НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Интеллектуальный анализ данных и основы машинного обучения

11 ноября 2017 г.

Что мы изучили

- Несколько методов классификации.
- Несколько моделей регрессии.
- Регуляризация и ее применение.

Что нам предстоит

- Обучение без учителя.
- Понижение размерности.
- Отбор признаков и выделение признаков.
- Ансамбли моделей.
- Еще раз о процессе анализа данных.

Напоминание: задачи в машинном обучении

- Обучение с учителем (supervised learning) – обучение, в котором есть некоторое число примеров с правильными ответами.
- Обучение без учителя (unsupervised learning) – обучение, в котором нет правильных ответов, только данные.
- Обучение с подкреплением (reinforcement learning) – обучение, в котором агент учится из собственных проб и ошибок.

Обучение без учителя

- Обучение без учителя – обучение, в котором у нас есть какие-то объекты с признаками, но нет “правильного” ответа:
  - сокращение размерности;
  - визуализация;
  - определение “степени похожести” объектов;
  - восстановление пропусков;
  - выявление аномалий;
  - поиск ассоциативных правил;
  - кластеризация.

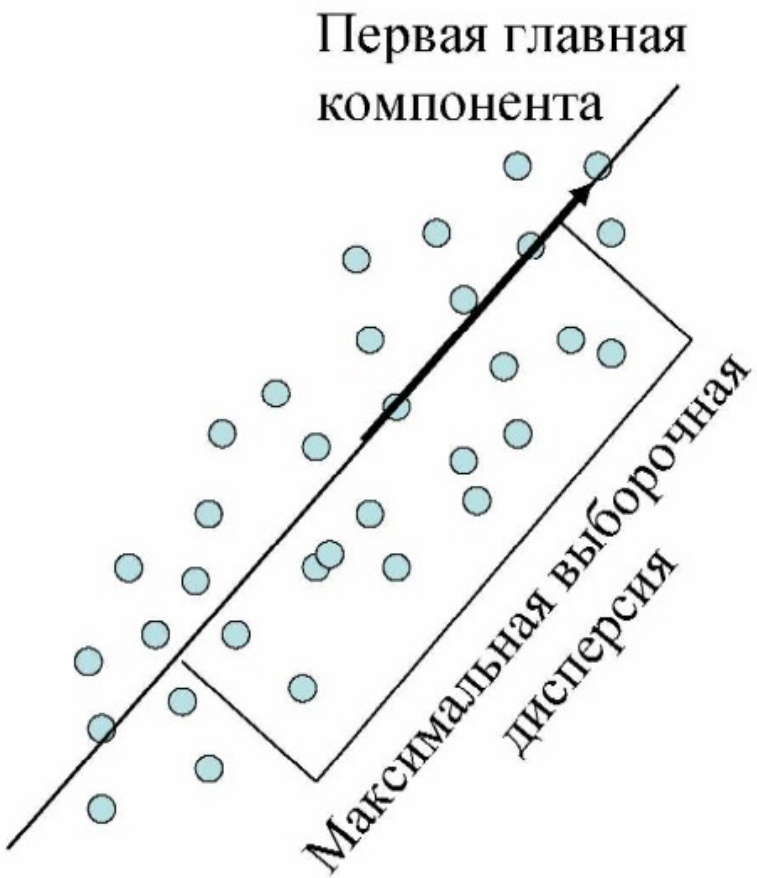
Сокращение размерности/визуализация

- Две связанные задачи.
- Хорошая визуализация позволяет увидеть некоторые закономерности.
- Многомерные данные сложно отобразить.
- Ищем шкалы (оси) в которых “различия” максимальны.

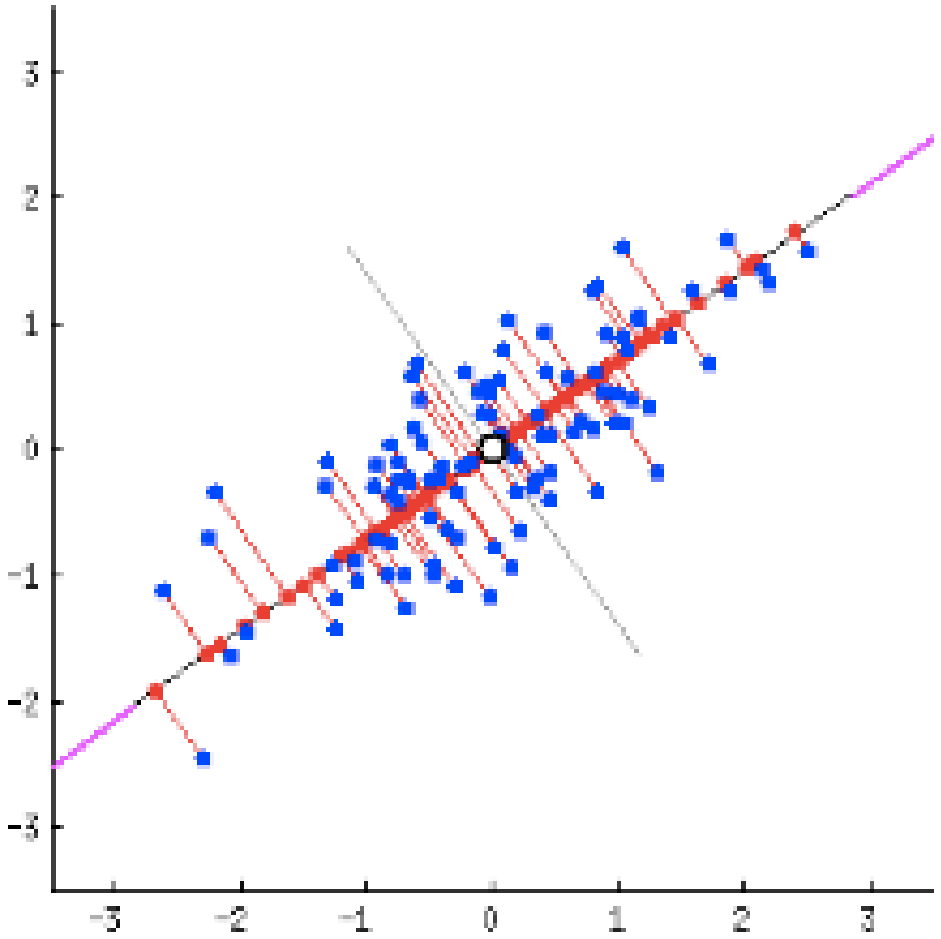
PCA

Самый популярный метод снижения размерности – анализ главных компонент (PCA).

- Ищем вектора, вдоль которых дисперсия данных максимальна.
- Переходим в базис на основе этих векторов.
- Обнуляем координаты при самых ”неважных”векторах.



PCA



Восстановление пропусков

- Часто данные могут быть не полны;
- Выкидывание данных с пропусками может существенно сократить базу для обучения;

Восстановление пропусков

- Часто данные могут быть не полны;
- Выкидывание данных с пропусками может существенно сократить базу для обучения;
- Можно свести задачу к задаче обучения с учителем;

Восстановление пропусков

- Часто данные могут быть не полны;
- Выкидывание данных с пропусками может существенно сократить базу для обучения;
- Можно свести задачу к задаче обучения с учителем;
- Можно свести задачу к задаче кластеризации;

Восстановление пропусков

- Часто данные могут быть не полны;
- Выкидывание данных с пропусками может существенно сократить базу для обучения;
- Можно свести задачу к задаче обучения с учителем;
- Можно свести задачу к задаче кластеризации;
- Классический пример восстановления пропусков — рекомендательная система.

Выявление аномалий

- Ищем данные, которые отлючаются от других;
- Например, количество пользователей сайта и их поведение;
- Отчеты о работе какой-либо системы;
- Необычные банковские операции;
- Иногда используется для предобработки данных.

Поиск ассоциативных правил

- Правила вида “если A, то скорее всего и B”;
- Пиво часто покупают с чипсами, а хлеб с молоком;
- Правила характеризуются достоверностью и поддержкой;
- Достоверность – это вероятность иметь B, если A уже есть;
- Поддержка – это вероятность встретить A и B вместе.
- Ассоциативные правила можно использовать для рекомендаций и поиска скрытых факторов.

Ансамбли моделей

Ансамбли моделей

- Одна голова хорошо, а две?
- *K* ближайших соседей.
- Если модели независимы, то простое голосование трех моделей превращает точность 0.6 в ???, а 0.9 в ???.

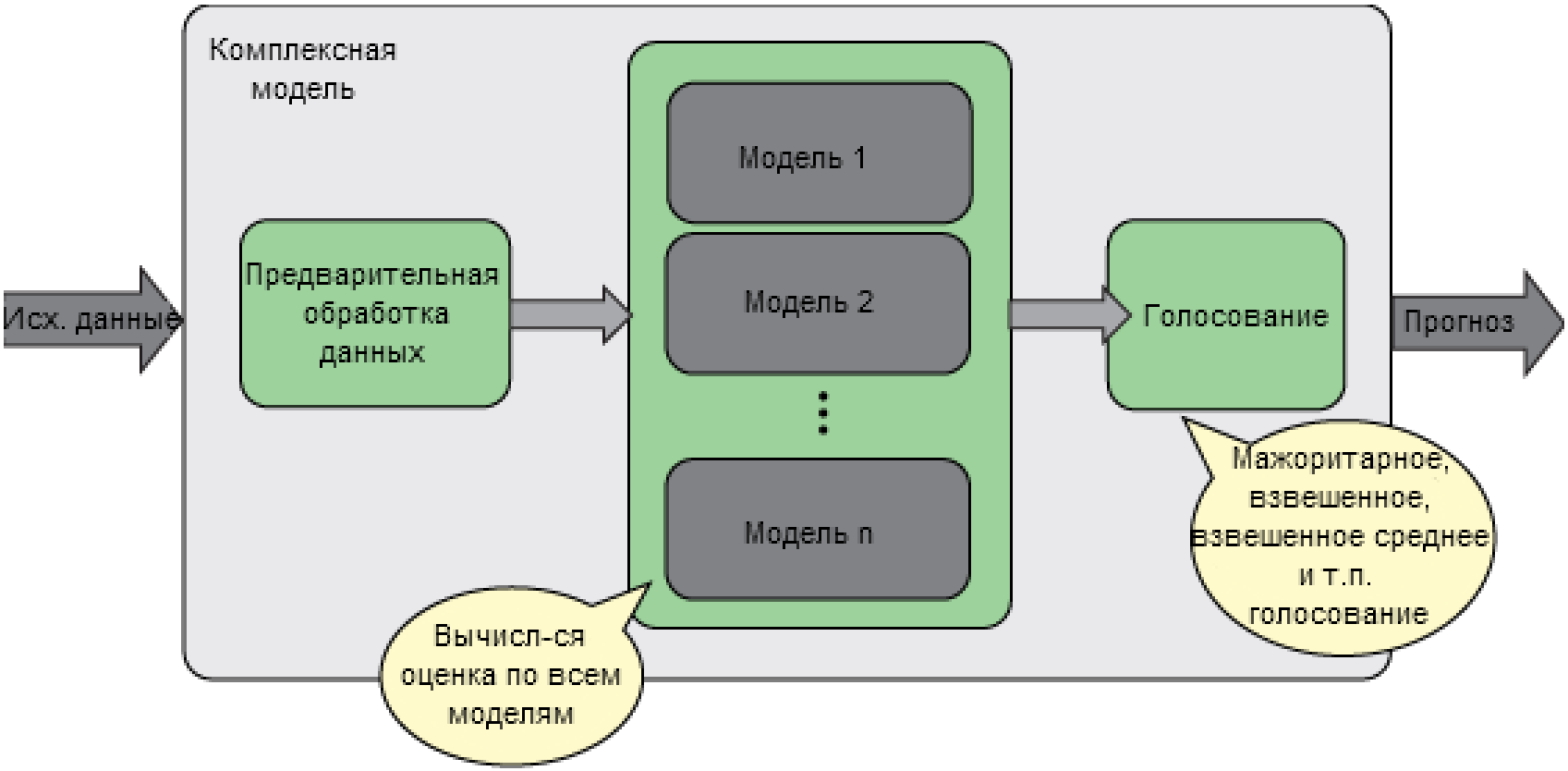
Ансамбли моделей

- Одна голова хорошо, а две?
- *K* ближайших соседей.
- Если модели независимы, то простое голосование трех моделей превращает точность 0.6 в 0.648, а 0.9 в 0.972.

Ансамбли моделей

- Одна голова хорошо, а две?
- *K* ближайших соседей.
- Если модели независимы, то простое голосование трех моделей превращает точность 0.6 в 0.648, а 0.9 в 0.972.
- То есть объединение моделей может улучшать результат.

Ансамбли моделей



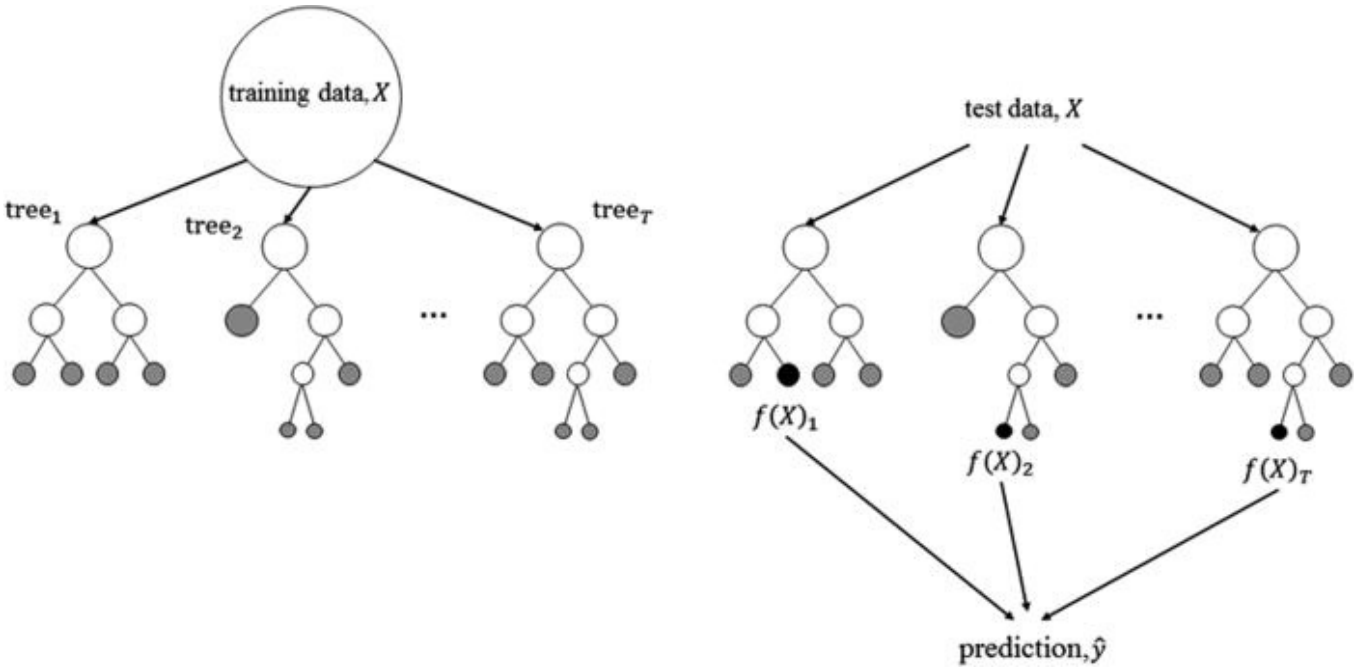
Ансамбли моделей

Можно объединять любые модели, просто устраивая голосование (для классификации) или беря взвешенную сумму (для регрессии), но есть несколько моделей построенных на идее ансамблей изначально:

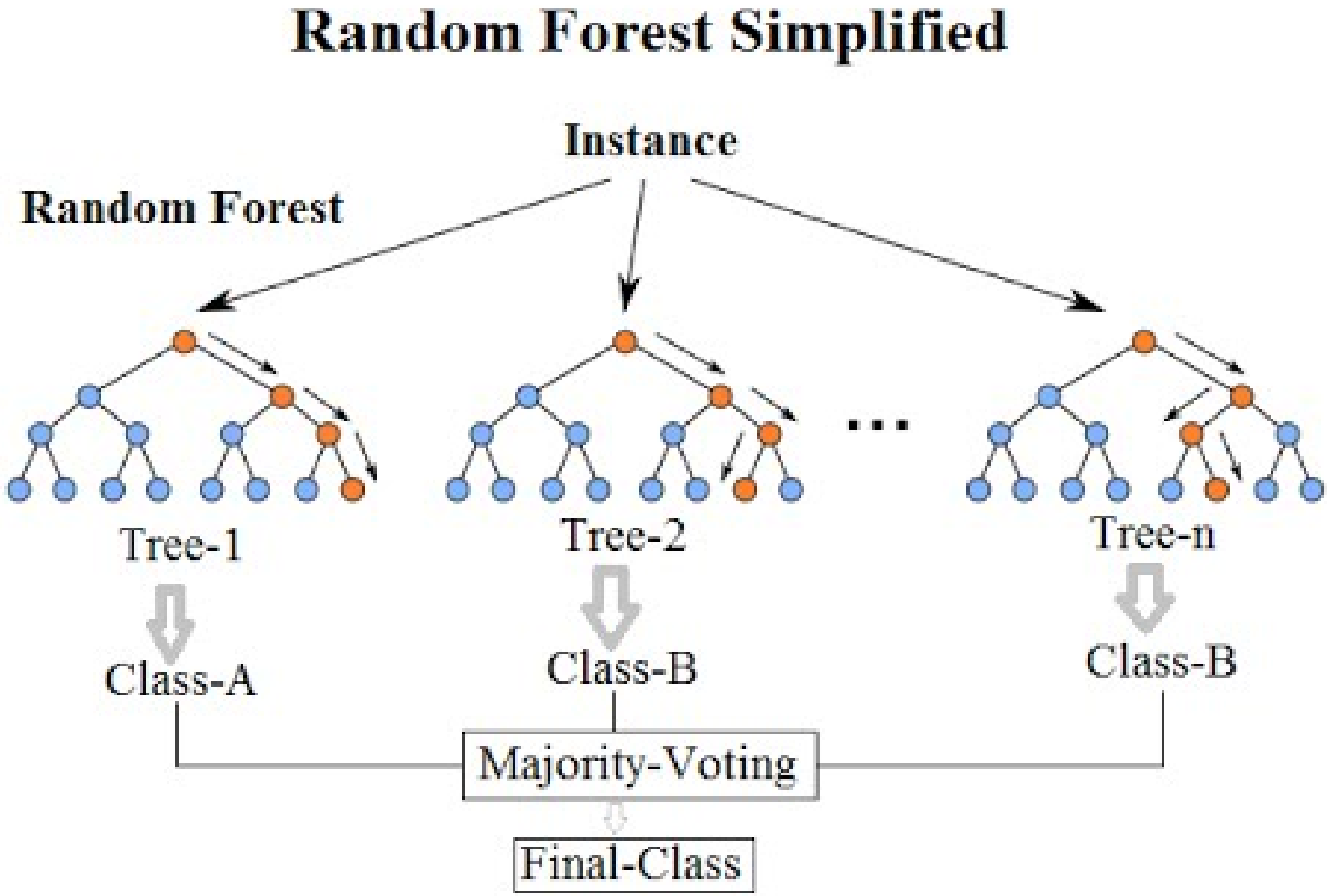
- случайный лес;
- градиентный бустинг.

Случайный лес

- Случайный лес – это ансамбль деревьев.
- Чтобы деревья были разными, они учатся не на всех признаках, а на случайном поднаборе.
- Это позволяет сделать модели «более независимыми».



Случайный лес



Случайный лес

- Достоинства:
  - Хорошо работает с большим числом признаков.
  - Не требует масштабирования признаков.
  - Хорошо параллелится.
  - Имеет встроенную “систему” оценки значимости параметров.
- Недостатки:
  - Большой размер модели.
  - Склонен к переобучению.

Градиентный бустинг

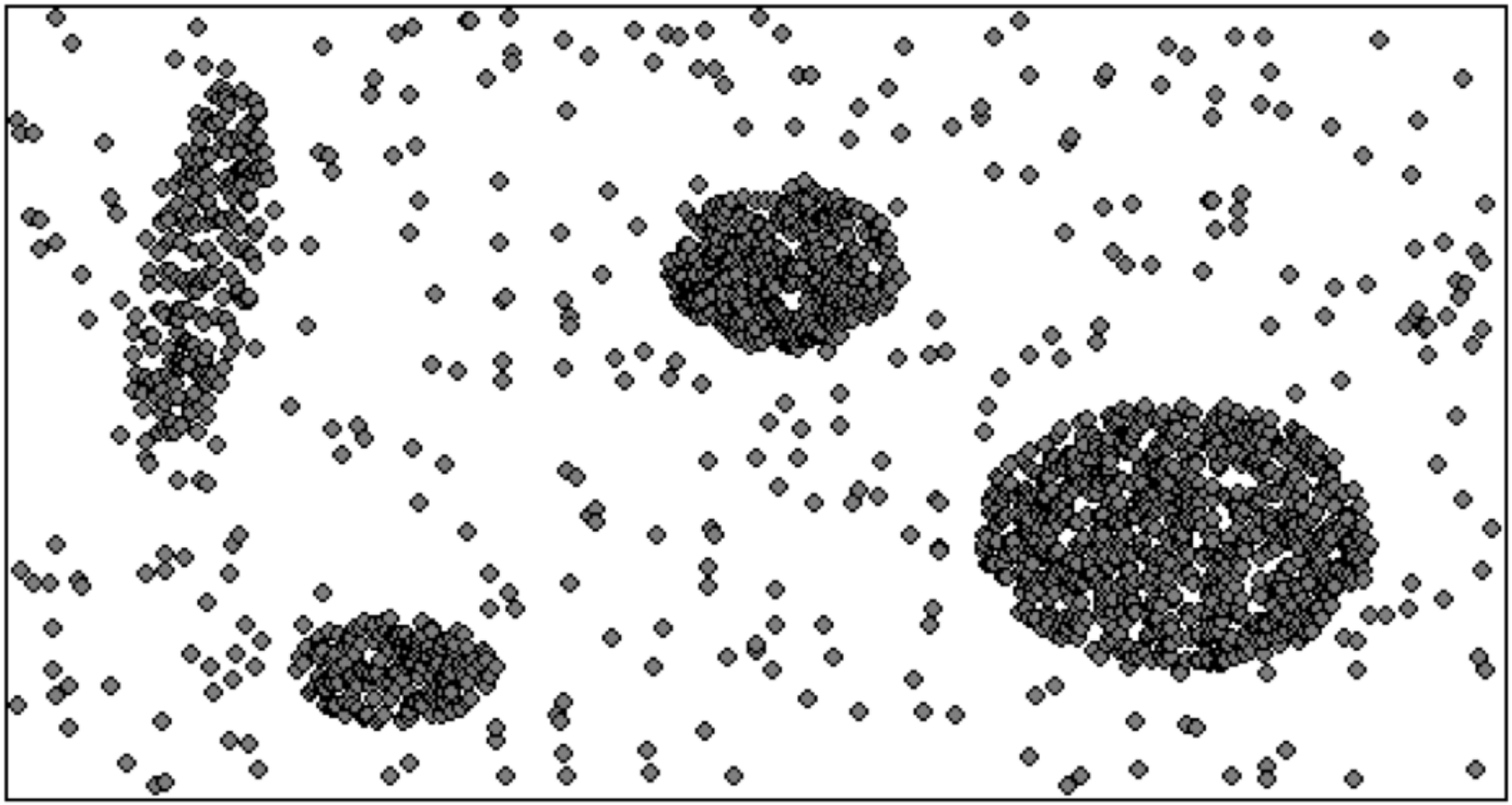
- Один из самых популярных алгоритмов.
- Весь Kaggle завален историями “как люди XGBoost’ы стекают”.
- Отличие от классических ансамблей в том, что новые подмодели строятся не для решения основной задачи, а для предсказания и исправления ошибки на уже построенной модели.

Кластеризация

Кластеризация

- Цель – объединение объектов в классы/группы
- Кластер – калька английского слова «cluster», которое переводится как “сгусток”, “гроздь (винограда)”, “скопление (звезд)” и т.п.
- Нужно, чтобы элементы внутри кластеров были как можно больше похожи друг на друга, а кластеры между собой отличались

Итоговая структура (дерево)



Виды кластеризации

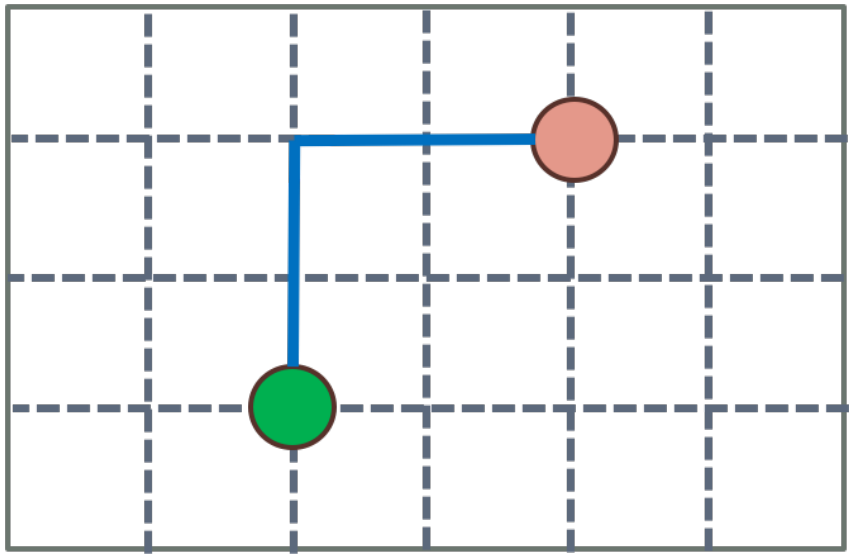
- Иерархическая, последовательно объединяем или разбиваем кластеры:
  - Агломеративная – считаем каждый объект кластером, на каждом шаге объединяем два ближайших кластера.
  - Разделительная – считаем все одним кластером, на каждом шаге делим какой либо кластер на два.
- Неиерархические – оптимизируем некую целевую функцию
  - Кластеризация на графах;
  - Алгоритм *k*-средних.
  - Алгоритмы основанные на EM.

Иерархическая кластеризация

- сначала каждая точка = кластер
- вычисляем расстояния, объединяем близкие точки в кластеры
- считаем расстояния между кластерами, объединяем близкие кластеры в большие кластеры
- ...

Меры сходства точек

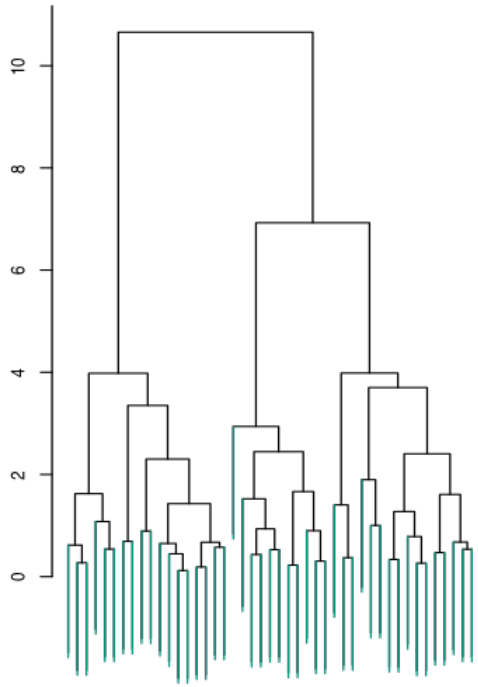
- Евклидово расстояние
- Квадрат Евклидова расстояния
- Блок (Манхеттен, сити-блок)
- и так далее…

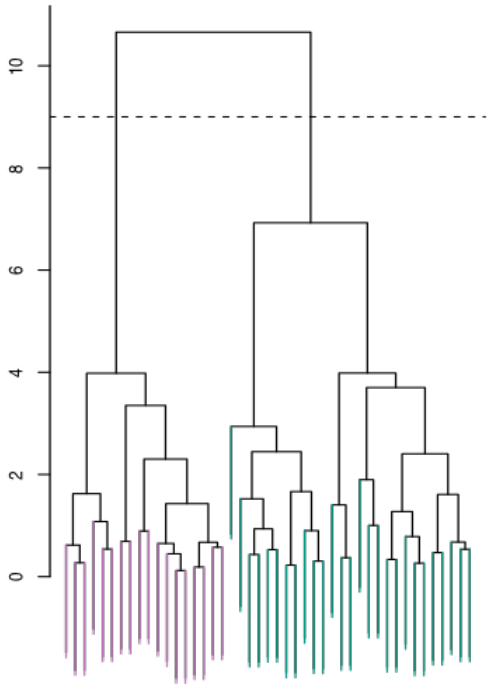


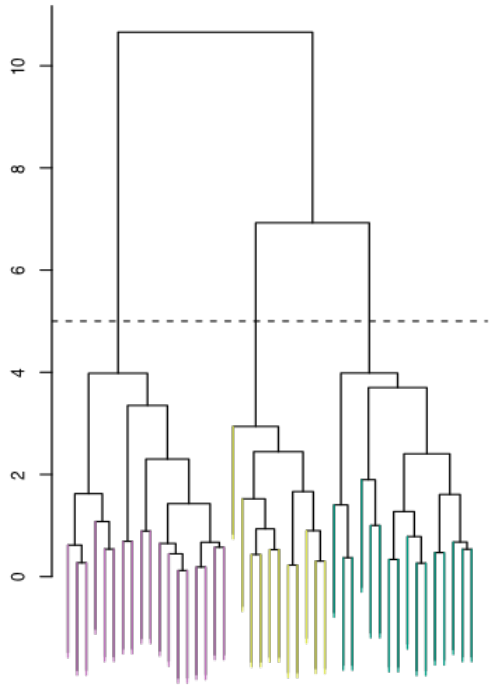
Меры сходства кластеров

- 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).

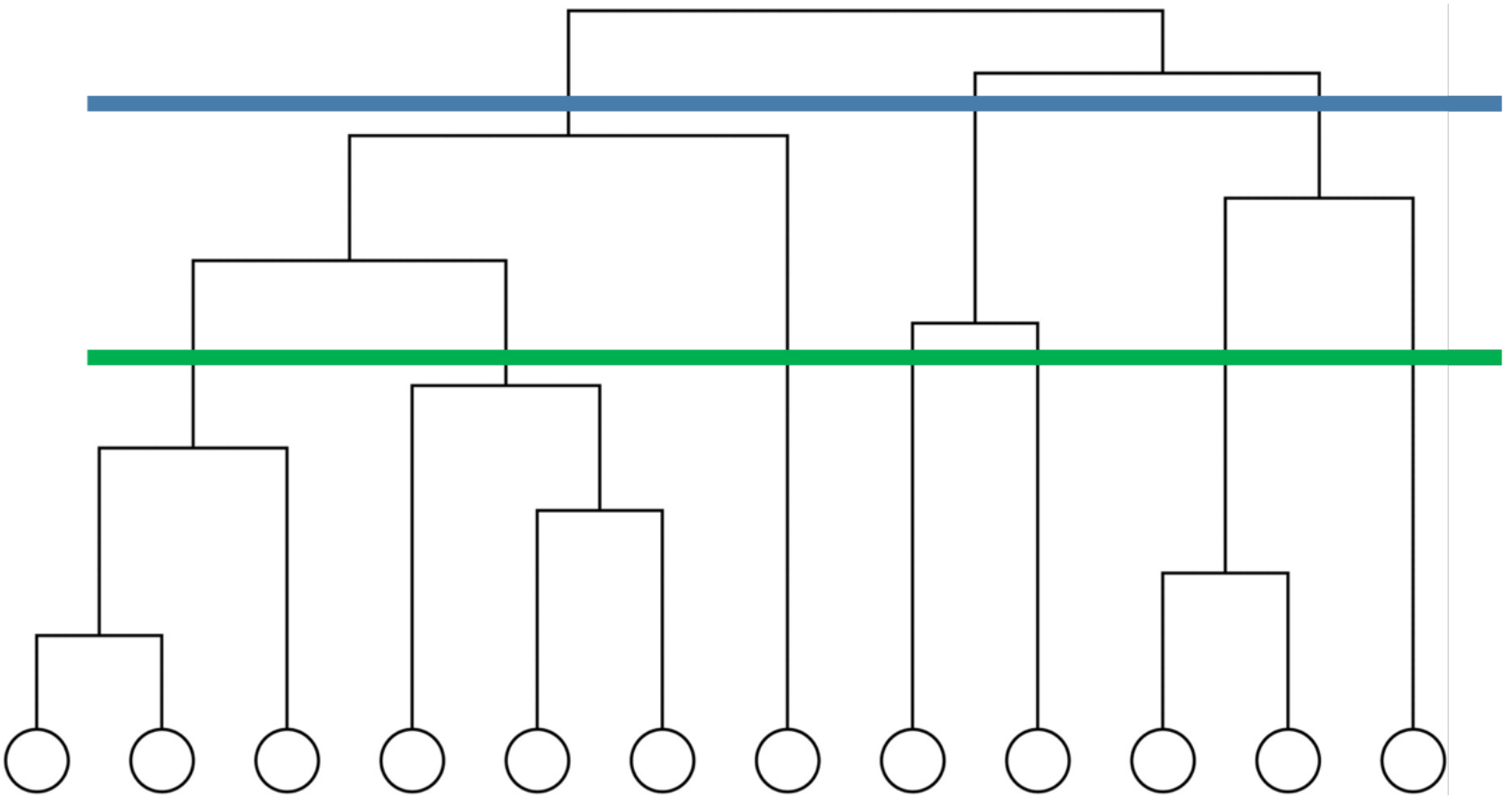
Итоговая структура (дерево)



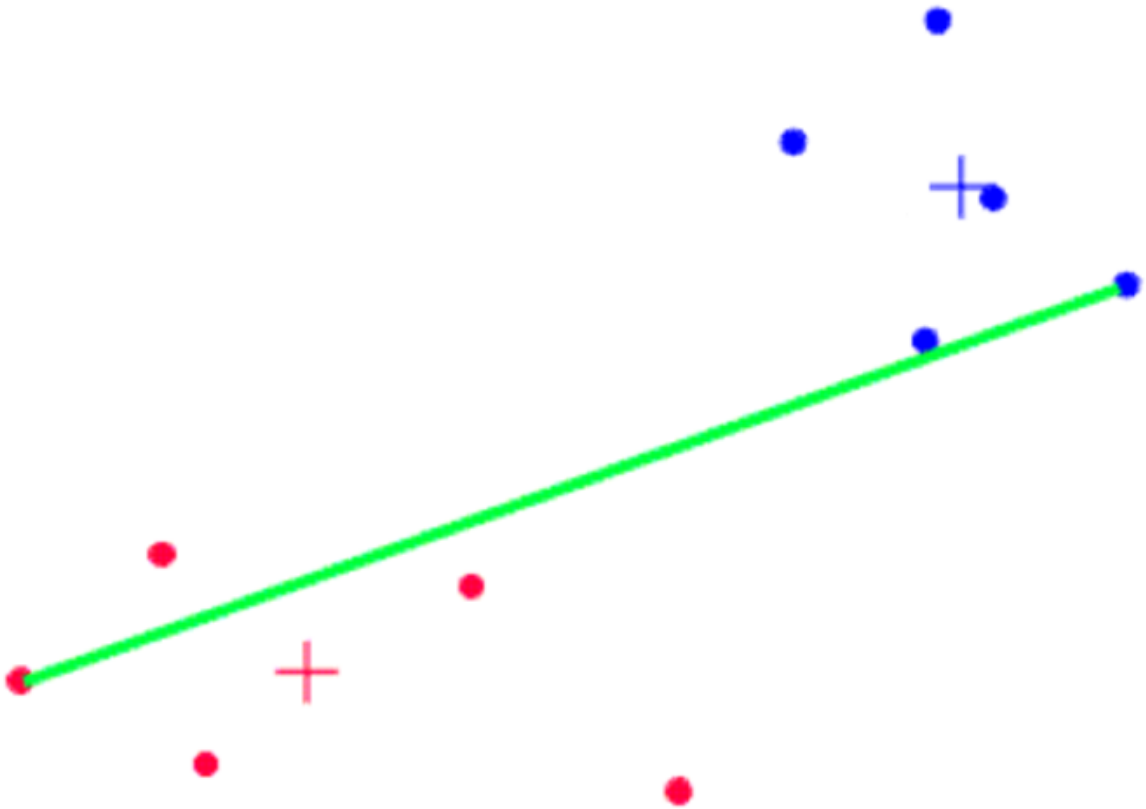




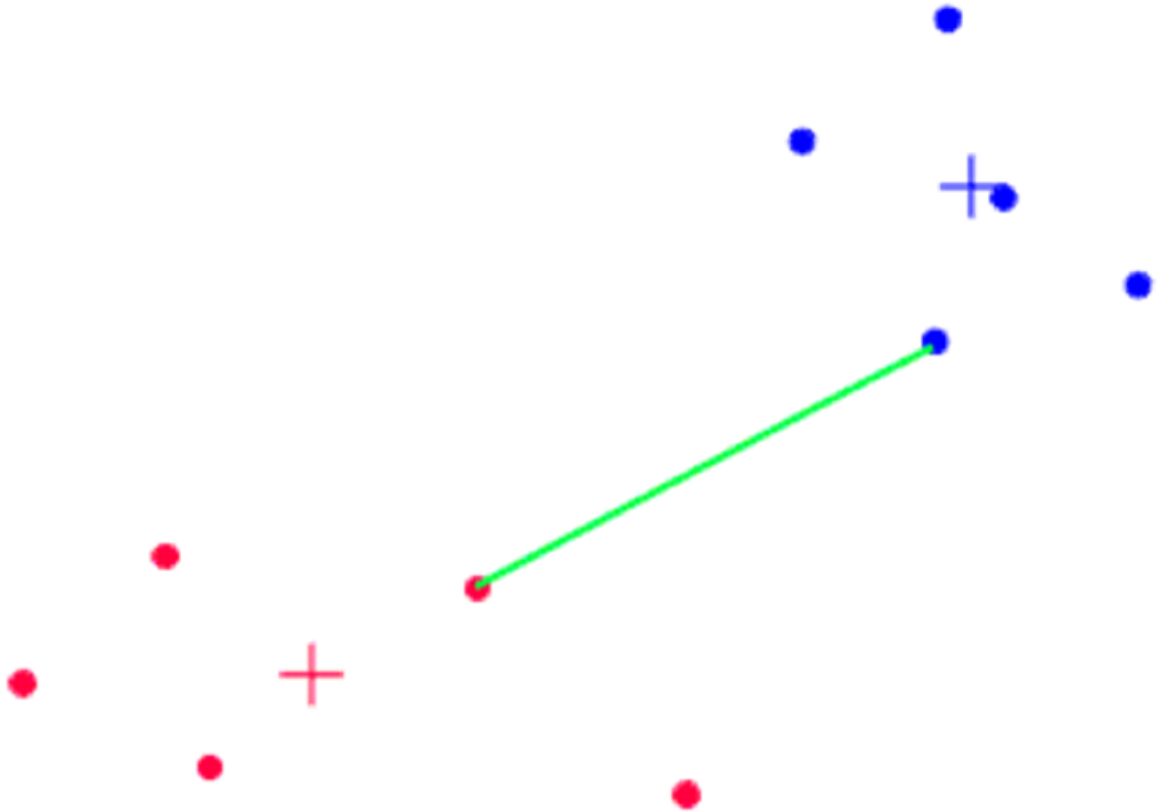
Итоговая структура (дерево)



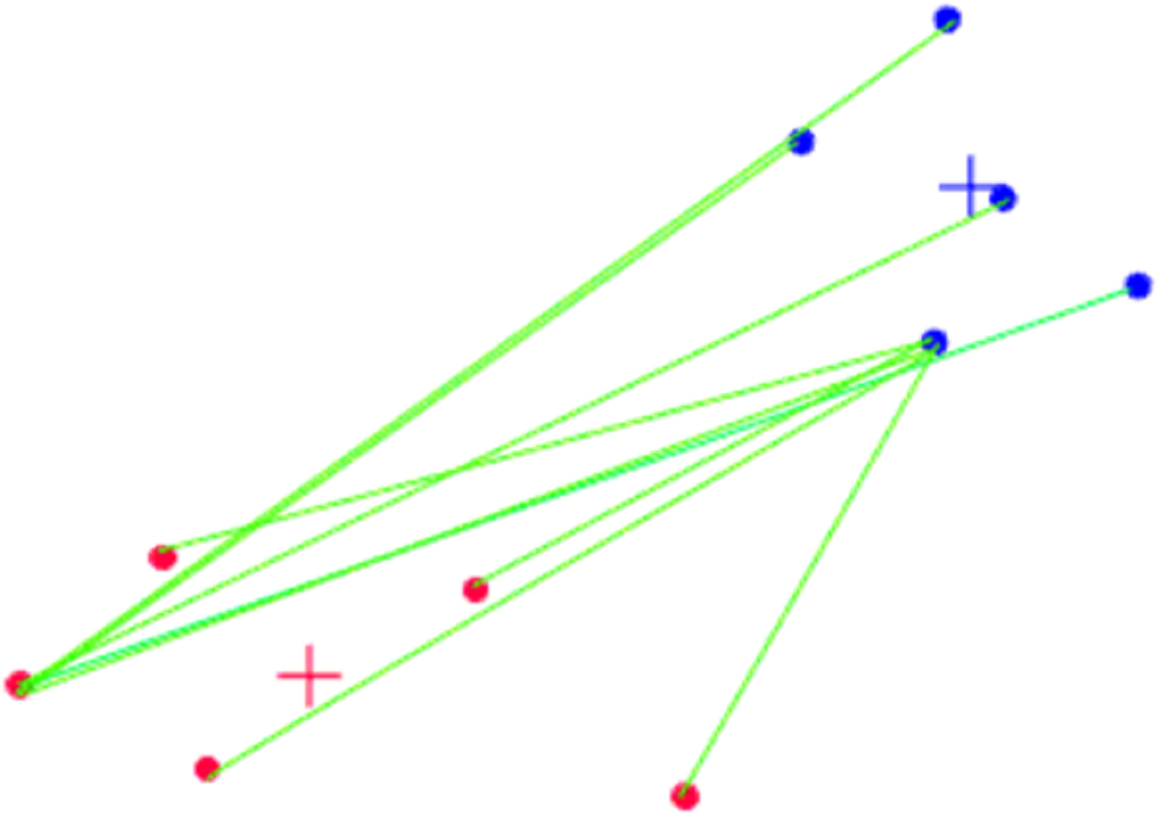
Метод дальнего соседа, максимального расстояния



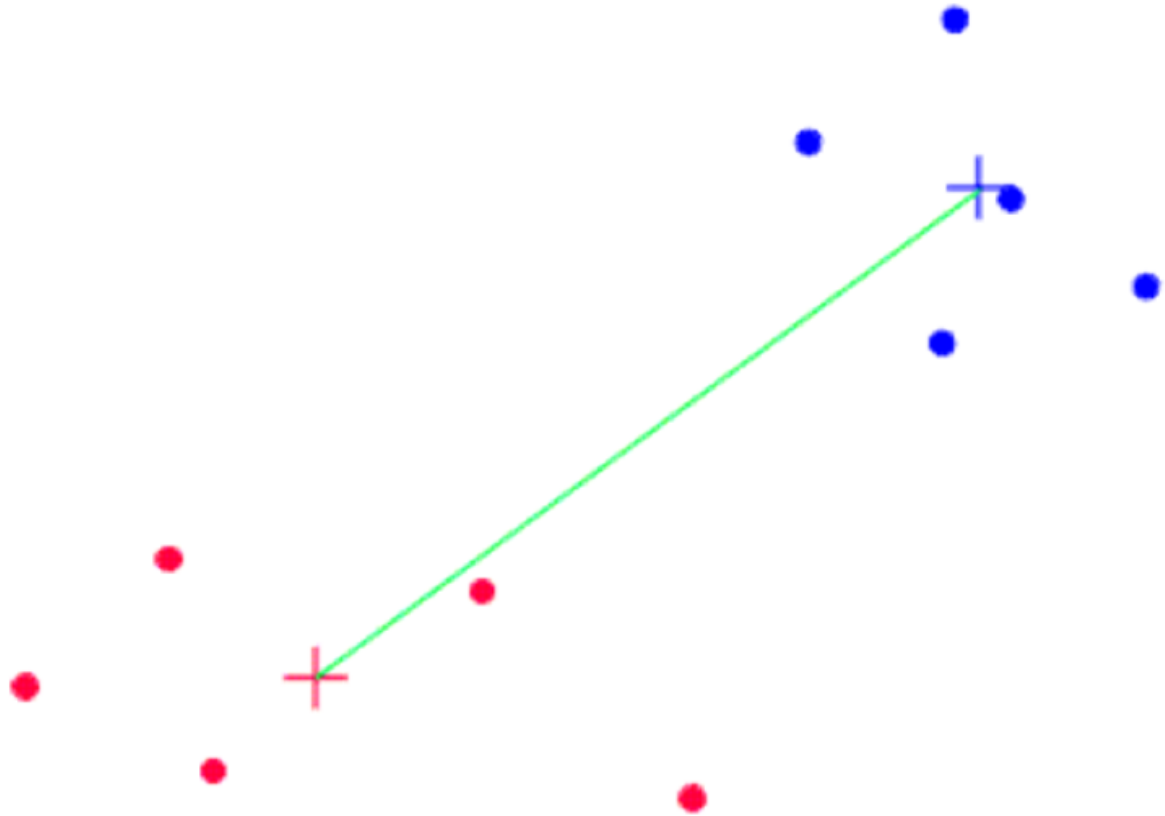
Метод ближайшего соседа



Среднее невзвешенное расстояние



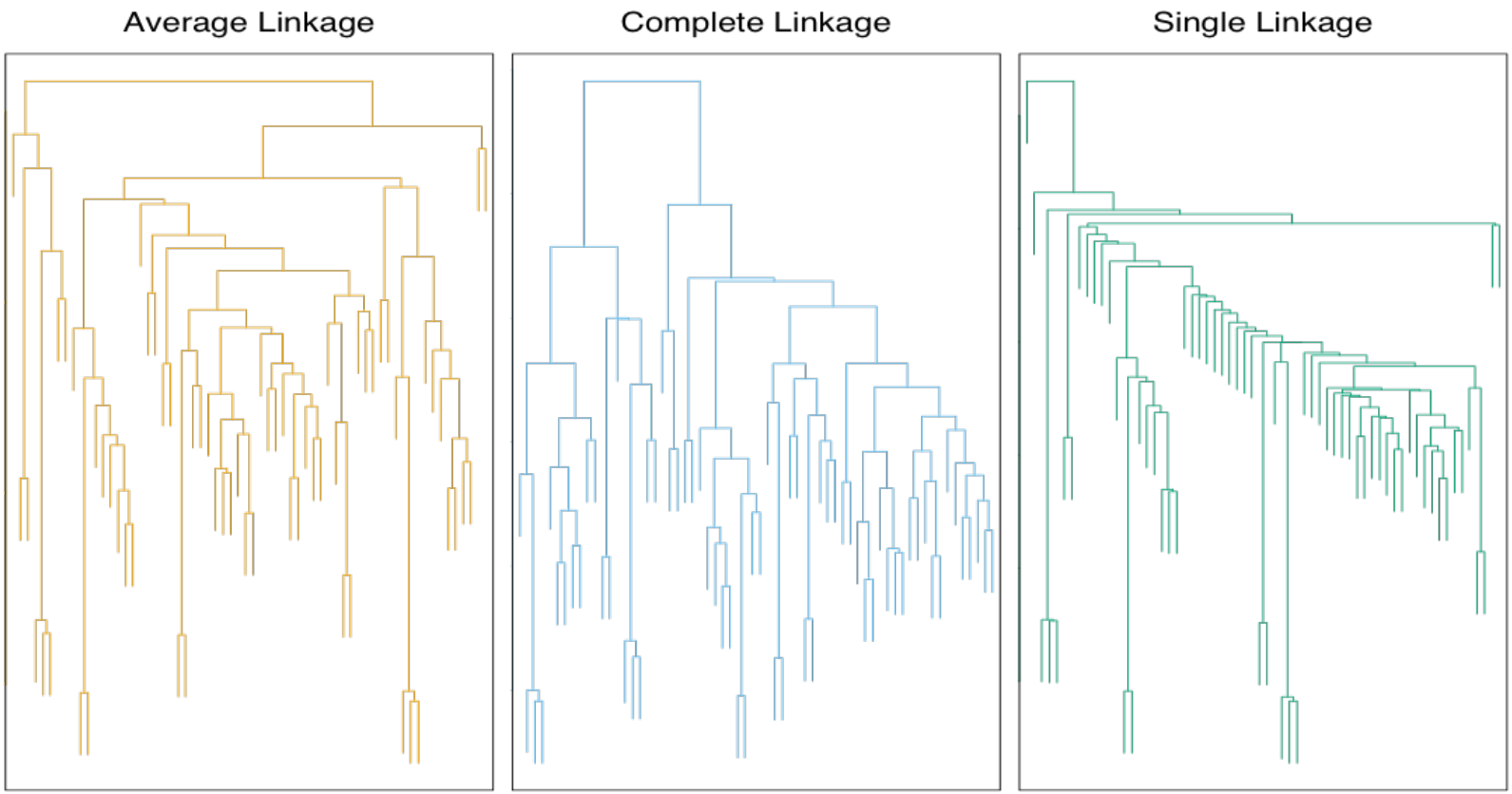
Центроидный метод



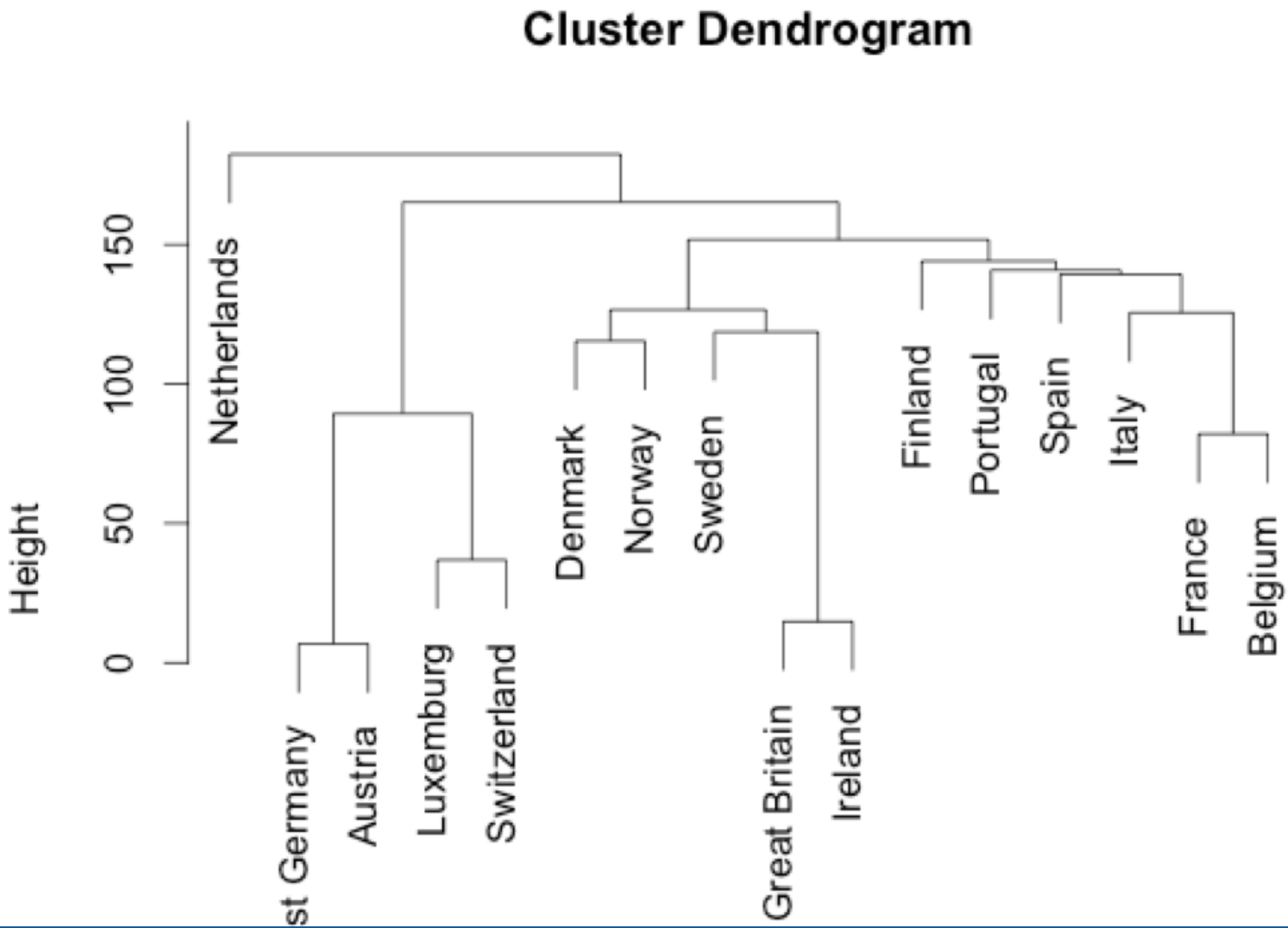
Метод Варда

- Сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных.
- Затем вычисляются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до этого кластерного среднего значения.
- Эти дистанции суммируются.
- Потом в один новый кластер объединяются те кластера, при объединении которых получается наименьший прирост общей суммы дистанций.

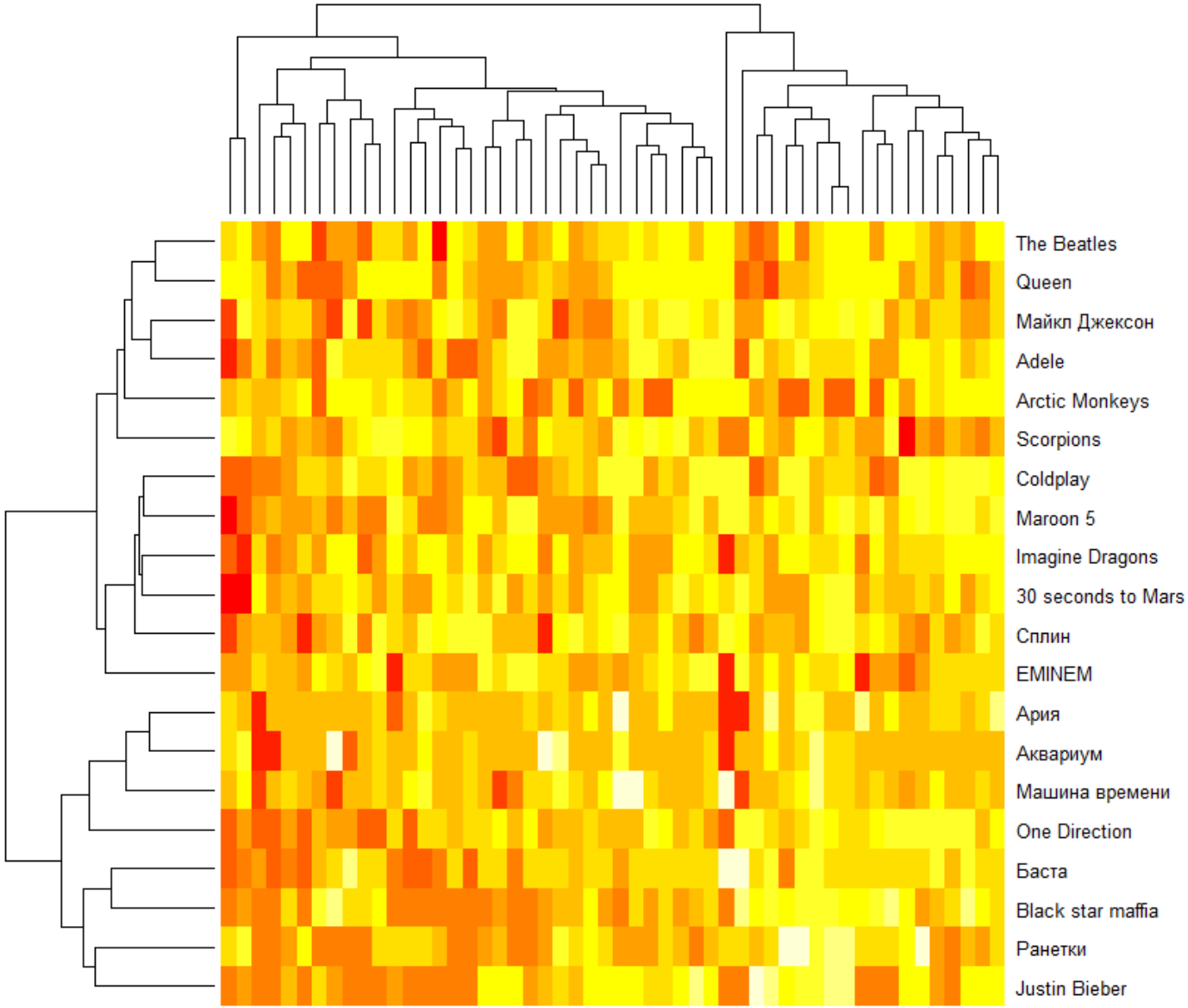
Выбор метода имеет значение



Иерархическая кластеризация: пример



Иерархическая кластеризация: пример 2



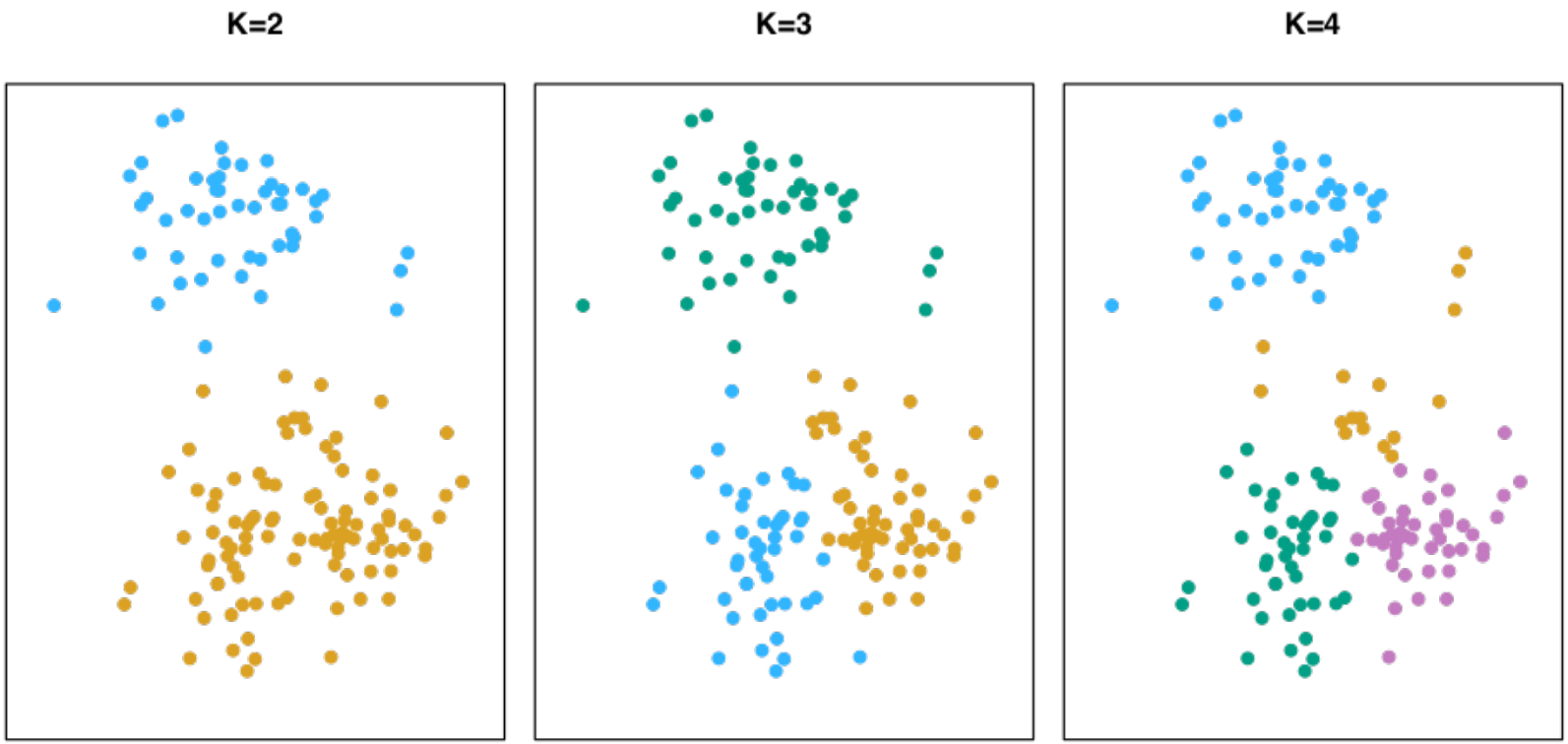
Метод k-средних

1. Выбрать k случайных центров
2. Приписать каждый объект из множества исходных данных кластеру исходя из того, какой центр к нему ближе
3. Пересчитать центры кластеров используя полученное распределение объектов
4. Если алгоритм не сошелся, то перейти к п. 2.

Типичные критерии схождения алгоритма это либо среднеквадратичная ошибка, либо отсутствие перемещений объектов из кластера в кластер.

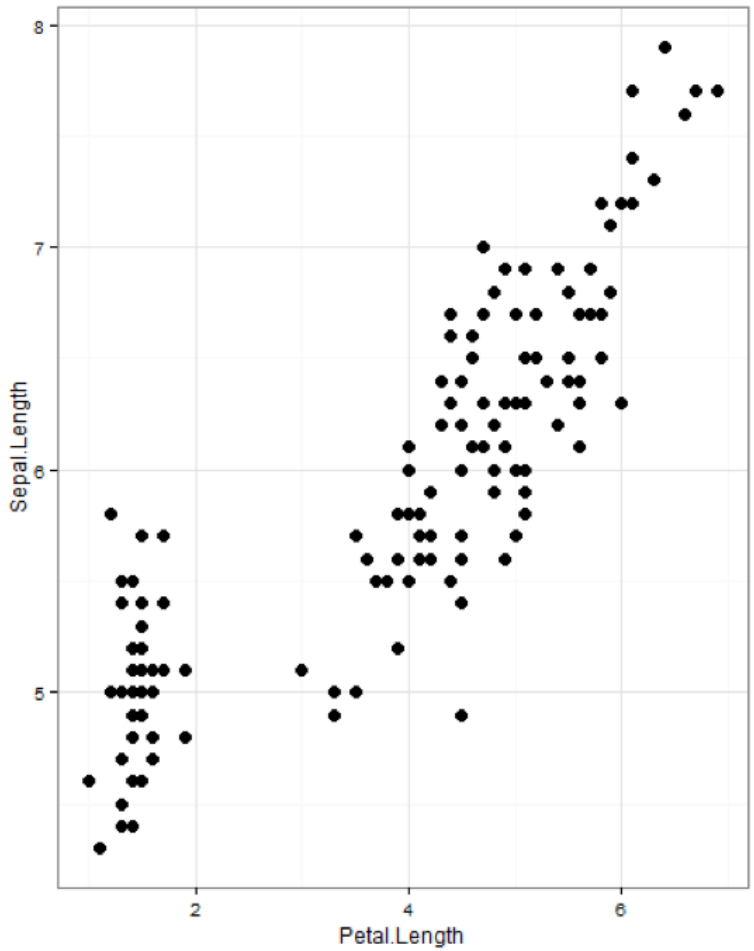
Метод k-средних

Выбор числа кластеров осуществляется заранее

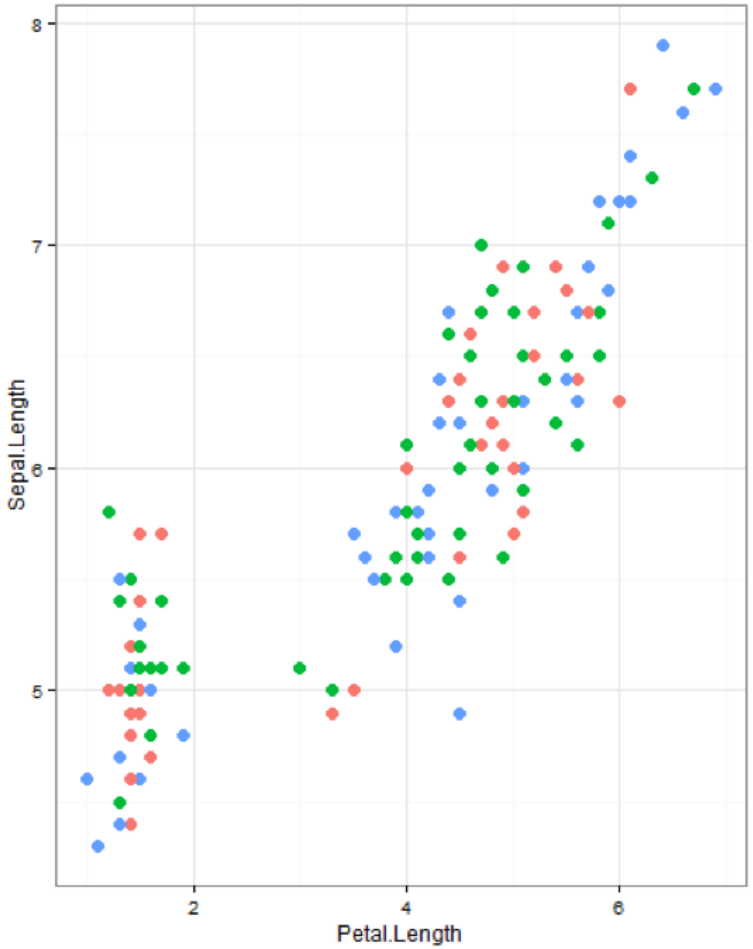


Шаг 1

Исходные данные

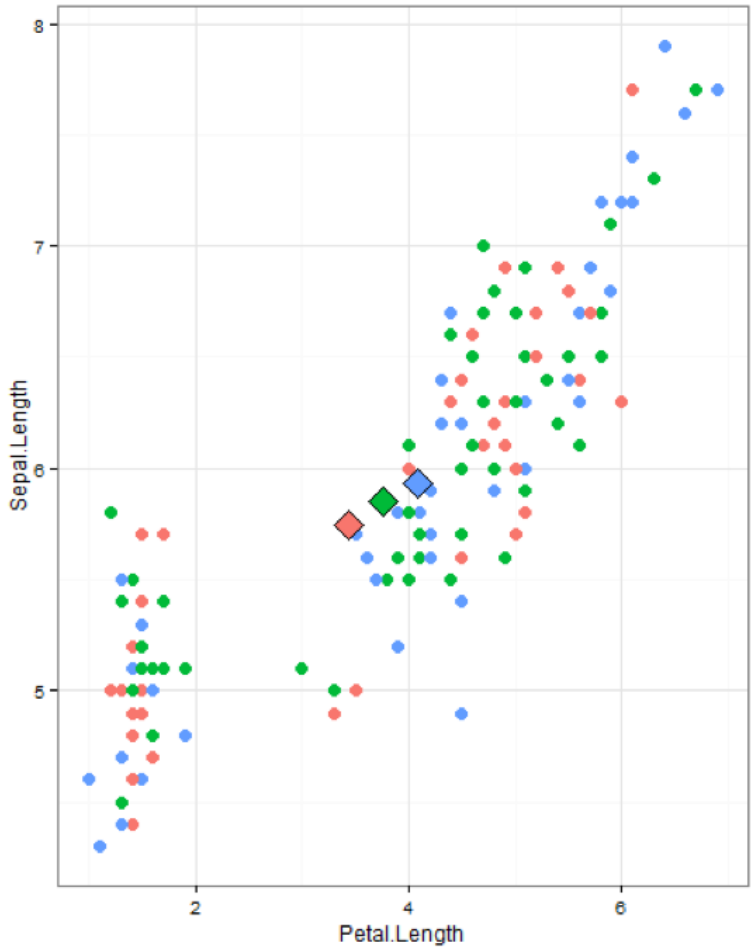


Случайное распределение точек

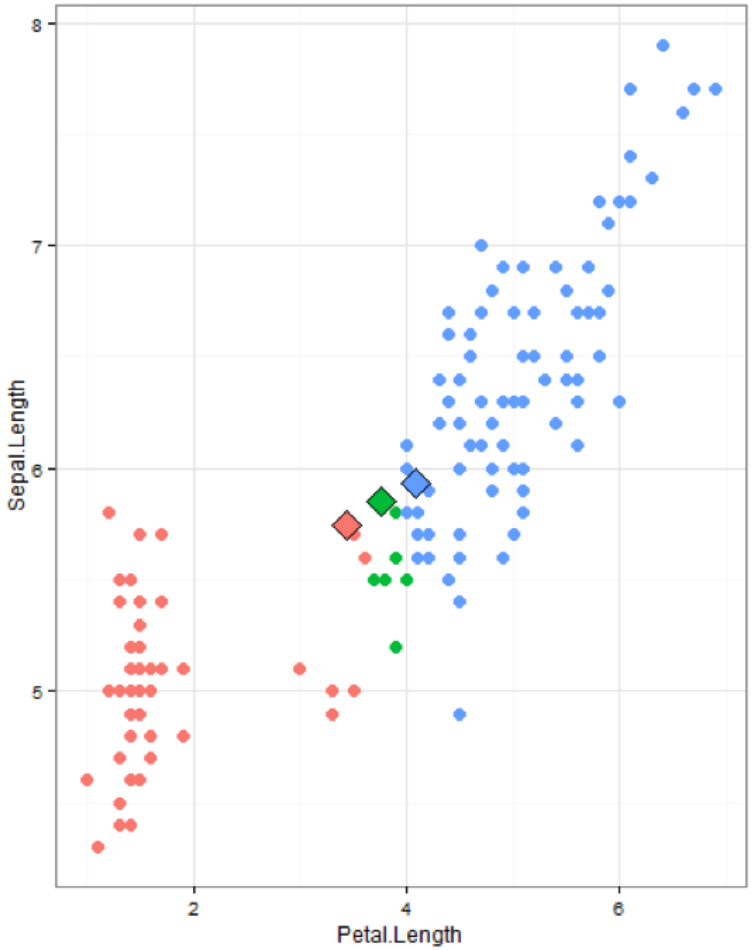


Шаг 2

Расчет центров кластеров

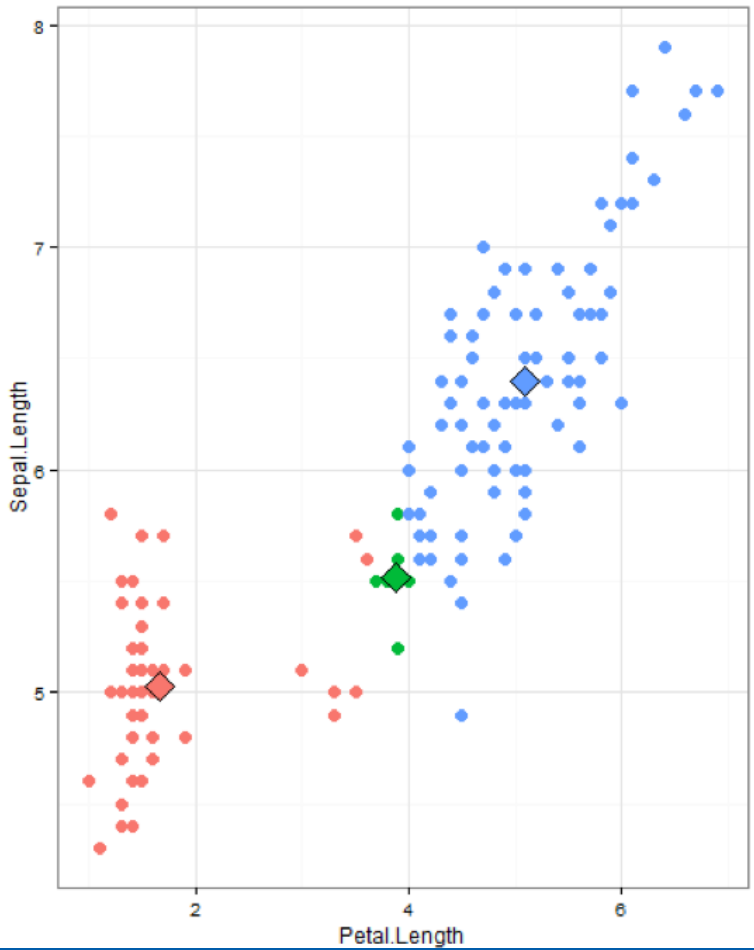


Перераспределение точек

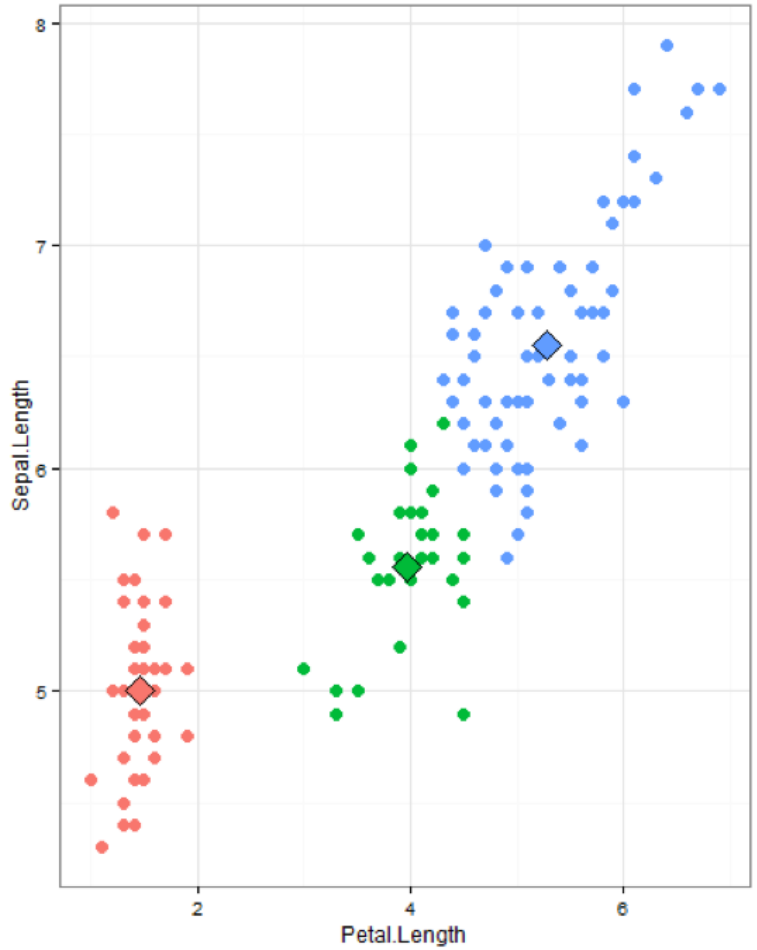


Шаг 3+

Расчет новых центров кластеров



... Окончательное разделение



Еще о кластеризации

- разные методы дают разные результаты
- в большей степени описательная, исследовательская техника, а не “абсолютная правда”
- результаты требуют дальнейшего исследования
- выбор методов отпределяется задачей
  - если мы хотим найти покупателей с одинаковым паттерном покупок, какую меру сходства мы возьмем?

Вместо заключения