# Кластерный анализ. Иерархическая кластеризация

# Кластерный анализ: терминология и постановка задачи

**Кластер** — калька слова «cluster», то есть «сгусток», «скопление)» и т.п.

Другие названия: обучение без учителя, распознавание образов без учителя.

Кластерный анализ:

1. разбивает набор объектов на группы (кластеры);
2. определяет число кластеров.

# Задача: сегментация рынка

По данным о покупателях (результаты опроса, поведение на сайте интернет-магазина) выявить и описать/понять рыночные сегменты.

Прежде, чем фирма определится, какие сегменты рынка создают для нее наибольшие возможности, надо решить, какие сегменты уже существуют.

# Задача: определение групп клиентов

Страховая компания интересуется группами, на которые разделяются потенциальные клиенты. Результаты классификации используются, чтобы для разных групп определять оптимальные цены на услуги, оптимальные тарифы, ...

Для разбиения потребителей на группы можно выбирать разные наборы характеристики объектов, например возраст, образование, место жительства, тип личности, и так далее.

Несложно разделить покупателей на сегменты по **одной** (или по каждой) характеристике. Кластерный анализ помогает выявить уже сложившееся разбиение потребителей на «группы со схожими потребностями в отношении конкретного товара или услуги, достаточными ресурсами, а также готовностью и возможностью покупать» учитывая **все** выбранные показатели одновременно.

# Задача: товарные группы для рекомендательной системы

На рынке присутствует большой выбор товаров схожего назначения под разными торговыми марками. Надо разбить товары на группы.

Иногда такое разбиение известно и получается без применения статистической техники. Например, компьютеры бывают «для дома», «для офиса», «серверы» и «специализированные».

Кластерный анализ применяется, если нет классификации, признанной всеми.

Важно! Результат будет зависеть от выбора набора показателей.

# Внимание! Возможна путаница

Разделение объектов на группы нередко называют **классификацией**. Однако в Data Mining и смежных дисциплинах классификация – это совсем другая задача.

- **Кластерный анализ**: какие группы объектов существуют? сколько этих групп?
- **Классификация** (обучение с учителем): к какой из заранее заданных групп следует отнести новый объект (наблюдение).

# Идея кластеризации: сведем задачу к геометрической

- Каждый объект — точка.
- Похожие объекты расположены «близко» друг к другу.
- Различающиеся объекты расположены «далеко».
- Скопления похожих точек образуют кластер.

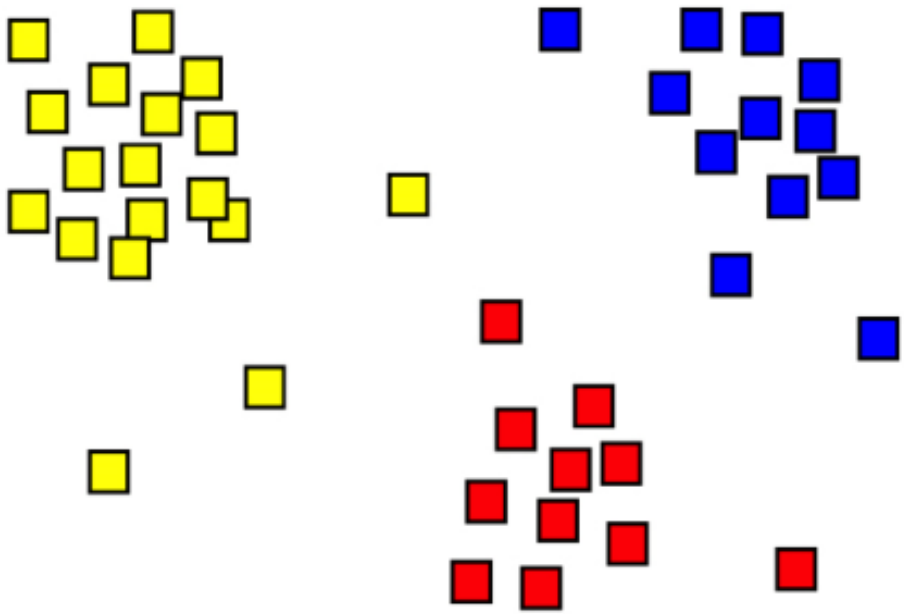


Figure 1: Три кластера

# Технические проблемы

- Как считать расстояние между точками (объектами)?
- Как считать расстояние между кластерами?

# Выбор меры расстояния между точками

# Расстояние между точками

- Евклидово расстояние
- Квадрат евклидова расстояния
- Расстояние городских кварталов (манхэттенское расстояние, расстояние таксиста)
- Расстояние Чебышева
- Степенное расстояние ...

# Меры расстояния между точками — 1

## Евклидово расстояние

Наиболее распространенная функция расстояния. Представляет собой геометрическое расстояние в многомерном пространстве:

$$d(x, x') = \sqrt{\sum_{i=1}^{n} (x_i - x'_i)^2}$$

## Квадрат евклидова расстояния

Применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

$$d(x, x') = \sum_{i=1}^{n} (x_i - x'_i)^2$$

# Меры расстояния между точками — 2

## Степенное расстояние

Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

$$d(x, x') = \sqrt[r]{\sum_{i=1}^{n}(x_i - x'_i)^p},$$

где $r$ и $p$ – параметры, определяемые пользователем. $p$ отвечает за постепенное взвешивание разностей по отдельным координатам, $r$ – за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – $r$ и $p$ — равны двум, то это расстояние совпадает с расстоянием Евклида.

# Расстояние Manhatten


Figure 2:

# Меры расстояния между точками — 3

## Расстояние городских кварталов (манхэттенское расстояние)

Это расстояние является средним разностей по координатам. Для этой меры влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат.

$$d(x, x') = \sum_{i=1}^{n} |x_i - x'_i|$$

## Расстояние Чебышева

Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они существенно различаются по какой-либо одной координате.

$$d(x, x') = \max(|x_i - x'_i|)$$

# Меры расстояния между точками — 4

## Расстояние Хэмминга

Число позиций, в которых соответствующие символы двух слов одинаковой длины различны

$$d(1011101, 1001001) = 2$$

$$d(2173896, 2233796) = 3$$

$$d(toned, roses) = ?$$

Для категориальных данных (мера несогласия).

# Когда какое расстояние использовать?

Начинаем с евклидова или с Манхэттен.

При выборе между евклидовым и Манхэттен спрашиваем себя: если точки различаются по одной из координат, то похожи или нет соотвествующие им объекты? Евклидово расстояние эту непохожесть подчеркнет, а Манхэттен сгладит.

# Выбор расстояния между кластерами

# Расстояние между кластерами

- Среднее невзвешенное расстояние (Average linkage clustering).
- Центроидный метод (Centroid Method).
- Метод дальнего соседа, максимального расстояния (Complete linkage clustering).
- Метод ближайшего соседа (Single linkage clustering).
- Метод Варда (Ward's method).

# Среднее невзвешенное расстояние

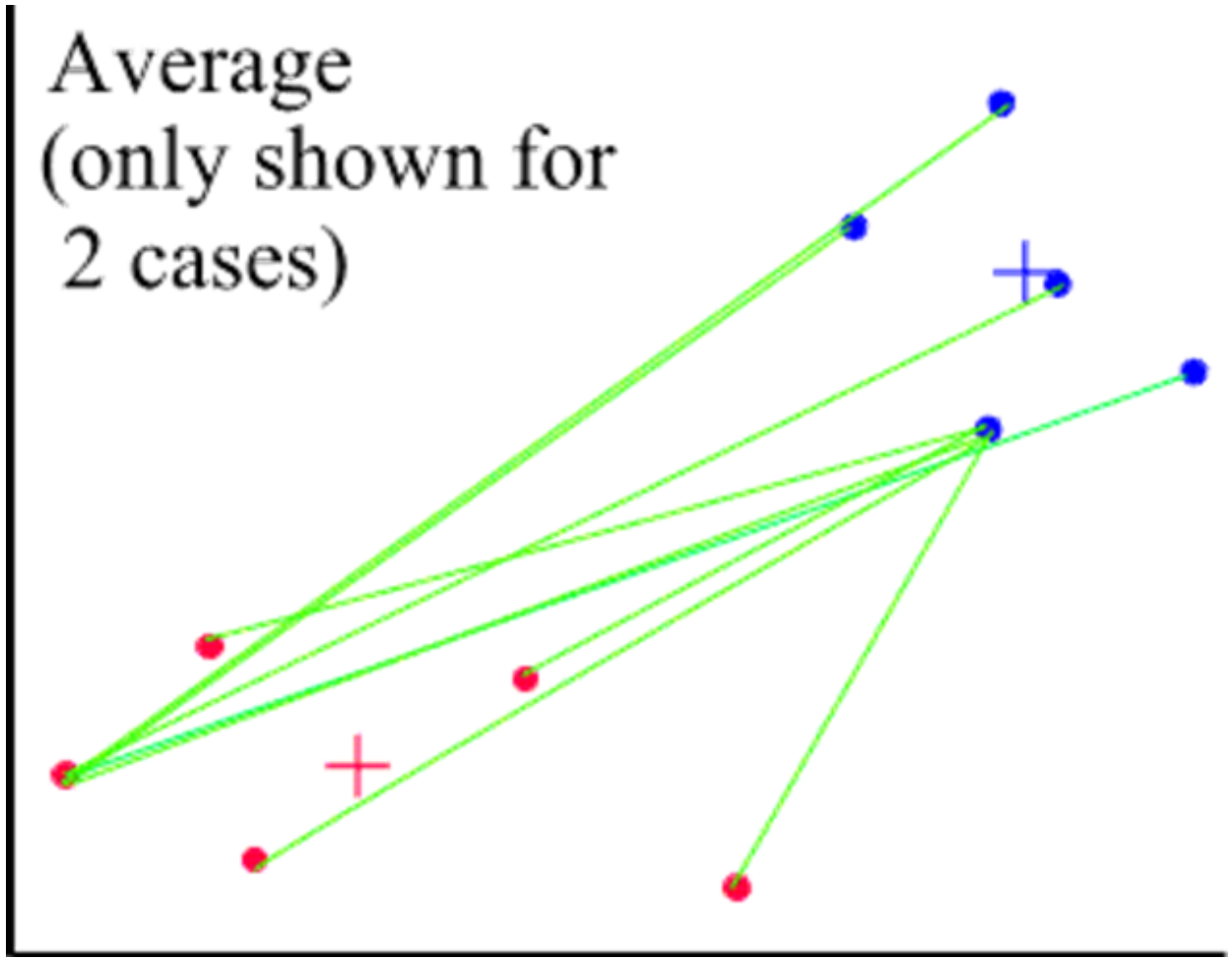
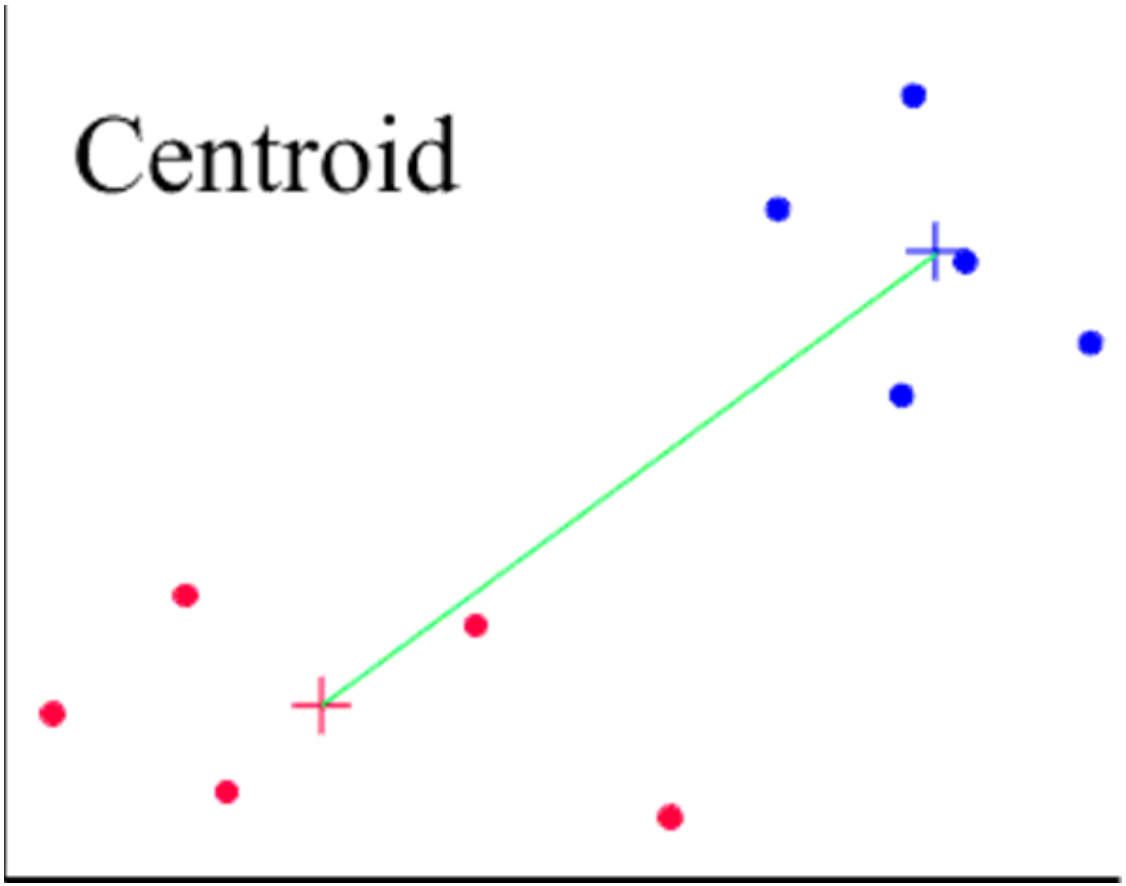



Figure 3: Невзвешенное попарное среднее. Расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них

# Центроидный метод




$$\bar{x}_j = \frac{1}{N_j}\sum_i^{N_j} x_i, \quad \bar{y}_j = \frac{1}{N_j}\sum_i^{N_j} y_i.$$

# Центроидный метод — 2

## Pro

- Вычислительная простота.
- Объем кластера не влияет.

## Contra

- Дендрограмма может иметь самопересечения.

# Метод дальнего соседа

Этот метод обычно работает очень хорошо, когда объекты происходят из отдельных групп. Если же кластеры имеют удлиненную форму или их естественный тип является «цепочечным» (ленточным), то этот метод непригоден.

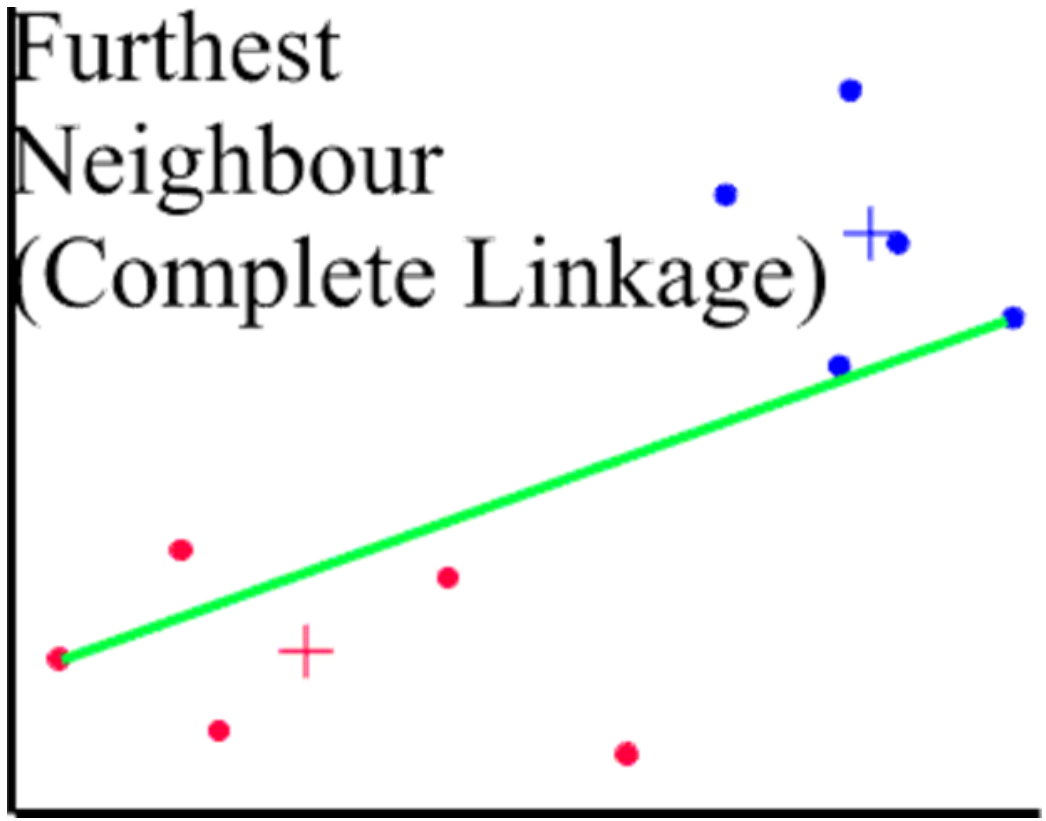



Figure 4: Полная связь. Расстояние между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах, т. е. между наиболее удаленными соседями

# Метод ближайшего соседа

Результирующие кластеры имеют тенденцию объединяться в цепочки.

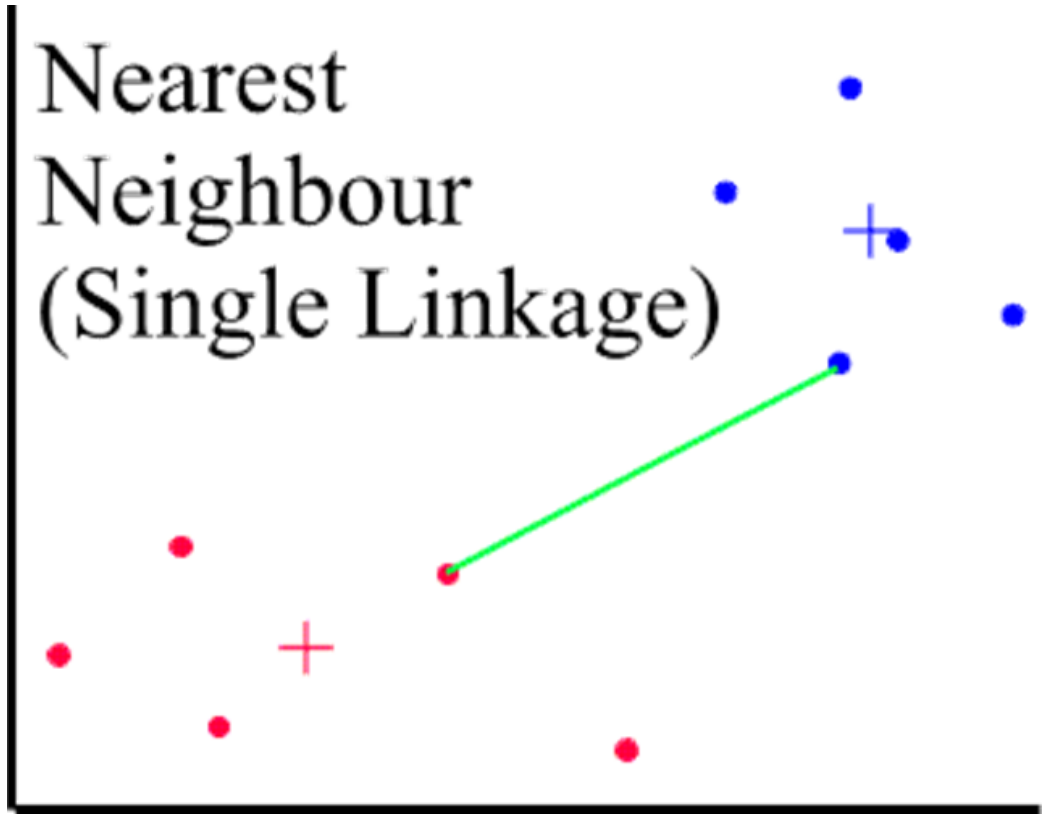



Figure 5: Одиночная связь. Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах

# Растояние Sørensen–Dice

Расстояние между сайтами.

$$Q = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

# Метод Варда (Ward)

В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения.

На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера.

Теперь можно использовать не только квадрат евклидова расстояния.

# Выраженные кластеры – все равно какой метод

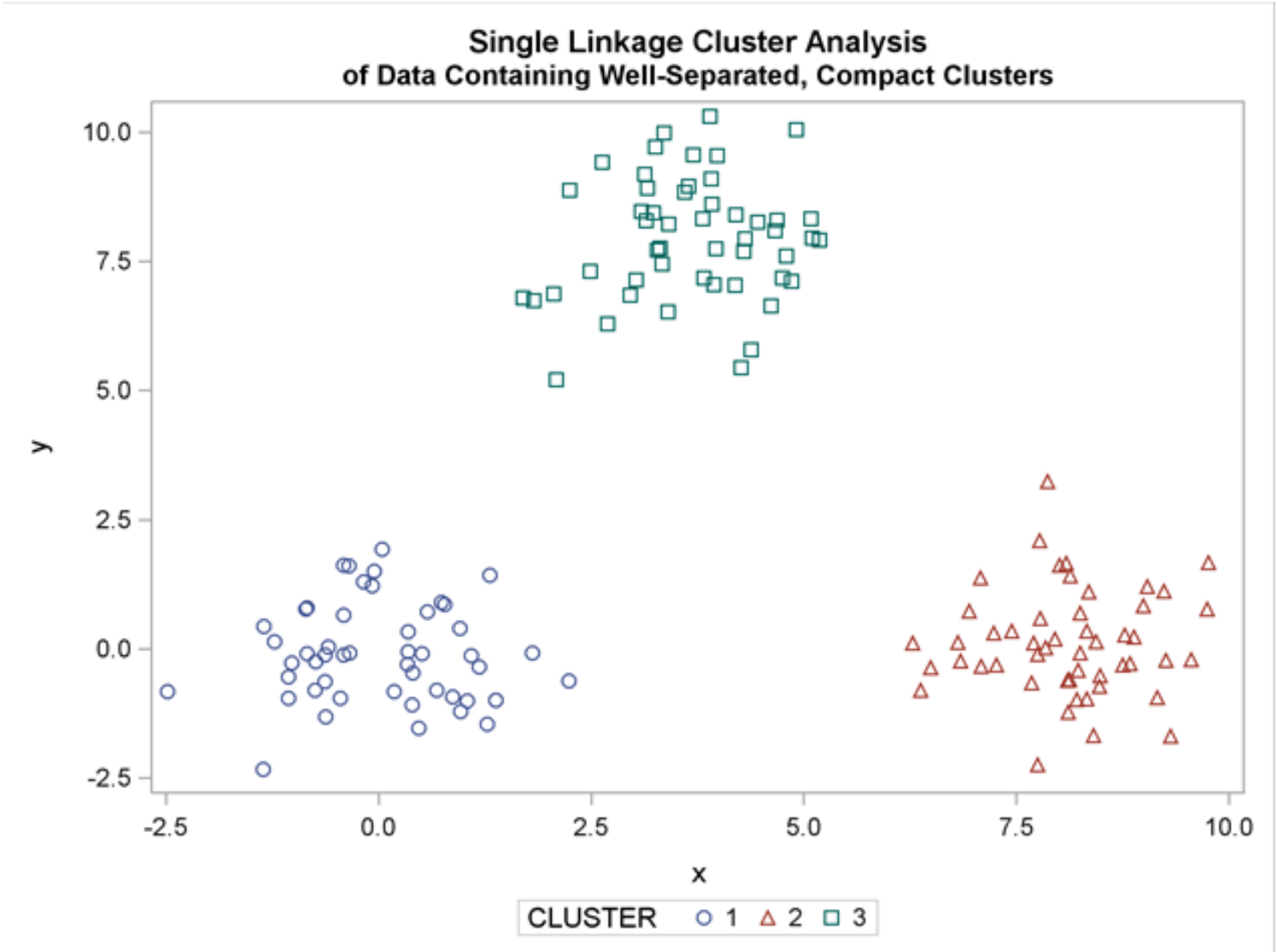



Figure 6:

# Ленточные кластеры: попарное среднее

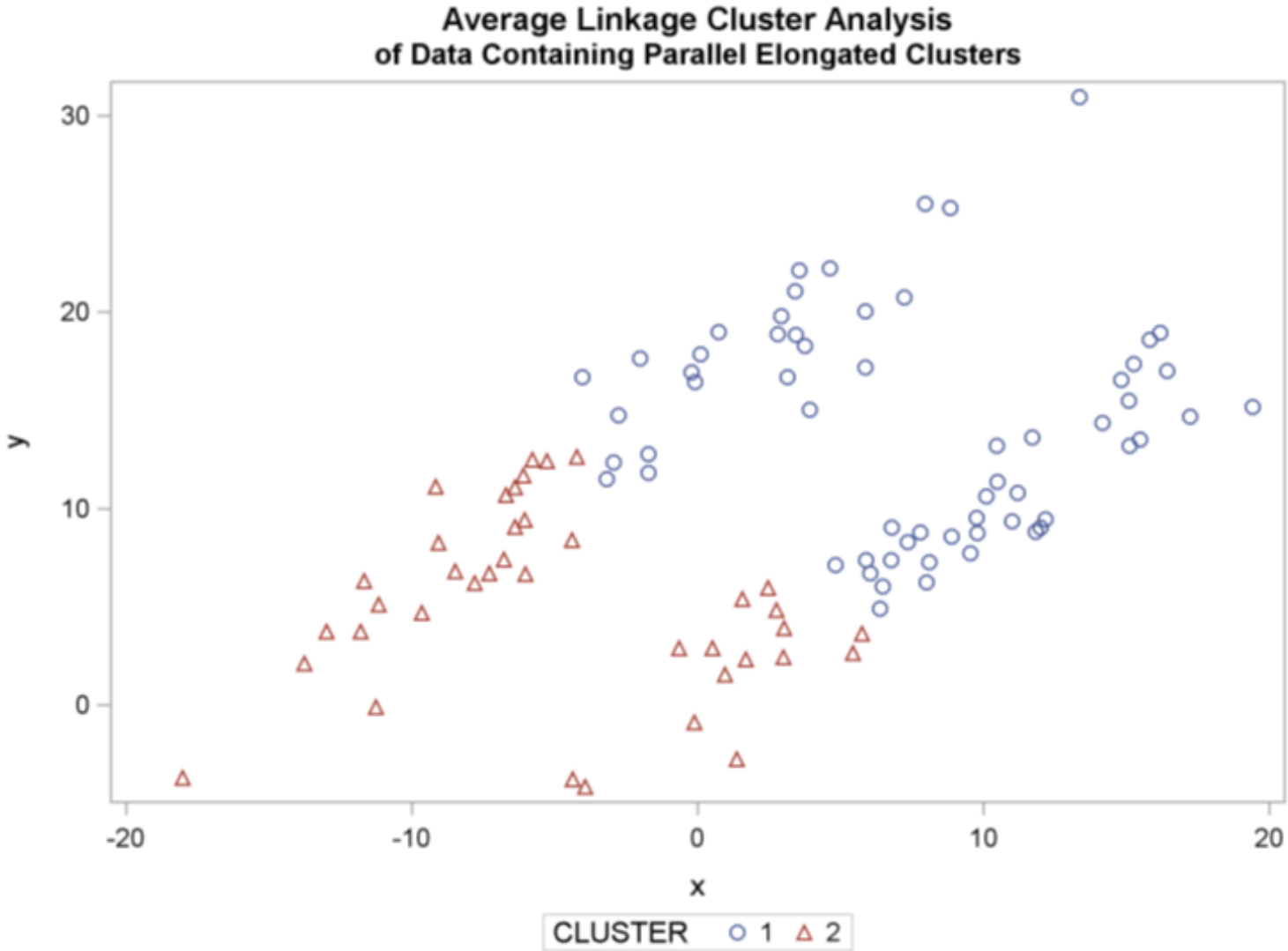



Figure 7:

# Еще ленточные кластеры...

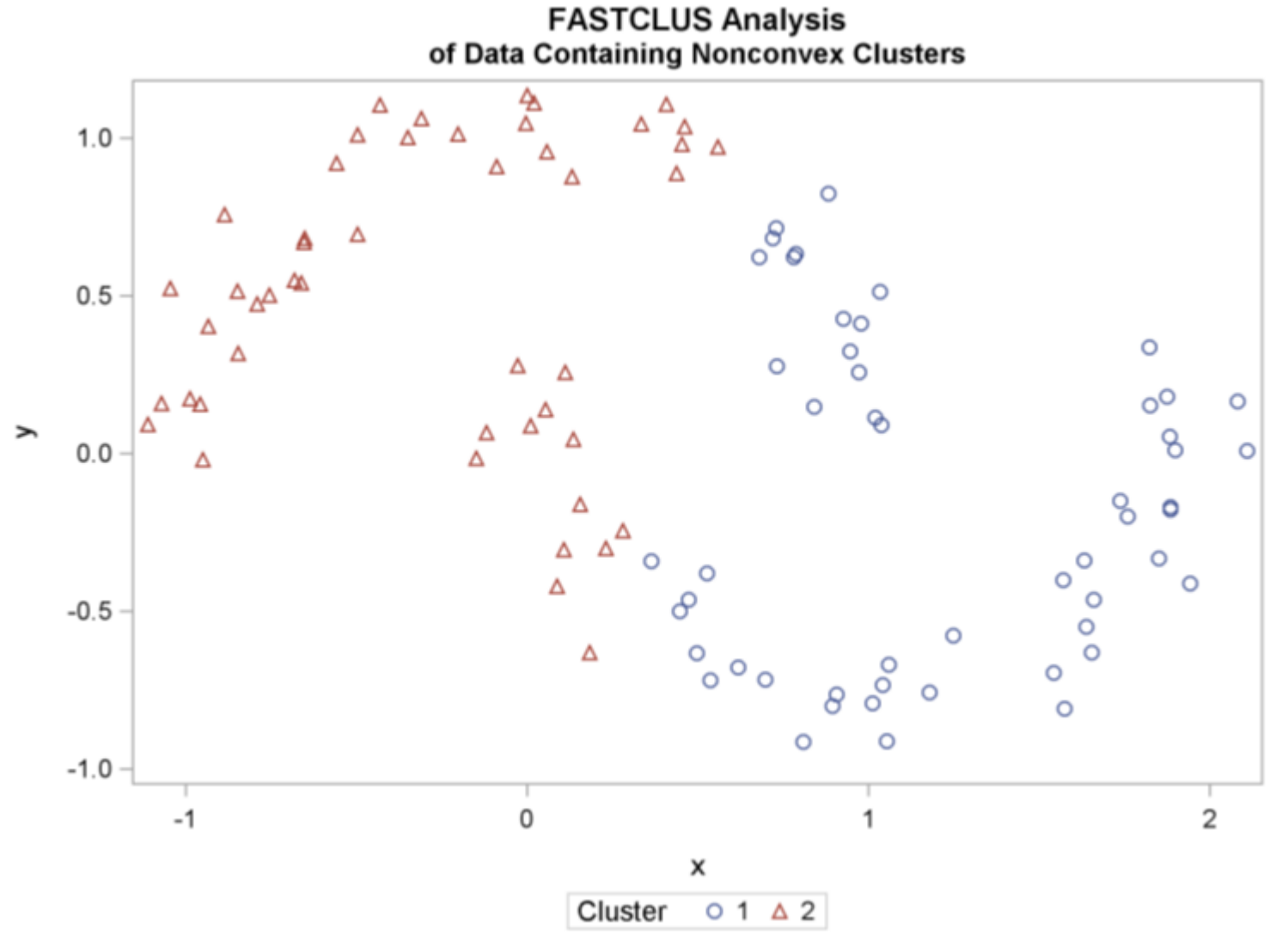



Figure 8:

# И снова неудача

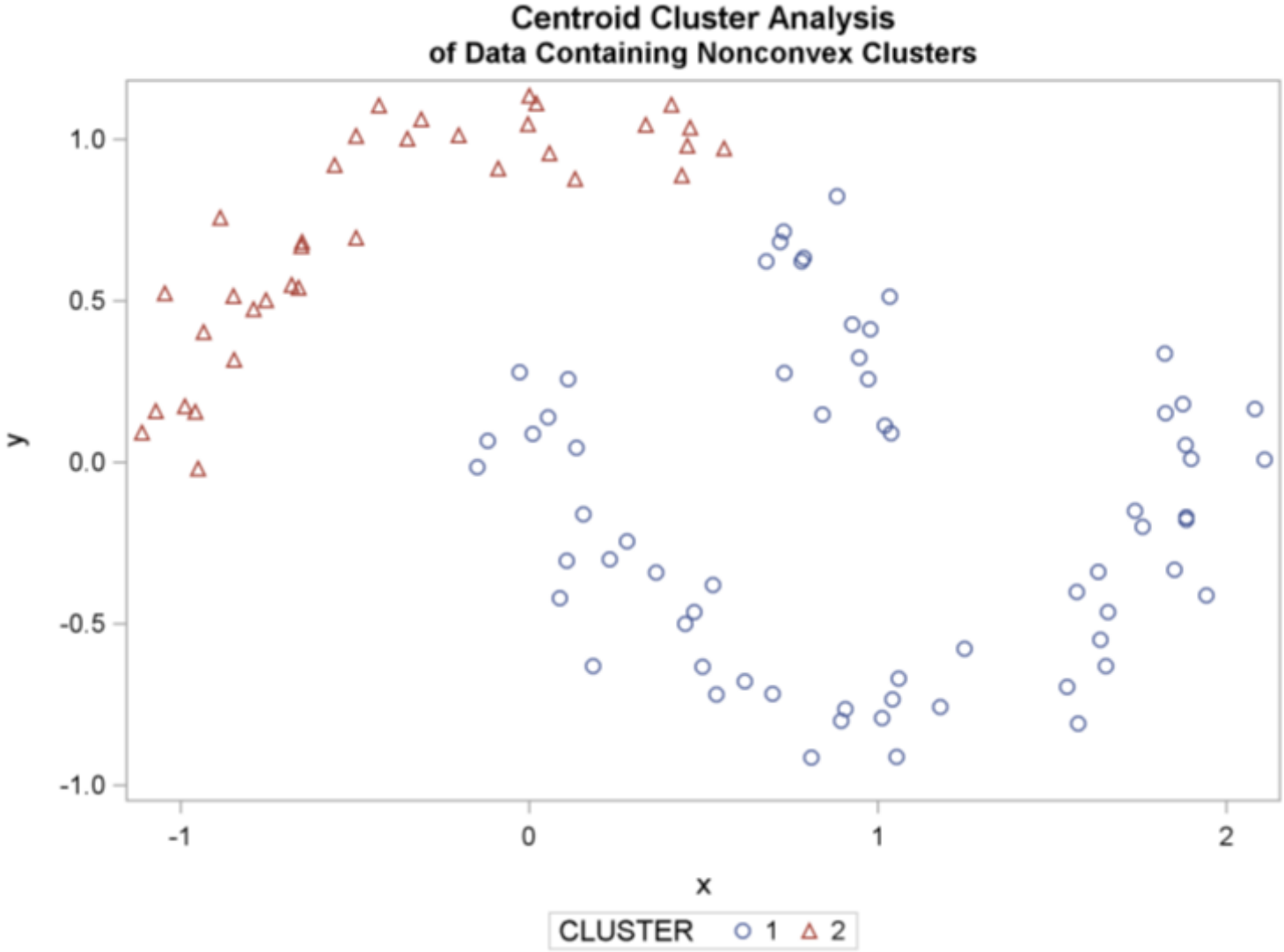



Figure 9:

# Метод ближайшего соседа

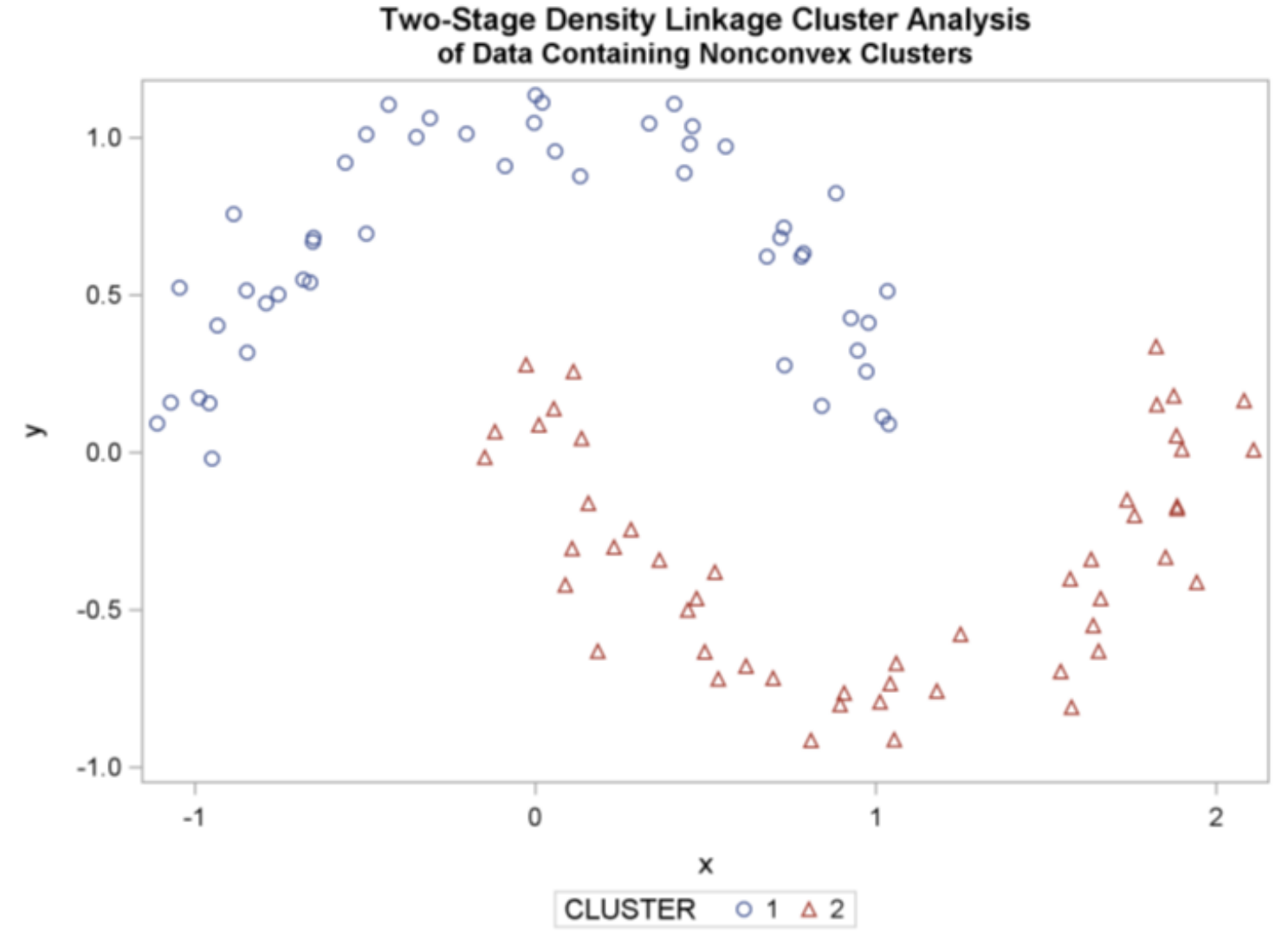



Figure 10:

# Бывает и так. . .

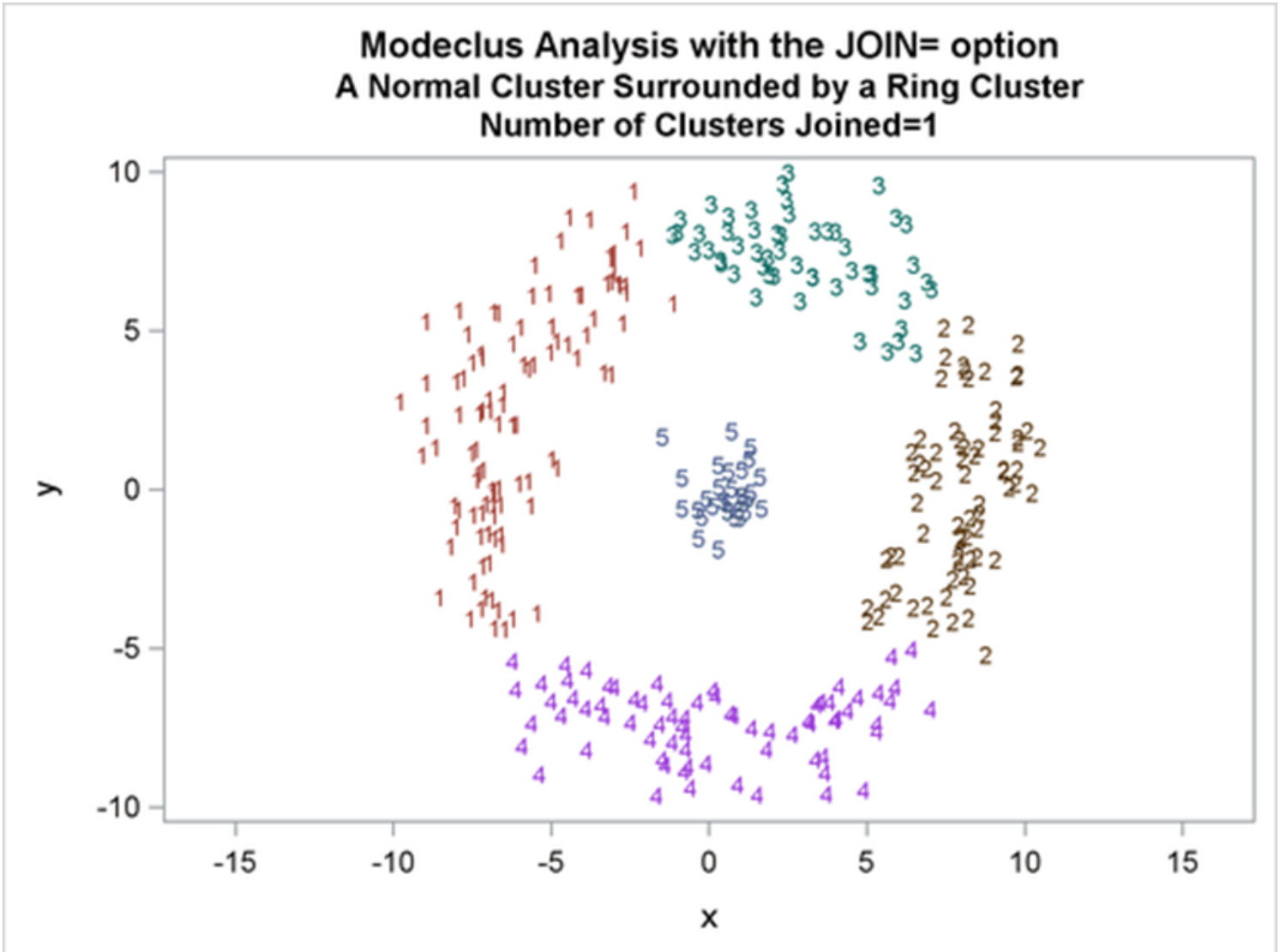



Figure 11: Спасти может расстояние ближайшего соседа или осмысление данных: может быть следует выполнить преобразование координат?

# Начинающим рекомендуем

- метод Варда (Ward's method) — плотные шаровидные скопления;
- метод ближайшего соседа (Single linkage clustering) — ленточные кластеры;
- среднее невзвешенное расстояние (Average linkage clustering) — плотные шаровидные скопления.

# Общие замечания о кластерном анализе

# Участие аналитика

1. Отбор переменных
2. Метод стандартизации
3. Расстояние между кластерами
4. Расстояние между объектами

# 1. Какие переменные будут использоваться при анализе

- Все? Избыточные переменные могут перетащить одеяло информативности на себя.
- Как влияет цвет глаз покупателя на средний объем выпиваемого пива?
- Распознавание танков.

# С другой стороны

Некоторые переменные очень важны, но данные о них нам не удается получить. Тогда информацию о них можно восстановить по другим переменным.

- Если нам неизвестны зарплаты/доходы покупателей, но для каждого из них известны профессия, образование и стаж работы, исключение этих трех переменных влечет за собой исключение из рассмотрения платежеспособности покупателей.
- Если классифицируются школы, и не включены ни переменная «число школьников», ни переменная «число учителей», то кластеры будут формироваться без учета размера школ.

# Вывод

- Правильный выбор переменных очень важен.
- Критерием при отборе переменных для анализа является в первую очередь ясность интерпретации полученного результата, во вторую – здравый смысл и интуиция исследователя.

# 2. Стандартизация переменных

Надо ли стандартизировать переменные? Правило для новичка:

- Если Вы не знаете, стандартизировать или нет, то стандартизируйте.

**Нужно стандартизировать:**

| | | |
|---|---|---|
| 5296782.7 | 0.5 | 1 |
| 7400381.4 | 0.7 | 0 |
| 9362870.2 | 0.1 | 0 |
| 7594038.5 | 0.4 | 0 |
| 6455034.1 | 0.4 | 1 |

Figure 12:

# Способы стандартизации

Если данные в столбцах несоизмеримы, то они должны стать соизмеримыми.

1. Максимальное значение = 1, минимальное = 0 (-1)

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

2. Среднее равно 0, выборочная дисперсия равна 1

$$z_i = \frac{x_i - \bar{x}}{sd(x)}$$

# Если кластеров нет

то они все равно будут найдены

Figure 13:

Если кластерный анализ дал вам решение, то это вовсе не означает, что в данных есть кластеры.

# Результаты кластерного анализа нуждаются в интерпретации

**Вопрос:** Какой вариант кластеризации даст лучшие результаты?

**Ответ:** Тот, который вы смогли понять и проинтерпретировать.

Кластерный анализ нацелен на то чтобы **лучше понять** имеющиеся данные. Однако иногда удается лишь **сократить размерность** данных.

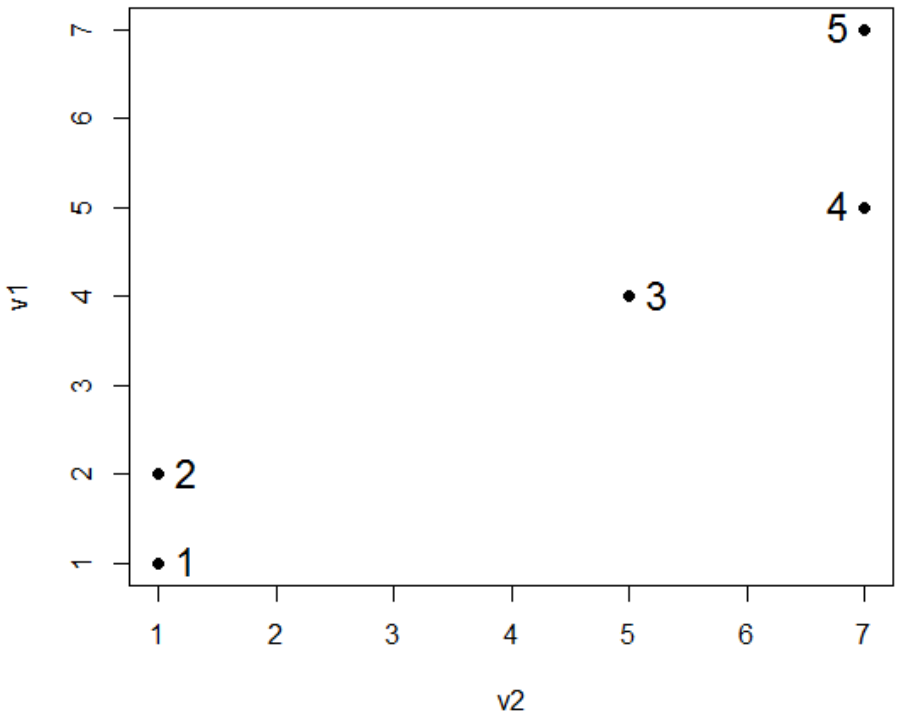
# Иерархический кластерный анализ

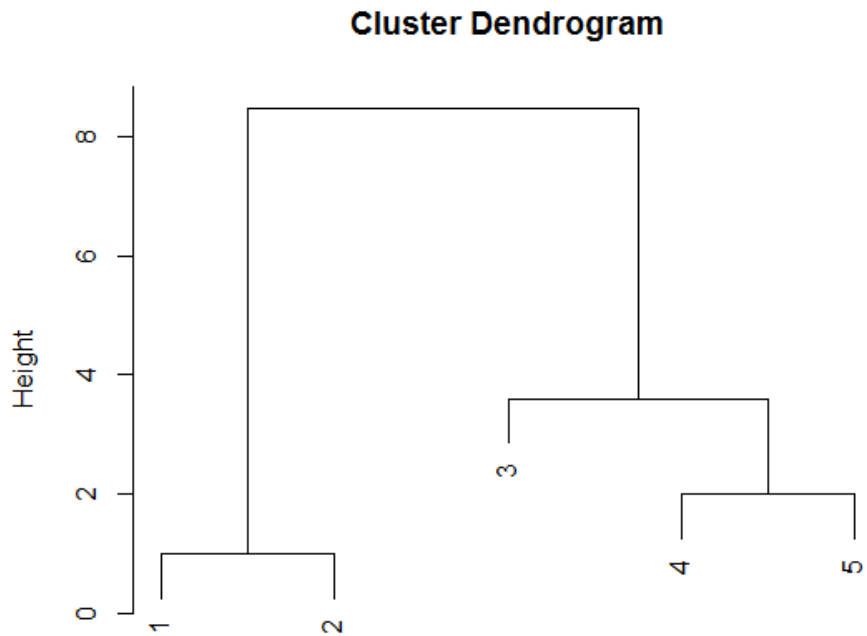
# Алгоритм иерархического кластерного анализа

1. Каждый объект — кластер.
2. Выбираем два кластера, расположенных ближе всего друг к другу и объединяем их.
3. Повторяем шаг 2. . . пока не объединим все кластеры.

Когда нужно остановить объединение? В этом поможет дендрограмма.

# Тестовые данные и дендрограмма



# Построение дендрограммы

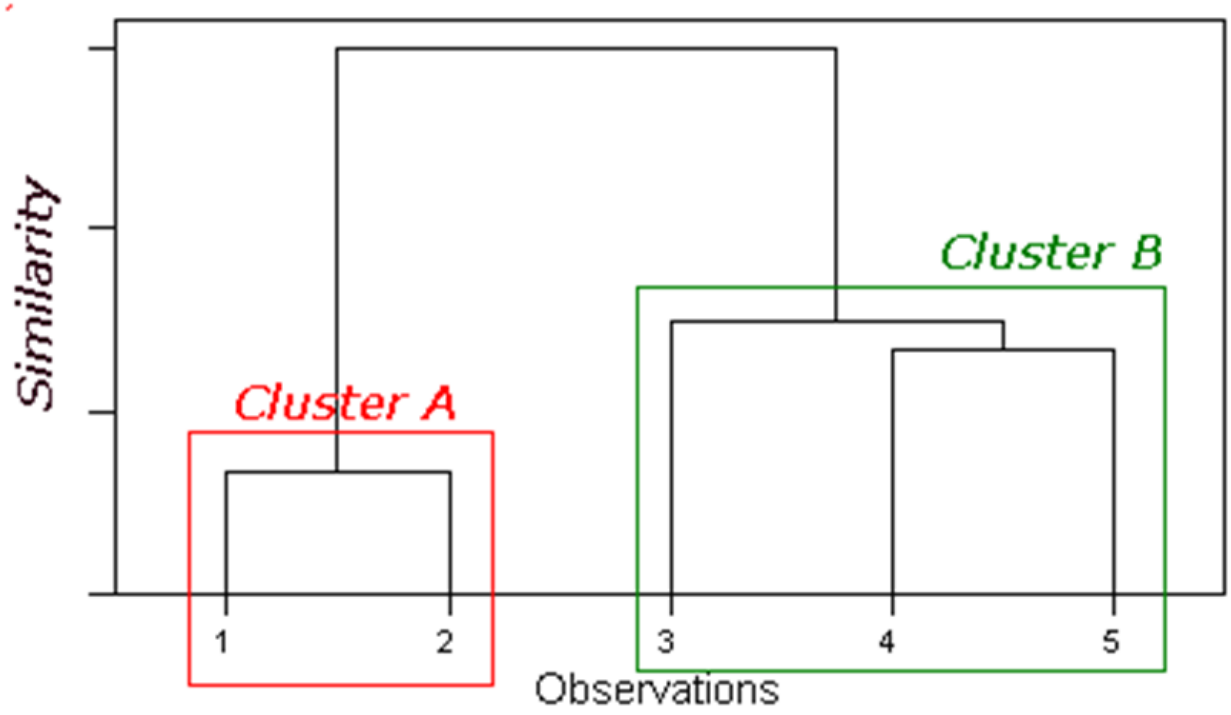



Figure 14: Дендрограмма описывает процесс объединения кластеров

1. Кластеру соответствует вертикальная линия.
2. Объединению кластеров соответствует горизонтальная линия.
3. Высота горизонтальной линии есть расстояние между кластерами в момент объединения.

# Где на дендрограмме кластеры?

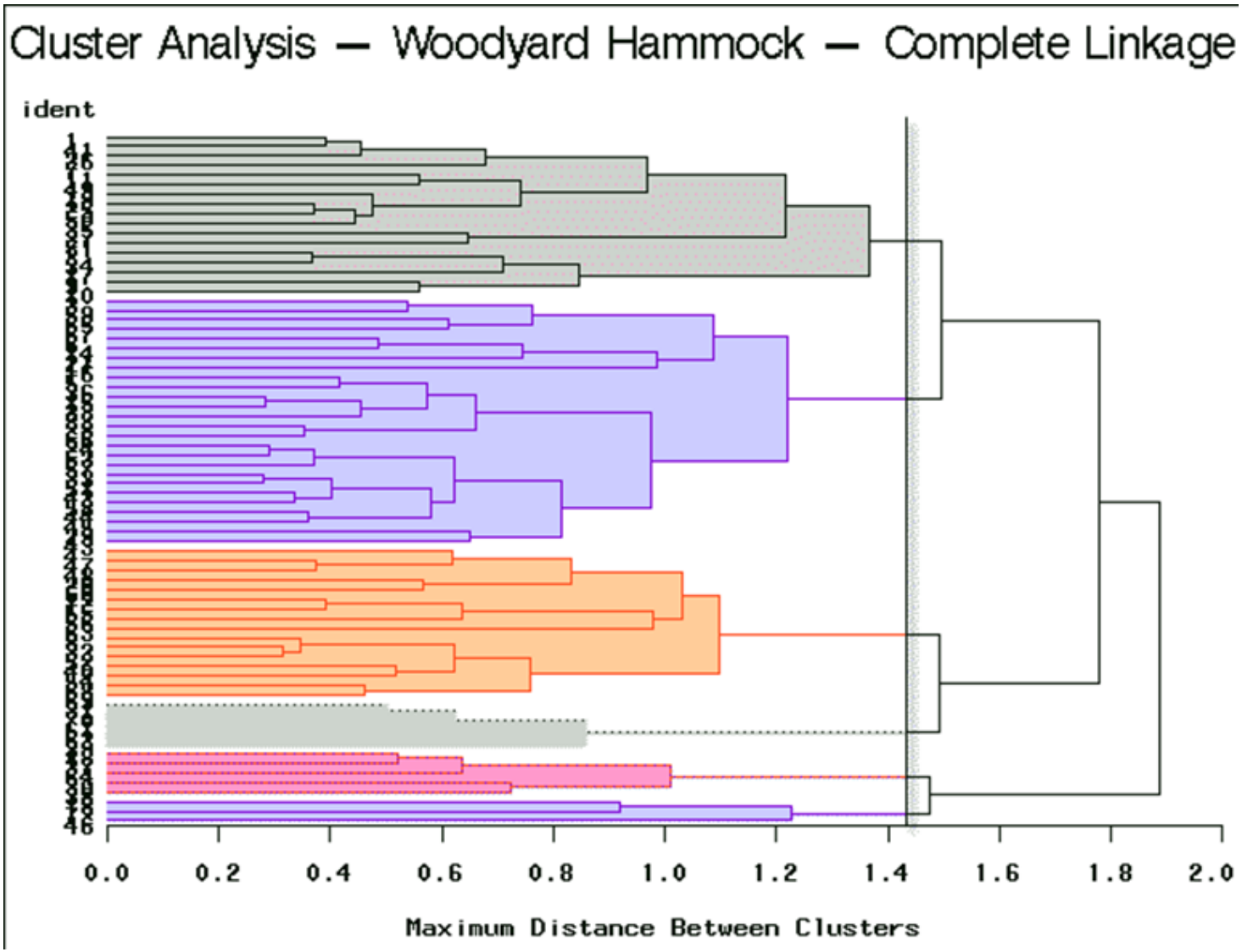



Figure 15:

# Где обрезать дендрограмму?

Поводом для обрезания может служить скачок в расстояниях между кластерами.

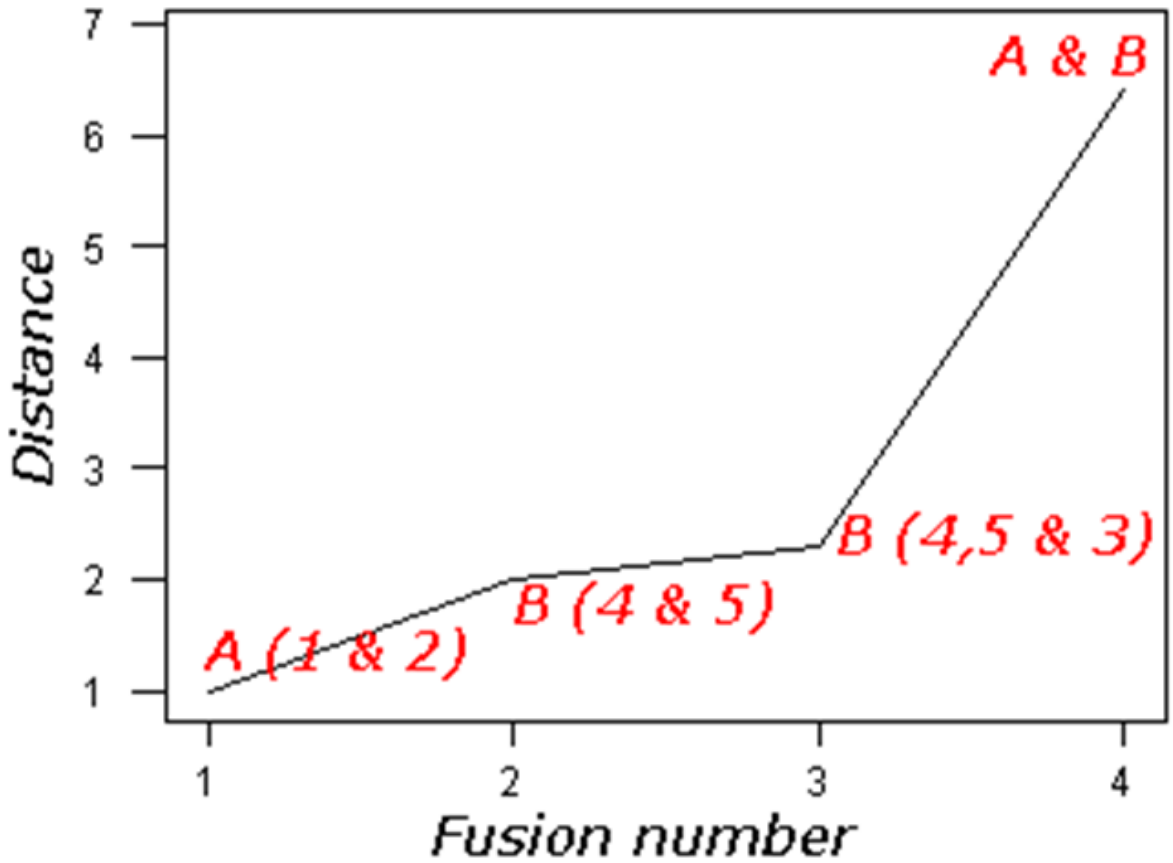
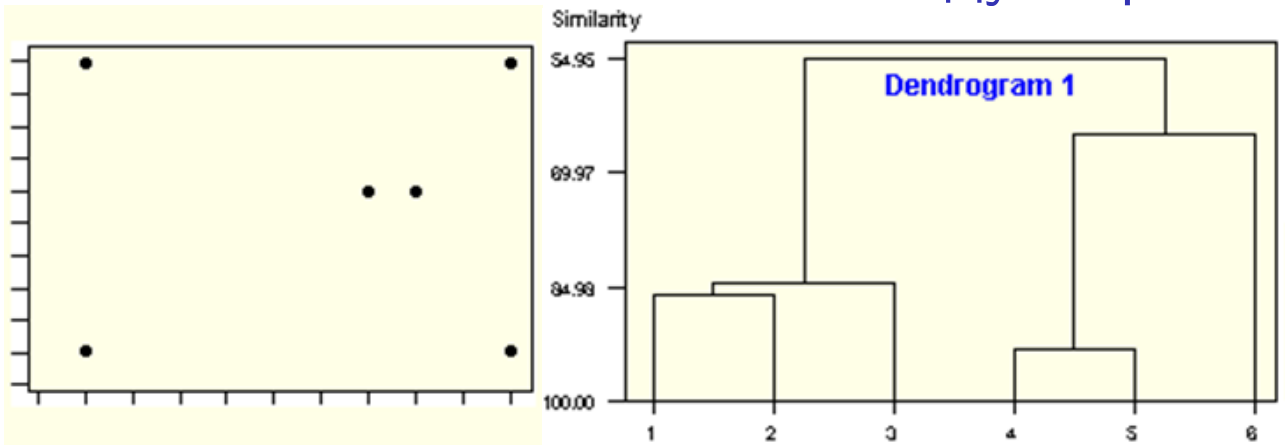
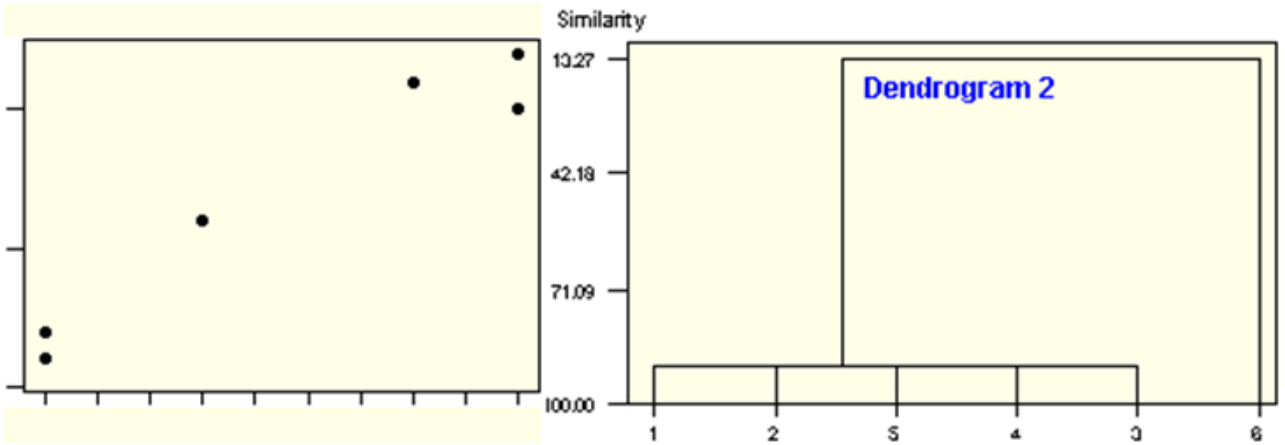
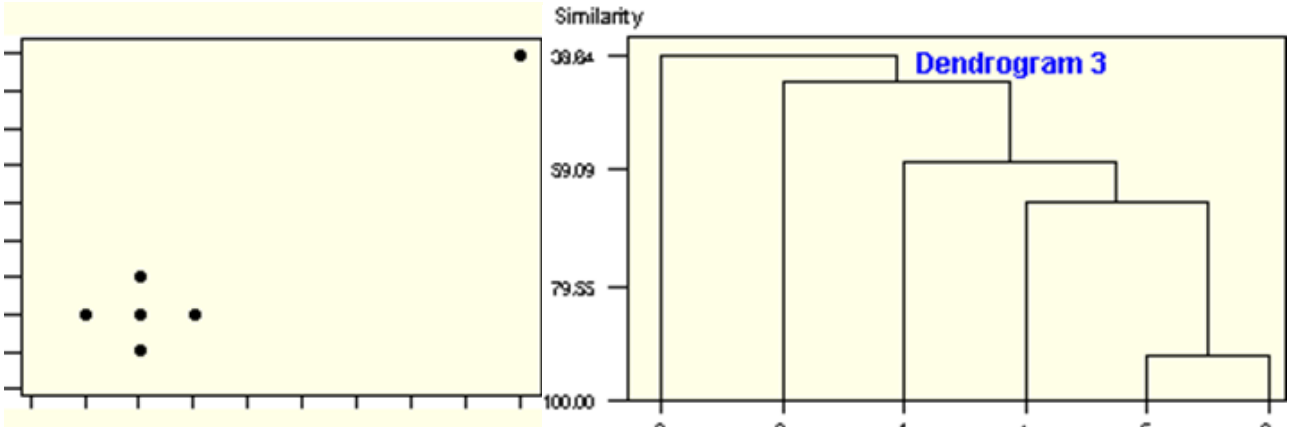



Figure 16: График «каменистая осыпь» (scree plot) / «локоть» (elbow)

# Упражнение: соответствие между парами



# Дополнительная информация

- *StatSoft. Электронный учебник по статистике. Кластерный анализ.*
  (http://statsoft.ru/home/textbook/modules/stcluan.html) — раздел, посвященный кластерному анализу из учебника по статистике от компании StatSoft, разработчика пакета STATISTICA. Внимания достоин весь учебник.