Современные методы анализа данных и машинного обучения

Тема 7. Лекция 10

Классическое машинное обучение. Обучение без учителя. Кластеризация

Юрий Саночкин

ysanochkin@hse.ru

НИУ ВШЭ, 2024

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

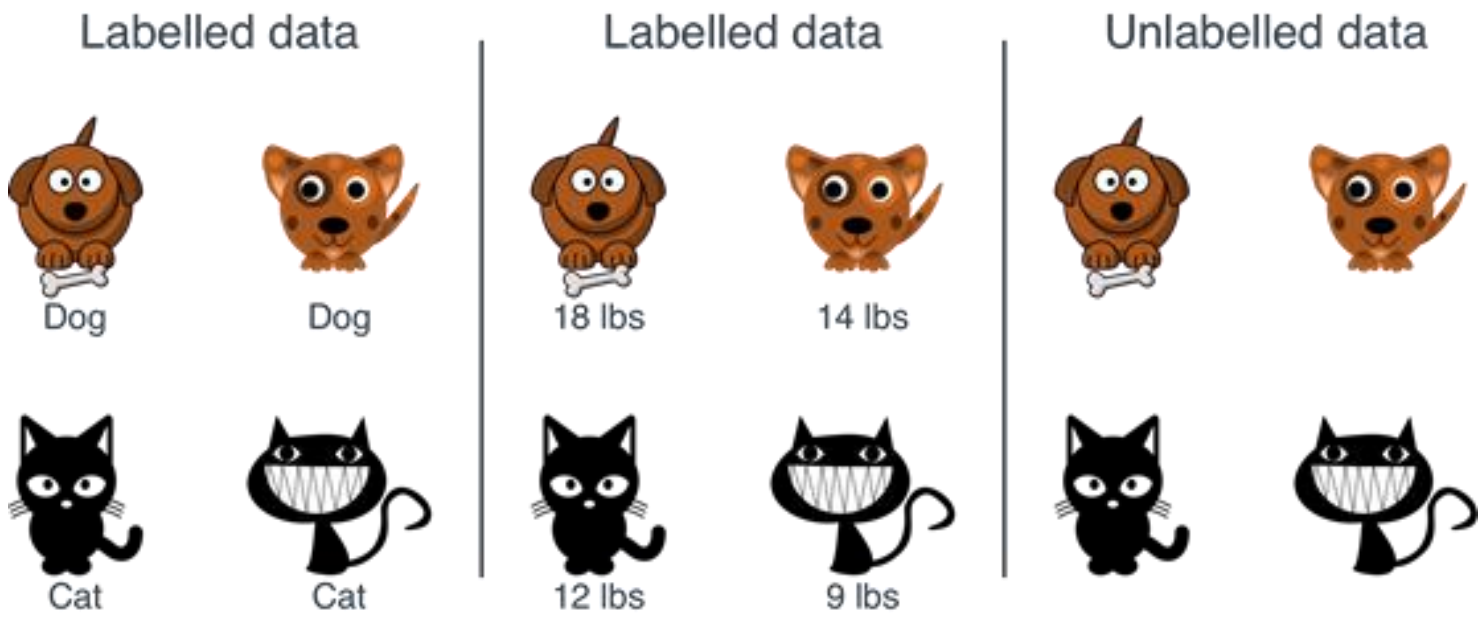
Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Размеченные (labelled) vs неразмеченные (unlabelled) данные



Классическое Обучение

Данные заранее категоризированы или численные

С учителем

Предсказать категорию

Классификация

«Разложи носки по цвету»

Предсказать значение

Регрессия

«Разложи галстуки по длине»

Данные никак не размечены

Без учителя

Разделить по схожести

Кластеризация

«Разложи похожие вещи по кучкам»

Найти зависимости

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Выявить последовательности

Ассоциация

«Найди какие вещи я часто ношу вместе»

Сегодня мы будем обсуждать задачи unsupervised learning

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»

Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Вот это у нас уже где-то было… Кстати где?

Классическое Обучение

Данные заранее категоризированы или численные

Данные никак не размечены

С учителем

Предсказать категорию

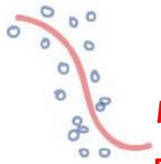
Предсказать значение

Классификация

«Разложи носки по цвету»

Регрессия

«Разложи галстуки по длине»



Без учителя

Разделить по схожести

Найти зависимости

Выявить последовательности

Кластеризация

«Разложи похожие вещи по кучкам»

Ассоциация

«Найди какие вещи я часто ношу вместе»

Уменьшение Размерности (обобщение)

«Собери из вещей лучшие наряды»

Ну а предметно мы с вами сегодня поговорим вот про этот блок

Задача кластеризации

- Что такое задача кластеризации?

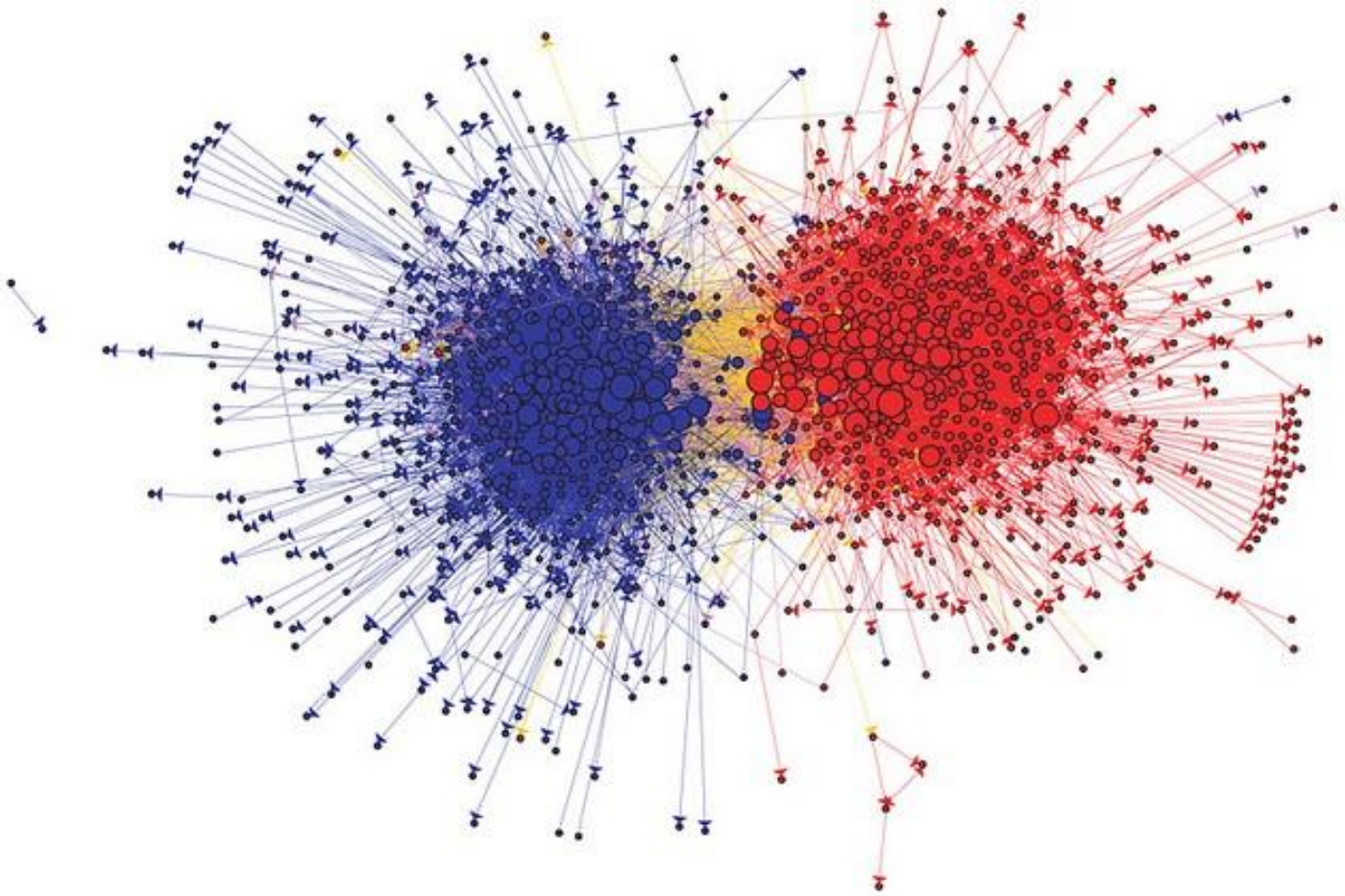
Задача кластеризации

- Что такое задача кластеризации?
- Говоря по простому — задача, где мы хотим разделить наши объекты на группы (сегменты), заранее не зная критерии и принципы разделения, но при этом так, чтобы объекты в группах были максимально похожи между собой

Задача кластеризации

- Что такое задача кластеризации?
- Говоря по простому — задача, где мы хотим разделить наши объекты на группы (сегменты), заранее не зная критерии и принципы разделения, но при этом так, чтобы объекты в группах были максимально похожи между собой
- Приведите примеры каких-нибудь задач кластеризации
  - Сегментация аудитории для таргетирования рекламы
  - Идентификация типов клеток в образце данных секвенирования
  - Поиск сообществ в социальном графе (из соцсети или из инсайдерской информации о структуре организации)
  - Задача разделения смеси распределений
  - И так далее

Задача кластеризации



Задача кластеризации

- В классических задачах unsupervised learning есть X, но нет обучающей выборки, т.е. мы не знаем и не имеем правильных ответов.

Задача кластеризации

- В классических задачах unsupervised learning есть X, но нет обучающей выборки, т.е. мы не знаем и не имеем правильных ответов.
- И как же мы тогда будем действовать, чтобы убедиться, что наш алгоритм делает что-то значимое, а не полную ерунду?

Задача кластеризации

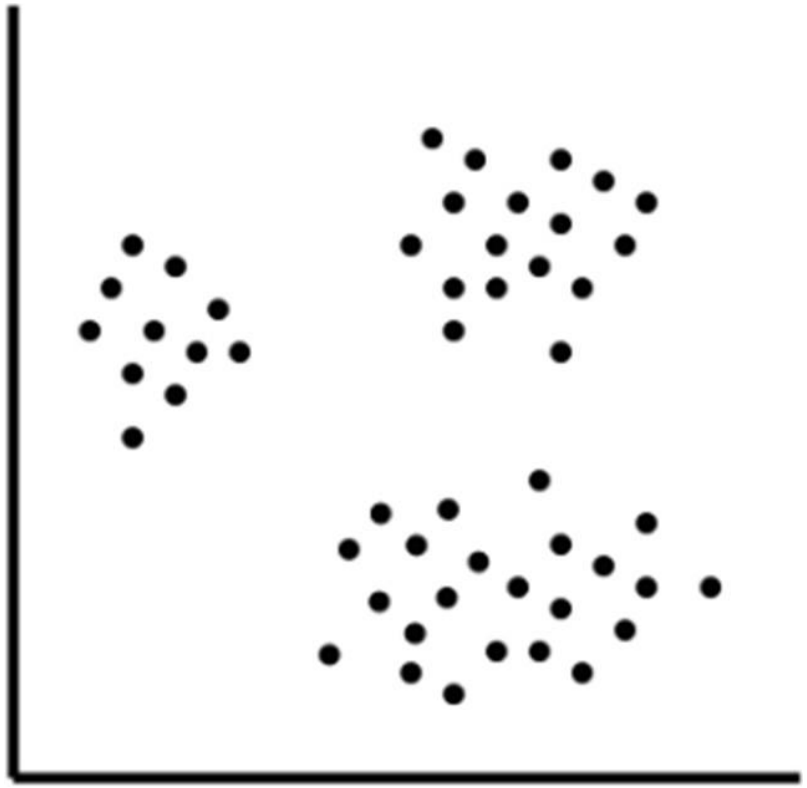
- В классических задачах unsupervised learning есть X, но нет обучающей выборки, т.е. мы не знаем и не имеем правильных ответов.
- И как же мы тогда будем действовать, чтобы убедиться, что наш алгоритм делает что-то значимое, а не полную ерунду?
- В таких задачах обычно минимизируют “энтропию” системы (или меру “хаоса”): ищут наиболее удачную расстановку меток с точки зрения разделимости наших объектов.

Задача кластеризации

- Рассмотрим основную идею задачи кластеризации:

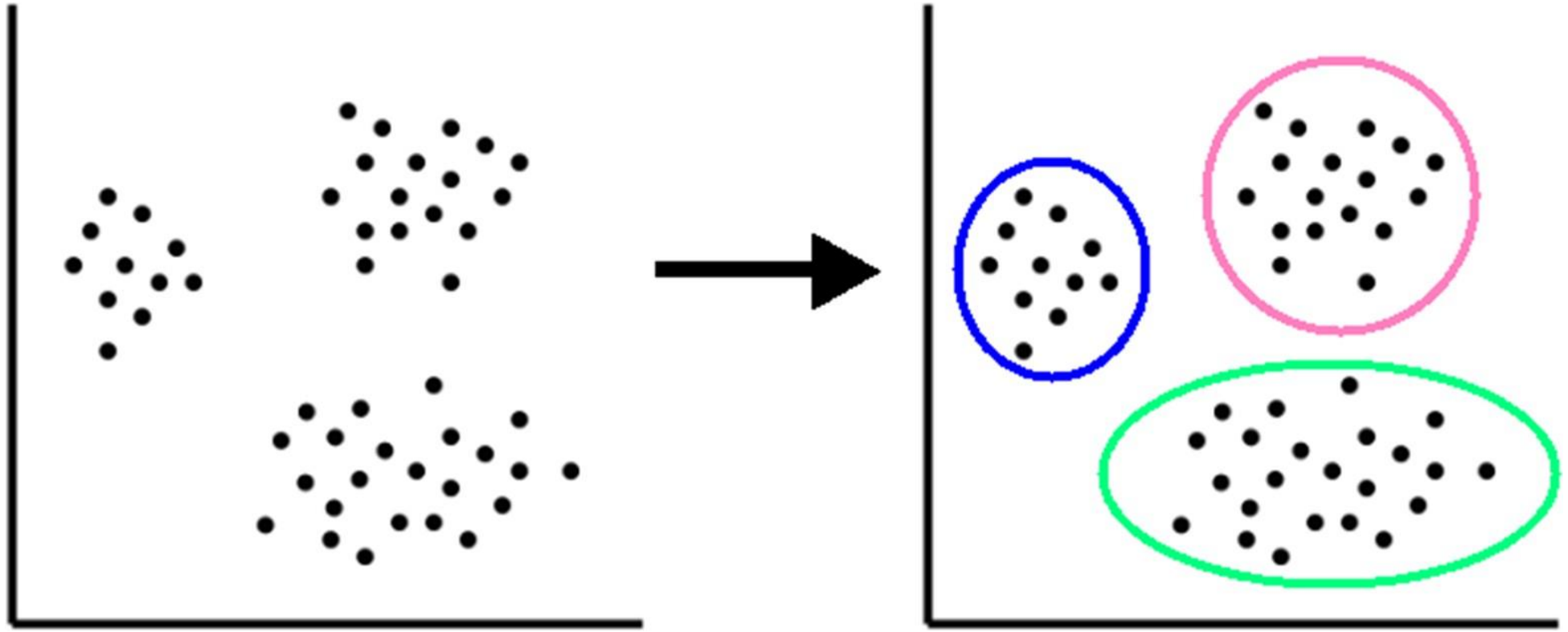
Задача кластеризации

- Рассмотрим основную идею задачи кластеризации:

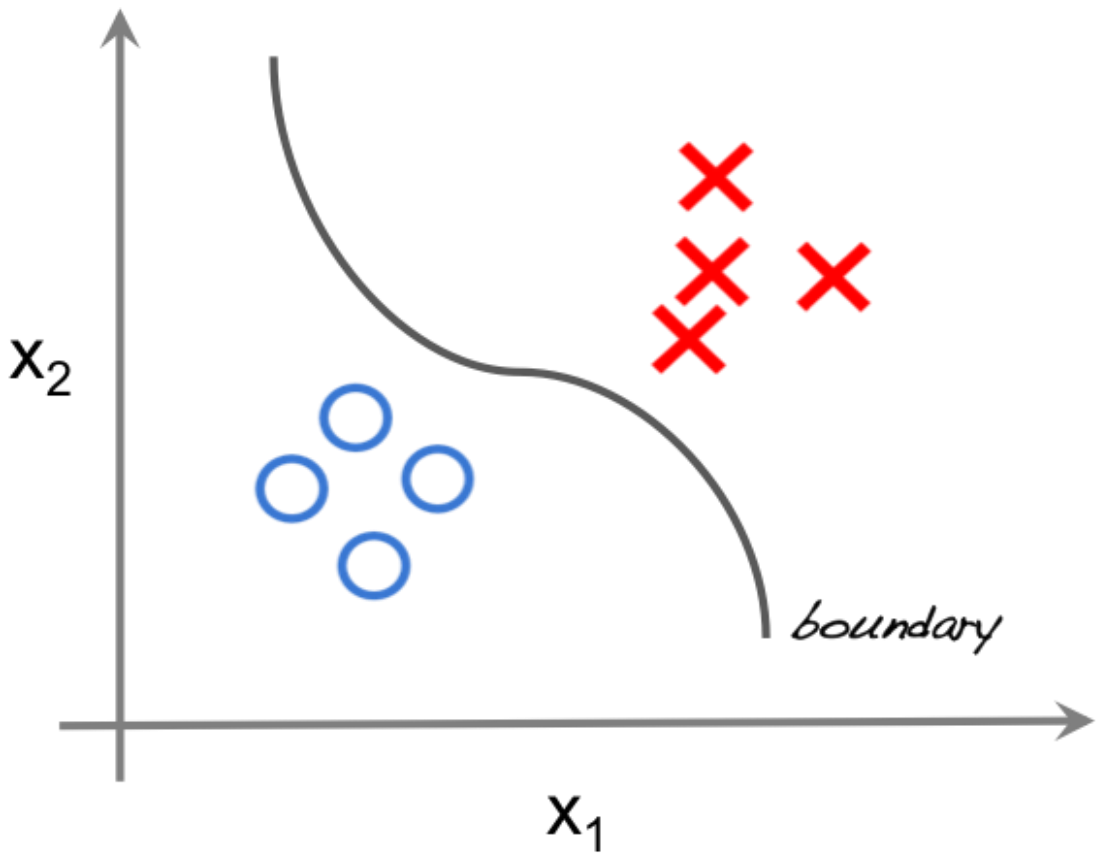


Задача кластеризации

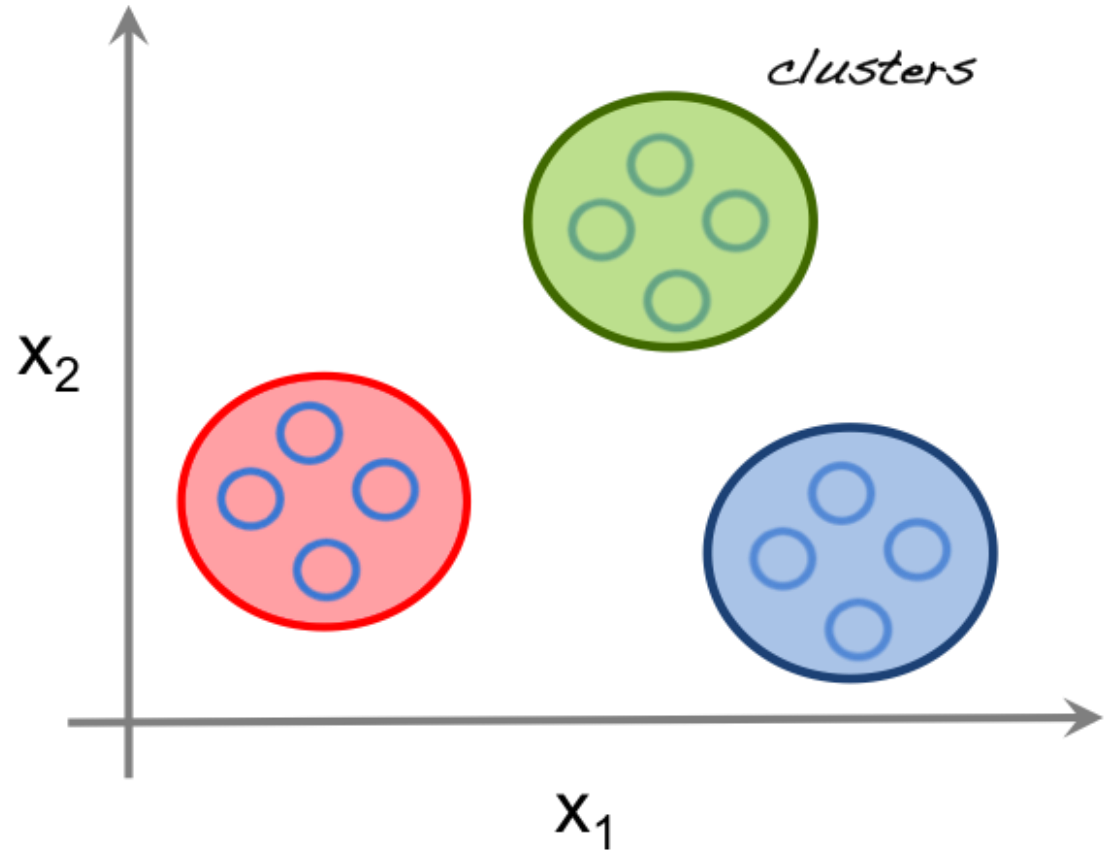
- Рассмотрим основную идею задачи кластеризации:



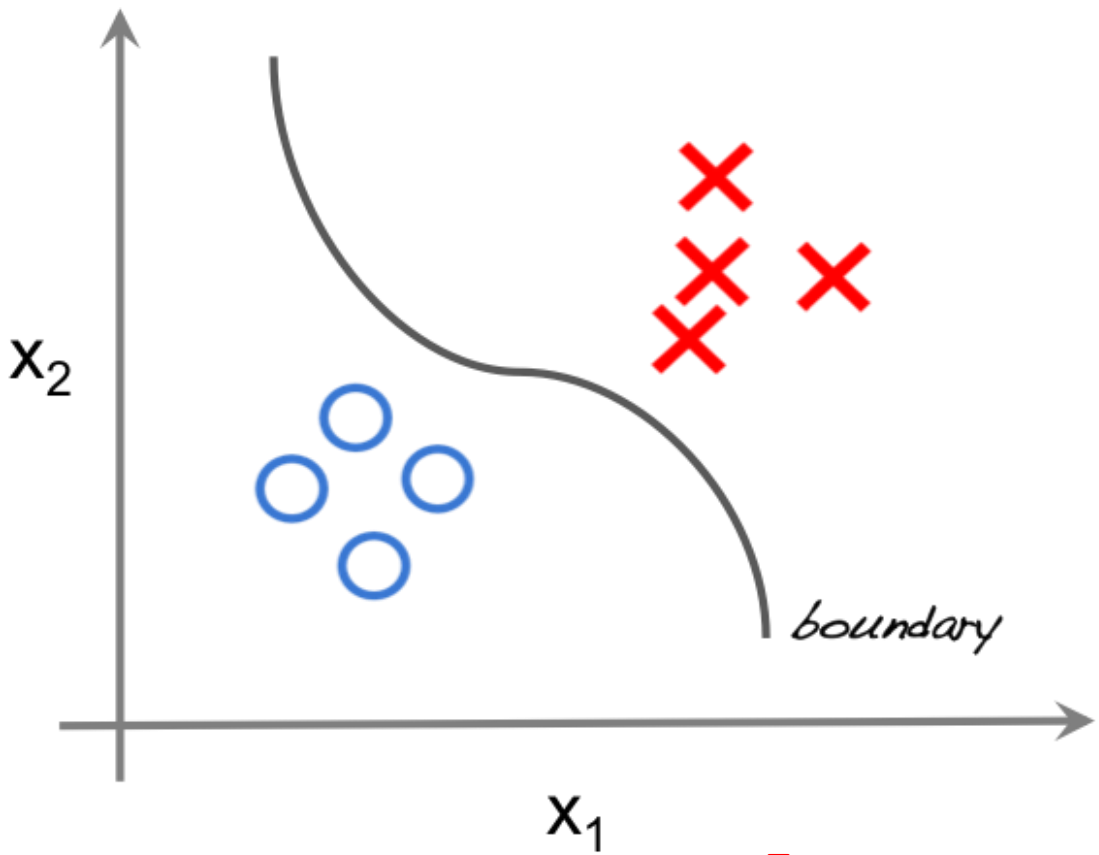
Supervised learning



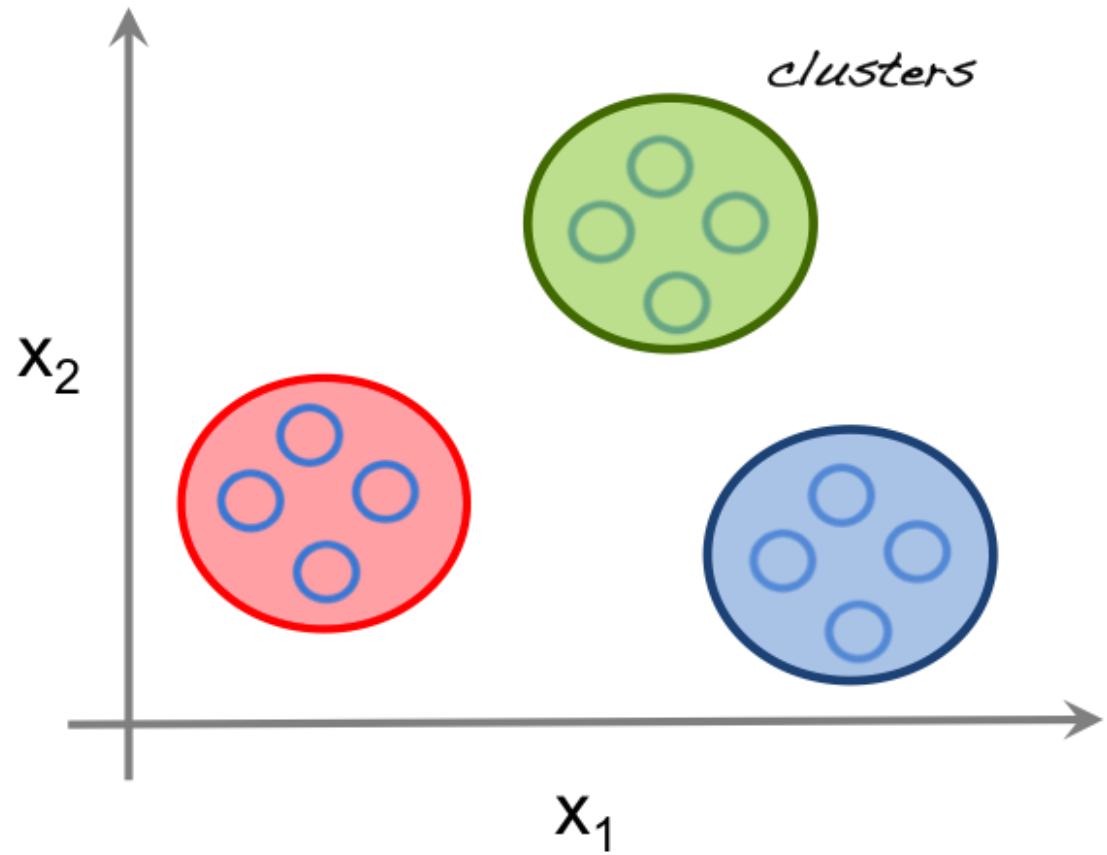
Unsupervised learning



Supervised learning



Unsupervised learning

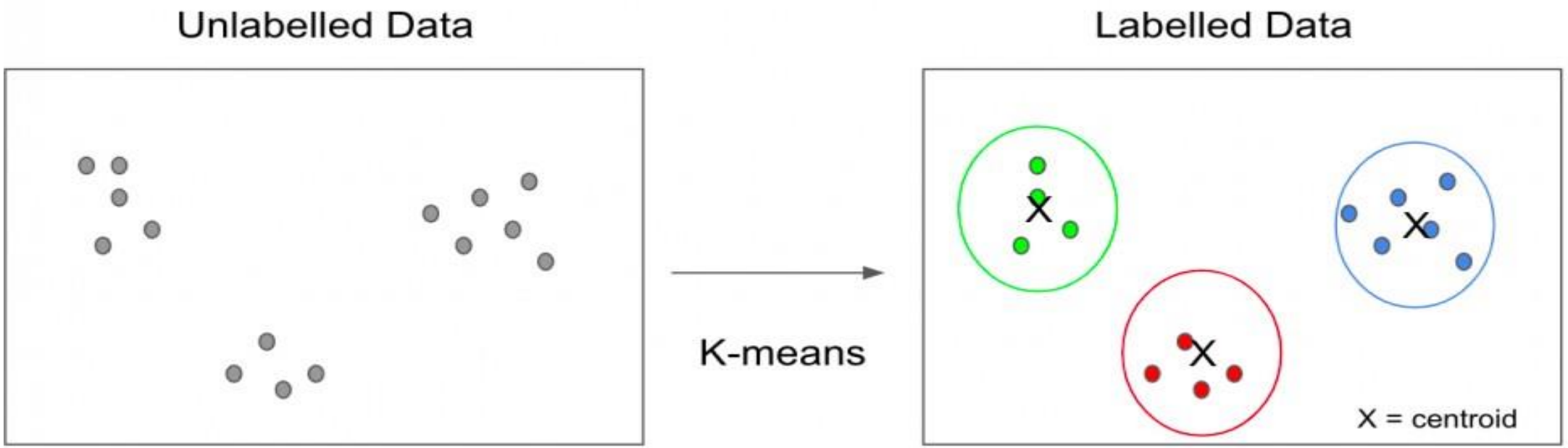


Прокомментируйте различие задач supervised
и unsupervised learning на этом примере

Алгоритм K-Means

Алгоритм K-Means

- Для начала идея алгоритма схематично:



Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся… (кстати, а напомните, что такое метрический алгоритм?)

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:
  - Случайно инициализируем k центроидов.

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:
  - Случайно инициализируем k центроидов.

Что подразумевается под центроидом?

Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:
  - Случайно инициализируем k центроидов.
  - Для каждой точки находим ближайший центроид; назначаем для неё соответствующую метку.

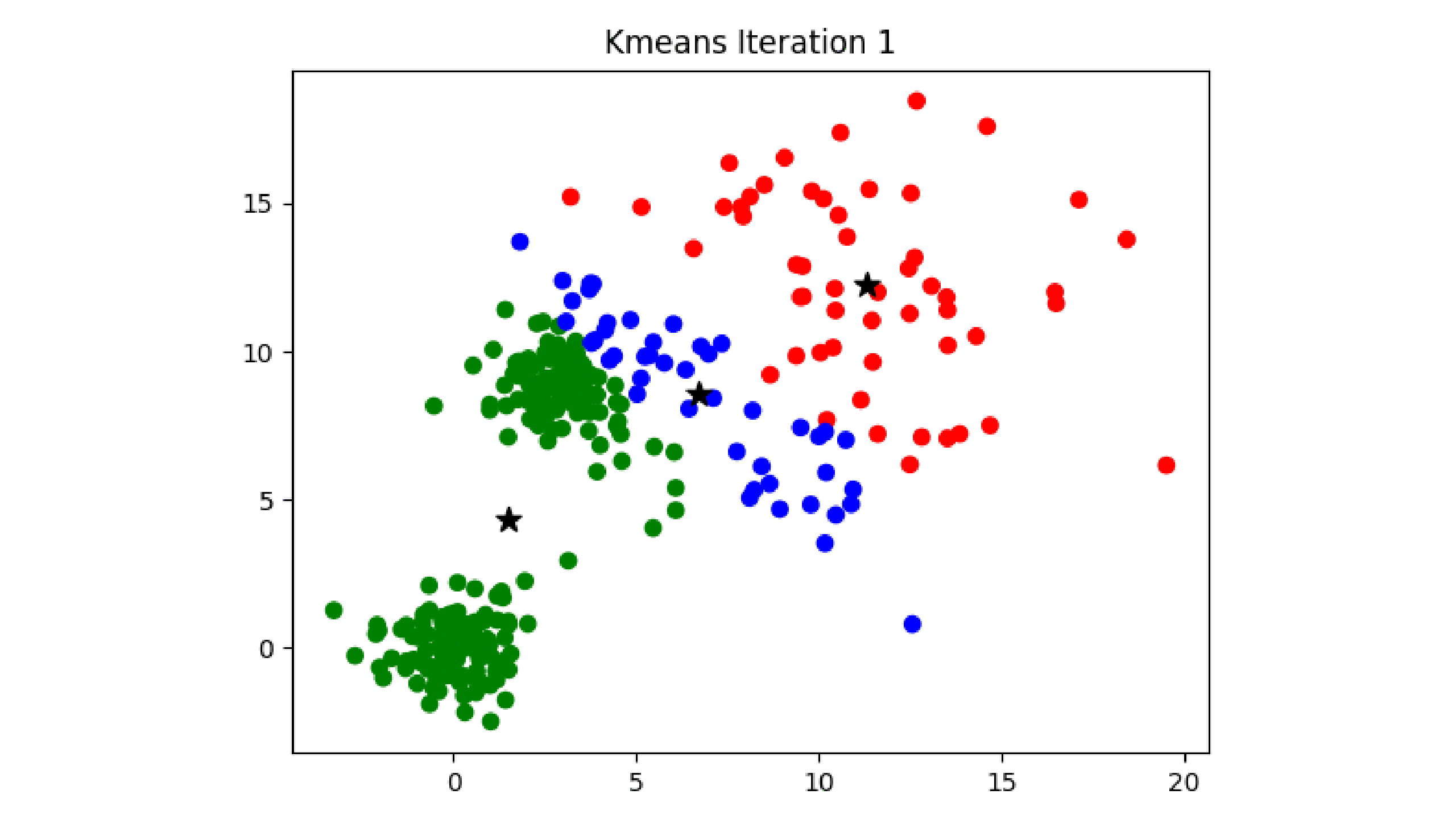
Алгоритм K-Means

- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:
  - Случайно инициализируем k центроидов.
  - Для каждой точки находим ближайший центроид; назначаем для неё соответствующую метку.
  - Пересчитываем позиции центроидов как центры масс соотв. кластеров.

Алгоритм K-Means

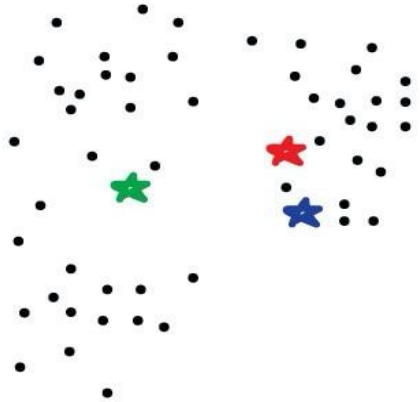
- K-Means — метрический алгоритм кластеризации, являющийся близким родственником для другого метрического алгоритма: KNN для классификации и регрессии.
- Концепция алгоритма заключается в следующем: для начала выбираем значение гиперпараметра k и метрику для расчета расстояний (пока что всё точно так же, как и в KNN).
- Затем обучаем алгоритм следующим образом:
  - Случайно инициализируем k центроидов.
  - Для каждой точки находим ближайший центроид; назначаем для неё соответствующую метку.
  - Пересчитываем позиции центроидов как центры масс соотв. кластеров.

Вопрос: мы обучили алгоритм; как теперь получить предсказание?

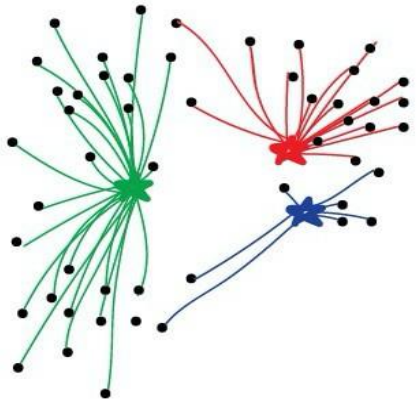

Kmeans Iteration 1
15
10
5
0
0
5
10
15
20

Ставим три ларька с шаурмой оптимальным образом

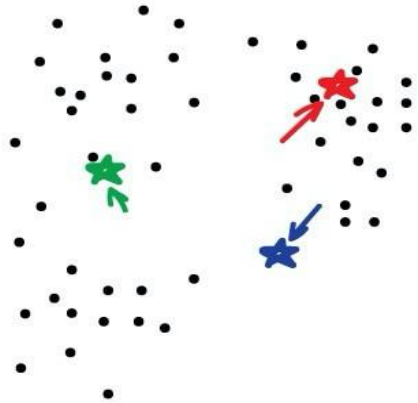
(иллюстрируя метод K-средних)



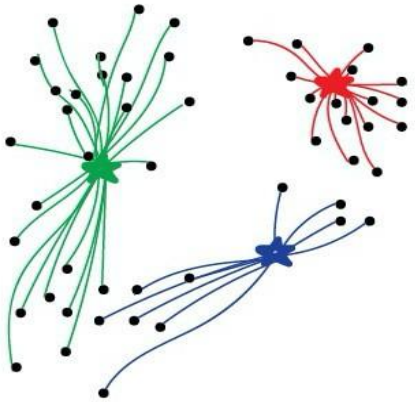
1. Ставим ларьки с шаурмой в случайных местах



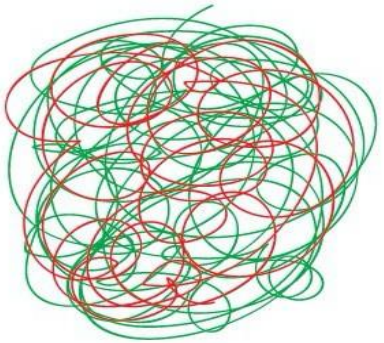
2. Смотрим в какой кому ближе идти



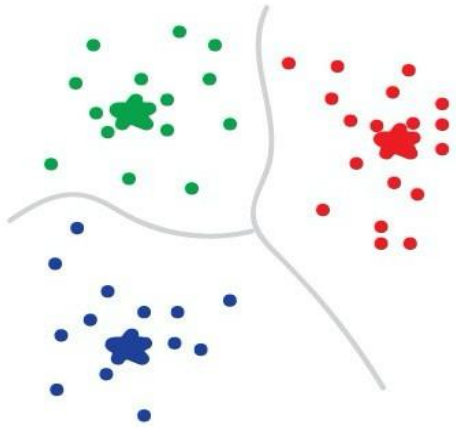
3. Двигаем ларьки ближе к центрам их популярности



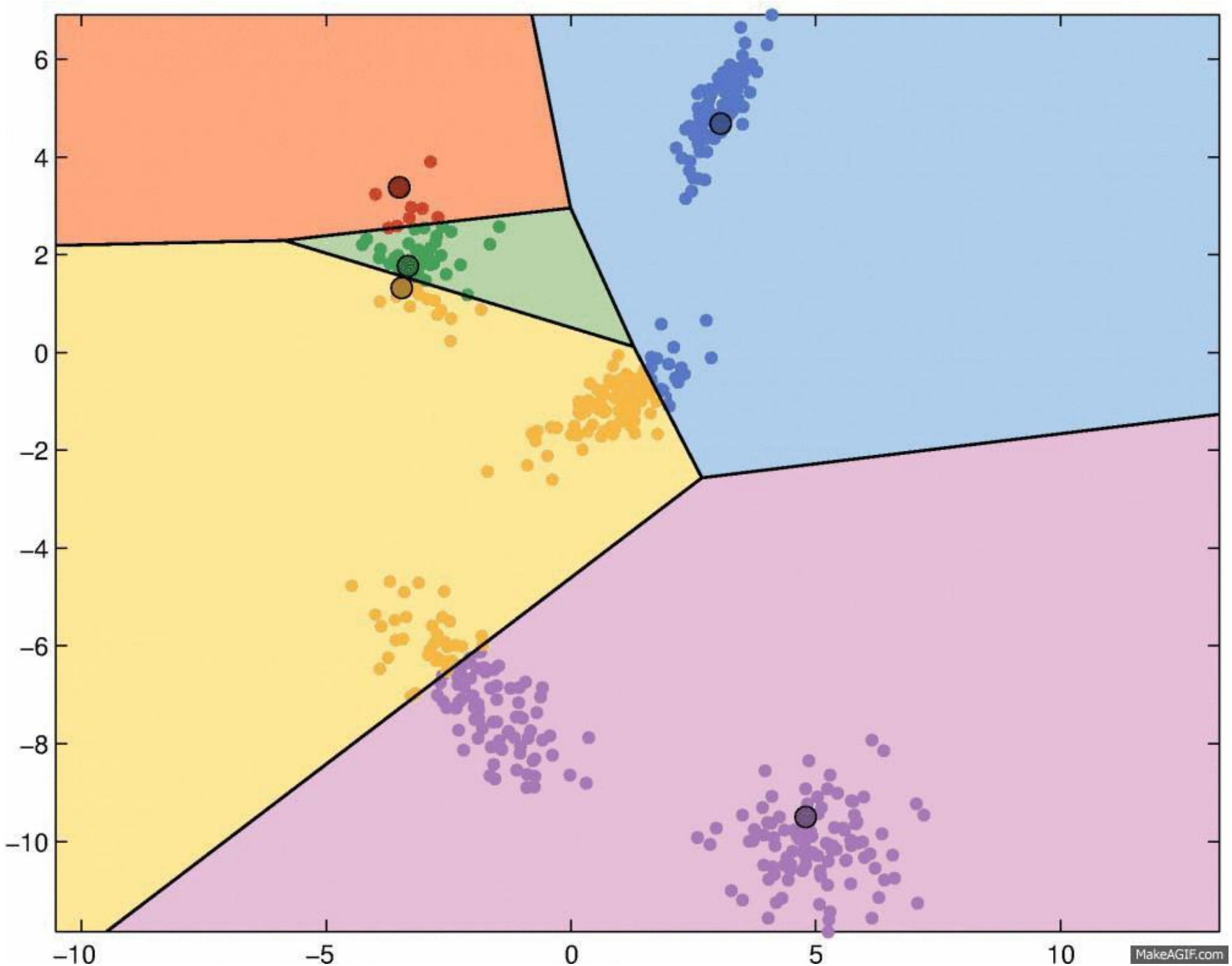
4. Снова смотрим и двигаем



5. Повторяем много раз



6. Готово, вы великолепны!



k-Means выстраивает приближение т.н. диаграммы Вороного по данным

Алгоритм K-Means

- В качестве примера области применения алгоритма можно привести, скажем, следующую задачу понижения разреженности данных.

Алгоритм K-Means

- В качестве примера области применения алгоритма можно привести, скажем, следующую задачу понижения разреженности данных.
- Вспомните матрицу “пользователи × контент”, которую мы обсуждали с вами, когда говорили про PCA.

Алгоритм K-Means

- В качестве примера области применения алгоритма можно привести, скажем, следующую задачу понижения разреженности данных.
- Вспомните матрицу “пользователи × контент”, которую мы обсуждали с вами, когда говорили про PCA.
- Такая матрица, куда пишут историю взаимодействия всех пользователей со всем контентом на платформе, — крайне разреженная: каждый пользователь взаимодействует лишь с очень малой долей контента.

Алгоритм K-Means

- В качестве примера области применения алгоритма можно привести, скажем, следующую задачу понижения разреженности данных.
- Вспомните матрицу “пользователи × контент”, которую мы обсуждали с вами, когда говорили про PCA.
- Такая матрица, куда пишут историю взаимодействия всех пользователей со всем контентом на платформе, — крайне разреженная: каждый пользователь взаимодействует лишь с очень малой долей контента.
- Обучать алгоритмы ML на таких данных чрезвычайно сложно.

Алгоритм K-Means

- Обучать алгоритмы ML на таких данных чрезвычайно сложно.

Алгоритм K-Means

- Обучать алгоритмы ML на таких данных чрезвычайно сложно.
- Однако мы можем значительно упростить себе задачу:

Алгоритм K-Means

- Обучать алгоритмы ML на таких данных чрезвычайно сложно.
- Однако мы можем значительно упростить себе задачу:
  - Сгруппируем пользователей при помощи алгоритма K-Means.
  - Проагрегируем предпочтения в пределах каждой полученной группы.
  - Вместо матрицы “пользователи × контент”, будем работать с матрицей “группа × контент”.
  - При необходимости сделать предсказание для нового пользователя — находим наиболее близкую к нему группу и берём предсказание для неё.

Алгоритм K-Means

- Обучать алгоритмы ML на таких данных чрезвычайно сложно.
- Однако мы можем значительно упростить себе задачу:
  - Сгруппируем пользователей при помощи алгоритма K-Means.
  - Проагрегируем предпочтения в пределах каждой полученной группы.
  - Вместо матрицы “пользователи × контент”, будем работать с матрицей “группа × контент”.
  - При необходимости сделать предсказание для нового пользователя — находим наиболее близкую к нему группу и берём предсказание для неё.
- Выглядит довольно мощно, неправда ли?

Алгоритм K-Means. Плюсы

- Простой, интерпретируемый алгоритм.
  - Хороший baseline, с которого можно начать.
- Даёт качественную кластеризацию при грамотном подборе метрики.
- Сложность предсказания (построенной и обученной модели) — $O(k \cdot \log k)$ в среднем; и $O(k^2)$ — в худшем случае. Нужно $O(k)$ дополнительной памяти.
  - Это практически мгновенно и не зависит от размера входных данных.
- Понятно, как устроено оптимальное решение
  - Диаграмма Вороного с k ячейками.
- Понятно, как пересчитать центроиды при поступлении новой точки.

Алгоритм K-Means. Минусы

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.
- Ответ сильно зависит от начальной инициализации.

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.
- Ответ сильно зависит от начальной инициализации.
  - С ней может и не повезти.

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.
- Ответ сильно зависит от начальной инициализации.
  - С ней может и не повезти.
- Итеративный процесс обучения.

Алгоритм K-Means. Минусы

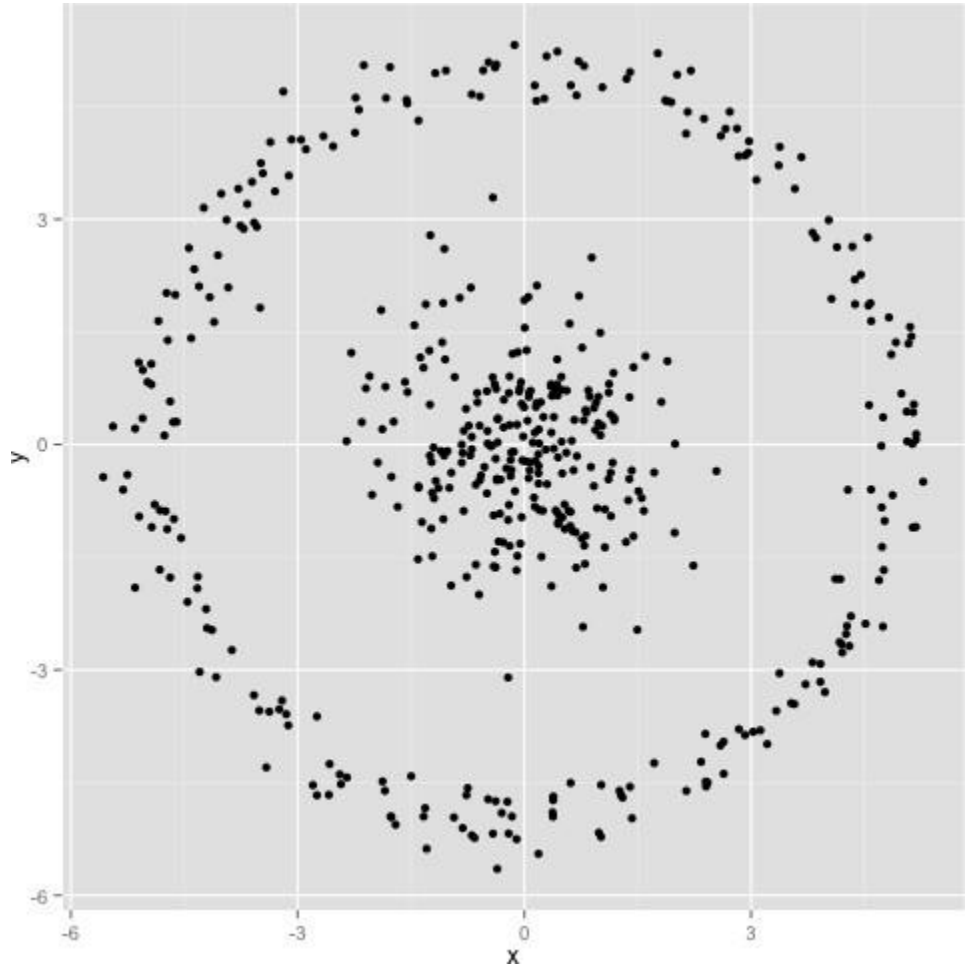
- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.
- Ответ сильно зависит от начальной инициализации.
  - С ней может и не повезти.
- Итеративный процесс обучения.
  - Непонятно, сколько итераций потребуется до сходимости.

Алгоритм K-Means. Минусы

- Часто сходу может быть абсолютно неясно, как подобрать k и правильную метрику.
  - Здесь это важно, т.к. евклидова метрика может быть адекватна локальной геометрии данных, но глобальную структуру она чаще всего описывает неправильно.
- Ответ сильно зависит от начальной инициализации.
  - С ней может и не повезти.
- Итеративный процесс обучения.
  - Непонятно, сколько итераций потребуется до сходимости.
- А еще…

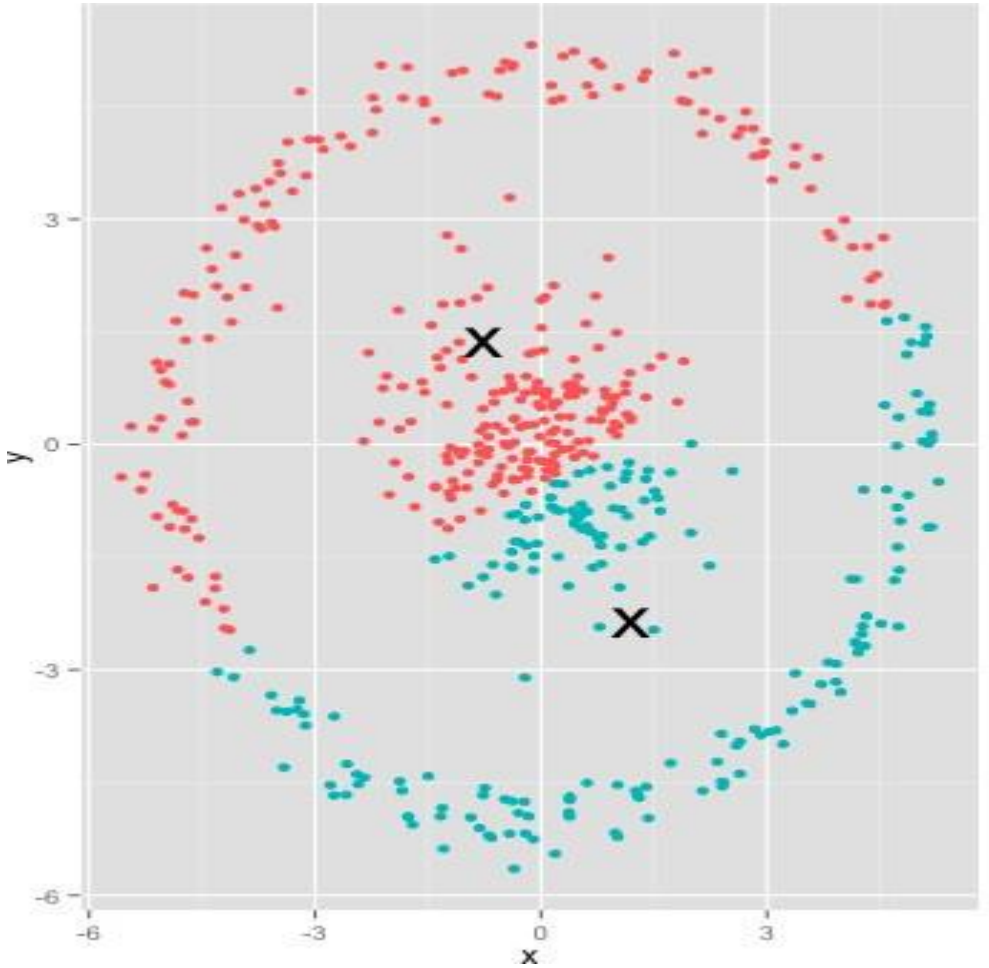
Алгоритм K-Means. Минусы

- Как K-Means разделит такие точки на кластеры при $k = 2$?



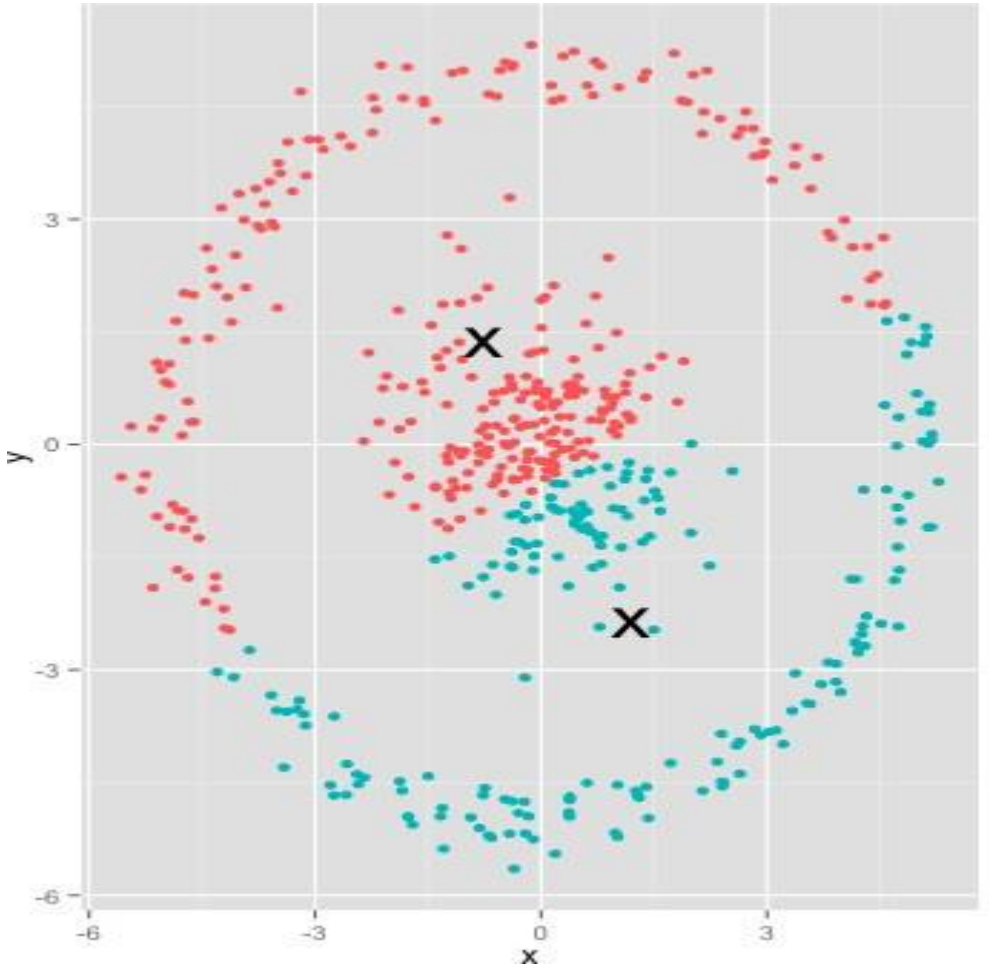
Алгоритм K-Means. Минусы

- Как K-Means разделит такие точки на кластеры при $k = 2$?



Алгоритм K-Means. Минусы

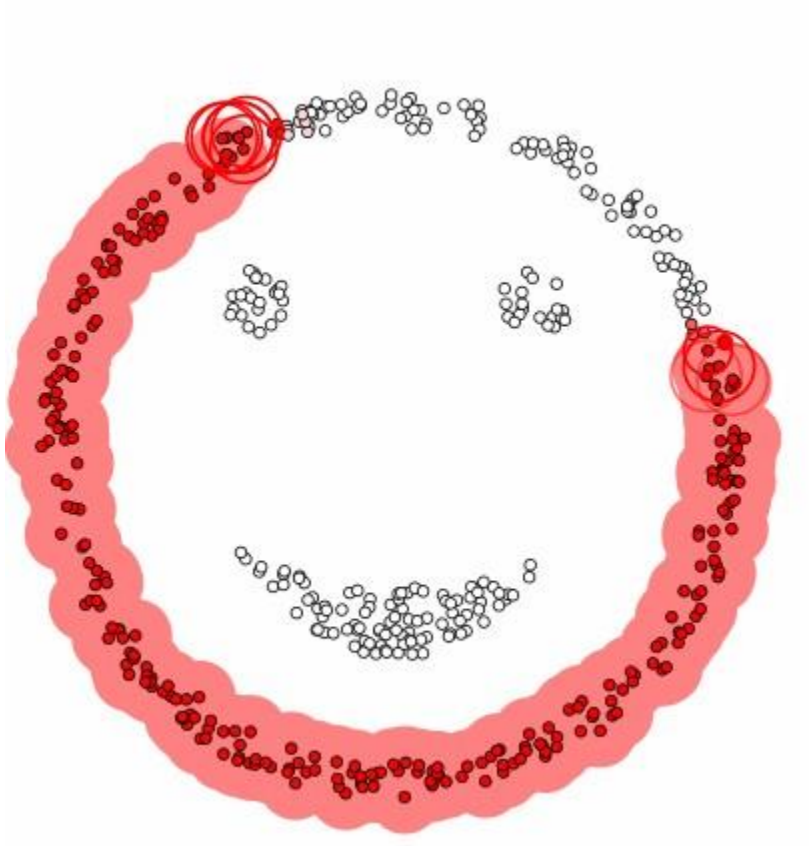
- Как K-Means разделит такие точки на кластеры при $k = 2$?
- Не самый лучший вариант…



Алгоритм DBSCAN

Алгоритм DBSCAN

- Для начала идея алгоритма схематично:



Алгоритм DBSCAN

- Концепция алгоритма заключается в следующем. Выбираем гиперпараметры: метрику, радиус окрестности вокруг точек, минимальное количество точек в пределах радиуса.
- Затем обучаем алгоритм следующим образом:

Алгоритм DBSCAN

- Концепция алгоритма заключается в следующем. Выбираем гиперпараметры: метрику, радиус окрестности вокруг точек, минимальное количество точек в пределах радиуса.
- Затем обучаем алгоритм следующим образом:
  - Выстраиваем окрестность вокруг каждой точки данных.

Алгоритм DBSCAN

- Концепция алгоритма заключается в следующем. Выбираем гиперпараметры: метрику, радиус окрестности вокруг точек, минимальное количество точек в пределах радиуса.
- Затем обучаем алгоритм следующим образом:
  - Выстраиваем окрестность вокруг каждой точки данных.
  - Перебираем окрестности в порядке убывания плотности.

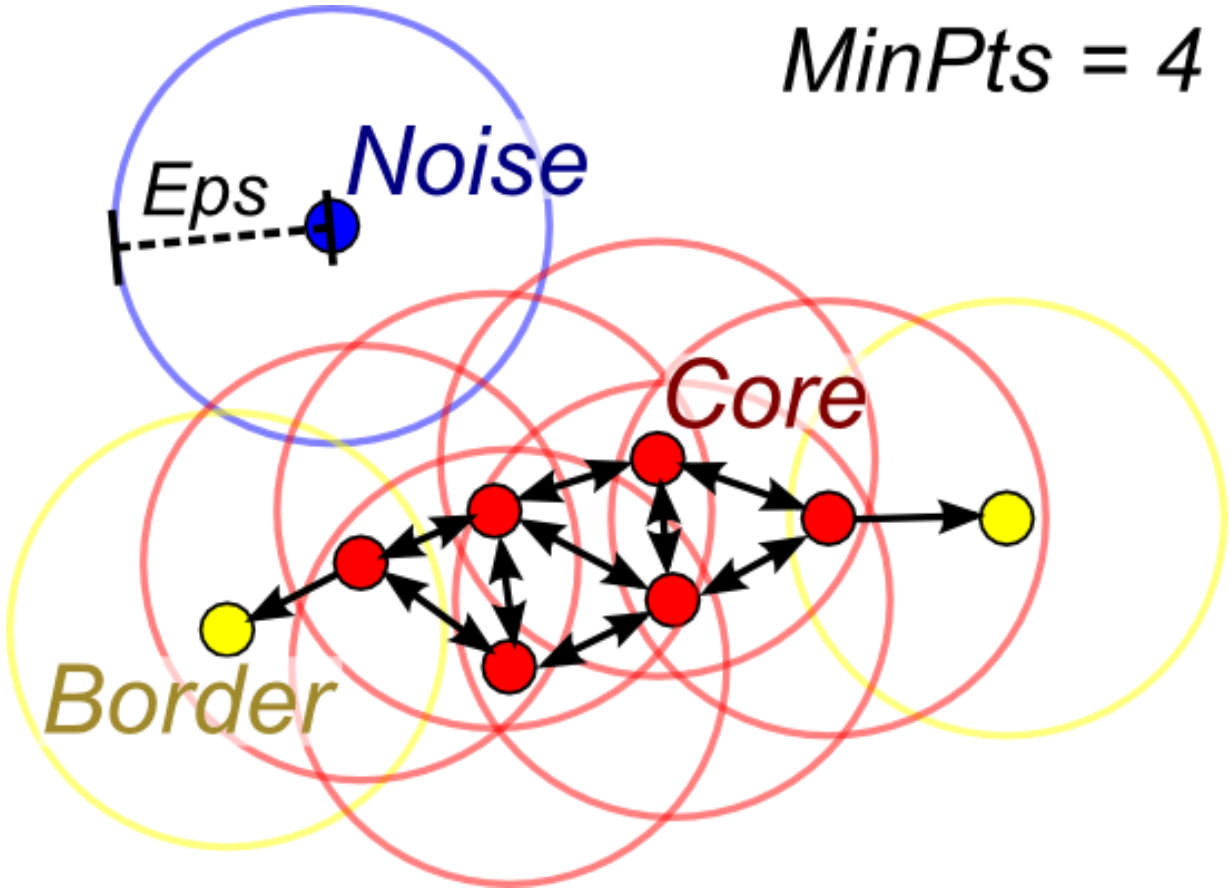
Алгоритм DBSCAN

- Концепция алгоритма заключается в следующем. Выбираем гиперпараметры: метрику, радиус окрестности вокруг точек, минимальное количество точек в пределах радиуса.
- Затем обучаем алгоритм следующим образом:
  - Выстраиваем окрестность вокруг каждой точки данных.
  - Перебираем окрестности в порядке убывания плотности.
  - Если в пределах окрестности содержится хотя бы MinPts точек, то классифицируем соотв. точку как core.

Алгоритм DBSCAN

- Концепция алгоритма заключается в следующем. Выбираем гиперпараметры: метрику, радиус окрестности вокруг точек, минимальное количество точек в пределах радиуса.
- Затем обучаем алгоритм следующим образом:
  - Выстраиваем окрестность вокруг каждой точки данных.
  - Перебираем окрестности в порядке убывания плотности.
  - Если в пределах окрестности содержится хотя бы MinPts точек, то классифицируем соотв. точку как core.
  - В противном случае — классифицируем точку либо как border — если в её окрестности есть хотя бы одна core-точка, — либо как noise иначе.

Алгоритм DBSCAN



Алгоритм DBSCAN

- Предсказание же алгоритмом получаем следующим образом:

Алгоритм DBSCAN

- Предсказание же алгоритмом получаем следующим образом:
  - Core- и border-точки в пределах одной окрестности соединяются рёбрами.

Алгоритм DBSCAN

- Предсказание же алгоритмом получаем следующим образом:
  - Core- и border-точки в пределах одной окрестности соединяются рёбрами.
  - Кластерами становятся компоненты связности полученного графа.

Алгоритм DBSCAN

- Предсказание же алгоритмом получаем следующим образом:
  - Core- и border-точки в пределах одной окрестности соединяются рёбрами.
  - Кластерами становятся компоненты связности полученного графа.

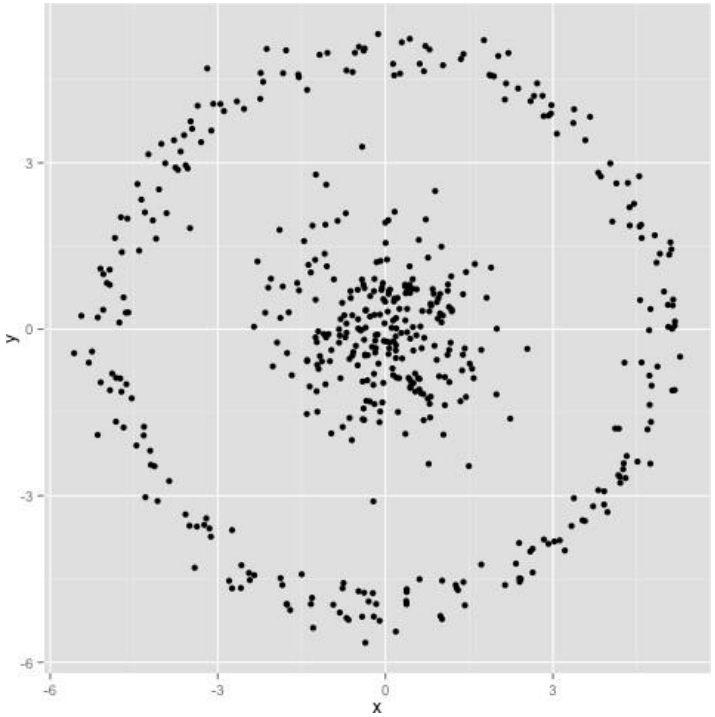
Ого, это что, дискретная математика??
Кто помнит, что такое компоненты связности графа?

Алгоритм DBSCAN

- Предсказание же алгоритмом получаем следующим образом:
  - Core- и border-точки в пределах одной окрестности соединяются рёбрами.
  - Кластерами становятся компоненты связности полученного графа.
  - Noise-точки рапортуются отдельно как независимый кластер шума (или же некластеризуемые объекты).

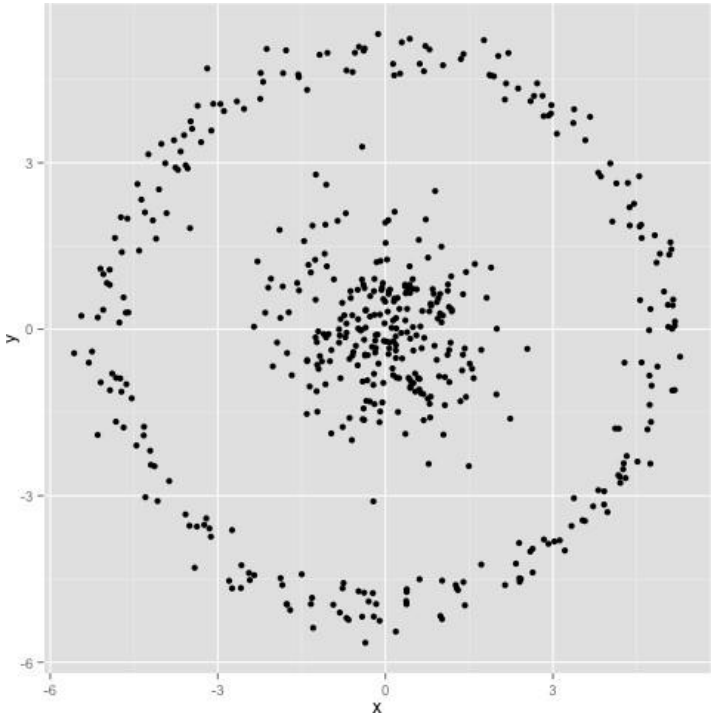
Сравнение K-Means и DBSCAN

Сравнение K-Means и DBSCAN

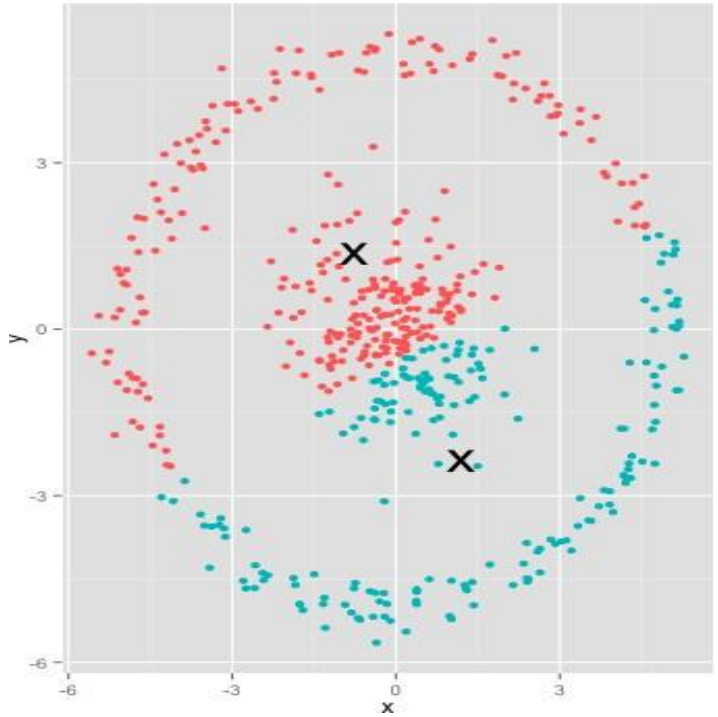


Unlabeled data

Сравнение K-Means и DBSCAN

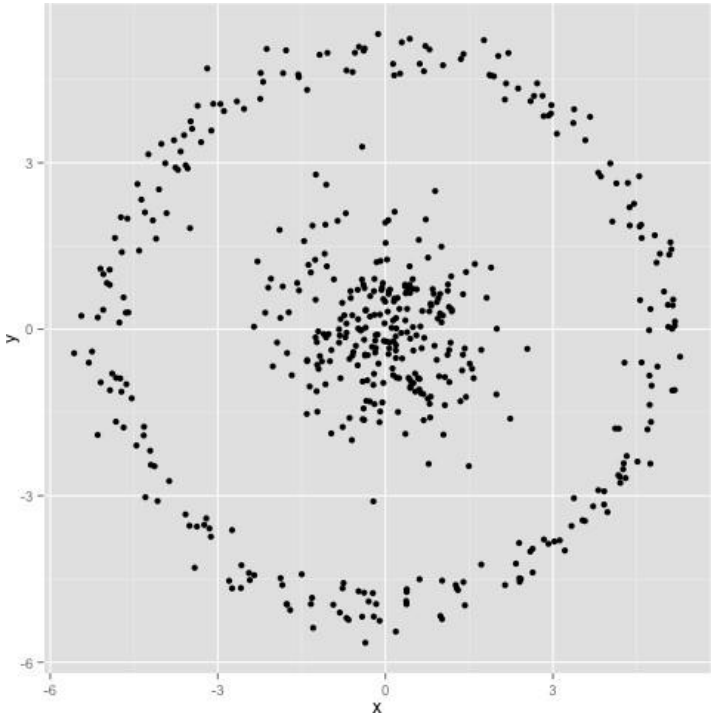


Unlabeled data

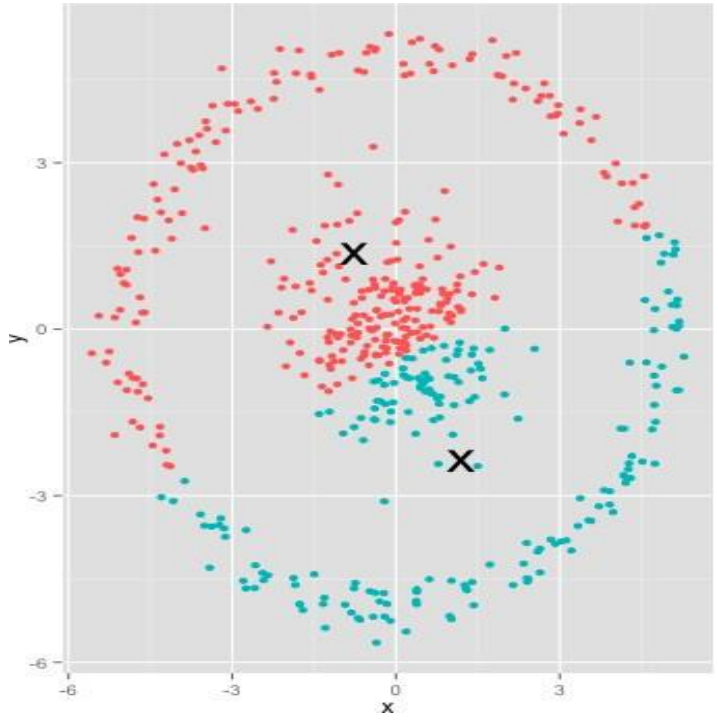


K-Means

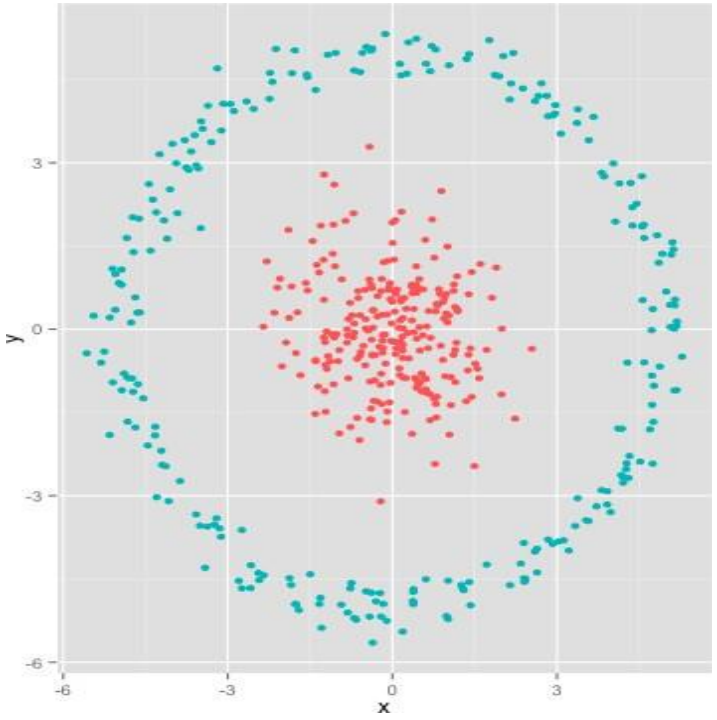
Сравнение K-Means и DBSCAN



Unlabeled data



K-Means



DBSCAN

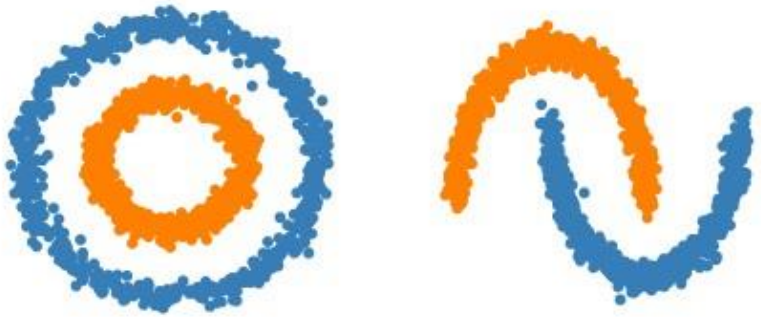
Сравнение K-Means и DBSCAN

DBSCAN

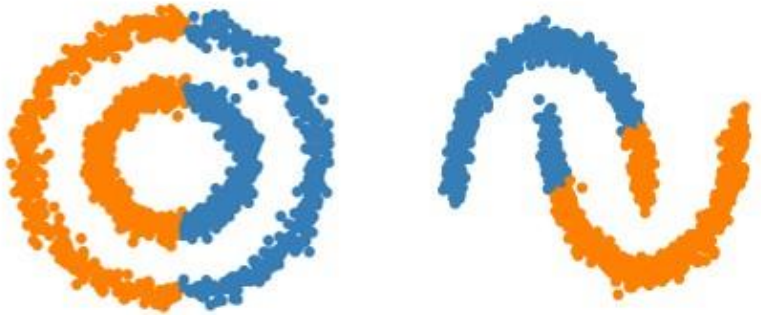
k-means

Сравнение K-Means и DBSCAN

DBSCAN



k-means



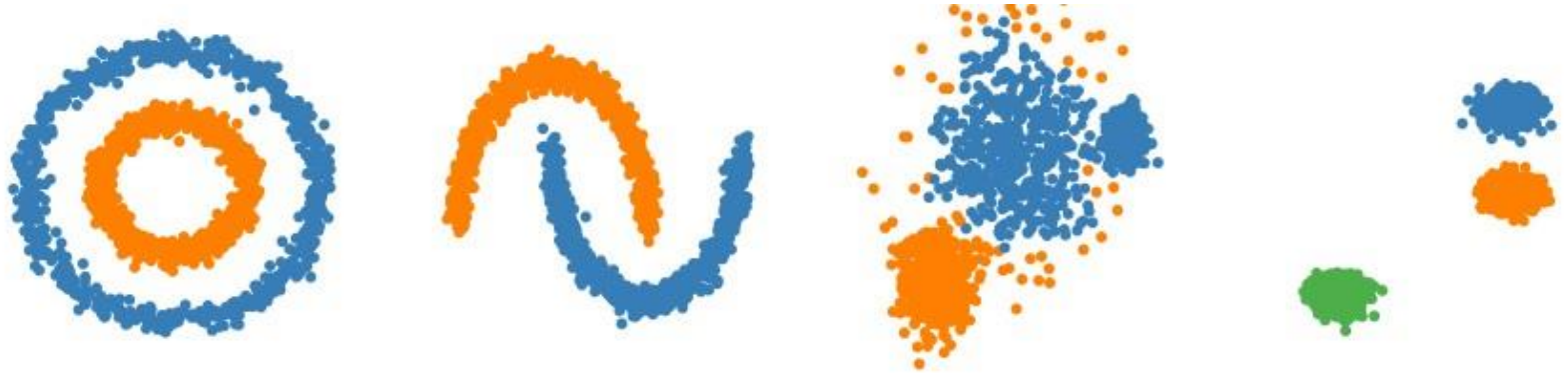
Сравнение K-Means и DBSCAN

DBSCAN

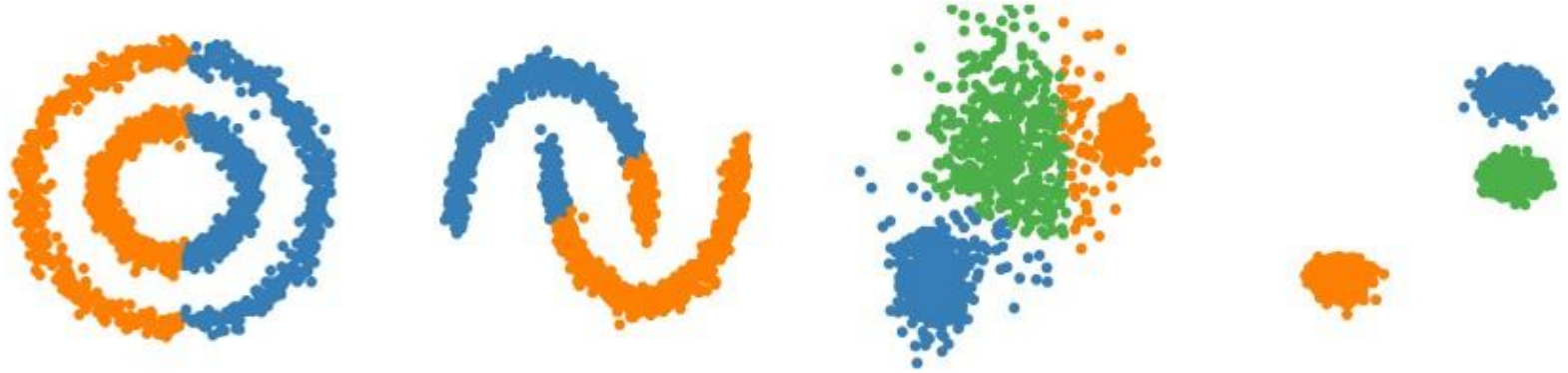
k-means

Сравнение K-Means и DBSCAN

DBSCAN



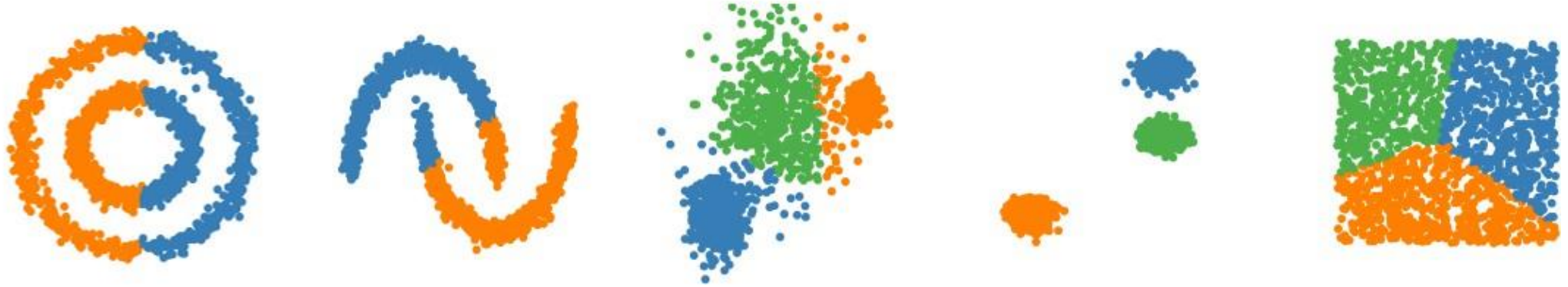
k-means



Сравнение K-Means и DBSCAN

DBSCAN

k-means



Алгоритм DBSCAN

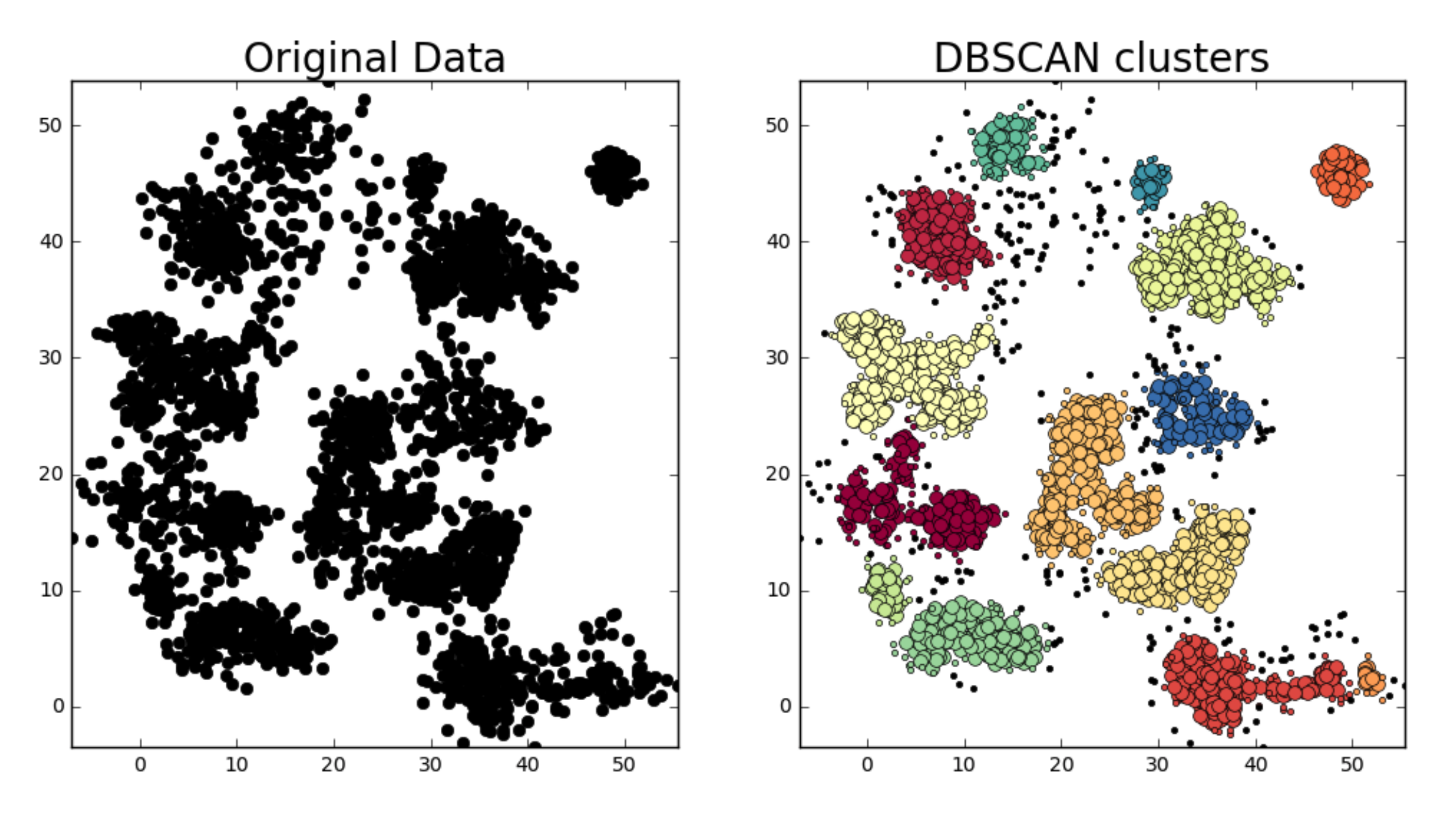
- Области применения DBSCAN все примерно те же, что и у K-Means.

Алгоритм DBSCAN

- Области применения DBSCAN все примерно те же, что и у K-Means.
- Особенно полезен DBSCAN может быть, когда эксперт по предметной области не может заранее оценить k — число кластеров, поскольку у K-Means это является обязательным гиперпараметром.

Алгоритм DBSCAN

- Области применения DBSCAN все примерно те же, что и у K-Means.
- Особенно полезен DBSCAN может быть, когда эксперт по предметной области не может заранее оценить k — число кластеров, поскольку у K-Means это является обязательным гиперпараметром.
- Также будет очень полезен при сложных, нелинейных зависимостях в данных.


Original Data
DBSCAN clusters
0
10
20
30
40
50

Алгоритм DBSCAN. Плюсы

- Простой, интерпретируемый алгоритм.
  - Результаты зачастую лучше, чем у K-Means.
- Улавливает более тонкие локальные особенности в данных.
  - Многообразия локально неотличимы от обычного n-мерного пространства, поэтому часто можно спокойно использовать евклидову метрику.
- Не требует заранее указывать количество кластеров.
  - Находит их все сам!
- Находит заодно и выбросы.
- Быстро обучается, не требует итеративного уточнения.

Алгоритм DBSCAN. Минусы

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.
- Трудно делать предсказания для новых точек.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.
- Трудно делать предсказания для новых точек.
  - Так как каждая новая точка изменяет плотность в окрестностях уже имеющихся точек.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.
- Трудно делать предсказания для новых точек.
  - Так как каждая новая точка изменяет плотность в окрестностях уже имеющихся точек.
- Требуются более сложные структуры данных, нежели в K-Means.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.
- Трудно делать предсказания для новых точек.
  - Так как каждая новая точка изменяет плотность в окрестностях уже имеющихся точек.
- Требуются более сложные структуры данных, нежели в K-Means.
- Не учитывается структура построенного графа.

Алгоритм DBSCAN. Минусы

- Нужно подбирать радиус окрестности, MinPts.
  - Часто не очень понятно, как это сделать из интуитивных соображений.
- Трудно делать предсказания для новых точек.
  - Так как каждая новая точка изменяет плотность в окрестностях уже имеющихся точек.
- Требуются более сложные структуры данных, нежели в K-Means.
- Не учитывается структура построенного графа.
  - При том, что в ней содержится немало полезной информации.