

Иерархический кластерный анализ

Аббакумов

Вадим Леонардович

Версия 08.2020

Кластерный анализ

- Кластерный анализ разделяет объекты на группы
 - объекты внутри группы похожи между собой
 - объекты из разных групп различаются
- Определение количества групп — часть задачи
- Группы, на которые разбита выборка, называются кластерами.

Происхождение термина

- Кластер – калька слова «cluster», «сгусток», «гроздь (винограда)», «скопление (звезд)» и т.п.

Кластерный анализ раньше назывался

- распознавание образов без учителя,
- стратификация,
- таксономия,
- автоматическая классификация.

Методы кластеризации

- **Метод к-средних**
- Самоорганизующиеся карты Кохонена (SOM)
- Смесь (нормальных) распределений
- Методы, использующие непараметрические оценки плотности
- **DBSCAN** и **OPTICS**
- **Иерархический кластерный анализ**
- Разрезание минимального остовного дерева (Minimum spanning tree)

Использование кластерного анализа

- Сегментация
- Сокращение размерности (вместо объектов кластера берется их типичный представитель)
- Выбросы

Пример:

Определение групп потребителей

- По данным о покупателях (результаты опроса, поведение на сайте)
 - выявить и описать/понять рыночные сегменты
 - выбрать сегменты, которые компания хочет обслуживать.
-

Сегментация

- Страховая компания разделяет потенциальных клиентов на группы
- По результатам кластеризации для разных сегментов определяются оптимальные цены на услуги, оптимальные тарифы
- Но ближайшая цель — индивидуальные тарифы
(«Вы не принимаете лекарства»)

Пример:

Потребительские сегменты

Покупателей можно группировать по разным наборам характеристик, например по возрасту, образованию, месту жительства, типу личности, и так далее.

Недостаточно разделить покупателей на сегменты по **одной** (или по каждой) характеристике.

Кластерный анализ разбивает потребителей на «*группы со схожими потребностями в отношении конкретного товара или услуги, достаточными ресурсами, а также готовностью и возможностью покупать*» учитывая все информативные показатели **одновременно**.

Пример: 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. Например, компьютеры бывают «для дома», «для офиса», «серверы» и «специализированные».

Кластерный анализ применяется, если нет классификации, признанной всеми.

Результат будет зависеть от выбранных показателей.

Пример

- Определение целевой аудитории баннерной рекламной компании в интернете.
- 100000 сайтов, 60+М кук
- Каждое посещение кукой сайта описывает ее интересы, ее текущее настроение...
- Надо отождествить схожие сайты
- Примеры кластеров

Другая задача классификации

- Machine Learning
- Классификация с учителем
- Распознавание образов

Отличие

- Заранее известно, к какому классу принадлежит каждое из наблюдений.
- группирующая переменная
- класс

Что тогда классифицировать?

- Надо придумать правило.
- Для классификации новых наблюдений.
- $y = f(x)$
- $p = f(x)$

Методы распознавания

- деревья классификации
 - К-го ближайшего соседа
 - Нейронная сеть / Deep Learning
 - SVM
 - Случайный лес
 - Gradient boosting machine и XGBoost
 - наивный байесовский классификатор
 - дискриминантный анализ
 - Логистическая регрессия

-
- Во избежание путаницы, я буду избегать термин классификация
 - кластеризация
 - распознавание
-

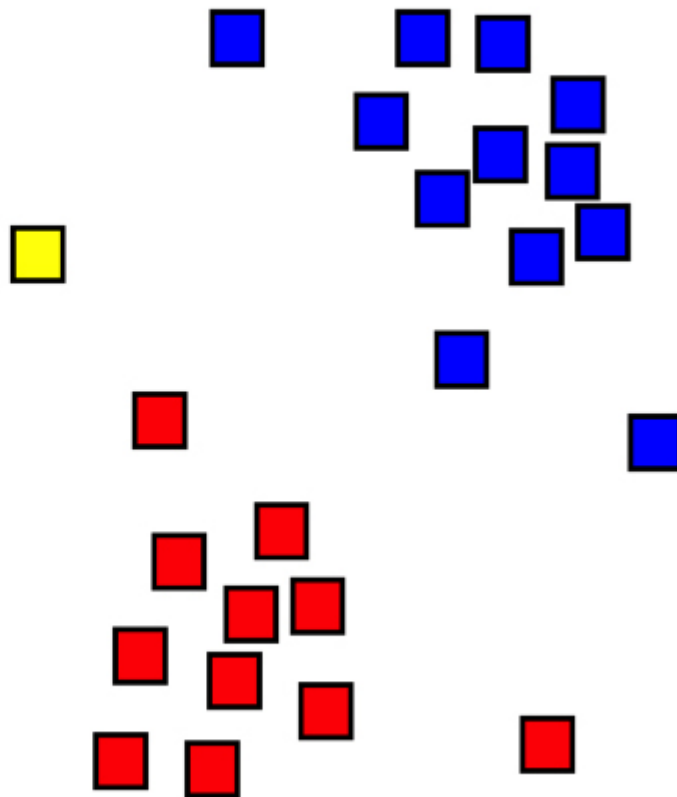
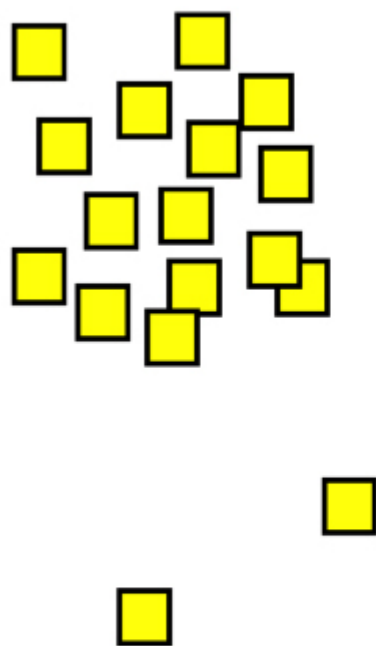
Иерархический кластерный анализ

- заранее неизвестно число кластеров (групп, на которые разбивается набор объектов).

Идея метода:

сведем задачу к геометрической

- Каждый объект – точка в R^k .
- Похожие объекты расположены «близко» друг к другу
- Различающиеся объекты расположены «далеко»
- Скопления точек – кластер.



-
- Если кластер содержит единственную точку, то эта точка — выброс
 - распознаем породы собак
 - фотография слоненка
-

-
- Разные методы дают разные кластеризации.
 - Если все методы дают одну и ту же кластеризацию?
-

Расстояние между объектами

- Евклидово расстояние
 - Квадрат Евклидова расстояния
 - Блок (Манхеттен, сити-блок)
 - и так далее...
-

Расстояние Евклида

- *Две точки*

(x_1, x_2, x_3)

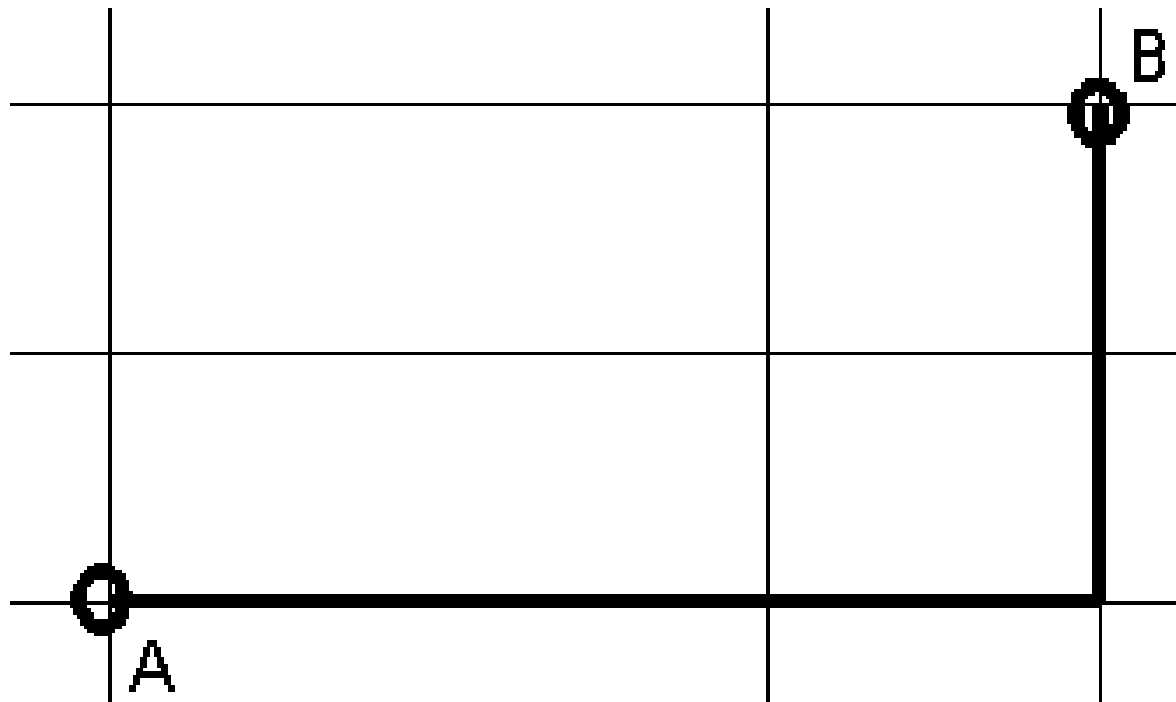
(y_1, y_2, y_3)

$$d_{xy} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Квадрат евклидова расстояния

не является расстоянием...

Расстояние Block
(Manhattan, таксиста).



Расстояние Block

(Manhattan, таксиста, Минковского при $p=1$).

$$X = (x_1, x_2, \dots, x_k)$$

$$Y = (y_1, y_2, \dots, y_k)$$

$$d_{XY} = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_k - y_k|$$

Расстояние Хэмминга

- Два слова одинаковой длины
- Расстояние - число позиций, в которых соответствующие символы различны
- $D(1011101, 1001001) =$
- $D(2173896, 2233796) =$
- $D(\text{toned}, \text{roses})$

Расстояние Sørensen–Dice

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

Мера Жаккарда (Жаккара, Джаккарда)

В 1901 году коэффициент флористической общности

$$Q = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Кластеризация текстов

- При анализе текстов (NLP: Natural language Processing) объектами могут быть книги, представленные множеством слов
- Нам надо измерять схожесть книг.
-
- Сначала упрощаем каждую книгу.
- Преобразуем: все существительные в именительном падеже, единственном числе, а все глаголы в неопределенной форме.
- Затем отбрасываем малоинформативные слова, например "и", "или", "этот"
- Как упрощать тексты это отдельная непростая наука, но идея, я надеюсь понятна
- После упрощений мы можем измерить схожесть двух книг, посчитав меру Жаккарда

Вопрос:

- Когда использовать евклидово расстояние, а когда расстояние Манхэттен?
-

Большое различие по одной координате

$$X = (1, 5, 6, 4)$$

$$Y = (11, 10, 7, 3)$$

$$\text{Dist}_M = 10 + 5 + 1 + 1$$

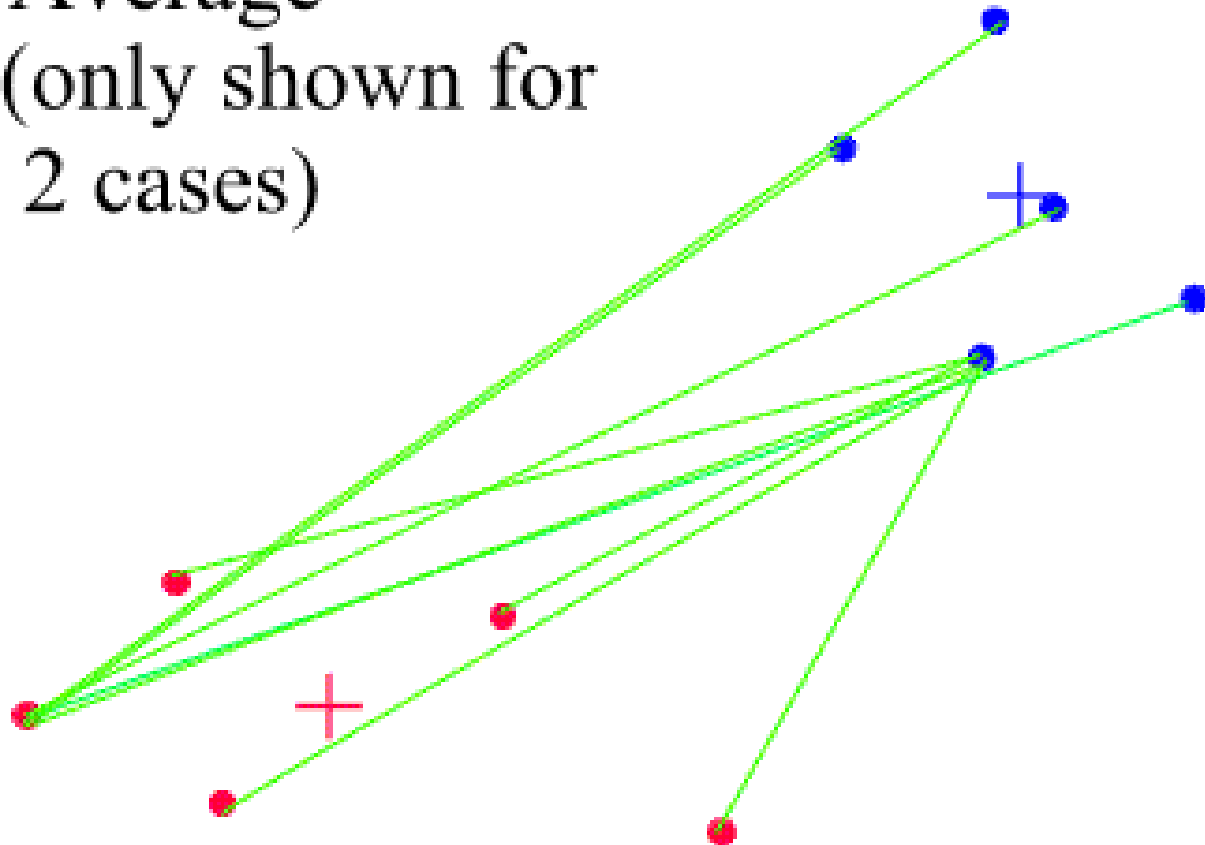
$$\text{Dist}_E = \text{sqrt}(100 + 25 + 1 + 1)$$

Расстояние между кластерами

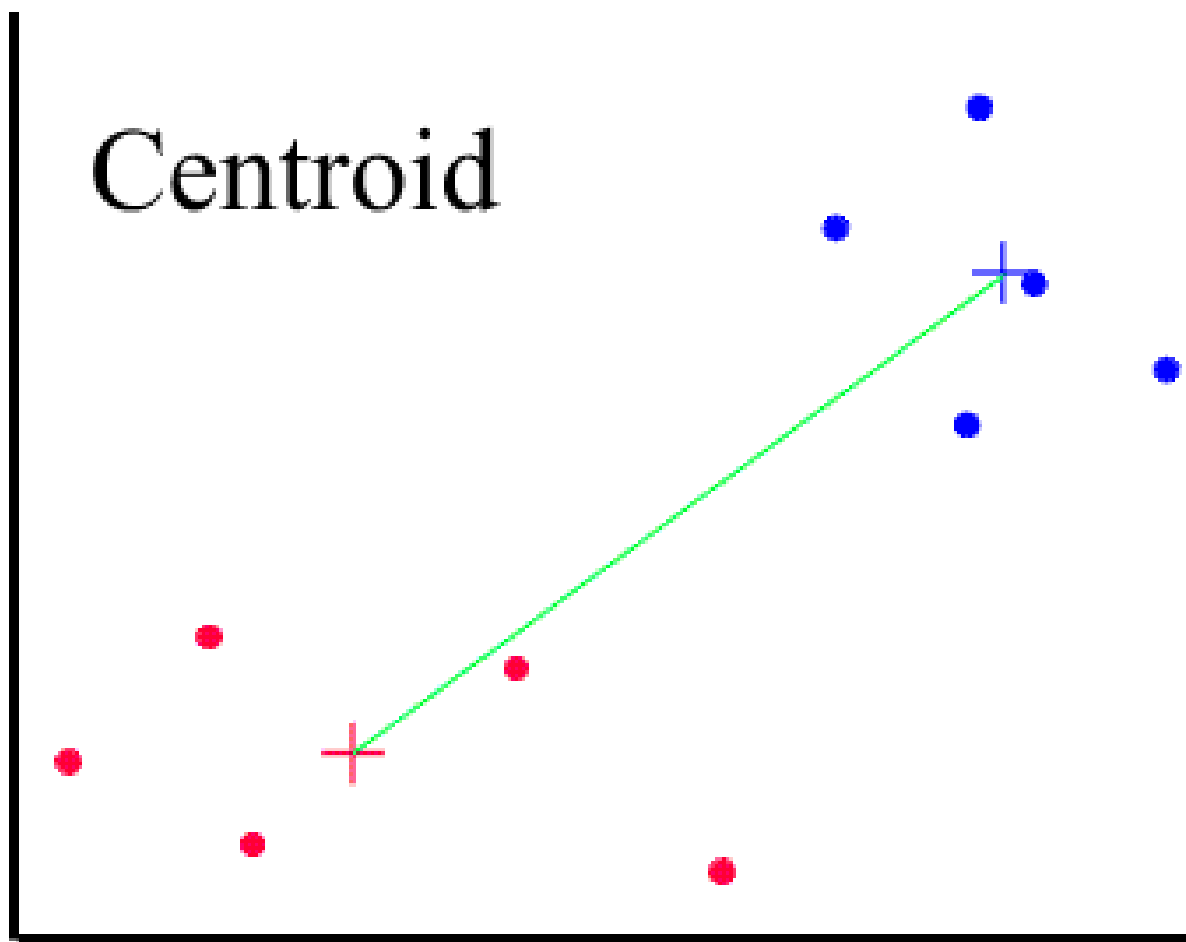
- Среднее невзвешенное расстояние (Average linkage clustering).
- Центроидный метод (Centroid Method).
- Метод дальнего соседа, максимального расстояния (Complete linkage clustering).
- Метод ближайшего соседа (Single linkage clustering).
- Метод Варда (Ward's method).

Среднее невзвешенное расстояние

Average
(only shown for
2 cases)



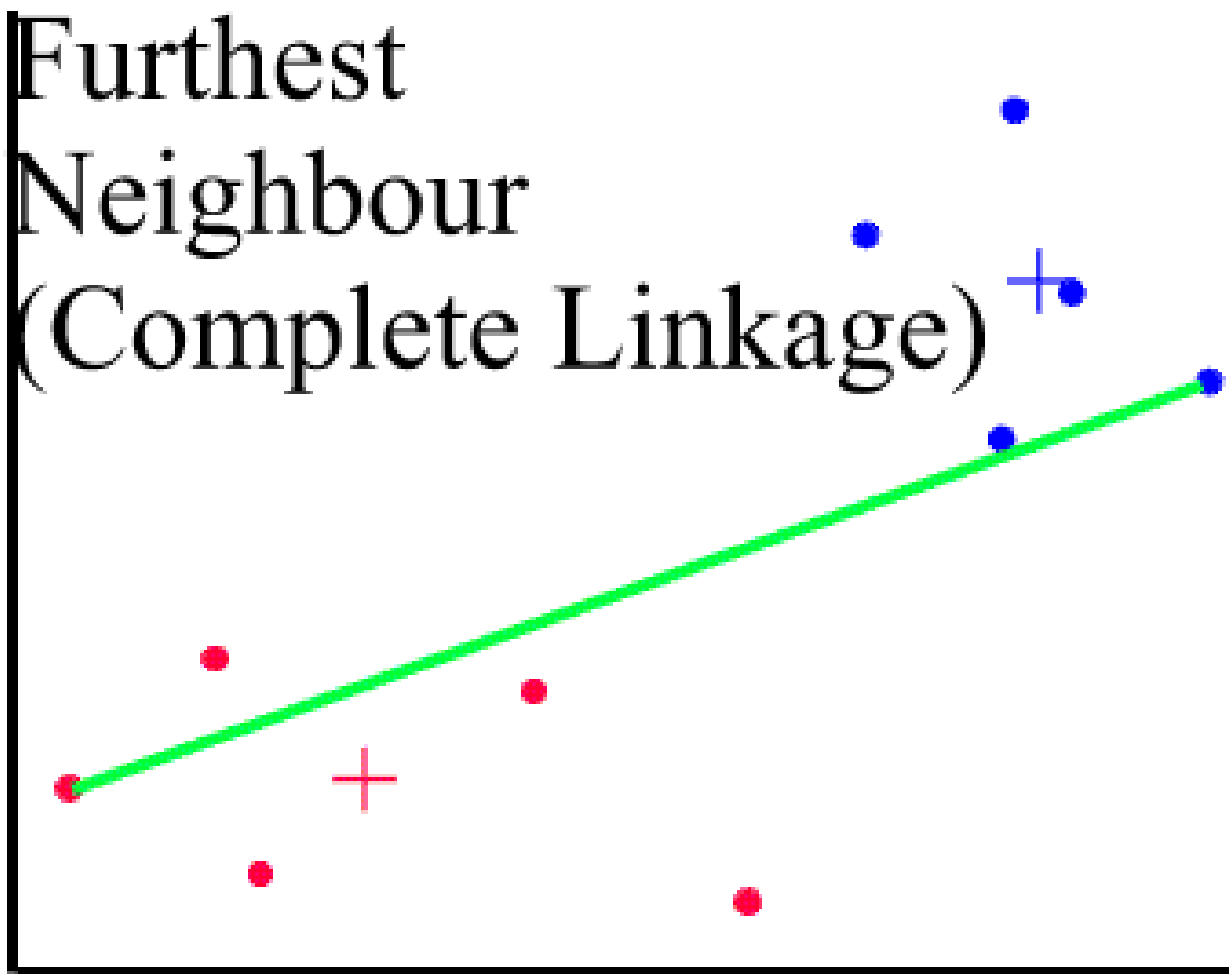
Центроидный метод



Центроидный метод

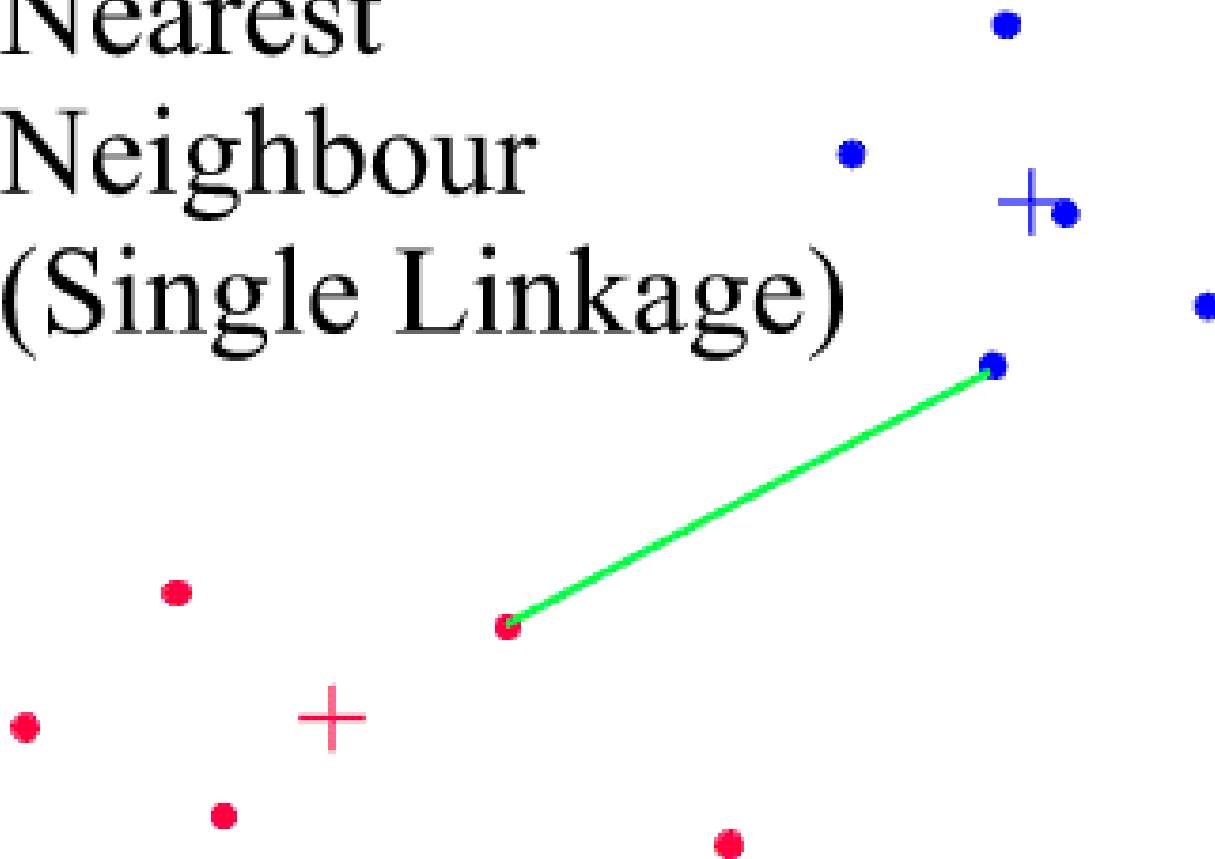
- Вычислительно прост
 - Объем кластера не влияет
 - Дендрограмма может иметь самопересечения
 - Выходит из употребления
-

Метод дальнего соседа



Метод ближнего соседа

Nearest
Neighbour
(Single Linkage)



Метод Варда (WARD)

- Предполагается использование квадрата евклидова расстояния, но это требование нарушается (в *Python* не реализовано)
- Выявляет шаровые скопления
- Объяснение в следующей теме «метод k-средних»

Начинающим рекомендую

- – метод Варда - шаровые скопления;
- – метод ближнего соседа (Complete linkage clustering) — ленточные кластеры;
- – среднее невзвешенное расстояние (Average linkage clustering) - шаровые скопления.

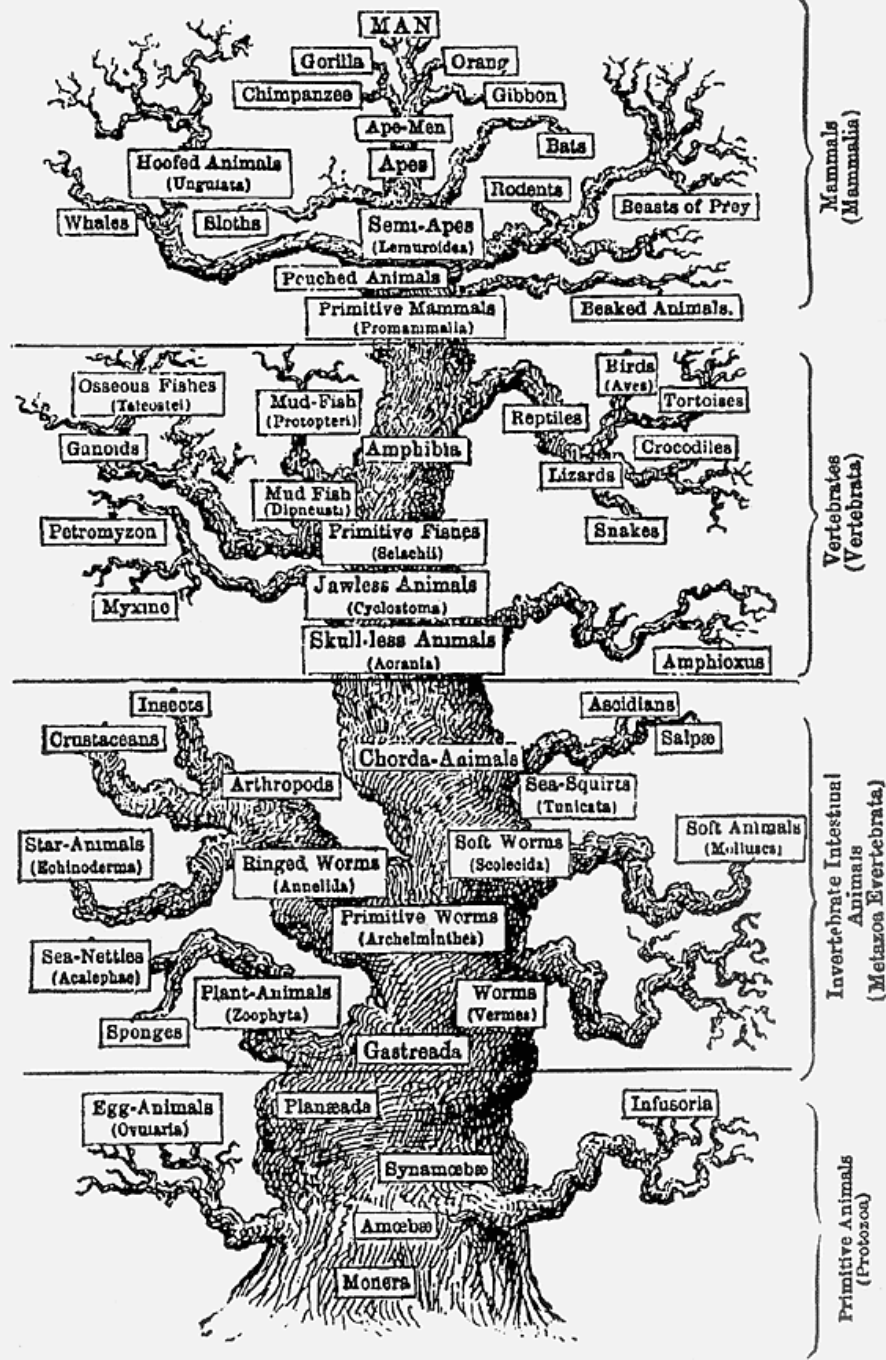
Алгоритм кластерного анализа

- Разберемся с процедурой иерархического кластерного анализа на примере

Алгоритм построения дендрограммы

Деревья — древний инструмент

- Ernst Haeckel
 - Tree of Life
 - The Evolution of Man (1879)
 -
 - Но он не был первым...
 - Древо Порфирия (300+ год)
-



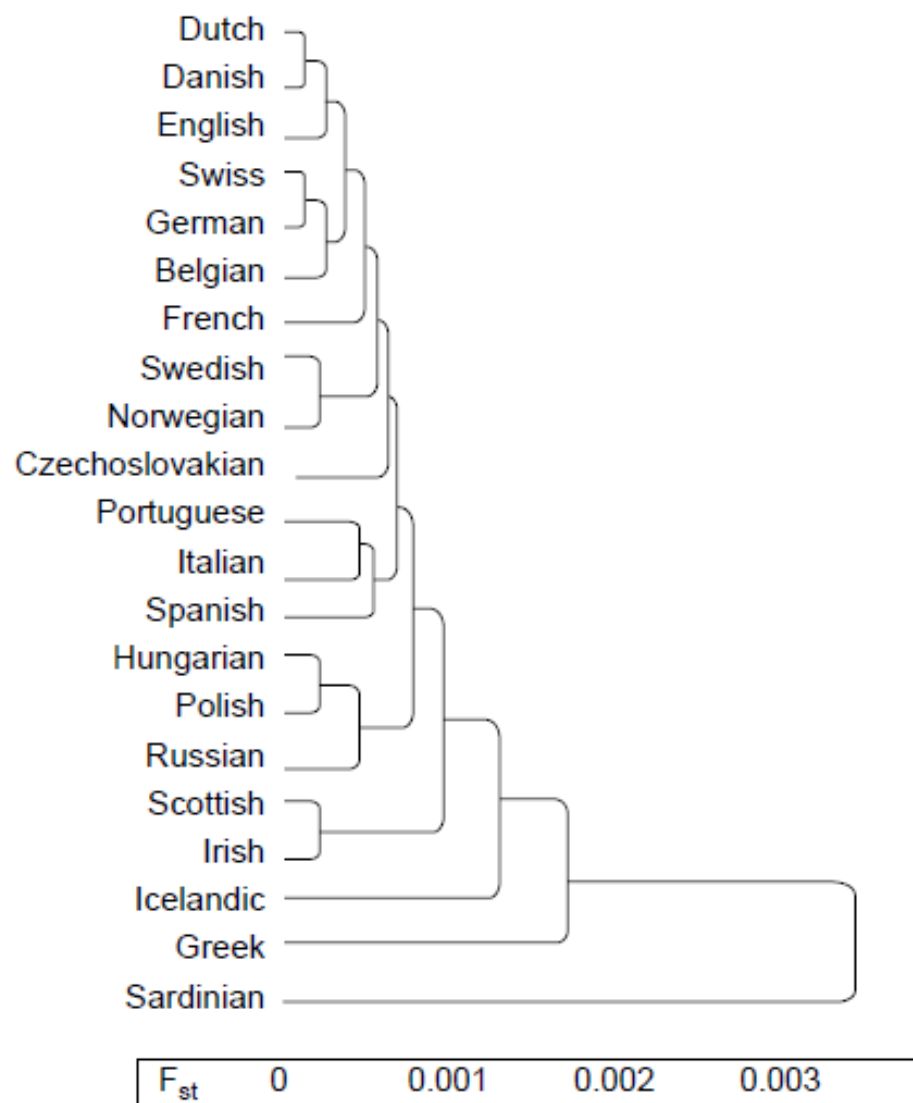
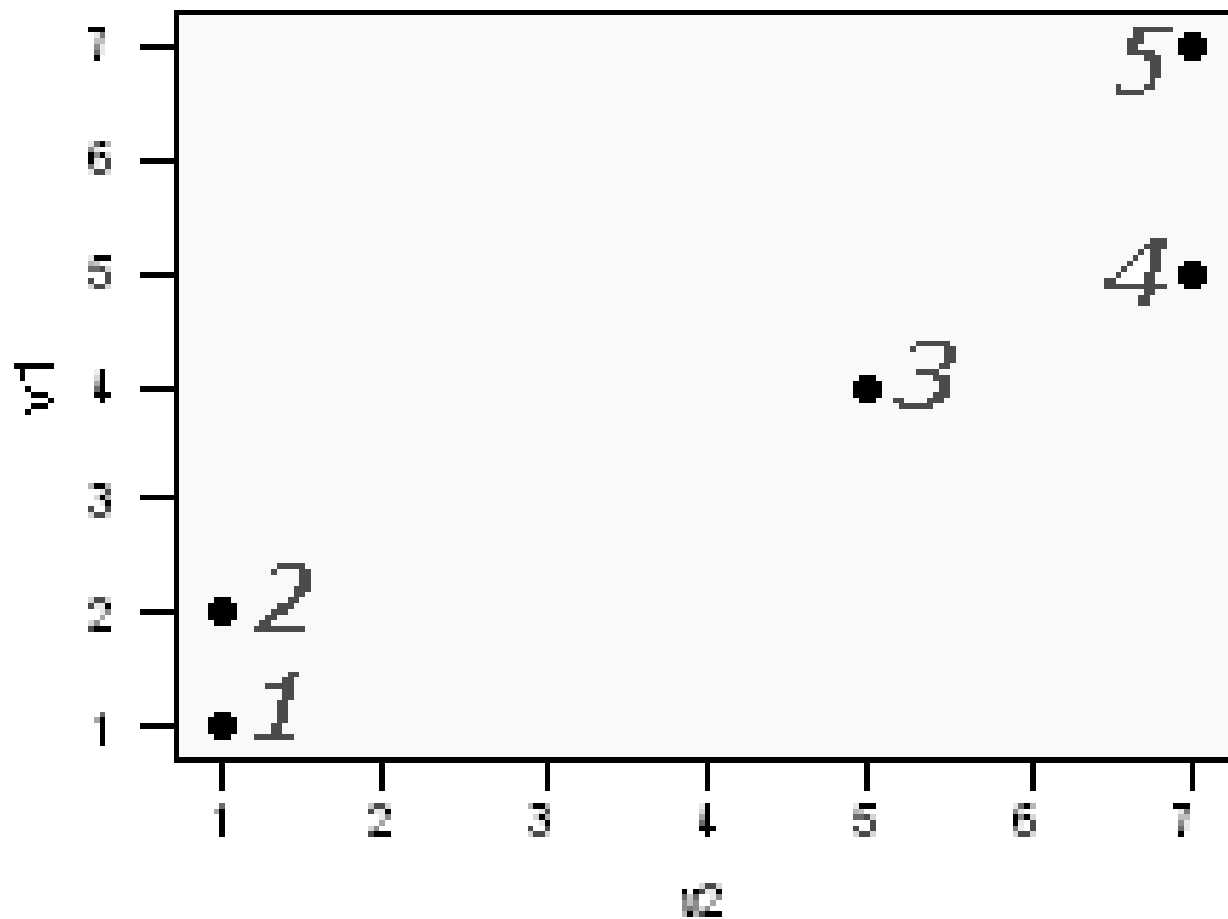
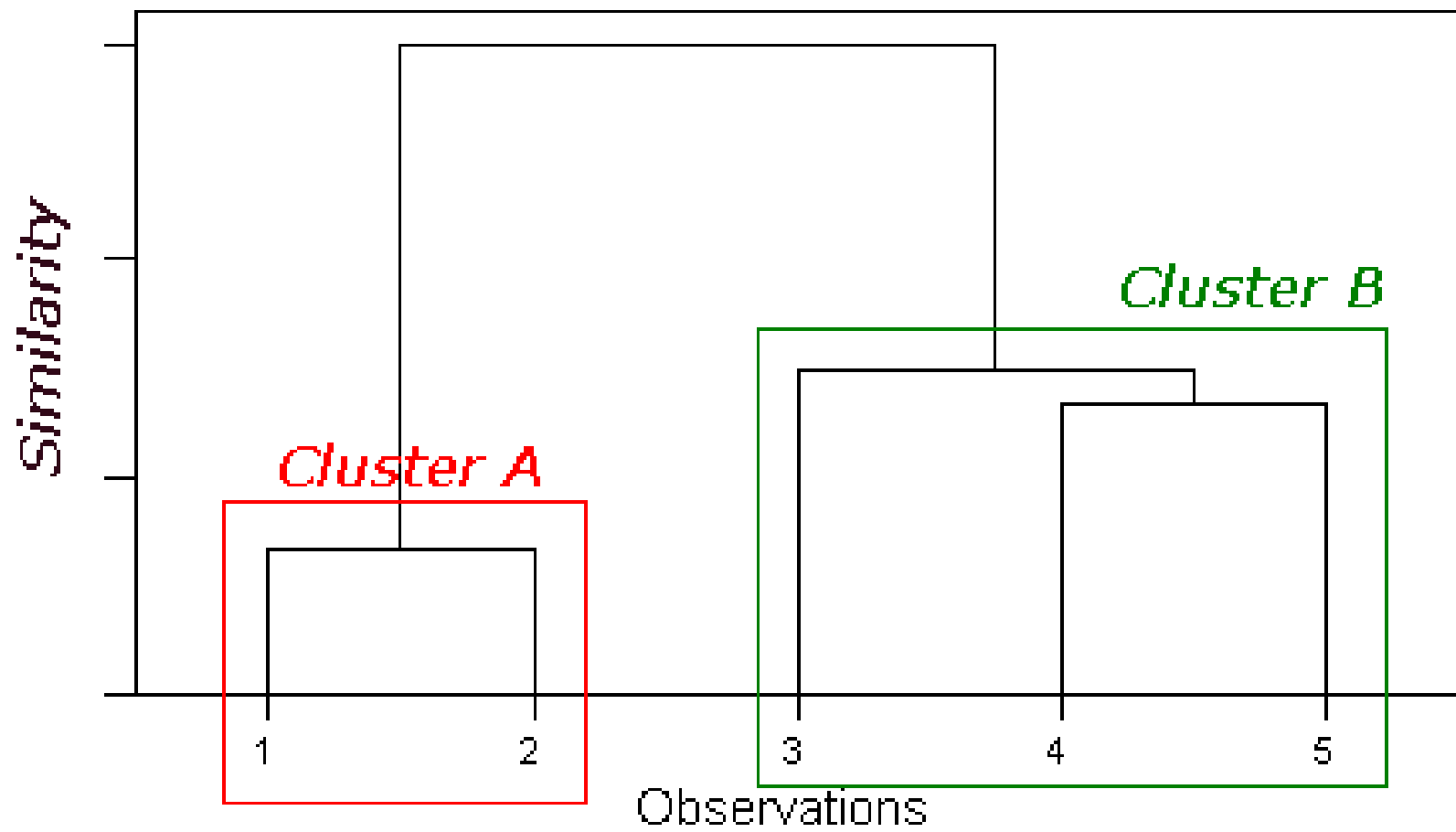
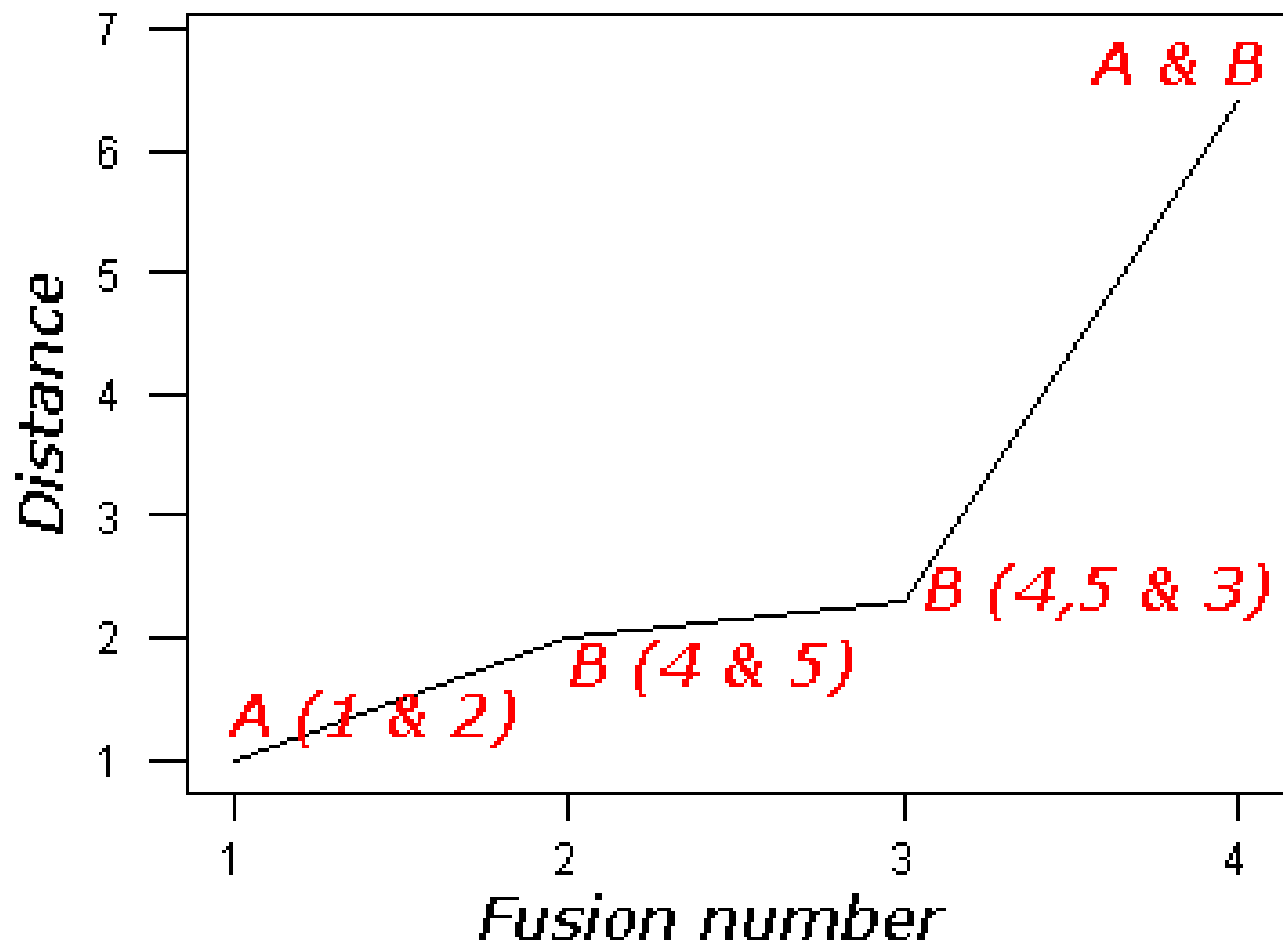


Fig. 5.2 Genetic tree of European populations from genetic distances ($= F_{st}$) between populations, based on 88 genetic polymorphisms from data in Cavalli-Sforza et al. (1994)



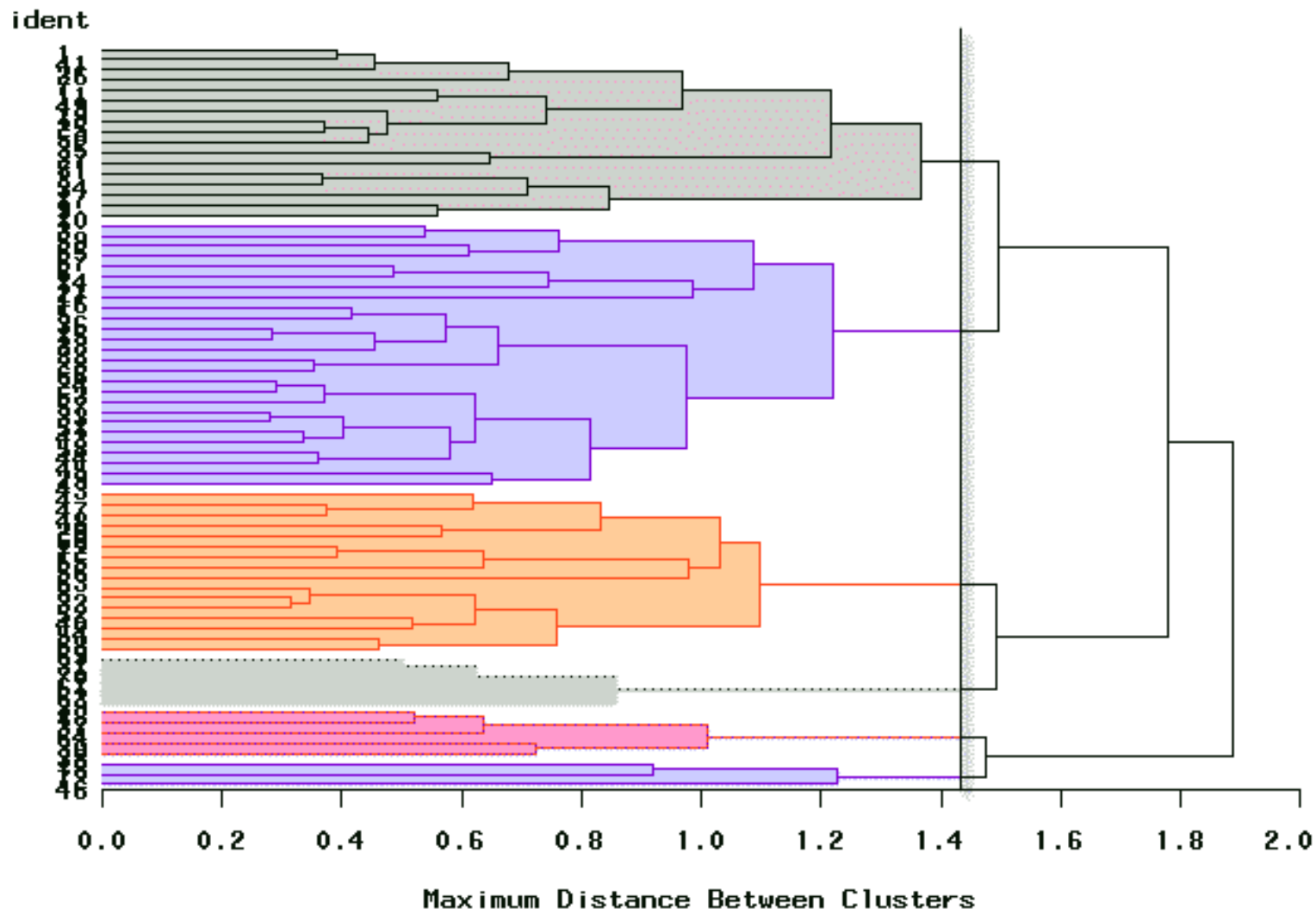


каменистая осыпь / ЛОКОТЬ



-
- Кеттел (Cattell) предложил график каменистая осыпь (scree plot) в 1966 году в факторном анализе
 - Потомки переименовали график в локоть (elbow plot)
-

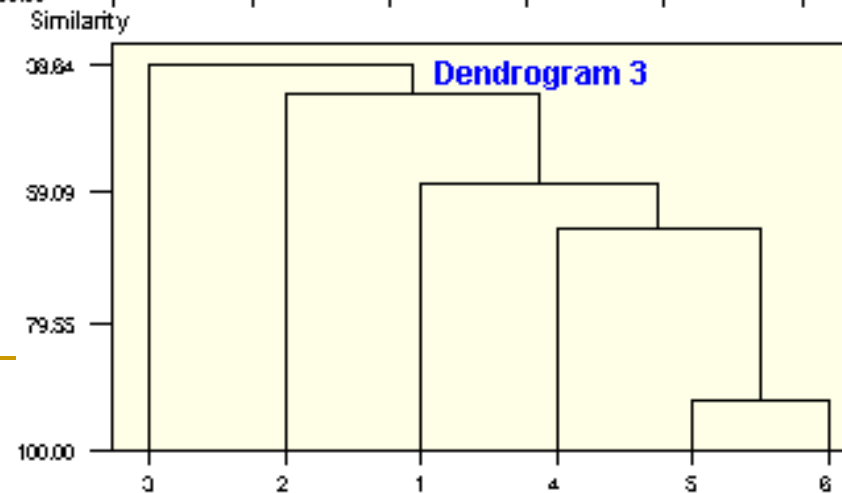
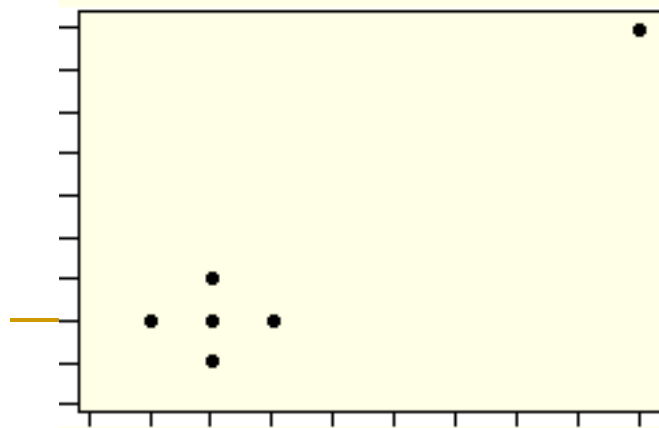
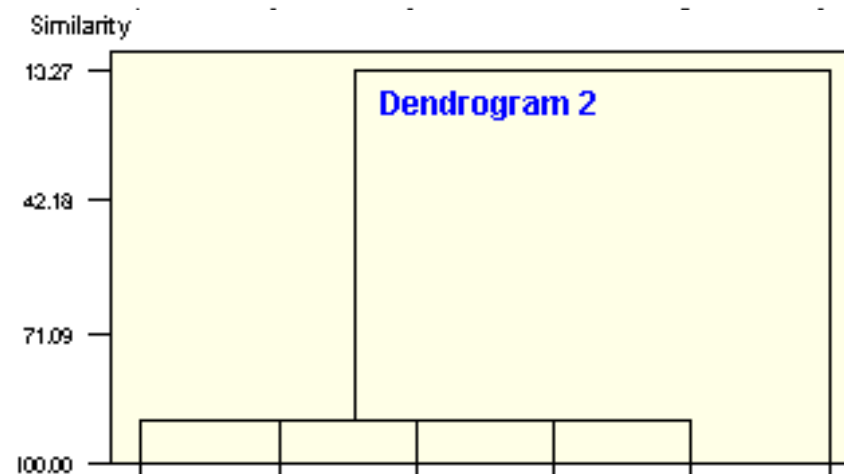
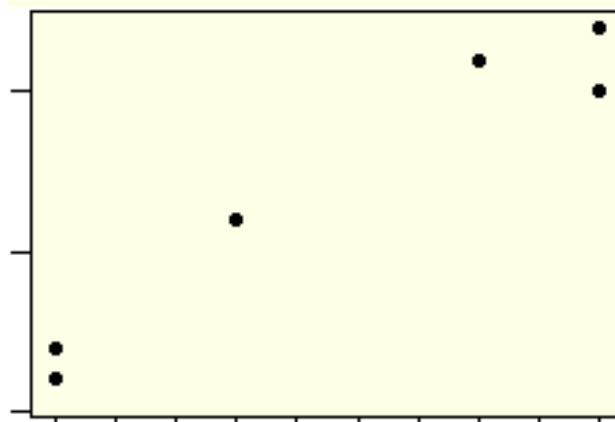
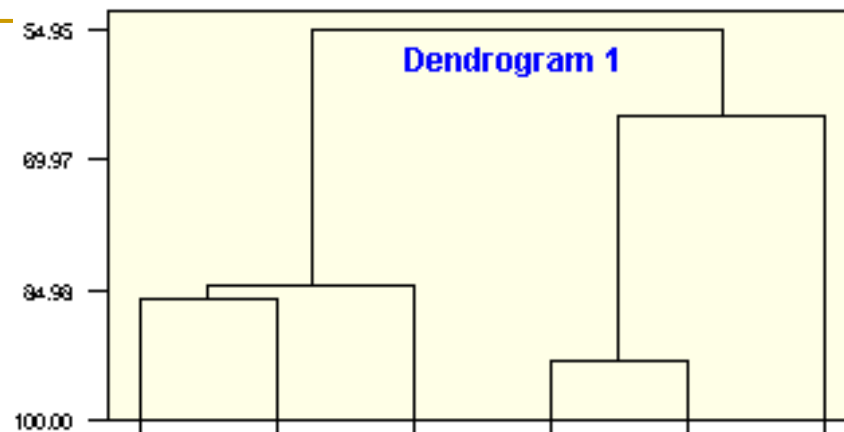
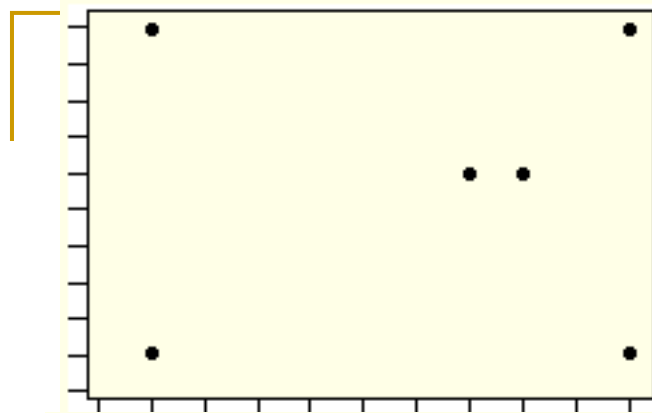
Cluster Analysis — Woodyard Hammock — Complete Linkage



-
- В многомерном случае диаграмму рассеивания не построить.
 - Вместо нее используем дендрограмму и каменистую осыпь
 - зачем нам каменистая осыпь, если у нас есть дендрограмма?
-

Упражнение

- Разбить на пары:
- Каждой диаграмме рассеивания поставить в соответствие дендрограмму



-
- Кластеризовали 100000 объектов, получили 21 кластер
 - Получим 20 кластеров, состоящих из одного наблюдения каждый. И один кластер, содержащий все остальные $(100000 - 20) = 99980$ наблюдений.
-

-
- Расстояние между кластерами по методу «центроид»
 - Почему на дендрограмме возможны самопересечения?
-

Участие аналитика

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

Отбор переменных

- 1. Какие переменные будут использоваться при анализе?
- Все?
- Три радиостанции одновременно...
- Как влияет цвет глаз покупателя на средний объем выпиваемого пива?
- Распознавание танков

С другой стороны

- если неизвестны доходы покупателей, но известны их профессия, образование и стаж работы, эти три переменные (surrogate variables) позволят распознать доходы.
- Перепись населения — разгадать доход
- Если классифицируются школы, и отсутствуют переменные «число школьников» и «число учителей», то размеры школ не будут учитываться при кластеризации .

Вывод

- Правильный выбор переменных очень важен.
 - Критерием при отборе переменных для анализа является
 - 1) ясность интерпретации результата,
 - 2) интуиция исследователя.
-

Надо ли стандартизировать переменные?

- Правило для новичка:
- если Вы не знаете, стандартизировать или нет, стандартизируйте.

Надо 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

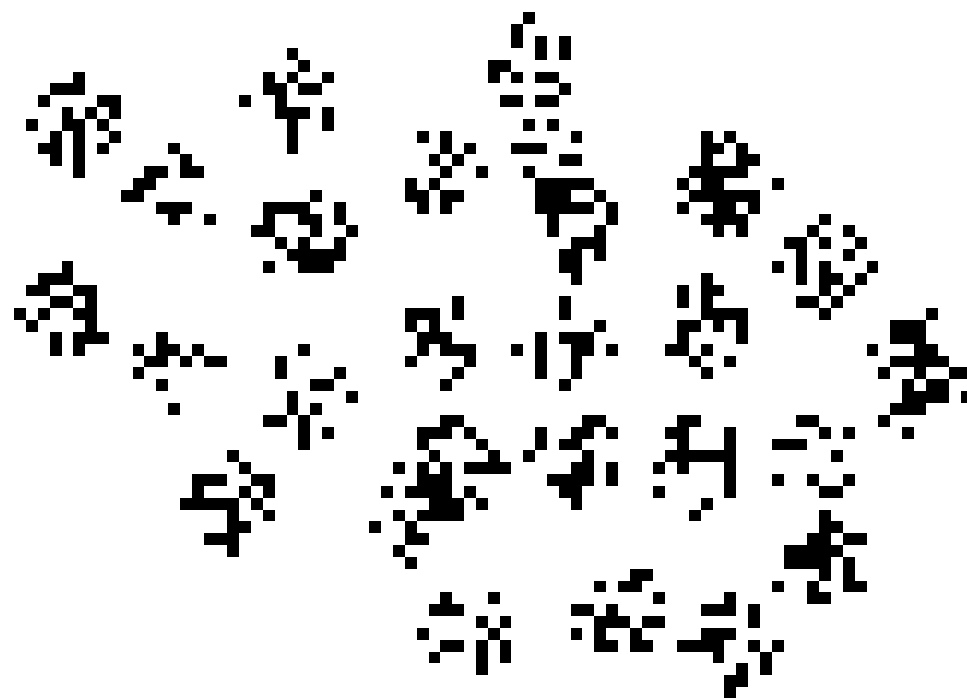
Стандартизация

- Для каждого столбца.
- Линейное преобразование
- 1. Максимальное значение = 1, минимальное = 0 (-1)
- 2. z-метки. Среднее равно 0, выборочная дисперсия равна 1.

-
- Иногда решением будет преобразование данных

Очень трудно искать черную кошку в темной комнате, особенно, если там ее нет (Может это сказал Конфуций, может нет)

Если кластеров нет, они все равно будут найдены



- Договор

провести удачно кластерный и
факторный анализ, проинтерпретировать
результаты

Результаты кластерного анализа нуждаются в интерпретации

- какой вариант кластеризации даст лучшие результаты?
- тот, который вы смогли понять и проинтерпретировать
- кластерный анализ завершён, когда мы смогли объяснить себе и заказчику, что общего у объектов в кластере и чем различаются объекты из разных кластеров между собой

Еще раз об участии аналитика

Иерархический кластерный анализ требует вдохновенного выбора формул для расстояния между объектами и расстояния между кластерами. Еще надо угадать число кластеров. Потом останется неясной геометрия кластеров. Таким образом, многое надо угадать и осмыслить. Не всегда это удастся.

Число кластеров: Silhouette

- $Dist(x_i, c_k)$ = среднее расстояние от $x_i \in c_k$ до других объектов из кластера c_k (компактность),
- $Dist(x_i, c_l)$ = среднее расстояние от $x_i \in c_k$ до объектов из ближайшего другого кластера c_l : $k \neq l$ (отделимость).
- $$Silhouette(x_i) = \frac{Dist(x_i, c_l) - Dist(x_i, c_k)}{\max(Dist(x_i, c_k), Dist(x_i, c_l))}$$
- Среднее по кластеру, по всей выборке

Другие методы определения числа кластеров

- Gordon Classification 2ed
- https://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B5_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8#.D0.A1.D0.B8.D0.BB.D1.83.D1.8D.D1.82_.28.D0.B0.D0.BD.D0.B3.D0.BB._Silhouette.29

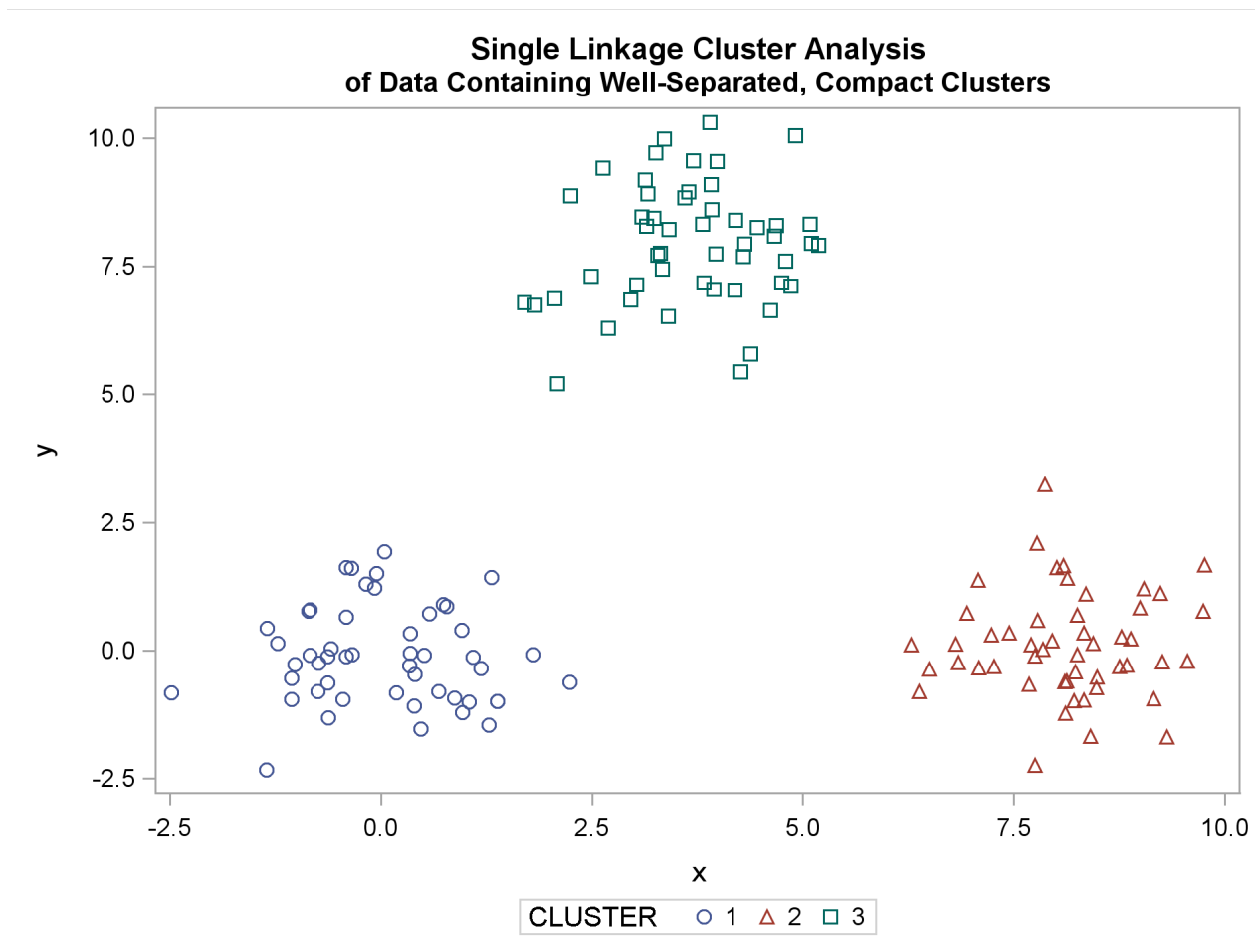
Автоматическое определение числа кластеров

- Разные методы дают разное число кластеров (один 2 кластера, другой 19)
- Мудрость толпы
- В R пакет/процедура Nbclust
- В Питоне аналогов пока нет

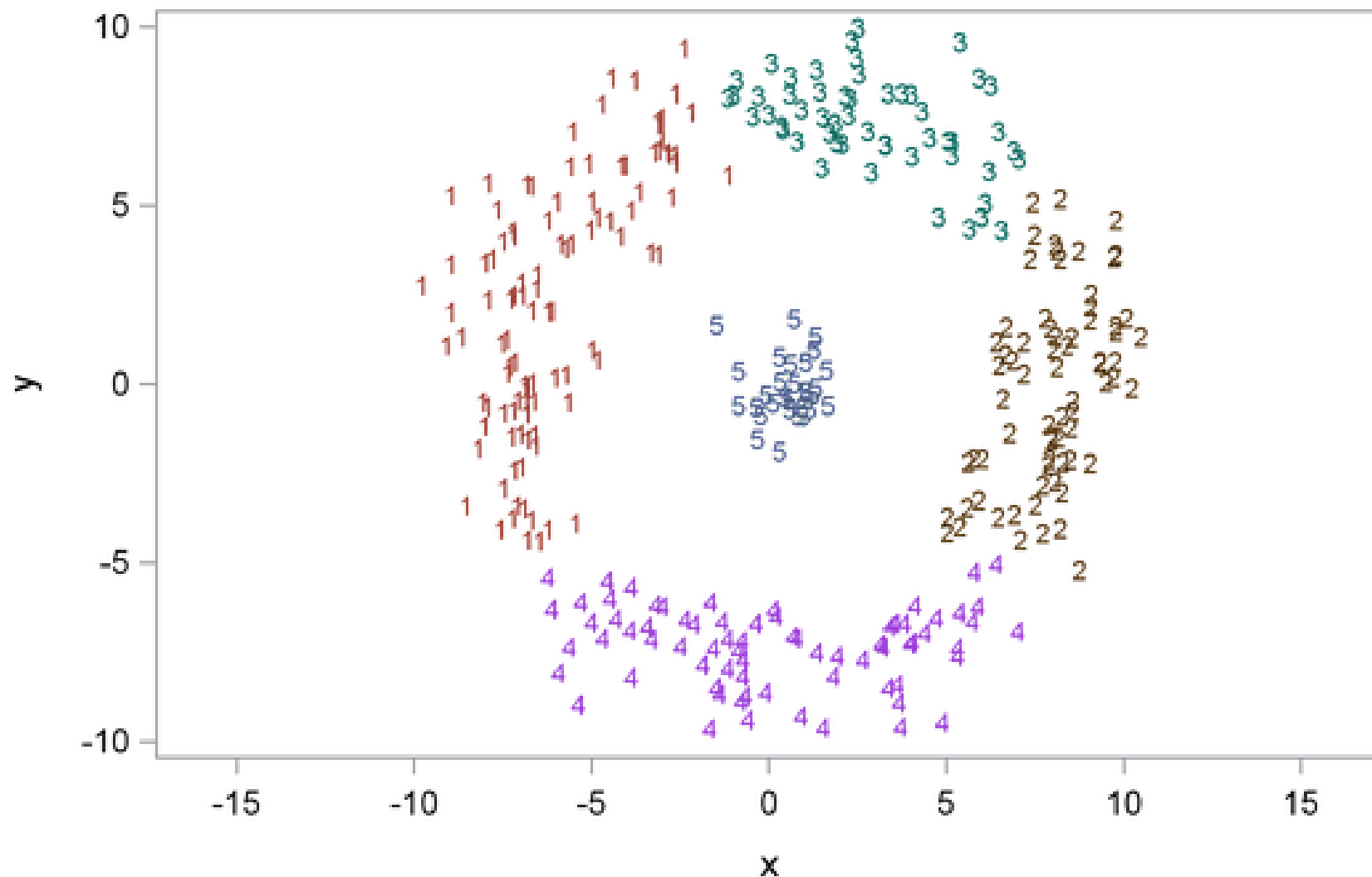
Типы кластеров

- Шаровые
 - Ленточные
 - ...
-
- Поможет выбор расстояния между кластерами

Выраженные кластеры – все равно какой метод

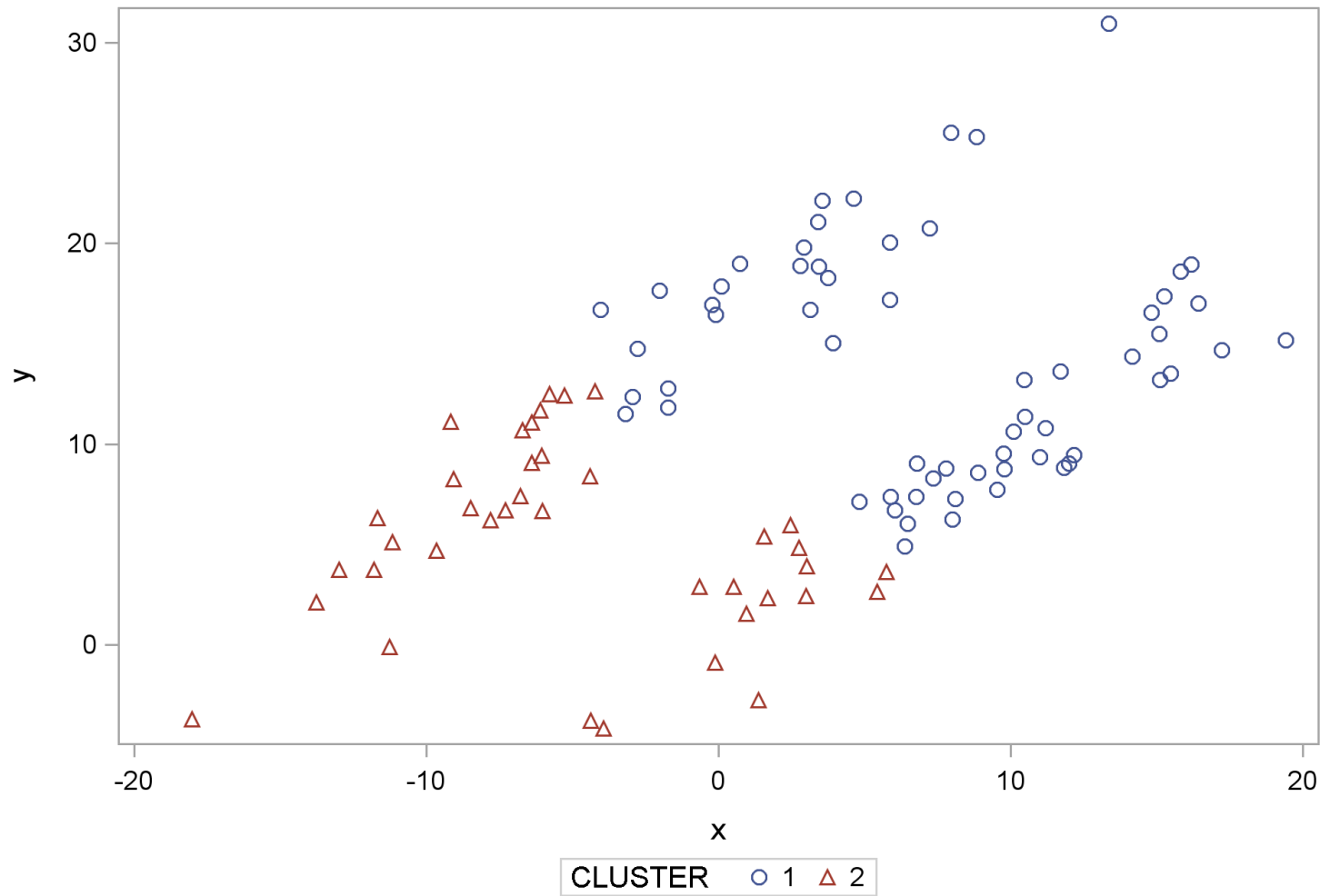


Modeclus Analysis with the JOIN= option
A Normal Cluster Surrounded by a Ring Cluster
Number of Clusters Joined=1



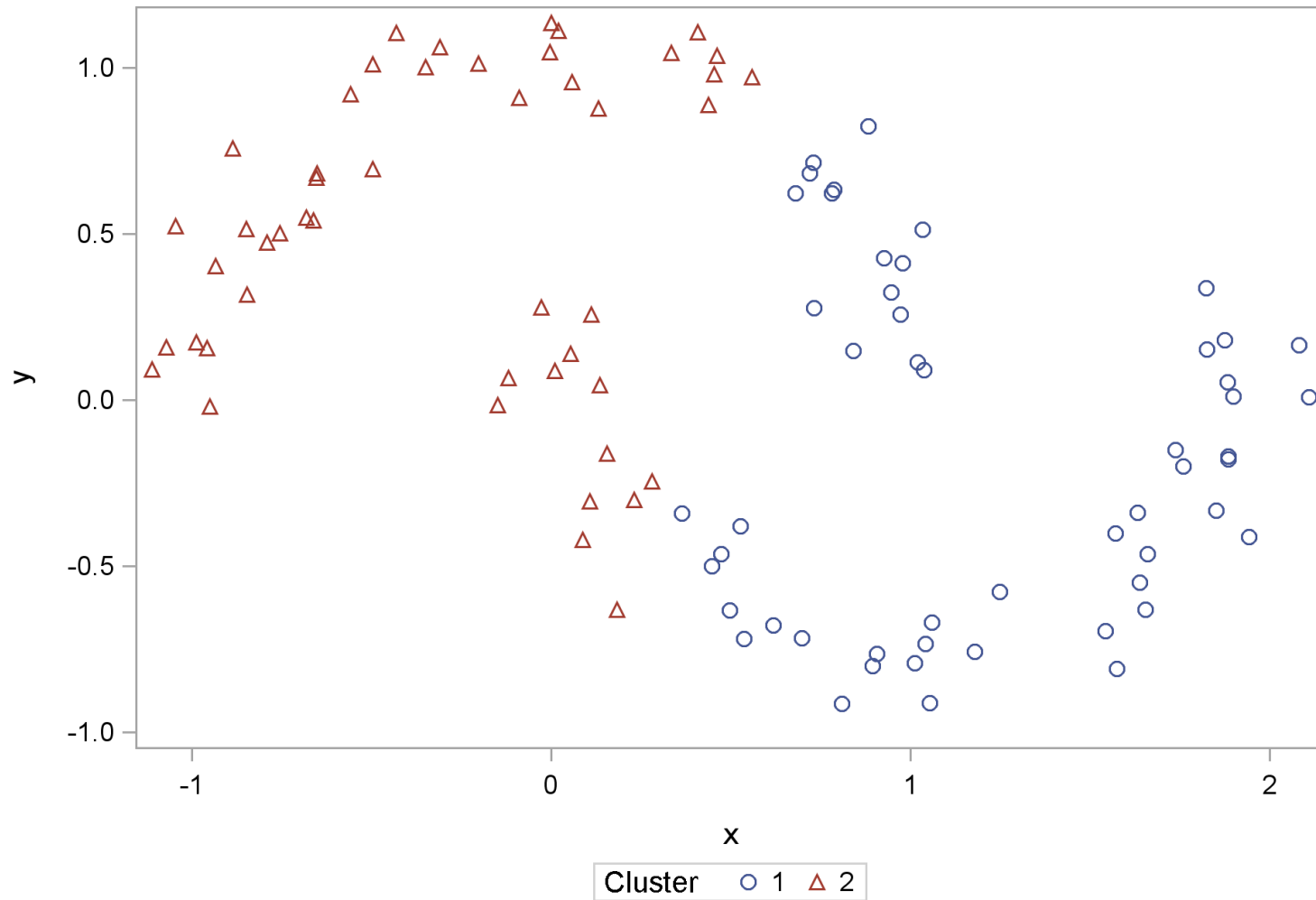
Какой метод будет лучше?

**Average Linkage Cluster Analysis
of Data Containing Parallel Elongated Clusters**



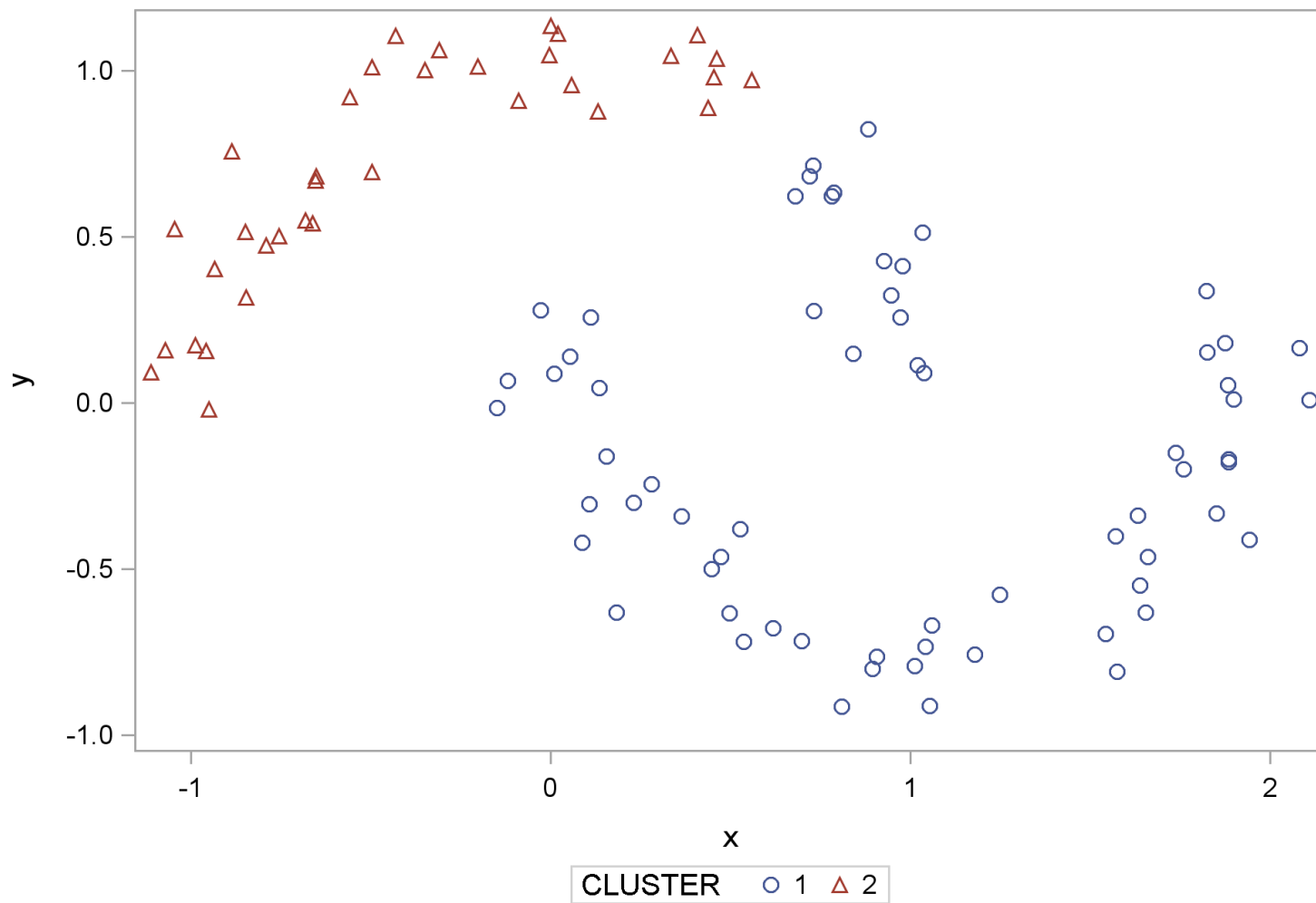
Неудача

FASTCLUS Analysis of Data Containing Nonconvex Clusters



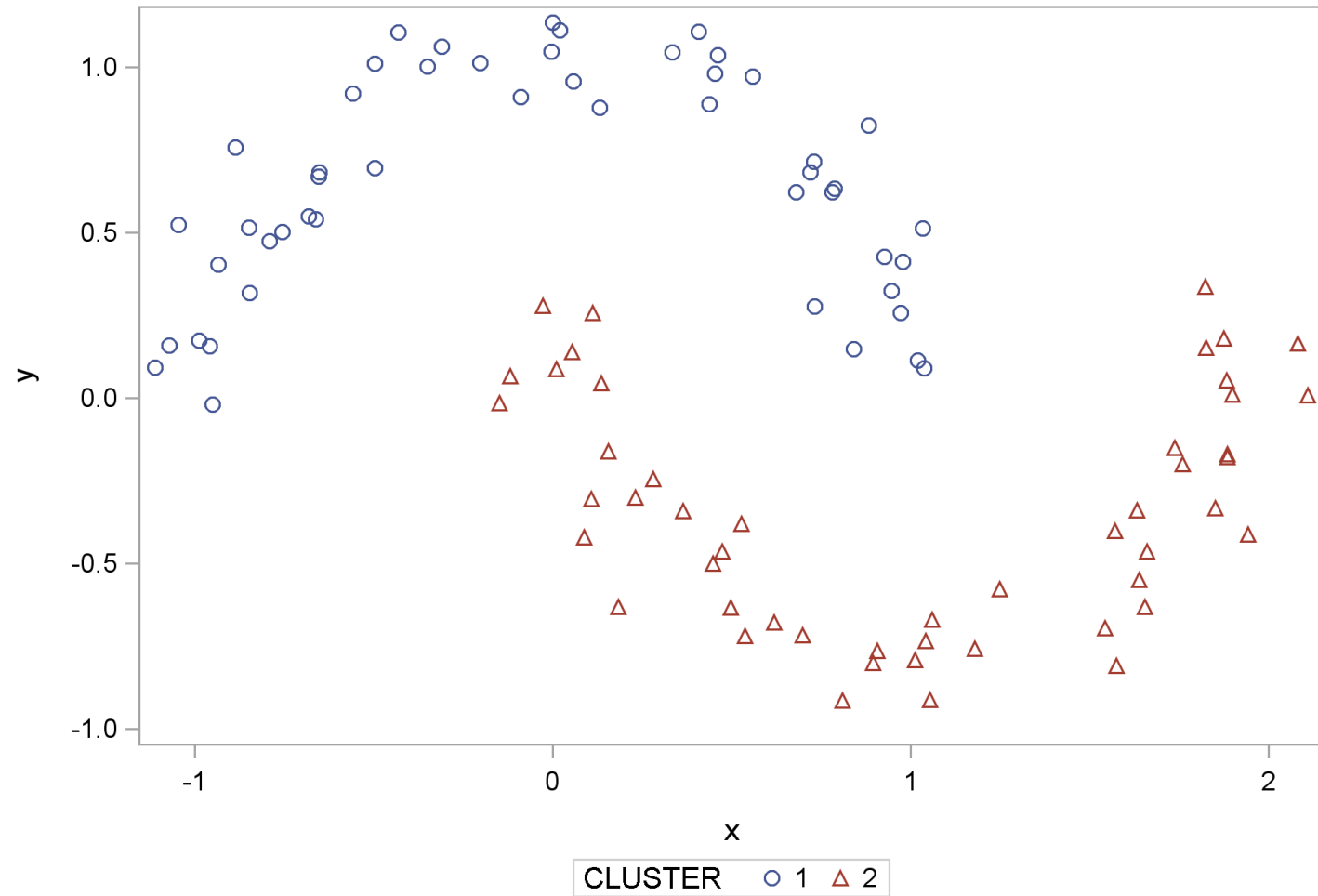
Неудача

**Centroid Cluster Analysis
of Data Containing Nonconvex Clusters**



Метод ближайшего соседа

**Two-Stage Density Linkage Cluster Analysis
of Data Containing Nonconvex Clusters**



Визуализация кластеров

- Проецируем точки на плоскость
 - Раскрашиваем точки из разных кластеров
 - Методы: факторный анализ, многомерное шкалирование, ...
-