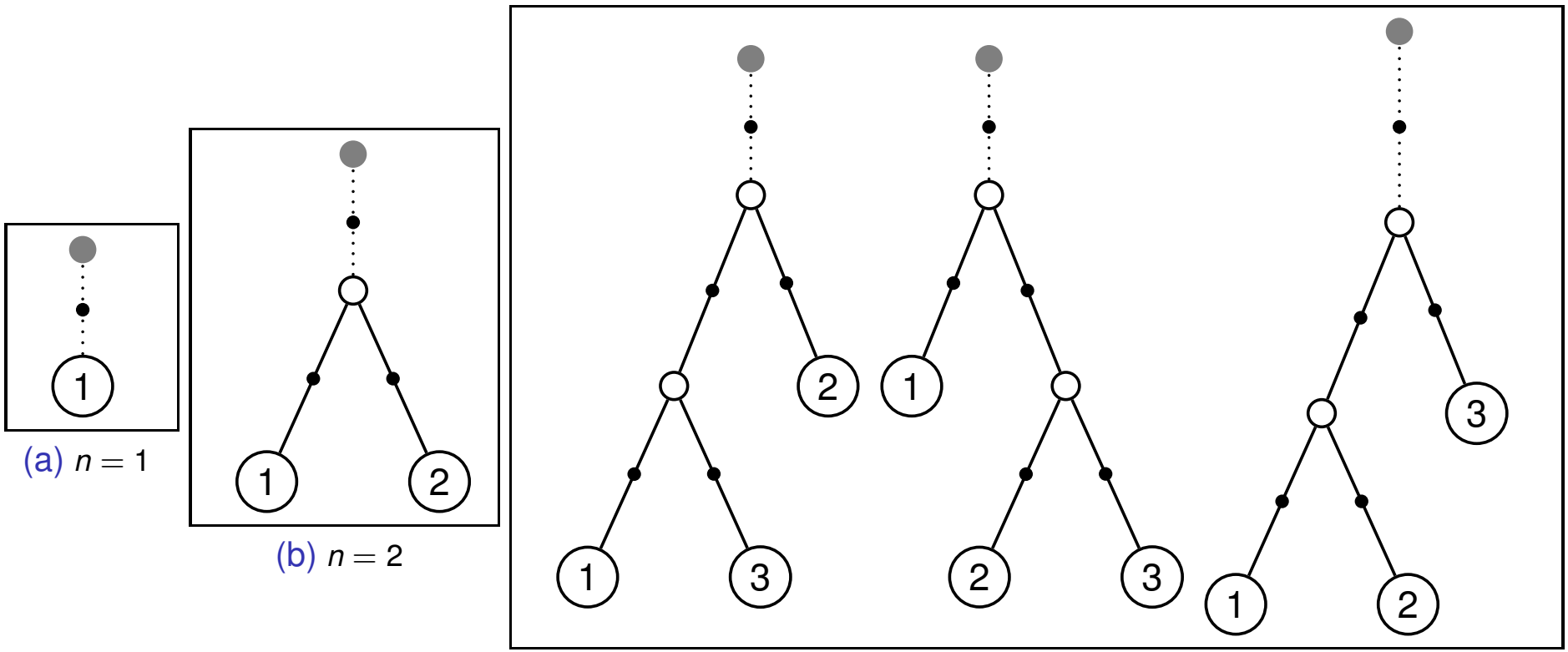
с условием $\mathcal{C}_{t-1} \subset \mathcal{C}_t$ для $t = 2, ..., 5$. Допустим также, что точки A и B объединяются раньше точек C и D.



Число иерархических кластеризаций

Общее число разных дендрограмм с n листьями задается формулой:

$$\prod_{m=1}^{n-1}(2m-1) = 1 \cdot 3 \cdot 5 \cdot \ldots \cdot (2n-3) = (2n-3)!!$$



Аггломеративная иерархическая кластеризация

В аггломеративной иерархической кластеризации в самом начале все точки находятся в отдельных кластерах. Для текущего разбиения на кластеры два ближайших кластера сливаются в один и так до тех пор, пока все точки не станут точками одного кластера.

Если дано семество кластеров $\mathcal{C} = \{C_1, ..., C_m\}$, то мы находим пару ближайших кластеров C_i и C_j и объединяем точки этих кластеров в новый кластер $C_{ij} = C_i \cup C_j$.

Далее мы одновляем множество кластеров, удаляя C_i и C_j и добавляя C_{ij}, а именно: $\mathcal{C} = (\mathcal{C} \setminus \{C_i, C_j\}) \cup \{C_{ij}\}$.

Процесс повторяется пока множество $\mathcal{C}$ не будет состоять из одного кластера. Если задано количество кластеров, равное k, то мы можем остановить процесс слияния, когда останется ровно k кластеров.

Алгоритм аггломеративной кластеризации

AgglomerativeClustering($\mathbf{D}, k$):

// Каждая точка в отдельном кластере

$\mathcal{C} \leftarrow \{C_i = \{\mathbf{x}_i\} \mid \mathbf{x}_i \in \mathbf{D}\}$

// Вычислить матрицу расстояний

$\Delta \leftarrow \{\delta(\mathbf{x}_i, \mathbf{x}_j) : \ \mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}\}$

repeat

  Найти ближайшую пару кластеров $C_i, C_j \in \mathcal{C}$

  $C_{ij} \leftarrow C_i \cup C_j$ // Объединить кластеры

  $\mathcal{C} \leftarrow \big(\mathcal{C} \setminus \{C_i, C_j\}\big) \cup \{C_{ij}\}$ // Обновить кластеризацию

  Обновить матрицу расстояний Δ для новой кластеризации $\mathcal{C}$

until $|\mathcal{C}| = k$

Расстояние между кластерами

- Расстояние между двумя точками (Евклидово расстояние или L_2-норма)

$$\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) = \left\|\overline{\mathbf{x}} - \overline{\mathbf{y}}\right\|_2 = \left(\sum_{i=1}^{d} \left(x_i - y_i\right)^2\right)^{1/2}.$$

Расстояние между кластерами можно вычислять различными методами:

- Метод одиночной связи (single link) – минимальное расстояние между точкой в C_i и точкой в C_j:

$$\delta\left(C_i, C_j\right) = \min_{\overline{\mathbf{x}}, \overline{\mathbf{y}}} \left\{\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) \mid \overline{\mathbf{x}} \in C_i, \overline{\mathbf{y}} \in C_j\right\}$$

- Метод полной связи (complete link) – максимальное расстояние между точками в двух кластерах:

$$\delta\left(C_i, C_j\right) = \max_{\overline{\mathbf{x}}, \overline{\mathbf{y}}} \left\{\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) \mid \overline{\mathbf{x}} \in C_i, \overline{\mathbf{y}} \in C_j\right\}$$

- Метод средней связи (group average) – среднее расстояние между точками в C_i и точками в C_j:

$$\delta\left(C_i, C_j\right) = \frac{1}{n_i n_j} \sum_{\overline{\mathbf{x}} \in C_i} \sum_{\overline{\mathbf{y}} \in C_j} \delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right)$$

Расстояние между кластерами (2)

- Расстояние между центрами (mean distance) – расстояние между кластерами определяется как расстояние между средними значениями или центрами кластеров:

$$\delta\left(C_i, C_j\right) = \delta\left(\overline{\mu}_i, \overline{\mu}_j\right) = \delta\left(\frac{1}{n_i}\sum_{\overline{\mathbf{x}}\in C_i}\overline{\mathbf{x}}, \frac{1}{n_j}\sum_{\overline{\mathbf{y}}\in C_j}\overline{\mathbf{y}}\right)$$

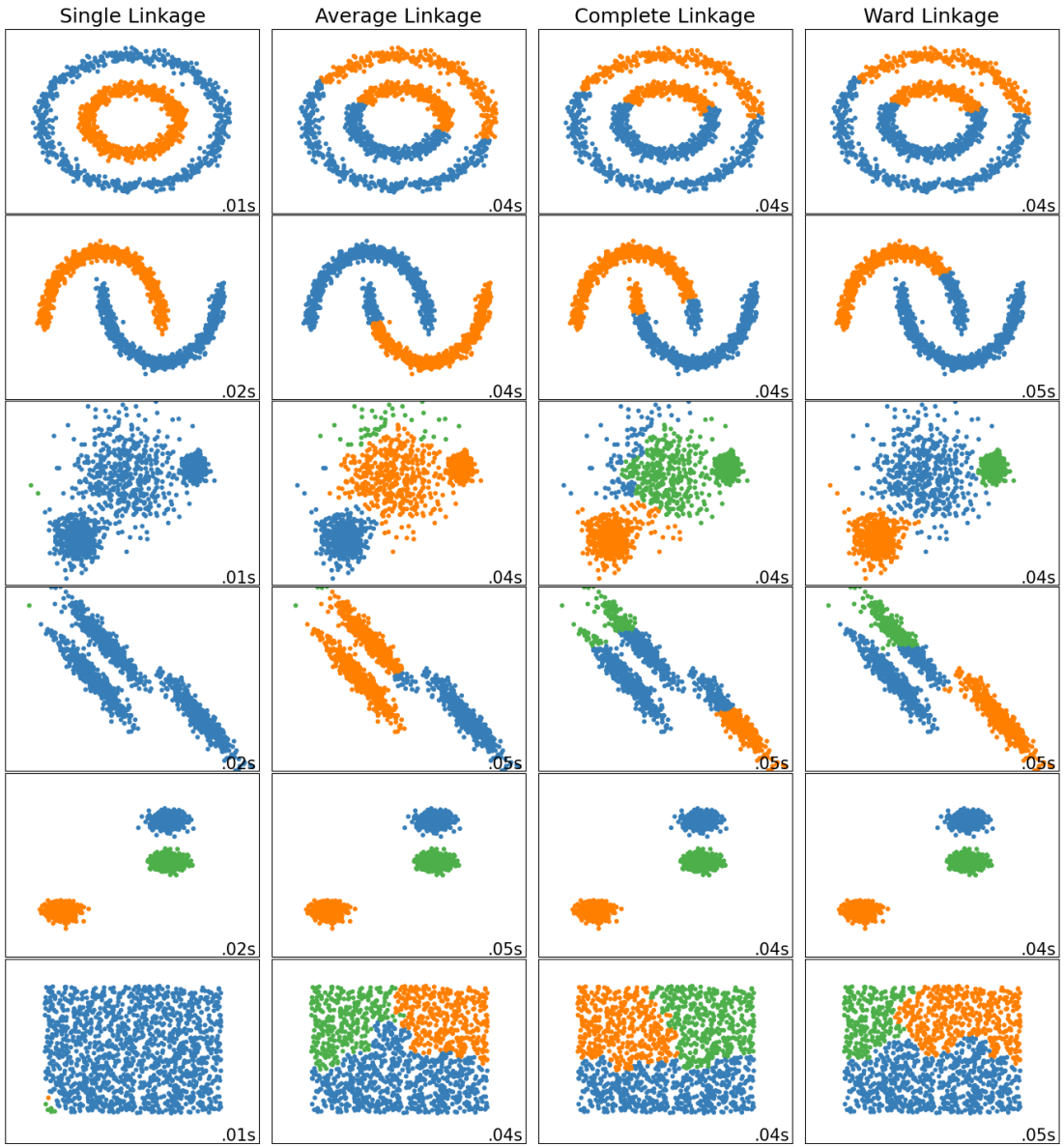
- Минимальная дисперсия или метод Уарда (Ward's measure) – расстояние между кластерами определяется как прирост суммы квадратичной ошибки (SSE), когда два кластера объединяются

$$\delta\left(C_i, C_j\right) = \triangle SSE_{ij} = SSE_{ij} - SSE_i - SSE_j,$$

где $SSE_i = \sum_{\overline{\mathbf{x}}\in C_i}\left\|\overline{\mathbf{x}} - \overline{\mu}_i\right\|^2$. После упрощения получаем

$$\delta\left(C_i, C_j\right) = \frac{n_i n_j}{n_i + n_j}\left\|\overline{\mu}_i - \overline{\mu}_j\right\|^2.$$

Качество иерархической кластеризации



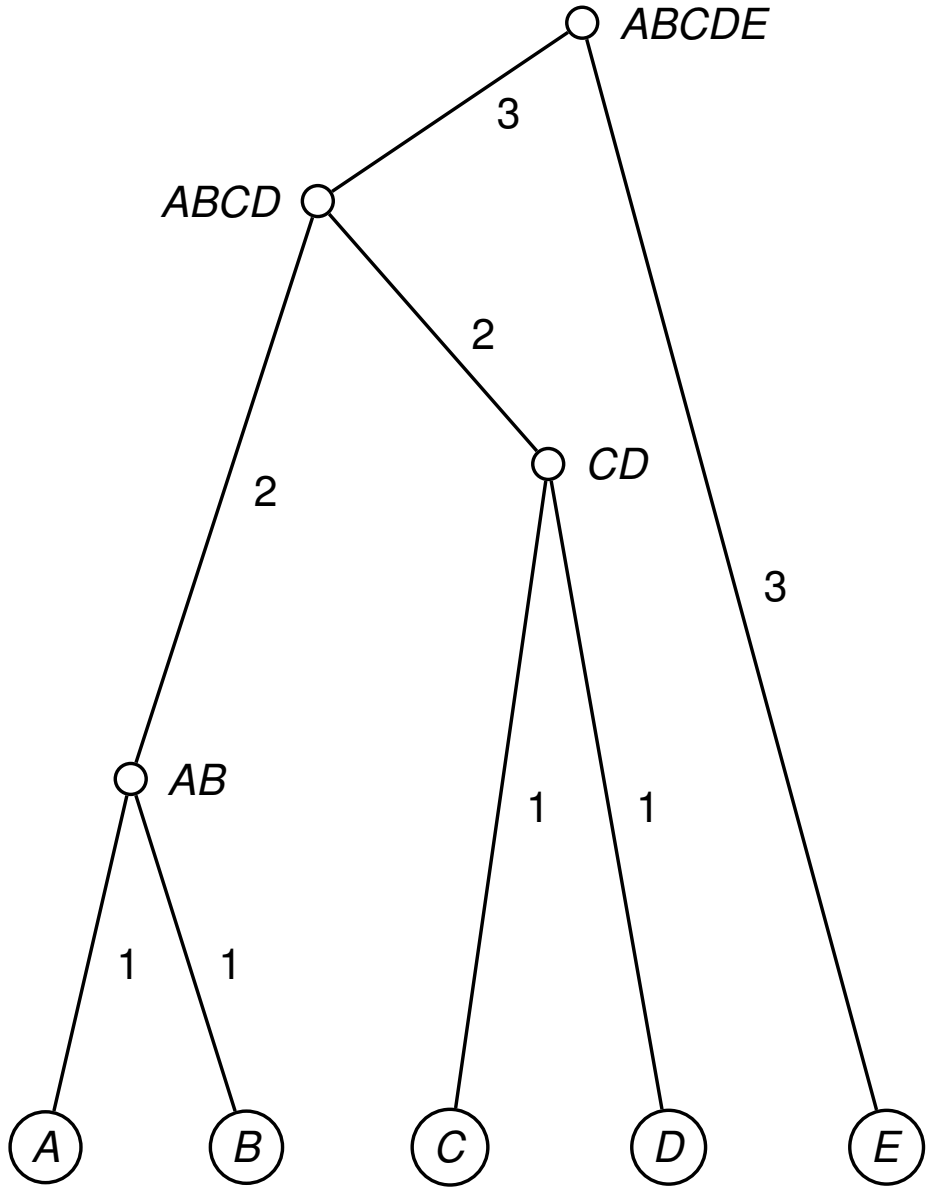
Кластеризация методом одиночной связи

δ	E
$ABCD$	(3)

δ	CD	E
AB	(2)	3
CD		3

δ	C	D	E
AB	3	2	3
C		(1)	3
D			5

δ	B	C	D	E
A	(1)	3	2	4
B		3	2	3
C			1	3
D				5



Формула Ланса-Уильямса

На каждом шаге кластеризации необходимо уметь быстро подсчитывать расстояние от образовавшегося кластера $C_{ij} = C_i \cup C_j$ до любого другого кластера C_r, где $r \neq i$, $r \neq j$, используя известные расстояния с предыдущих шагов. Это легко сделать при помощи формулы, предложенной Лансом и Уильямсом в 1967 году:

$$\delta(C_{ij}, C_r) = \alpha_i \delta(C_i, C_r) + \alpha_j \delta(C_j, C_r) + \beta\, \delta(C_i, C_j) + \gamma\, (\delta(C_i, C_r) - \delta(C_j, C_r))$$

Каждая из указанных выше функций расстояния удовлетворяет формуле Ланса-Уильямса со своими коэффициентами:

Расстояние	α_i	α_j	β	γ
Single link	$1/2$	$1/2$	0	$-1/2$
Complete link	$1/2$	$1/2$	0	$1/2$
Group average	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Mean distance	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Ward's measure	$\frac{n_i+n_r}{n_i+n_j+n_r}$	$\frac{n_i+n_r}{n_i+n_j+n_r}$	$\frac{-n_r}{n_i+n_j+n_r}$	0

Меры качества кластеризации

Для оценки эффективности кластеризации традиционно выделяют два типа мер: внешние меры, использующие дополнительную (внешнюю) информацию (метки классов) о настоящем распределении объектов по классам, и внутренние меры, использующие только информацию о самой кластеризации.

Пусть $\mathbf{D} = \{\overline{\mathbf{x}}_i\}_{i=1}^{n}$ – набор данных, состоящий из n точек в d-мерном пространстве, изначально разделенный на k классов. Пусть $y_i \in \{1, 2, ..., k\}$ обозначает первоначальную метку класса для каждой точки $\overline{\mathbf{x}}_i \in \mathbf{D}$.

Допустим, что разбиение на классы задано в виде $\mathcal{T} = \{T_1, T_2, ..., T_k\}$, где класс T_j состоит из всех точек с меткой j, т.е. $T_j = \{\overline{\mathbf{x}}_i \in D \mid y_i = j\}$.

Пусть $\mathcal{C} = \{C_1, ..., C_r\}$ обозначает кластеризацию того же набора данных на r кластеров, полученную при помощи некоторого алгоритма кластеризации, и пусть $\hat{y}_i \in \{1, 2, ..., r\}$ обозначает полученную алгоритмом метку кластера для $\overline{\mathbf{x}}_i \in \mathbf{D}$.

Внешние меры оценки качества кластеризации

Внешние меры качества кластеризации определяют показатели, в которой точки из одного и того же класса относятся к одному кластеру, и степень, в которой точки из разных классов группируются в разных кластерах.

Все внешние меры полагаются на таблицу (матрицу) сопряженности (contingency table) $\mathbf{N}$ размером $k \times r$, которая связывает первоначальное разбиение на классы $\mathcal{T}$ и кластеризацию $\mathcal{C}$ следующими соотношениями

$$\mathbf{N}(i, j) = n_{ij} = |T_i \cap C_j|$$

Счетчик n_{ij} обозначает число точек, являющихся общими для класса T_i и кластера C_j.

Таблица сопряженности $\mathbf{N}$ может быть вычислена по разбиению на классы $\mathcal{T}$ и кластеризации $\mathcal{C}$ за время $O(n)$ путем увеличения счетчика $n_{y_i \hat{y}_i}$ для каждой точки $\overline{\mathbf{x}}_i \in \mathbf{D}$ с меткой класса y_i и меткой кластера $\hat{y}_i$.

Пусть $n_i = |T_i|$ – число точек в классе T_i и $m_j = |C_j|$ – число точек в кластере C_j.

Таблица сопряженности

Таблицу сопряженности $\mathbf{N}$ наглядно можно представить в виде:

$\mathcal{T} \setminus \mathcal{C}$	C_1	C_2	$\dots$	C_r	$\sum$
T_1	n_{11}	n_{12}	$\dots$	n_{1r}	n_1
T_2	n_{21}	n_{22}	$\dots$	n_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T_k	n_{k1}	n_{k2}	$\dots$	n_{kr}	n_k
$\sum$	m_1	m_2	$\dots$	m_r	n

Здесь $n_{ij} = |T_i \cap C_j|$, $n_i = |T_i|$, $m_j = |C_j|$, n – число точек в наборе данных $\mathbf{D}$.

Таблица сопряженности реализована в библиотеке scikit-learn:

contingency_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)

Внешняя мера оценки: чистота

Чистота (purity) кластера ставит в соответствие кластеру C_j показатель максимальной доли точек из одного класса:

$$p_j = \frac{1}{m_j} \max_{i=\overline{1,k}} n_{ij},\ j = \overline{1,r}$$

Чистота кластеризации $\mathcal{C}$ определяется как взвешенная сумма показателей чистоты кластеров

$$p = \sum_{j=1}^{r} \frac{m_j}{n} p_j = \frac{1}{n} \sum_{j=1}^{r} \max_{i=\overline{1,k}} n_{ij},$$

где дробь $\frac{m_j}{n}$ обозначает долю точек, относящихся к кластеру C_j.

Чистота находится в интервале $[0, 1]$, причём значение 1 отвечает оптимальной кластеризации. При $r < k$ (число кластеров меньше числа классов) чистота не может быть равна 1, так как по крайней мере в одном кластере должны быть точки из разных классов.

Набор «Ирисы»: кластеризация Complete Link

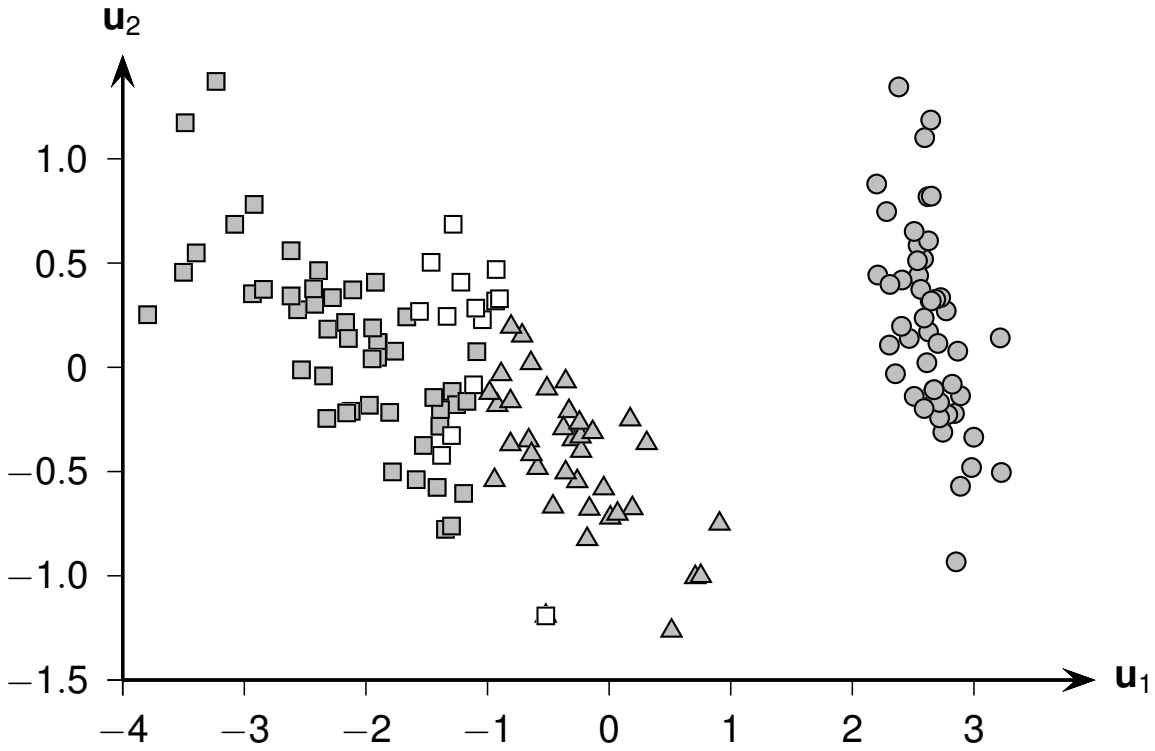


Таблица сопряженности:

Классы\кластеры	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)
T_1 (iris-setosa)	50	0	0
T_2 (iris-virginica)	0	1	49
T_3 (iris-versicolor)	0	36	14

Набор «Ирисы»: чистота кластеризации

Классы\кластеры	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)
T_1 (iris-setosa)	50	0	0
T_2 (iris-virginica)	0	1	49
T_3 (iris-versicolor)	0	36	14
$\sum$	50	37	63

Для набора «Ирисы показатели чистоты кластеров равны:

$$p_1 = \frac{50}{50} = 1,\ p_2 = \frac{36}{37} = 0.97,\ p_3 = \frac{49}{63} = 0.78$$

Чистота кластеризации $\mathcal{C}$ определяется как взвешенная сумма показателей чистоты кластеров

$$p = \frac{1}{150}(50 + 36 + 49) = \frac{135}{150} = 0.9$$

Точность и полнота кластера

Для кластера C_j обозначим через i_j класс, который содержит максимальное количество точек из C_j, т.е. $i_j = \max_{i=\overline{1,k}} n_{ij}$. Точность (precision) кластера C_j – то же самое, что его чистота

$$prec_j = \frac{1}{m_j} \max_{i=\overline{1,k}} n_{ij} = \frac{n_{i_j j}}{m_j}, \; j = \overline{1,r}$$

Точность измеряет долю точек $n_{i_j j}$ в кластере C_j.

Полнота (recall) кластера C_j определяется как

$$recall_j = \frac{n_{i_j j}}{\left|T_{i_j}\right|} = \frac{n_{i_j j}}{n_{i_j}}, \; j = \overline{1,r},$$

где $n_{i_j} = \left|T_{i_j}\right|$. Полнота измеряет долю точек $n_{i_j j}$ в классе T_{i_j}.

F-мера кластеризации

F-мера F_j – это гармоническое среднее точности и полноты для каждого кластера C_j:

$$F_j = \frac{2}{\frac{1}{prec_j} + \frac{1}{recall_j}} = \frac{2n_{i_j j}}{n_{i_j} + m_j}, \; j = \overline{1, r}$$

F-мера кластеризации $\mathcal{C}$ определяется как среднее значение показателей F-мер кластеров:

$$F = \frac{1}{r} \sum_{j=1}^{r} F_j$$

Таким образом, F-мера пытается сбалансировать показатели точности и полноты по всем кластерам. Для идеальной кластеризации при $r = k$ достигается максимальное значение F-меры, равное 1.

Набор «Ирисы»: расчет F-меры (1)

Полная таблица сопряженности:

	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)	n_i
T_1 (iris-setosa)	50	0	0	50
T_2 (iris-virginica)	0	1	49	50
T_3 (iris-versicolor)	0	36	14	50
m_j	50	37	63	$n = 150$

Для кластера C_1

$$prec_1 = \frac{50}{50} = 1,\ recall_1 = \frac{50}{50} = 1 \Rightarrow F_1 = 1$$

Для кластера C_2

$$prec_2 = \frac{36}{37} = 0.97,\ recall_2 = \frac{36}{50} = 0.72 \Rightarrow F_2 = \frac{2 \cdot 0.97 \cdot 0.72}{0.97 + 0.72} = 0.83$$

Набор «Ирисы»: расчет F-меры (2)

Полная таблица сопряженности:

	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)	n_i
T_1 (iris-setosa)	50	0	0	50
T_2 (iris-virginica)	0	1	49	50
T_3 (iris-versicolor)	0	36	14	50
m_j	50	37	63	$n = 150$

Для кластера C_3

$$prec_3 = \frac{49}{63} = 0.78,\ recall_3 = \frac{49}{50} = 0.98 \Rightarrow F_3 = \frac{2 \cdot 0.78 \cdot 0.98}{0.78 + 0.98} = 0.87$$

F-мера кластеризации:

$$F = \frac{1}{3}(1. + 0.83 + 0.87) = \frac{2.7}{3} = 0.9$$

Условная энтропия (1)

Энтропия кластеризации $\mathcal{C}$ определяется как:

$$H(\mathcal{C}) = -\sum_{j=1}^{r} p_{C_j} \log_2 p_{C_j}, \; p_{C_j} = \frac{m_j}{n}$$

Энтропия разбиения на классы $\mathcal{T}$ определяется как:

$$H(\mathcal{T}) = -\sum_{i=1}^{k} p_{T_i} \log_2 p_{T_i}, \; p_{T_i} = \frac{n_i}{n}$$

Условная энтропия $\mathcal{T}$ относительно кластера C_j равна

$$H(\mathcal{T} \mid C_j) = -\sum_{i=1}^{k} \left(\frac{n_{ij}}{n_i} \right) \log_2 \left(\frac{n_{ij}}{n_i} \right)$$

Условная энтропия (2)

Тогда условная энтропия $\mathcal{T}$ относительно кластеризации $\mathcal{C}$ определяется как взвешенная сумма

$$H(\mathcal{T} \mid \mathcal{C}) = \sum_{j=1}^{r} \frac{m_j}{n} H(\mathcal{T} \mid C_j) = -\sum_{j=1}^{r} \sum_{i=1}^{k} p_{ij} \log_2 \left(\frac{p_{ij}}{p_{C_j}} \right),$$

где $p_{ij} = \frac{n_{ij}}{n}$ – вероятность того, чтобы точка в кластере j также принадлежит классу i. Можно показать, что

$$H(\mathcal{T} \mid \mathcal{C}) = H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C}),$$

где $H(\mathcal{C}, \mathcal{T}) = -\sum_{j=1}^{r} \sum_{i=1}^{k} p_{ij} \log_2 p_{ij}$ – совместная энтропия $\mathcal{C}$ и $\mathcal{T}$.

Для идеальной кластеризации условная энтропия равна нулю.

Набор «Ирисы»: условная энтропия

Полная таблица сопряженности:

	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)	n_i
T_1 (iris-setosa)	50	0	0	50
T_2 (iris-virginica)	0	1	49	50
T_3 (iris-versicolor)	0	36	14	50
m_j	50	37	63	$n = 150$

$$H(\mathcal{T} \mid C_1) = -\frac{50}{50}\log_2\frac{50}{50} = 0$$

$$H(\mathcal{T} \mid C_2) = -\frac{1}{37}\log_2\frac{1}{37} - \frac{36}{37}\log_2\frac{36}{37} = 0.18$$

$$H(\mathcal{T} \mid C_3) = -\frac{49}{63}\log_2\frac{49}{63} - \frac{14}{63}\log_2\frac{14}{63} = 0.76$$

$$H(\mathcal{T} \mid \mathcal{C}) = \frac{50}{150}0 + \frac{37}{150}0.18 + \frac{63}{150}0.76 = 0.36$$

Парные меры

Пусть дана кластеризация $\mathcal{C}$ и разбиение на классы $\mathcal{T}$ и пусть $\hat{y}_i$ обозначает метку кластера, а y_i – истинную метку класса. Рассмотрим всевозможные пары точек $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}$ и подсчитаем следующие величины:

- $TP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j, \hat{y}_i = \hat{y}_j\}|$ (True Positives) – точки принадлежат одному кластеру и одному классу
- $FN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j, \hat{y}_i \neq \hat{y}_j\}|$ (False Negatives) – точки принадлежат разным кластерам, но одному классу
- $FP = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j, \hat{y}_i = \hat{y}_j\}|$ (False Positives) – точки принадлежат одному кластеру, но разным классам
- $TN = |\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j, \hat{y}_i \neq \hat{y}_j\}|$ (True Negatives) – точки принадлежат разным кластерам и разным классам

Позитивным событием считаем попадание точек $\mathbf{x}_i, \mathbf{x}_j$ в один кластер.

Так как всего возможны $N = \begin{pmatrix} n \\ 2 \end{pmatrix} = \frac{n(n-1)}{2}$ пар точек, справедливо равенство

$$N = TP + FN + FP + TN$$

Формулы для расчета парных мер

$$TP = \sum_{i=1}^{k}\sum_{j=1}^{r}\binom{n_{ij}}{2} = \frac{1}{2}\left(\left(\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2\right) - n\right)$$

$$FN = \sum_{i=1}^{k}\binom{n_i}{2} - TP = \frac{1}{2}\left(\sum_{i=1}^{k} n_i^2 - \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2\right)$$

$$FP = \sum_{j=1}^{r}\binom{m_j}{2} - TP = \frac{1}{2}\left(\sum_{j=1}^{r} m_j^2 - \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2\right)$$

$$TN = N - (TP + FN + FP) = \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2 - \sum_{j=1}^{r} m_j^2 + \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2\right)$$

Индексы Rand, Жаккара, Фоулкса – Мэллоуса

Индекс Rand оценивает, насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом:

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

Индекс Жаккара (Jaccard Index) не учитывает пары элементов находящиеся в разные классах и разных кластерах (TN):

$$\text{Jaccard} = \frac{TP}{TP + FP + FN}$$

Область значений обоих индексов от 0 до 1, где 1 — полное совпадение кластеров с заданными классами, а 0 — отсутствие совпадений.
Индекс Фоулкса – Мэллоуса (Fowlkes-Mallows Index) используется для определения сходства между двумя кластерами.

$$\text{FM} = \sqrt{\frac{TP}{TP + TN}\frac{TP}{TP + FP}}$$

Более высокое значение индекса означает большее сходство между кластерами.

Набор «Ирисы»: парные меры

Полная таблица сопряженности:

	C_1 (кружки)	C_2 (треугольники)	C_3 (квадраты)	n_i
T_1 (iris-setosa)	50	0	0	50
T_2 (iris-virginica)	0	1	49	50
T_3 (iris-versicolor)	0	36	14	50
m_j	50	37	63	$n = 150$

$$TP = \frac{1}{2}\left(50^2 + 1^2 + 36^2 + 49^2 + 14^2 - 150\right) = 3122$$

$$FN = \frac{1}{2}\left(3 \cdot 50^2 - 50^2 - 1^2 - 36^2 - 49^2 - 14^2\right) = 553$$

$$FP = \frac{1}{2}\left(50^2 + 37^2 + 63^2 - 50^2 - 1^2 - 36^2 - 49^2 - 14^2\right) = 722$$

$$TN = N - (TP + FN + FP) = 11175 - 3122 - 553 - 722 = 6778$$

Набор «Ирисы»: индексы парных мер

$$TP = 3122, FN = 553, FP = 722, TN = 6778$$

$$Rand = \frac{TP + TN}{TP + TN + FP + FN} = \frac{3122 + 6778}{11175} = 0.89$$

$$Jaccard = \frac{TP}{TP + FP + FN} = \frac{3122}{3122 + 722 + 553} = 0.71$$

$$FM = \sqrt{\frac{TP}{TP + TN}\frac{TP}{TP + FP}} = \sqrt{\frac{3122}{3122 + 6778}\frac{3122}{3122 + 722}} = 0.51$$