Введение в многомерный статистический анализ
2021-2022 учебный год

Автор лекции: Тамбовцева А.А.

Лекция 1. Иерархический кластерный анализ: методы кластеризации количественных данных

1.1 Постановка задачи

Классический кластерный анализ решает задачу **классификации без обучения** на данных, распределение которых нам неизвестно. Другими словами, классический кластерный анализ распределяет имеющиеся в массиве данных наблюдения на группы, не предлагая при этом алгоритм для предсказания класса, к которому будет отнесено то или иное наблюдение[1]. Если нас интересует задача классификации без обучения на данных, распределение которых нам известно, вместо кластерного анализа можно рассмотреть расщепление смесей, а для классификации с обучением пригодятся различные разновидности логистической регресссии, решающие деревья и дискриминантный анализ.

Для реализации кластерного анализа на входе необходимо иметь p-мерный массив данных (датафрейм из n наблюдений и p переменных):

$$\begin{array}{cccc} X_1 & X_2 & \dots & X_p \\ x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots \\ x_{j1} & x_{j2} & \dots & x_{jp} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{array}$$

Если речь идёт об иерархическом кластерном анализе, заранее знать количество кластеров, которое мы хотим получить, необязательно, однако если мы говорим о кластеризации методом k-средних, количество желаемых кластеров является необходимым параметром. О различиях этих видов кластерного анализа мы поговорим чуть позже, пока стоит отметить, что в любом случае априорная информация о количестве кластеров будет полезна.

Итак, на входе мы имеем p-мерный массив данных X, а на выходе получаем правило, которое позволяет наилучшим в определённом смысле образом разбить имеющиеся наблюдения на однородные в определённом смысле группы. Что это за «определённый смысл», и почему всё так неоднозначно? Иерархический кластерный анализ не предлагает единственного решения, некоторого оптимального разбиения на кластеры, поэтому именно на исследователя возлагается задача выбрать наилучший вариант классификации. Тем не менее, однородность групп достигается довольно понятным образом, в кластерном анализе реализуется довольно распространённый для сравнения групп принцип: внутригрупповой разброс значений должен быть минимальным, а межгрупповой разброс – довольно существенным.

[1] В машинном обучении существуют другие методы кластеризации, например, классификация методом k-ближайших соседей, которые позволяют предсказывать, к какому кластеру отнести новое наблюдение, но здесь мы такие варианты не рассматриваем.

Пример 1. Чтобы понять, какая идея стоит за алгоритмами реализации кластерного анализа, давайте рассмотрим следующую задачу. Пусть у нас есть небольшой двумерный массив данных, где X – время, на которое преподаватель опаздывает на пару (в минутах), а Y – время, на которое преподаватель опаздывает на личные встречи (в минутах):

id	X	Y
A	2	6
B	2	8
C	8	2
D	10	3
E	5	5

Наша задача – понять, каким образом разделить этих преподавателей на группы. Пока, исходя из содержательных соображений, можно предположить, что всех людей можно разделить на три группы: те, кто старается не опаздывать на работу, но может опаздывать на личные встречи; те, кто спокойно опаздывает на работу, но не допускает опоздания на личные встречи; те, кто опаздывает всегда и везде.

Для наглядности построим **диаграмму рассеивания** на основе имеющихся данных.

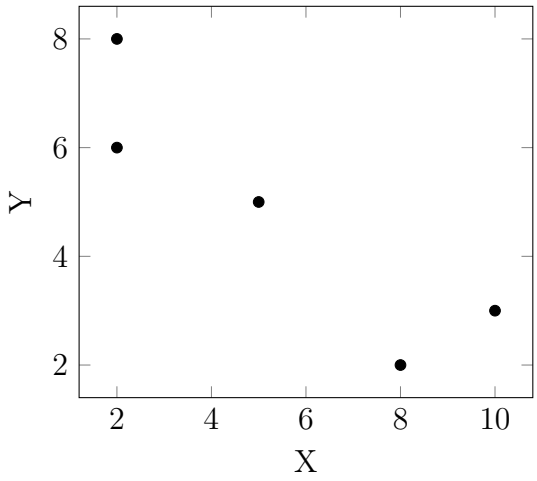

Рис. 1: Диаграмма рассеивания X vs Y

Возникает логичный вопрос: что нам понадобится, чтобы поделить все эти наблюдения на группы? Во-первых, способ измерения различий между этими наблюдениями, а во-вторых, алгоритм, который позволит на основе выявленных различий закреплять наблюдения за определёнными группами. Итак, мы подошли к самому главному – к параметрам кластеризации.

1.2 Иерархический кластерный анализ: параметры кластеризации

Прежде чем описывать параметры кластеризации, давайте договоримся, что в этой лекции мы говорим об **иерархическом кластерном анализе**. В основе данного вида кластерного анализа лежат два предположения:

1. На самом первом шаге кластерного анализа количество кластеров совпадает с количеством наблюдений (имеем n кластеров, состоящих ровно из одного наблюдения).

2. Количество кластеров заранее неизвестно, мы объединяем точки в кластеры до тех пор, пока не получим один большой кластер. Так, на первом шаге иерархического кластерного анализа у нас есть n кластеров, на втором шаге $(n-1)$ кластеров, на третьем уже $(n-2)$ кластеров, и так далее, а на последнем шаге остаётся один кластер. Другими словами, мы выстраиваем некоторую иерархию из кластеров, вложенных друг в друга, а потом решаем, на каком делении, более детальном (много маленьких кластеров) или более общем (мало больших кластеров), стоит остановиться.

У иерархического кластерного анализа есть **два параметра кластеризации**:

1. Метрика: мера расстояния между точками (наблюдениями).
2. Метод агломерации или метод агрегирования: алгоритм, который позволяет решать, каким образом объединять точки в кластеры на основе выбранной метрики[2].

Остановимся на метриках и сформулируем **свойства метрики**.

Пусть i, j, k – некоторые точки, а $d(i,j)$ – метрика, расстояние от точки i до точки j. Тогда:

1. $d(i,i) = 0$;
2. $d(i,j) \geqslant 0$;
3. $d(i,j) = d(j,i)$;
4. $d(i,k) \leqslant d(i,j) + d(j,k)$.

Переводя эти свойства на менее формальный язык, получим довольно логичные утверждения: расстояние от точки до самой себя равно 0, расстояние не бывает отрицательным, расстояние от точки i до точки j – то же самое, что расстояние от точки j до точки i, расстояние от точки i до точки k меньше суммы расстояния от точки i до точки j и расстояния от точки j до точки k (неравенство треугольника).

Давайте рассмотрим основные **виды метрик**, которые часто используют при кластеризации данных, измеренных **в количественной шкале**.

Для определённости давайте зафиксируем, что в p-мерном пространстве у нас есть две точки $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ и $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$. Если обозначения кажутся не совсем понятными, посмотрите, как у нас записан массив X в постановке задачи (x_i и x_j – просто строки, соответствующие i-тому и j-тому наблюдению).

Итак, виды метрик для данных в количественной шкале:

1. Евклидово расстояние, также обозначается L2, одно из самых распространённых расстояний:
$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{p} (x_{ip} - x_{jp})^2}.$$

[2]На самом первом шаге иерархического кластерного анализа необходимости в методе агломерации нет, мы просто однозначным образом определяем расстояние между точками. Начиная со второго шага, однозначность теряется – как определить расстояние между группами точек? Можно посчитать расстояния между каждой точкой в первой группе и каждой точкой во второй группе и выбрать из них минимальное, а можно – максимальное. Наконец, можно, рассчитав расстояния между всеми парами точек, усреднить их... Способов много, это мы увидим далее.

2. Квадрат евклидова расстояния, также обозначается L2 squared, необходимо для некоторых методов агломерации, в частности, для метода Уорда (Варда):

$$d(x_i, x_j) = \sum_{p=1}^{p} (x_{ip} - x_{jp})^2.$$

3. Манхэттенское расстояние, оно же блочное расстояние, также обозначается L1:

$$d(x_i, x_j) = \sum_{p=1}^{p} |x_{ip} - x_{jp}|.$$

4. Расстояние Чебышёва:

$$d(x_i, x_j) = \max\{|x_{ip} - x_{jp}|\}.$$

Конечно, перечисленными выше расстояниями список метрик не исчерпывается, их гораздо больше. Но далеко не все метрики активно используются в кластерном анализе. Тем не менее, хотелось бы выделить группу метрик, которые помимо «естественных» геометрических координат используют информацию о распределении данных. Так, например, расстояние Махаланобиса использует ковариационную матрицу, содержащую информацию о связи между случайными векторами и их дисперсии.

Для того, чтобы реализовать иерархический кластерный анализ, нам необходимо определиться с метрикой и получить **матрицу расстояний** – матрицу, состоящую из расстояний между всеми парами точек. Как можно догадаться, эта матрица будет квадратной, симметричной, а на главной диагонали будут находиться 0 (вспомним свойства метрики).

Пример 2. Построим матрицу расстояний для нашего массива данных по опаздывающим преподавателям. Сколько расстояний нам придётся посчитать для пяти точек? Всего 10 расстояний, так как число различных пар из n точек вычисляется так[3]:

$$\frac{n(n-1)}{2}.$$

Приведём вычисления для каждой пары точек[4]:

$d(A, B) = \sqrt{(2-2)^2 + (6-8)^2} = \sqrt{4} = 2;$

$d(A, C) = \sqrt{(2-8)^2 + (6-2)^2} = \sqrt{52} \approx 7.2;$

$d(A, D) = \sqrt{(2-10)^2 + (6-3)^2} = \sqrt{73} \approx 8.5;$

$d(A, E) = \sqrt{(2-5)^2 + (6-5)^2} = \sqrt{10} \approx 3.2;$

$d(B, C) = \sqrt{(2-8)^2 + (8-2)^2} = \sqrt{72} \approx 8.5;$

$d(B, D) = \sqrt{(2-10)^2 + (8-3)^2} = \sqrt{89} \approx 9.4;$

$d(B, E) = \sqrt{(2-5)^2 + (8-5)^2} = \sqrt{18} \approx 4.2;$

$d(C, D) = \sqrt{(8-10)^2 + (2-1)^2} = \sqrt{5} \approx 2.2;$

[3]Вспоминаем задачу о числе рукопожатий или количестве рёбер в полном графе.

[4]Строго говоря, перед вычислением расстояний мы должны шкалировать наши данные – из каждого значения в столбце вычесть среднее столбца и поделить на его стандартное отклонение. Но давайте пока пренебрежём этой операцией, учитывая, что X и Y приведены в одних и тех же единицах измерения.

$d(C, E) = \sqrt{(8-5)^2 + (2-5)^2} = \sqrt{18} \approx 4.2$;

$d(D, E) = \sqrt{(10-5)^2 + (3-5)^2} = \sqrt{29} \approx 5.4$.

Заполним матрицу расстояний:

$$D = \begin{bmatrix} & \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{A} & 0 & 2 & 7.2 & 8.5 & 3.2 \\ \boldsymbol{B} & 2 & 0 & 8.5 & 9.4 & 4.2 \\ \boldsymbol{C} & 7.2 & 8.5 & 0 & 2.2 & 4.2 \\ \boldsymbol{D} & 8.5 & 9.4 & 2.2 & 0 & 5.4 \\ \boldsymbol{E} & 3.2 & 4.2 & 4.2 & 5.4 & 0 \end{bmatrix}$$

Это был довольно трудоёмкий процесс, но этой матрицы нам будет достаточно, чтобы реализовать кластерный анализ, когда определимся с методом агломерации.

С метриками разобрались, теперь перейдём к **методам агломерации** или **методам агрегирования**. Можно выделить следующие основные методы агломерации.

1. Метод ближнего соседа, он же метод одиночной связи (*single linkage*).

   Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между ближайшими точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем выбираем минимальное из посчитанных – это и будет расстояние между кластерами A и B.

   Особенности. Имеет недостаток: склонен образовывать кластеры, состоящие из одного наблюдения (монокластеры).

2. Метод дальнего соседа, он же метод полной связи (*complete linkage*).

   Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между дальними точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем выбираем максимальное из посчитанных – это и будет расстояние между кластерами A и B.

   Особенности. Вполне надёжный метод, используется в качестве метода агломерации по умолчанию функцией `hclust()` в R.

3. Метод средней связи (*average linkage*).

   Реализация. Расстояние между двумя кластерами A и B определяется как среднее расстояние между точками этих кластеров. Считаем расстояния между всеми парами точек, одна точка в паре из кластера A, вторая – из кластера B, затем считаем среднее арифметическое – это и будет расстояние между кластерами A и B.

   Особенности. Особых примет нет, тоже вполне надёжный метод.

4. Метод центроидной связи (*centroid linkage*).

   Реализация. Расстояние между двумя кластерами A и B определяется как расстояние между центроидами (центрами тяжести) кластеров. Центроид – средний вектор кластера, его координаты считаются как средние арифметические соответствующих переменных[5].

[5] Допустим, у нас есть кластер из трёх точек $A = (2, 3)$, $B = (2, 6)$, $C = (5, 3)$. Тогда первая координата центра тяжести равна $(2+2+5)/3 = 3$, а вторая координата равна $(3+6+3)/3 = 4$. Итого получаем центроид $O = (3, 4)$.

Особенности. Имеет недостаток: может вызывать инверсию – ситуацию, когда на последующем шаге кластеризации объединение в кластеры происходит на расстоянии меньшем, чем на предыдущем шаге. Инверсия противоречит самой логике иерархического кластерного анализа: для минимальной потери информации об исходных наблюдениях (а мы теряем её, переходя к группам), мы должны на каждом шаге кластеризации объединять точки в более крупные кластеры, которые в большей степени удалены друг от друга, а здесь мы «скачем» от большего расстояния к меньшему.

5. Метод Уорда, он же метод Варда (*Ward's linkage*).

   Реализация. На каждом шаге обновления кластеров точка присоединяется к тому кластеру, присоединение к которому приводит к минимально возможному увеличению внутригрупповой дисперсии этого кластера. Внутригрупповую дисперсию можно определить следующим образом:

$$SS = \sum_{i \in G} (x_i - \bar{x}_G)^2,$$

   где G – кластер, а $\bar{x}_G$ – средний вектор, центроид этого кластера. Соответственно, на каждом шаге для каждого кластера оценивается текущее значение SS, возможное значение SS в случае добавления точки в кластер, и точка добавляется к тому кластеру, где изменение SS минимально.

   Особенности. Считается одним из самых эффективных методов[6] агломерации, требует использования только одной метрики – квадрата евклидова расстояния.

1.3 Иерархический кластерный анализ: построение дендрограммы

Итак, чтобы реализовать иерархический кластерный анализ, нам необходимо определиться с метрикой и методом агломерации. Допустим, мы выбрали евклидово расстояние и метод ближнего соседа[7].

Запустим кластерный анализ и построим **дендрограмму** – график, который визуализирует результаты иерархического кластерного анализа и позволяет увидеть все возможные варианты кластеризации, от наиболее детальной, где много маленьких кластеров, до наиболее общей, где мало больших кластеров.

По горизонтальной оси на дендрограмме отмечаются сами наблюдения, по вертикальной – расстояния между наблюдениями или кластерами на момент объединения их в более крупный кластер.

Пример 3. Построим дендрограмму для нашего примера с пятью точками. Так как точек пять, мы завершим иерархическую кластеризацию после пяти шагов.

Шаг 1. На первом шаге у нас 5 кластеров, каждый кластер состоит из одной точки. В иерархическом кластерном анализе всегда первый шаг будет таким, вне зависимости от выбранного способа агломерации.

Итого получаем 5 кластеров: A, B, C, D, E.

[6] Это неслучайно, запомните идею о минимизации внутригрупповой дисперсии, она немного в ином виде появится при обсуждении кластеризации методом k-средних, а он считается более точным, чем классический иерархический кластерный анализ.

[7] Да, он неидеальный, но довольно простой, поймём его, поймём по аналогии метод дальнего соседа и метод средней связи.

Шаг 2. Теперь объединим в кластер те точки, которые ближе всего друг к другу. Это точки A и B, расстояние между ними 2. Соединим эти точки, а на вертикальной оси отметим расстояние 2.

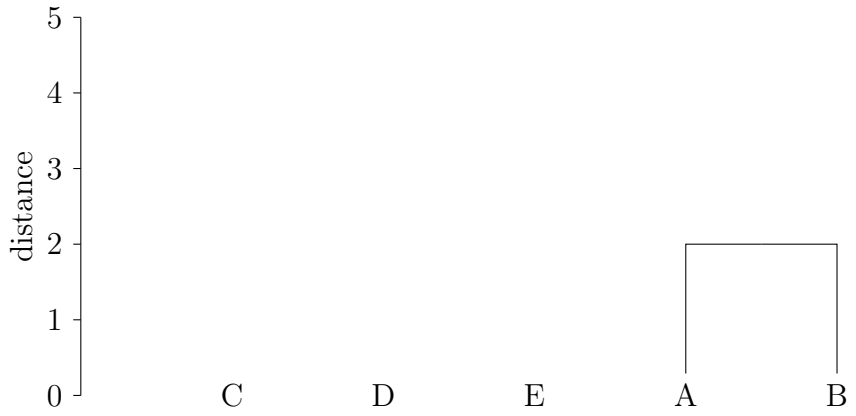

Рис. 2: Дендрограмма: шаг 2

По итогу этого шага получаем уже 4 кластера: $A + B$, C, D, E.

Шаг 3. Нужно снова объединить точки в более крупные кластеры. Для этого необходимо определить попарные расстояния между точками C, D и E, а также расстояние между кластером $A + B$ и точкой C, между кластером $A + B$ и точкой D, между кластером $A + B$ и точкой E.

С отдельными точками всё понятно, а с кластером $A + B$ всё сложнее. Возникает неоднозначность: как посчитать расстояние между группами точек? Можно взять расстояние между самыми близкими точками в группах, а можно – между самыми дальними, можно определить расстояние между центрами тяжести групп... Именно на это влияет выбор метода агломерации.

Мы выбрали метод ближнего соседа, поэтому нам надо посчитать расстояние от всех точек кластера $A + B$ до всех точек кластера E и выбрать минимальное:

$$d(A, E) = 3.2$$

$$d(B, E) = 4.2$$

$$d(A + B, E) = \min(d(A, E), d(B, E)) = 3.2$$

Как можно догадаться, при выборе метода дальнего соседа мы будем брать максимальное значение из посчитанных, а при выборе метода средней связи – среднее расстояние.

Проделаем такую операцию для остальных точек и получим матрицу расстояний:

$$D = \begin{bmatrix} & \boldsymbol{A+B} & \boldsymbol{C} & \boldsymbol{D} & \boldsymbol{E} \\ \boldsymbol{A+B} & 0 & 7.2 & 8.5 & 3.2 \\ \boldsymbol{C} & 7.2 & 0 & 2.2 & 4.2 \\ \boldsymbol{D} & 8.5 & 2.2 & 0 & 5.4 \\ \boldsymbol{E} & 3.2 & 4.2 & 5.4 & 0 \end{bmatrix}$$

На этом шаге самые близкими оказались точки C и D, соединим их на расстоянии 2.2.

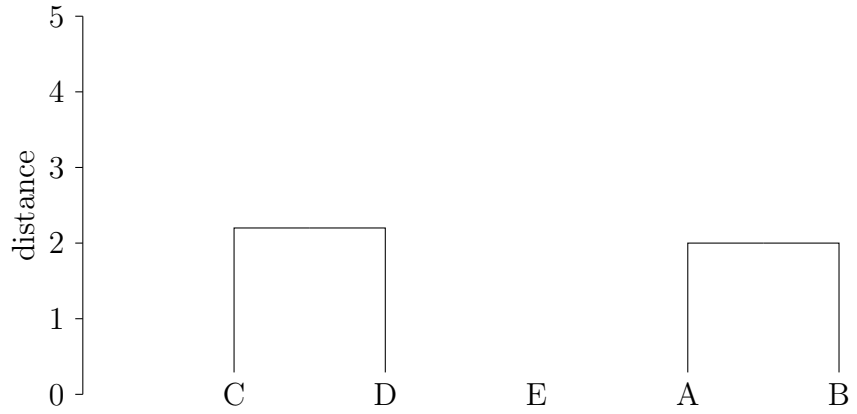

Рис. 3: Дендрограмма: шаг 3

Получаем три кластера: $A + B$, $C + D$, E.

Шаг 4. По той же схеме строим новую матрицу расстояний:

$$D = \begin{bmatrix} & \boldsymbol{A+B} & \boldsymbol{C+D} & \boldsymbol{E} \\ \boldsymbol{A+B} & 0 & 7.2 & 3.2 \\ \boldsymbol{C+D} & 7.2 & 0 & 4.2 \\ \boldsymbol{E} & 3.2 & 4.2 & 0 \end{bmatrix}$$

Исходя из полученных значений, точку E нужно присоединить к кластеру $A + B$. Обновим нашу дендрограмму.

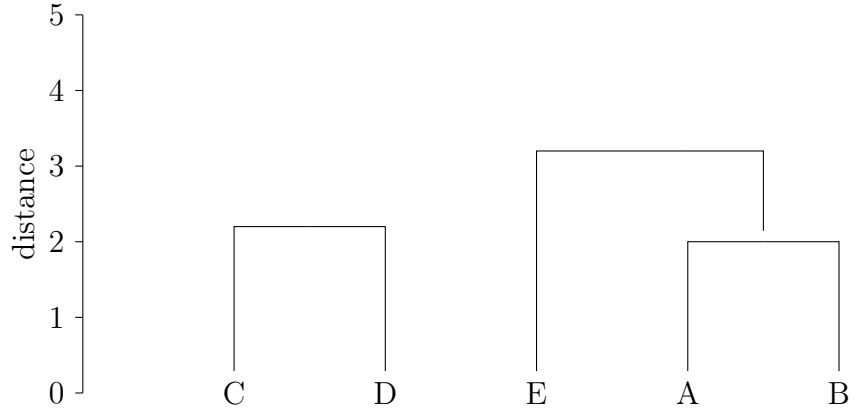

Рис. 4: Дендрограмма: шаг 4

Итого получаем два кластера: $A + B + E$, $C + D$.

Шаг 5. Осталось объединить всё в один большой кластер. Финальный штрих – соединяем все ветви на расстоянии 4.2 (можете посчитать и проверить самостоятельно). Завершим построение дендрограммы!

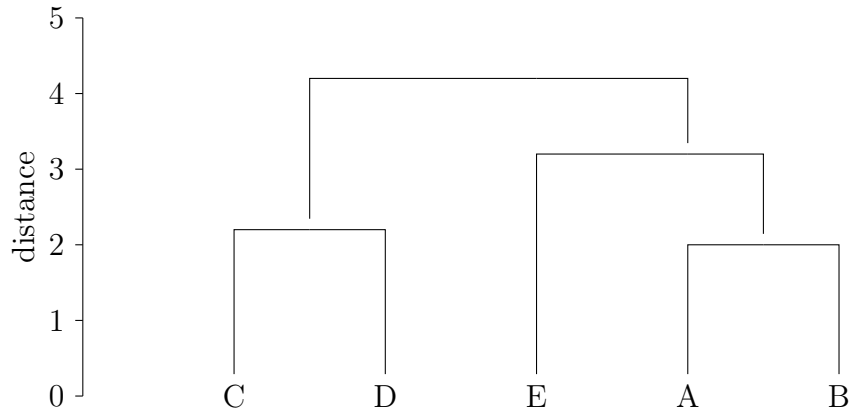

Рис. 5: Дендрограмма: шаг 5

1.4 Подводя итоги иерархического кластерного анализа: выбор числа кластеров

Итак, благодаря дендрограмме мы увидели все возможные варианты деления наших наблюдений на группы. Но сколько кластеров выбрать? Ответ на этот вопрос довольно простой, но при этом многозначный:

1. стоит взять столько кластеров, сколько можем содержательно проинтерпретировать;
2. стоит взять столько кластеров, сколько является разумным с точки зрения выраженности межгрупповых различий.

В то время как первый критерий выбора зависит исключительно от наших экспертных, во многом субъективных, знаний, выполнение второго мы можем проверить формально, используя известные статистические методы для сравнения двух и более групп (критерий Стьюдента для двух выборок, критерий Уилкоксона, ANOVA, критерий Краскелла-Уоллиса и другие).

Наш пример с пятью наблюдениями, конечно, слишком игрушечный, чтобы всерьёз рассуждать, сколько кластеров здесь выбрать, но явно выбор будет стоять между двумя и тремя группами. Если мы хотим получить более общую классификацию и избежать слишком маленьких кластеров из одного человека, то стоит выбрать две группы: $C + D$ и $A + B + E$. Если для нас выделяющиеся наблюдения играют важную роль, то деление на три кластера, где точка E обособлена, будет в приоритете.