Кластеризация

"Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Обучение «без учителя»

- *Регрессия и классификация – методы обучения «с учителем». При обучении «с учителем» у нас обычно есть набор характеристик $X_1,X_2,...,X_p$, измеренных в n наблюдениях, и отклик Y, также измеренный в тех же n наблюдениях. Цель – предсказать Y, используя $X_1,X_2,...,X_p$.*
- *Обучение «без учителя» -- набор статистических инструментов, в которых входными данными является набор характеристик $X_1,X_2,...,X_p$, измеренных в n наблюдениях. Здесь мы не заинтересованы в прогнозировании, потому что нет ассоциированной переменной отклика Y. Целью является обнаружение новых неизвестных знаний об измерениях $X_1,X_2,...,X_p$. Примерами методов обучения «без учителя» являются кластеризация, поиск ассоциативных правил и анализ главных компонент.*
- *Кластеризация – широкий класс методов для обнаружения неизвестных подгрупп в данных.*
- *Поиск ассоциативных правил – поиск закономерностей в транзакционных данных.*
- *Анализ главных компонент – процесс , при котором вычисляются главные компоненты данных, а затем используются для понимания данных.*

Трудности обучения «без учителя»

- *В методах обучения «без учителя» нет простой цели – прогнозирования поведения отклика Y. Обучение «без учителя» часто выполняется как часть объясняющего анализа данных. Также нет методов проверки результатов этих методов.*
- *Но методы обучения «без учителя» являются очень важными.*
- *Например, исследователь, занимающийся проблемой рака, может искать подгруппы среди образцов опухолей или среди генов, которые помогут ему лучше понять природу рака.*
- *Поисковая система может определять, какие результаты выводить конкретному пользователю, основываясь на истории запросов пользователей с похожими шаблонами поиска.*

Кластеризация

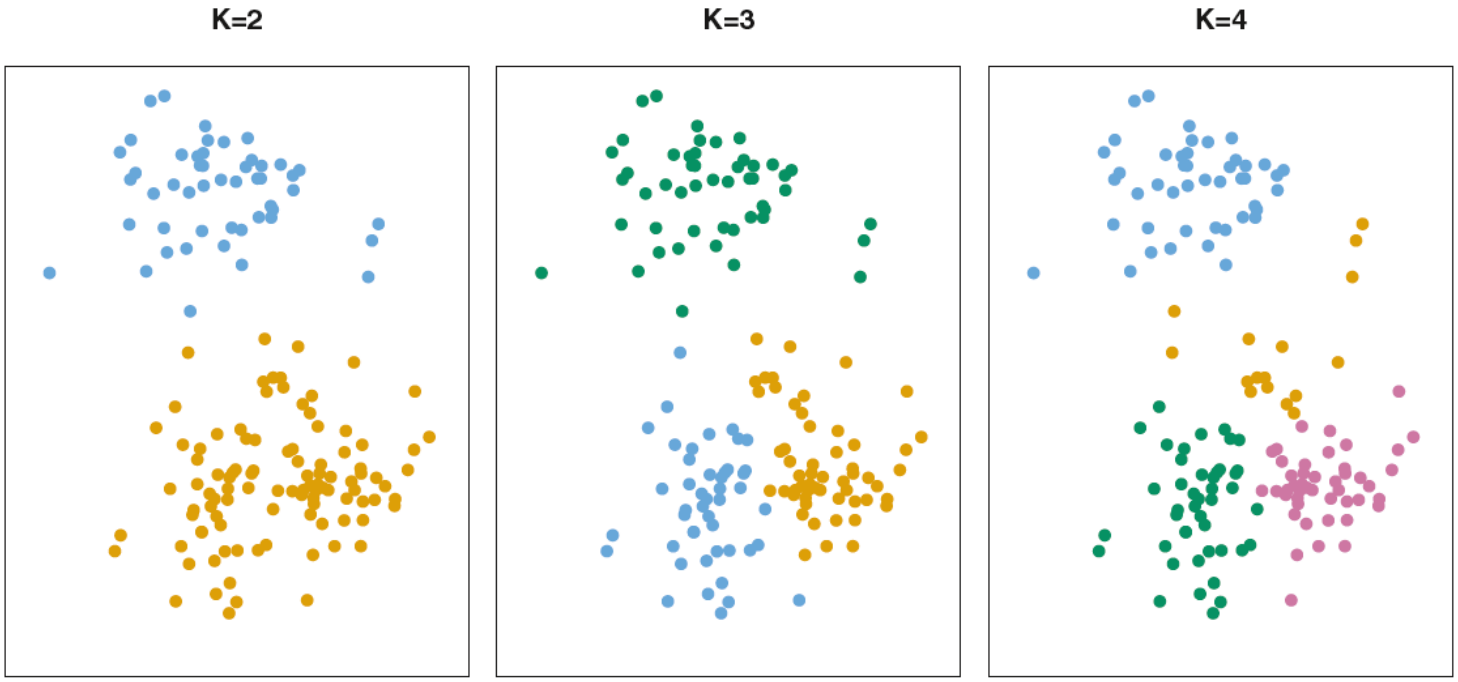
- *Кластеризацией называют очень широкое множество методик для нахождения подгрупп или кластеров в множестве данных. Когда мы кластеризуем наблюдения в множестве данных, мы ищем такое их разбиение на отдельные группы, что наблюдения в каждой группе похожи, а в разных группах сильно отличаются. Поэтому нужно определить, что для двух или большего наблюдений обозначает* ***быть похожими*** *или* ***быть различными****. Часто это определяется областью определения, и основывается на изучаемых данных. Но очень важно то, что кластеризация ищет* ***однородные*** *подгруппы наблюдений.*

Основные методы кластеризации

- *Метод кластеризации К-средних (K-means) и иерархическая кластеризация являются наиболее популярными методами. При кластеризации методом К-средних, мы ищем разделение множества наблюдений на предопределенное число кластеров. С другой стороны, в иерархической кластеризации мы не знаем заранее, сколько кластеров хотим получить. Фактически, алгоритм заканчивается дерево-подобным визуальным представлением наблюдений, называемым дендрограммой, которая позволяет нам увидеть разбиения, полученные для каждого возможного числа кластеров, от 1 до n. Каждый из подходов имеет свои преимущества и недостатки.*
- *Можно кластеризовать наблюдения на основе характеристик, т.е. находить подгруппы среди наблюдений. Также можно кластеризовать нахарактеристики на основе наблюдений, т.е. находить подгруппы среди характеристик. Далее будем обсуждать кластеризацию наблюдений, т.к. второй вариант достигается транспонированием матрицы данных.*

Кластеризация К-средних

- *Кластеризация К-средних – простой метод разделения множества данных на К различных непересекающихся кластеров. Для выполнения кластеризации сначала нужно определить желаемое число кластеров К, затем алгоритм К-средних будет относить каждое наблюдение в точности к одному из К кластеров.*



Математические основания для кластеризации

- *Пусть $C_1, \ldots, C_K$ – множества, содержащие индексы наблюдений в каждом кластере. Эти множества удовлетворяют двум свойствам:*
- *1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. Т.е. каждое наблюдение принадлежит по крайней мере одному из K кластеров.*
- *2. $C_k \cap C_{k'} = \emptyset$ для всех $k \neq k'$. Другими словами, кластеры не пересекаются.*
- *Например, если i-е наблюдение принадлежит k-му кластеру, то $i \in C_k$. Идея, которая стоит за кластеризацией K-средних состоит в том, что хорошей является та кластеризация, для которой внутри-кластерная вариация настолько мала, насколько это возможно. Внутри-кластерная вариация для кластера C_k это мера $W(C_k)$ количества, на которое наблюдения внутри кластера отличаются друг от друга. Следовательно, нам нужно решить задачу*

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \qquad (1)$$

Внутри-кластерная вариация

- *Есть много возможных способов определения этого понятия, но пока самым общим является использование квадрата Евклидова расстояния, т.е. внутри-кластерную вариацию определяют как*

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \qquad (2)$$

где $|C_k|$ обозначает число наблюдений в k-м кластере. Объединение (1) и (2) дает задачу оптимизации, которая определяет кластеризацию К-средних

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \qquad (3)$$

Алгоритм K-средних

Следующий алгоритм находит локальный минимум задачи (3) – достаточно хорошее решение.

1. *Произвольным образом присваиваем номер от 1 до K каждому из наблюдений. Это послужит начальным разделением на классы.*
2. *Повторяем, пока разбиение на классы не перестанет меняться:*
   (a) Для каждого из K кластеров вычисляем кластерный центроид. Центроид k-го кластера – это вектор средних p характеристик для наблюдений в k-м кластере.
   (b) Относим каждое наблюдение к тому кластеру, чей центроид ближайший (где ближайший определяется с использованием Евклидова расстояния).

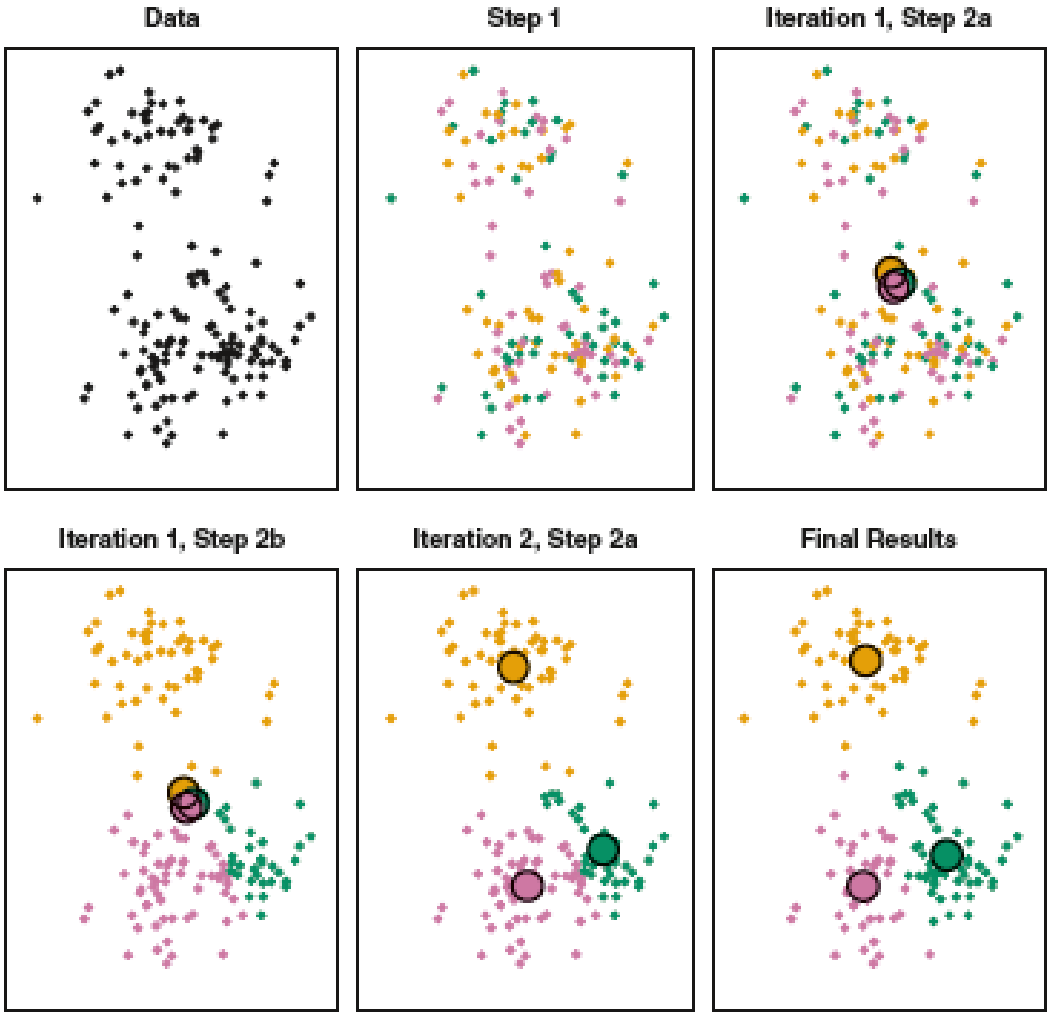
Алгоритм К-средних

- *Приведенный алгоритм гарантирует уменьшение целевой функции (3) на каждом шаге. Это происходит в силу следующего тождества:*

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

- *где* $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ *- это среднее характеристики j в классе* C_k*. Это значит, что при использовании алгоритма К-средних целевая функция (3) будет невозрастать, и когда результат перестанет меняться, будет достигнут локальный минимум.*

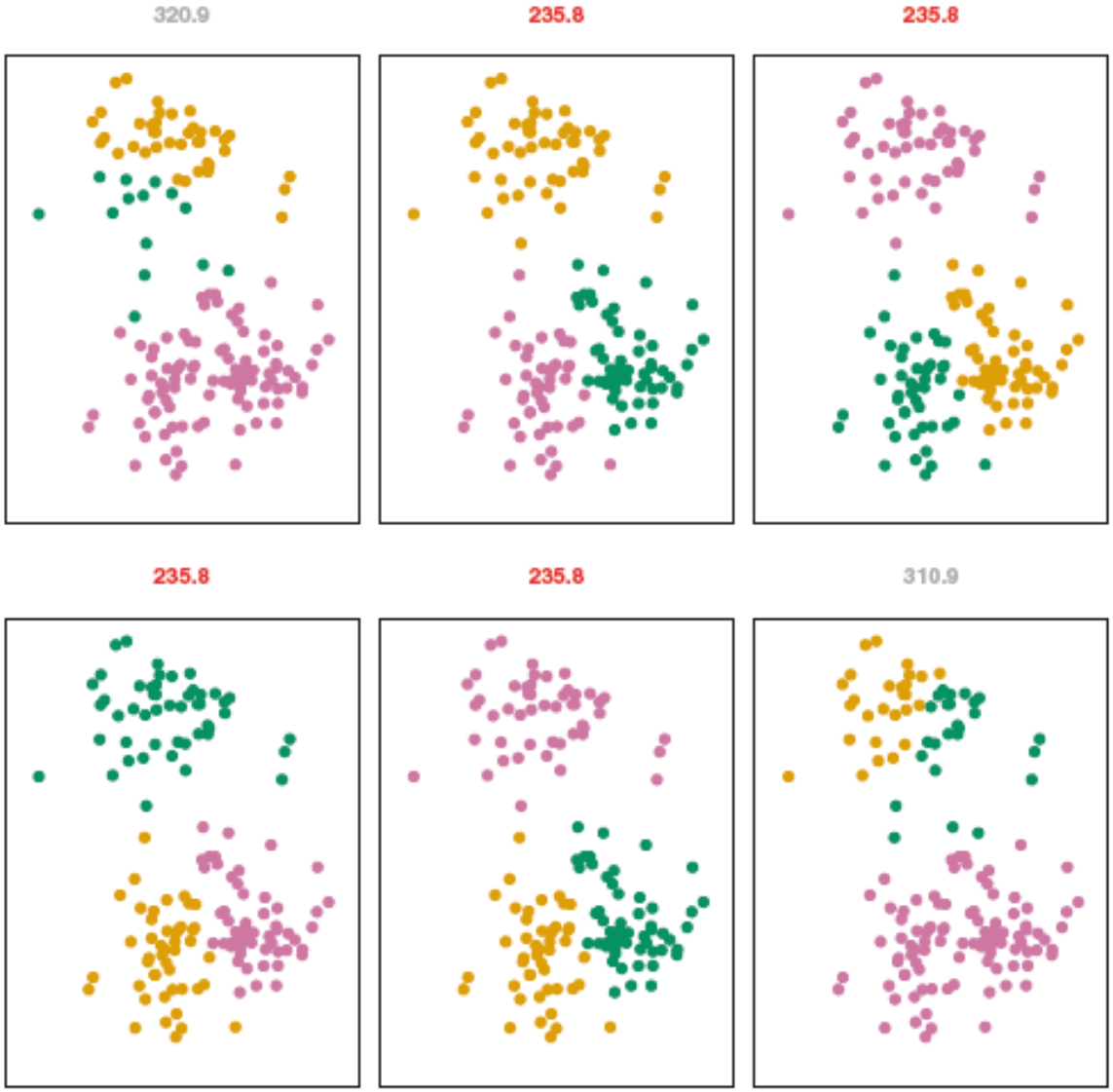
Процесс кластеризации для K=3. В правом нижнем углу – результат кластеризации после 10 итераций.



Выбор начальных кластеров

- *Так как алгоритм К-средних находит локальный, а не глобальный минимум, его результат зависит от выбора начальных (случайных) кластеров. Поэтому важно запустить алгоритм несколько раз для различных случайных начальных конфигураций. Затем выбирается наилучшее решение, т.е. то, для которого значение целевой функции (3) будет минимальным.*

Различные итоговые кластеры

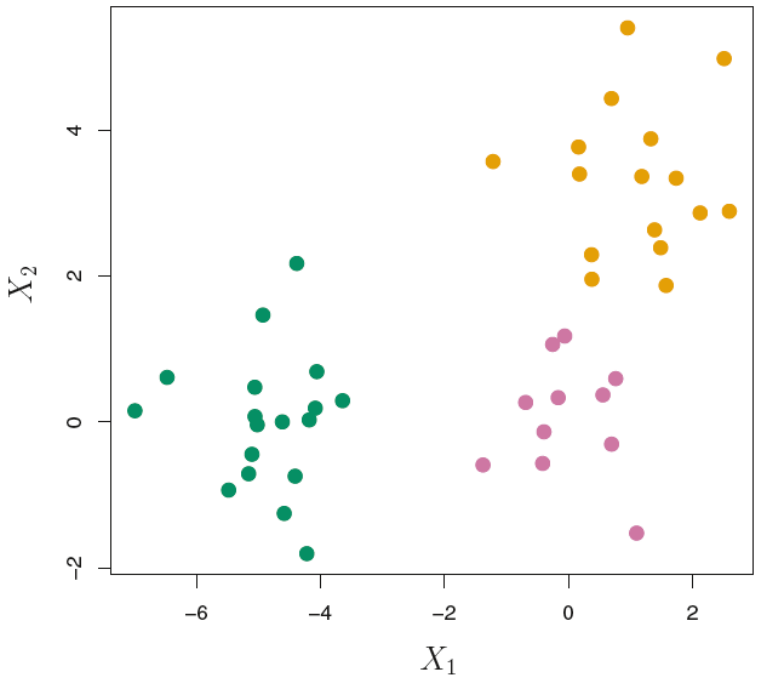


На картинке – итоги кластеризации К-средних для К=3, выполненная 6 раз с разными начальными кластерами. Над каждым рисунком – значение целевой функции. Найдены три локальных минимума, один из которых выражается в наименьшем значении целевой функции, и дает наилучшее разделение на кластеры.

Иерархическая кластеризация

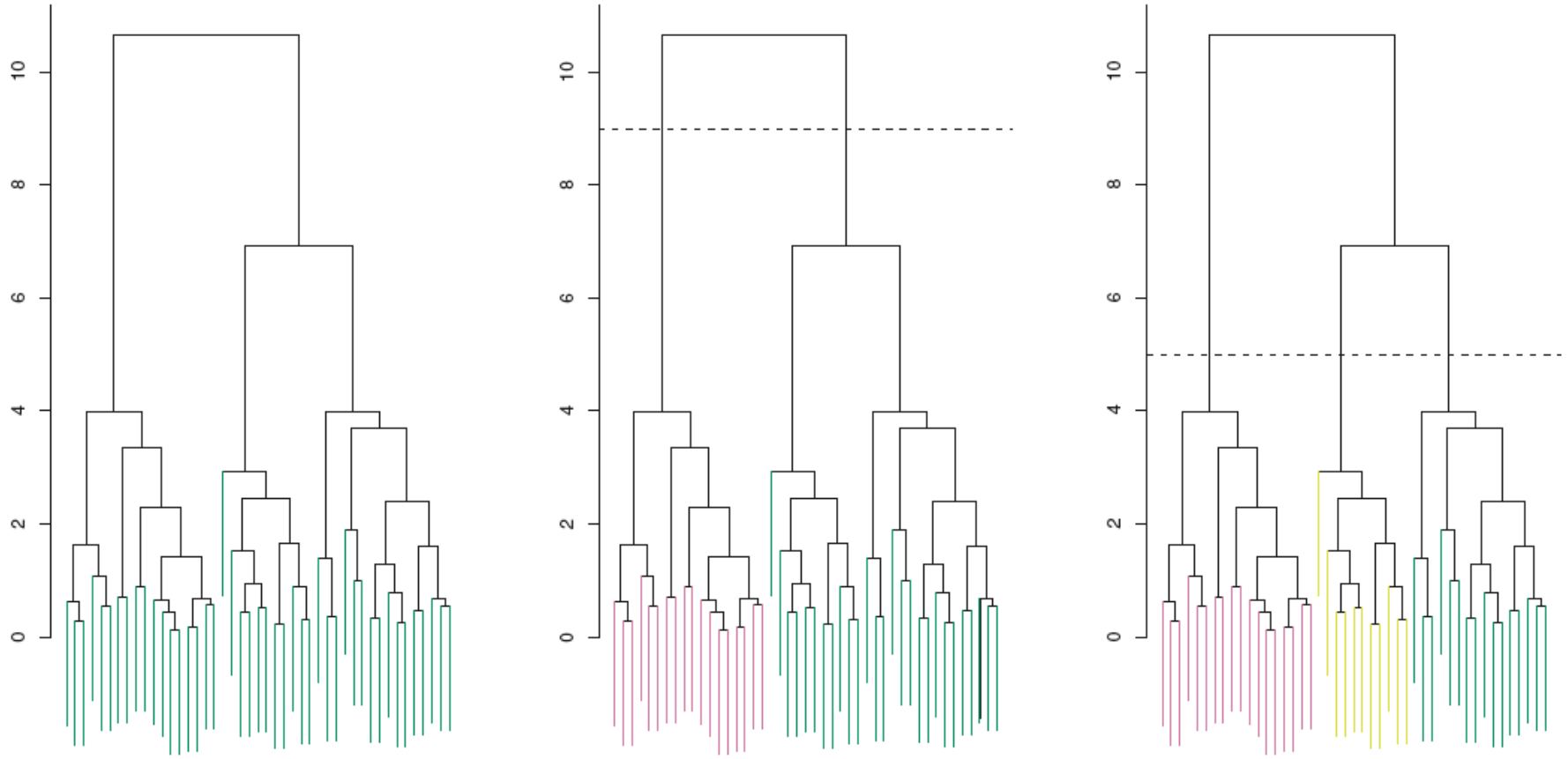
- *Потенциальный недостаток кластеризации К-средних – предопределение числа кластеров К. Иерархическая кластеризация – это альтернативный подход, который не требует выбора К. Результат иерархической кластеризации представляется в древообразном представлении наблюдений, называемом дендрограммой. Будем рассматривать агломеративную кластеризацию. Это наиболее общий тип иерархической кластеризации. Дендрограмма строится, начиная с листьев, и объединяет кластеры к корню дерева.*

Интерпретация дендрограммы



- *Рассмотрим искусственно сгенерированные 45 наблюдений в 2-мерном пространстве, в реальности разделенные на 3 класса.*
- *Предположим, что мы наблюдаем данные без метки класса и применяем для них метод иерархической кластеризации.*

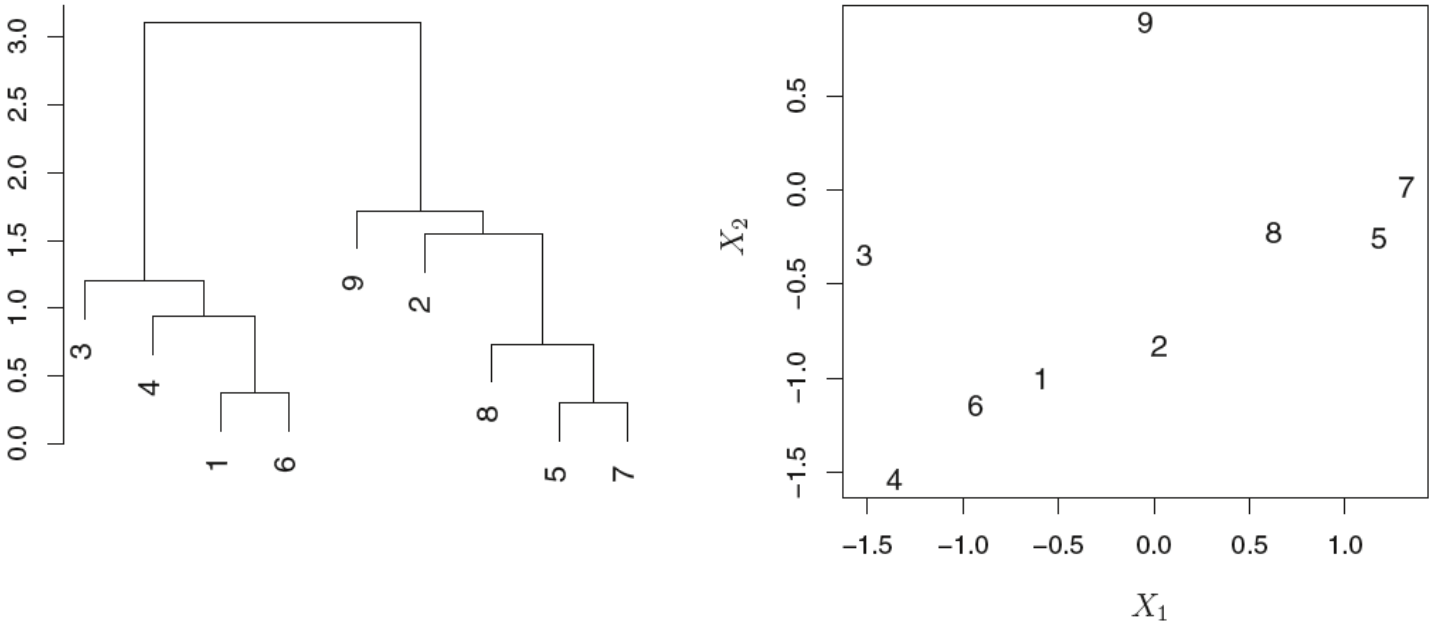
Рисунок 1. Дендрограмма



Интерпретация дендрограммы

- *В левой части рис. 1 каждый лист дендрограммы представляет одно из 45 исходных наблюдений. Но по мере того, как мы продвигаемся вверх к корню, некоторые листья начинают сливаться в ветви. Они соответствуют наблюдениям, которые похожи друг на друга. При продвижении вверх по дереву ветви сливаются с листьями или с другими ветвями. Чем раньше (ниже на дереве) происходит слияние, тем больше похожи группы наблюдений друг на друга. С другой стороны, наблюдения, которые сливаются позже (возле верхушки дерева), могут сильно отличаться. Это утверждение можно сформулировать точно: для любых двух наблюдений можно найти точку на дереве, где ветви , содержащие эти наблюдения, впервые сливаются. Высота этой точки слияния, измеренная по вертикальной оси, указывает, насколько различаются эти наблюдения. Таким образом, наблюдения, которые сливаются в самом низу дерева, очень похожи друг на друга, в то время как наблюдения, которые сливаются близко к верху дерева, будут скорее всего сильно различаться.*

Интерпретация простой дендрограммы



- *Слева на рисунке изображена дендрограмма, построенная с использованием Евклидова расстояния с полной связью. Наблюдения 5 и 7 подобны, также, как и наблюдения 1 и 6. Однако наблюдение 9 не более подобно наблюдению 2, чем наблюдениям 8, 5 и 7, несмотря на то, что наблюдения 9 и 2 близки по горизонтали. Это происходит потому, что все наблюдения 2, 8, 5 и 7 сливаются с наблюдением 9 на одинаковой высоте, примерно 1.8.*
- *Справа на рисунке изображены исходные данные для дендрогаммы. На этом рисунке видно, что наблюдение 9 не ближе к 2, чем к 8, 5 и 7.*

Определение кластеров на основе дендрограммы

- *Разрежем дендрограмму по горизонтали, как это показано в центре и справа на рис. 1. Отдельные множества наблюдений, которые находятся снизу от разреза могут рассматриваться в качестве кластеров. В центре рис. 1 разрез дендрограммы на высоте 9 дает 2 кластера, показанных разными цветами. Справа разрез дендрограммы на высоте 5 дает 3 кластера. Можно сделать и другие разрезы дендрограммы. 1 кластер получится, если не разрезать, а на высоте 0 будет n кластеров, т.е. каждое измерение будет отдельным кластером. Другими словами, высота разреза дендрограммы играет ту же роль, что и К в кластеризации К-средних: контролирует число полученных кластеров.*

Выбор числа кластеров

- *Таким образом, одна дендрограмма может использоваться для получения любого числа кластеров. На практике, часто приемлемое число кластеров выбирается визуально на основе высоты слияния и желаемого числа кластеров. В случае рис.1 таким будет выбор 2 или 3 кластеров. Однако, не всегда выбор так очевиден.*
- *Например, предположим, что множество наблюдений соответствует группе людей 50/50 состоящей из мужчин и женщин, и содержащей поровну американцев, японцев и французов. Можно представить сценарий, при котором наилучшее деление на 2 группы разделит этих людей по гендерному признаку, а лучшее деление на 3 группы разделит их по национальности. Но в этой ситуации кластеры не будут объединяться, в том смысле, что лучшее деление на 3 группы не происходит из лучшего деления на 2 группы и деления одной из этих групп. Следовательно, эта ситуация не может быть хорошо представлена иерархической кластеризацией. Поэтому иерархическая кластеризация иногда может давать худшие (менее точные) результаты, чем кластеризация К-средних для данного числа кластеров.*

Алгоритм иерархической кластеризации

- *Начнем с определения меры отличия (несхожести) между каждой парой наблюдений. Чаще всего используется Евклидово расстояние. Алгоритм продолжается итеративно. Начиная с дна дендрограммы каждое из n наблюдений рассматриваются свой собственный кластер. Два кластера, которые наиболее подобны, сливаются, так что становится n-1 кластер. Затем опять наиболее подобные 2 кластера сливаются, и становится n-2 кластера. Алгоритм продолжается, пока не получится 1 кластер, и дендрограмма не станет завершенной.*

Алгоритм иерархической кластеризации

1. *Начинаем с* n *наблюдений и меры (например, Евклидова расстояния) всех* $C_n^2 = n(n-1)/2$ *попарных расстояний между кластерами. Рассматриваем каждое наблюдение как отдельный кластер.*
2. *Для* $i = n, n-1, \ldots, 2$:
   (a) Рассматриваем все попарные межкластерные расстояния между i *кластерами и определяем пару кластеров, которые наименее неподобны (т.е. наиболее подобны, с наименьшим расстоянием). Объединяем эти два кластера. Расстояние между этими двумя кластерами указывает на высоту в дендрограмме, на которой это слияние должно быть расположено.*
   (b) Вычисляем новые попарные межкластерные расстояния между $i-1$ *оставшимися кластерами.*

Межкластерное расстояние

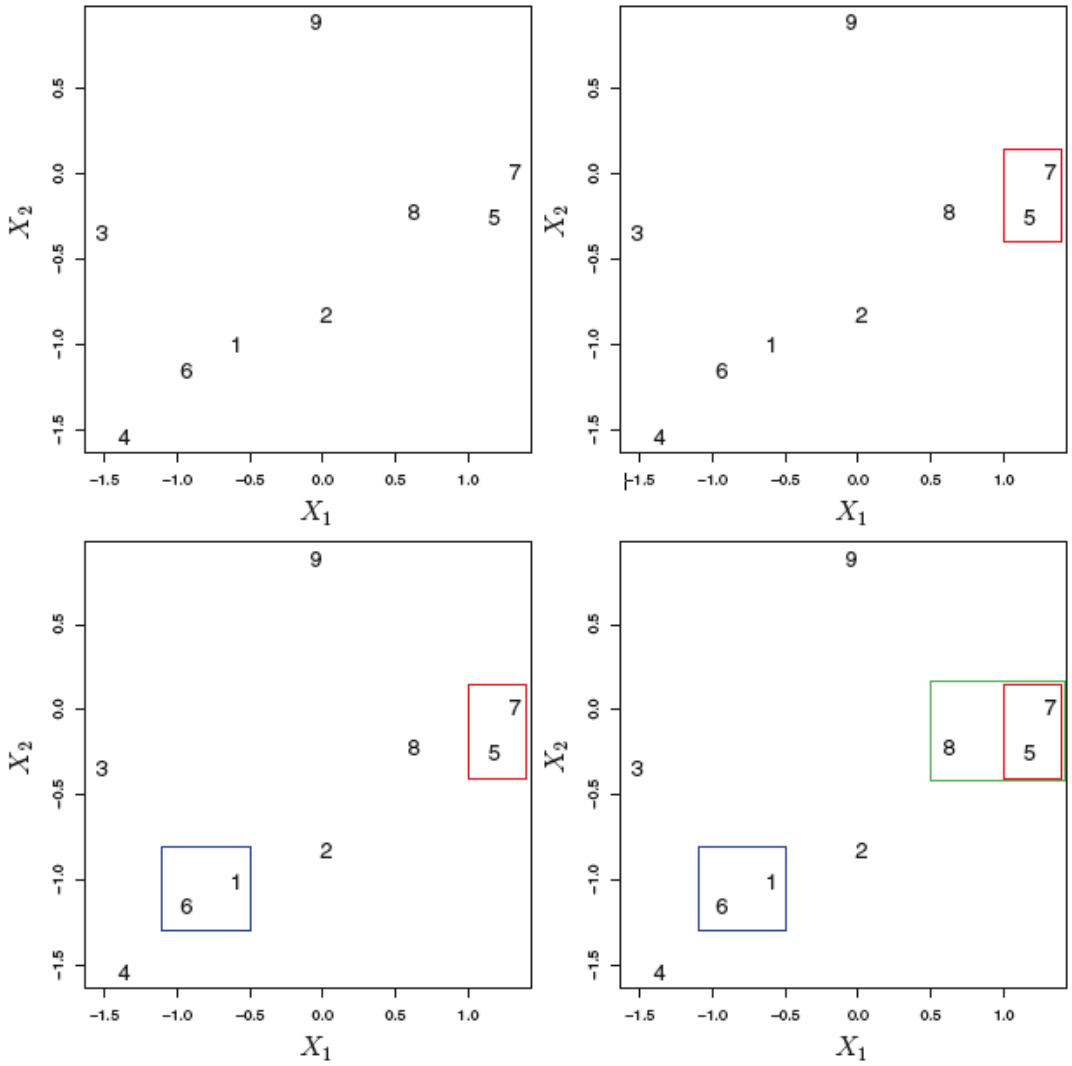
- *Как определить расстояние между кластерами, содержащими несколько наблюдений?*
- *Концепция расстояния между парами наблюдений должна быть расширена на пары групп наблюдений. Это расширение достигается с помощью разработки понятия связи (linkage), которое определяем расстояние между группами наблюдений. Четыре наиболее общих типа связей -- полная, средняя, одиночная и центроидная – приведены в следующей таблице.*

Связь	*Описание*
Полная	*Максимальное межкластерное расстояние. Вычисляются все попарные расстояния между наблюдениями кластера A и наблюдениями кластера B, и записывается наибольшее из этих расстояний.*
Одиночная	*Минимальное межкластерное расстояние. Вычисляются все попарные расстояния между наблюдениями кластера A и наблюдениями кластера B, и записывается наименьшее из этих расстояний. Одиночная связь может приводить к расширенным, протяженным кластерам, в которых одиночные наблюдения объединяются по одному за раз.*
Средняя	*Среднее межкластерное расстояние. Вычисляются все попарные расстояния между наблюдениями кластера A и наблюдениями кластера B, и записывается среднее этих расстояний.*
Центроидная	*Расстояние между центроидом кластера A (вектором средних длины p) и центроидом кластера B. Центроидная связь может приводить к нежелательным инверсиям.*

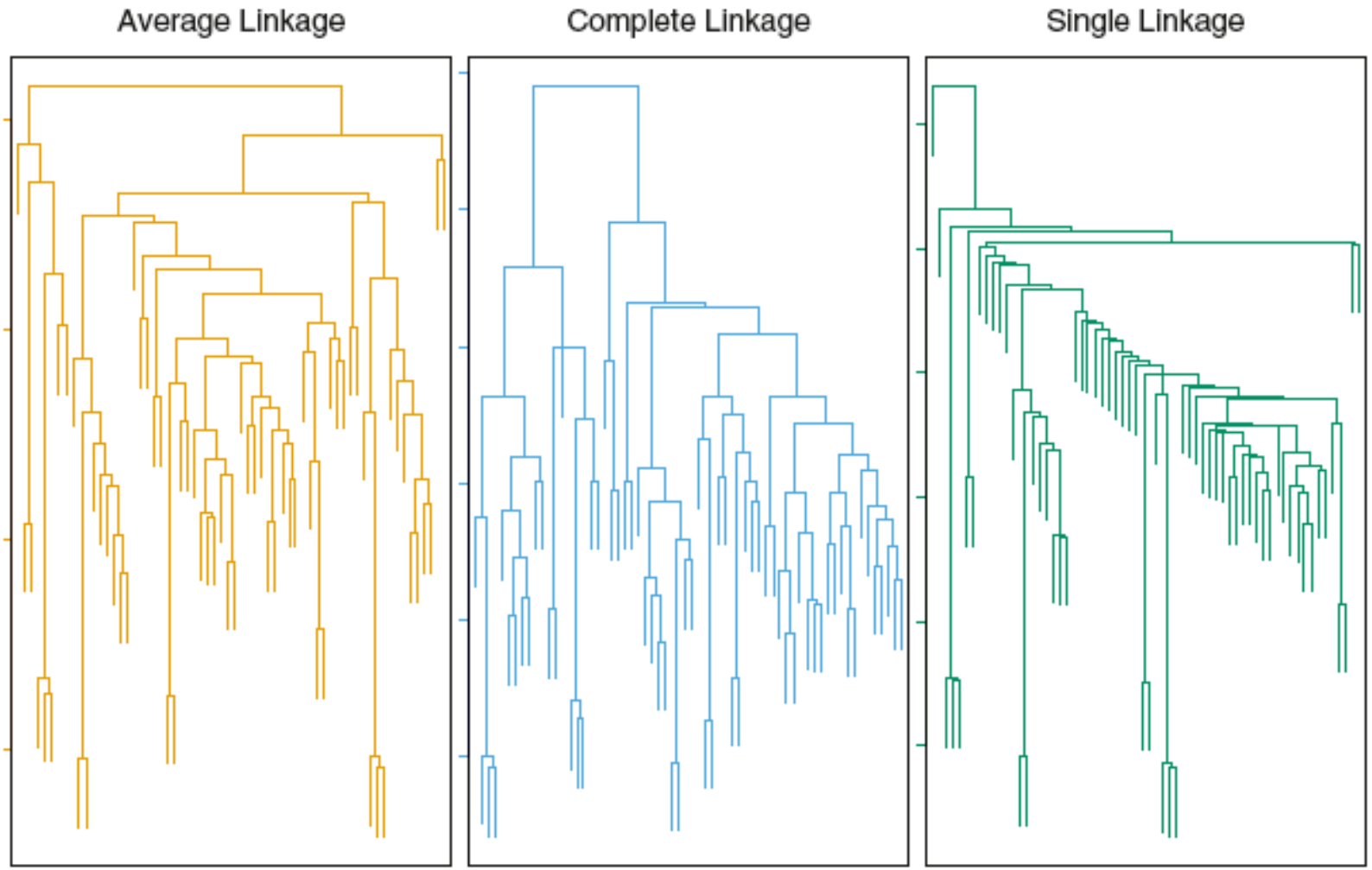
Выбор типа связи

- *Средняя, полная и одиночная связи наиболее популярны у статистиков. Средней и полной связям отдается предпочтение перед одиночной связью, т.к. они приводят к более сбалансированным дендрограммам. Центроидная связь часто используется в работе с генами, но имеет существенный недостаток – появление инверсий. Инверсия – это объединение двух кластеров на высоте более низкой, чем каждый из этих отдельных кластеров в дендрограмме. Это приводит к трудностям с визуализацией и интерпретацией дендрограммы. Расстояния, вычисляемые на шаге 2(b) алгоритма иерархической кластеризации, зависят как от от типа используемой связи, так и от выбора меры расстояния (несхожести). Следовательно, итоговая дендрограмма очень сильно зависит от типа используемой связи.*

Иллюстрация первых шагов алгоритма иерархической кластеризации, использующего Евклидово расстояние и полную связь



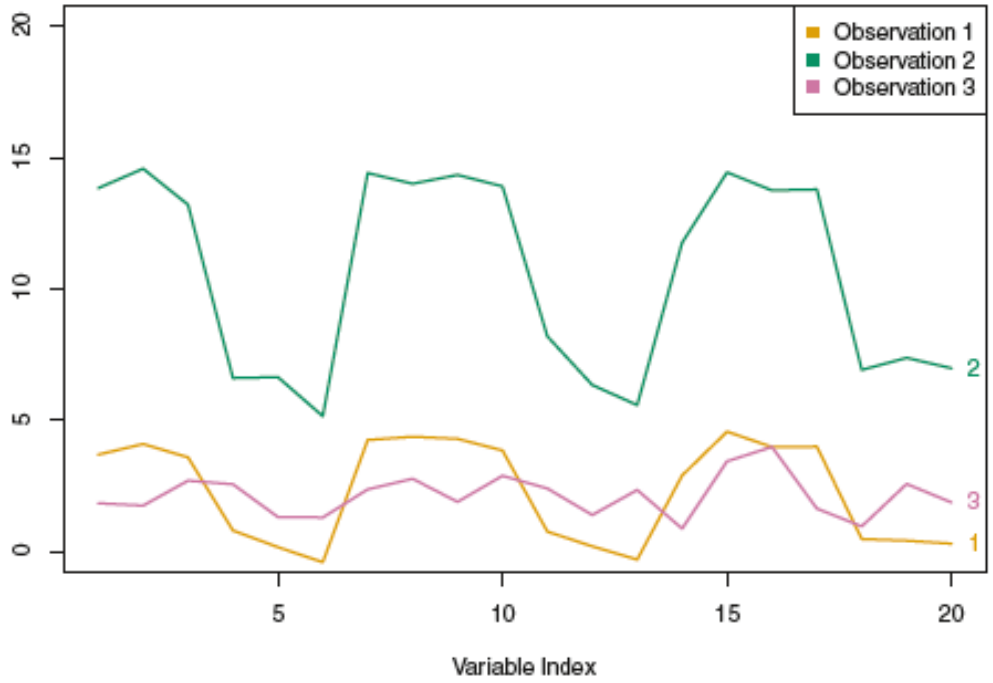
Применение полной, средней и одиночной связи к данным из примера. Средняя и полная связи приводят к более сбалансированным кластерам.



Выбор меры расстояния (несхожести)

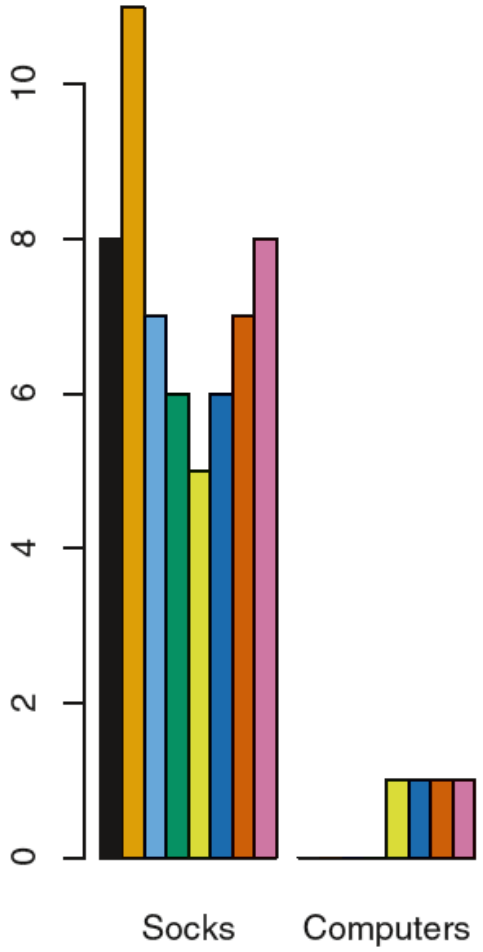
- *До сих пор в примерах использовалось Евклидово расстояние в качестве меры несхожести. Но иногда предпочтительнее использовать другие меры. Например, расстояние, основанное на корреляции, считает, что два наблюдения подобны, если их характеристики высоко коррелированы, даже если наблюдаемые значения находятся далеко в терминах Евклидова расстояния. Это необычное использование корреляции, которая обычно вычисляется между переменными, а здесь вычисляется между профилями наблюдений для каждой пары наблюдений. Расстояние, основанное на корреляции, фокусируется на форме профилей наблюдений, а не на их величинах.*
- *В общем, нужно очень внимательно относиться к типу кластеризуемых данных и вопросу, на который мы хотим получить ответ с помощью кластеризации. Это поможет выбрать тип расстояния, используемого в иерархической кластеризации.*

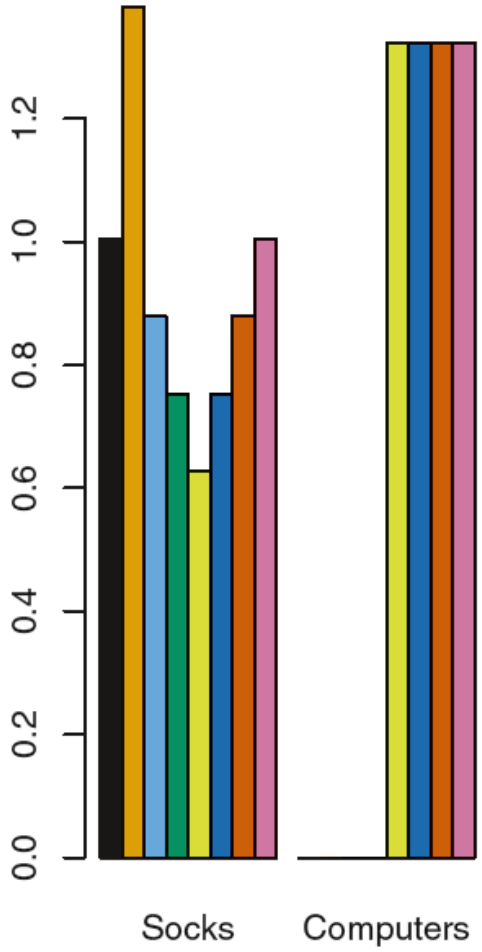
Использование Евклидова и основанного на корреляции расстояния

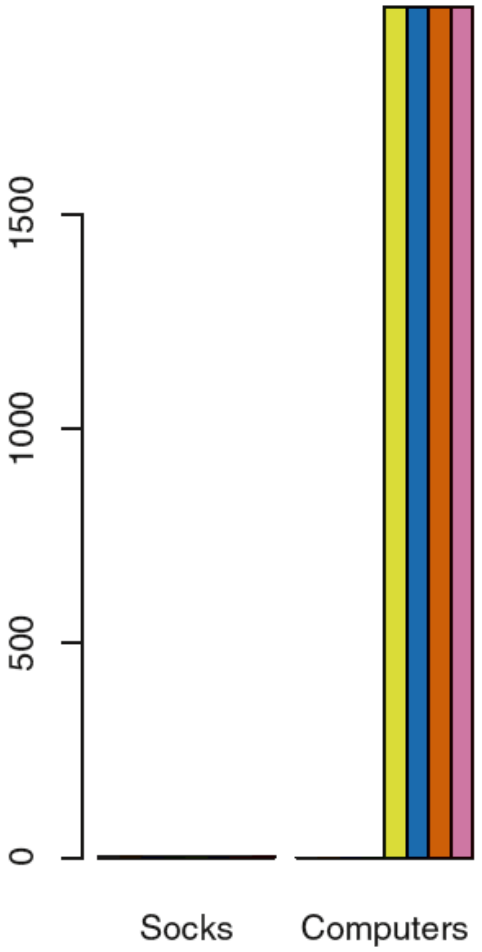


- *Показаны 3 наблюдения из 20 переменных. Наблюдения 1 и 3 имеют похожие значения для каждой переменно, поэтому между ними будет маленькое Евклидово расстояние. Но они слабо коррелируют, поэтому между ними будет большое расстояние, основанное на корреляции. С другой стороны, наблюдения 1 и 2 имеют совершенно различные значения для каждой переменной, поэтому Евклидово расстояние между ними будет большим. Но они сильно коррелируют, поэтому расстояние, основанное на корреляции, будет маленьким.*

Пример


0
2
4
6
8
10
Socks
Computers


0.0
0.2
0.4
0.6
0.8
1.0
1.2
Socks
Computers


0
500
1000
1500
Socks
Computers

Интерпретация примера

- *Эклектичный онлайновый розничный продавец продает 2 вида товаров: носки и компьютеры.*
- *Слева изображено количество пар носков и компьютеров, приобретенных 8 онлайновыми покупателями. Покупатели показаны разным цветом. Если использовать Евклидово расстояние для межкластерного расстояния, не изменяя переменные, то число купленных компьютеров почти не будет влиять на результат. Это может быть нежелательно, т.к. (1) компьютеры более дорогие, чем носки, и продавец может быть более заинтересован в том, чтобы покупатели покупали компьютеры, а не носки, и (2) большая разница в числе носков, купленных двумя покупателями, может быть менее информативно в отношении общих предпочтений покупателя, чем маленькая разница в числе купленных компьютеров.*
- *В центре показаны те же самые данные после масштабирования каждой переменной на СКО. Теперь число купленных компьютеров будет иметь значительно больший эффект на межкластерные расстояния.*
- *Справа изображены те же данные, но теперь ось y представляет сумму в долларах, потраченную каждым покупателем на носки и на компьютеры. Так как компьютеры значительно дороже носков, теперь история покупок компьютеров будет влиять на полученные межкластерные расстояния.*

Маленькие решения с большими последствиями

- *Чтобы применять алгоритм кластеризации, нужно сначала принять несколько решений.*

1. *Нужно ли сначала стандартизировать наблюдения или характеристики? Например, возможно переменные нужно центрировать, чтобы у них было нулевое мат. ожидание, и масштабировать, чтобы получить единичное СКО.*
2. *В случае иерархической кластеризации*

- *Какую меру расстояния использовать?*
- *Какой тип связи использовать?*
- *Где разрезать дендрограмму, чтобы получить кластеры?*

3. *В случае кластеризации К-средних, какое число кластеров мы будем искать?*

Каждое из этих решений сильно влияет на результат. На практике пробуют несколько вариантов, а затем выбирают тот, которые дает наиболее полезное или объяснимое решение. Нет единственного правильного ответа – каждое решение может раскрывать какие-нибудь интересные факты в данных.