

# Машинное обучение с учителем и без на примере классификации и кластеризации

Урок 6






# План курса

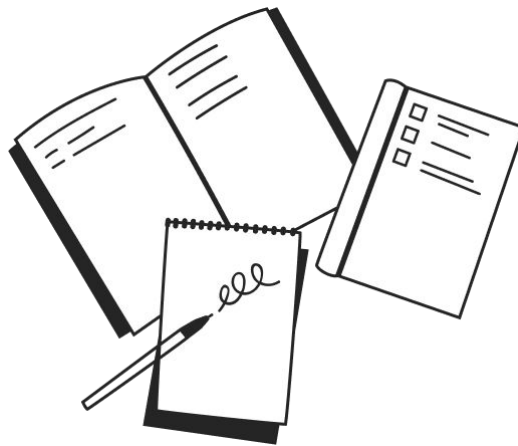
- 1 Data Science. Обзор.
- 2 Аналитика данных и ETL
- 3 Визуализация данных.

- 4 Основы статистики.
- 5 Машинное обучение. Базовое представление
- 6 Машинное обучение с учителем и без на примере классификации и кластеризации.



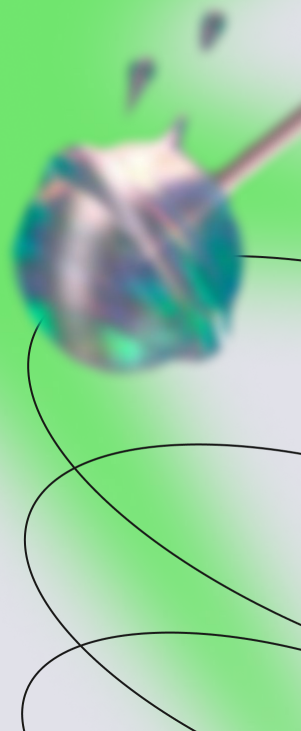
## Что будет на уроке сегодня

-  Машинное обучение с учителем и без
-  Классификация
-  Кластеризация





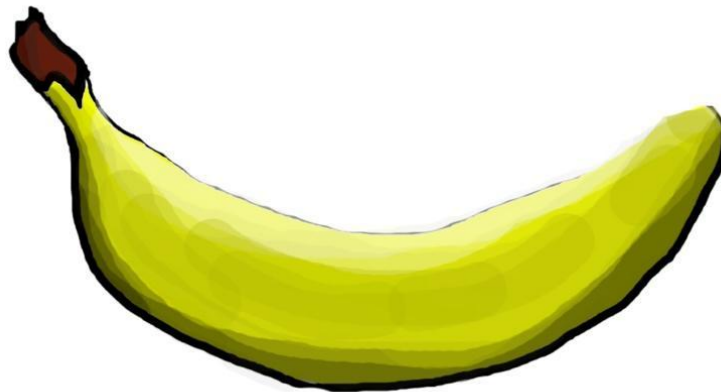
# Машинное обучение с учителем





## Обучение с учителем

**Обучение с учителем**, как следует из названия, предполагает присутствие руководителя в качестве учителя. По сути, контролируемое обучение — это когда мы обучаем или тренируем машину, используя данные, которые хорошо размечены





## Категории обучения с учителем

- 📌 **Классификация** : проблема классификации возникает, когда выходная переменная представляет собой категорию, такую как «красный» или «синий», «заболевание» или «отсутствие заболевания».
- 📌 **Регрессия** : проблема регрессии возникает, когда выходная переменная представляет собой реальное значение, например «доллары» или «вес».

### Типы обучения с учителем:

- Регрессия
- Логистическая регрессия
- Классификация
- Наивные байесовские классификаторы
- K-NN (k ближайших соседей)
- Деревья решений
- Метод опорных векторов



## Преимущества

- 📌 Обучение под наблюдением позволяет собирать данные и выводить данные из предыдущего опыта.
- 📌 Помогает оптимизировать критерии эффективности с помощью опыта.
- 📌 Контролируемое машинное обучение помогает решать различные типы реальных вычислительных задач.
- 📌 Оно выполняет задачи классификации и регрессии.
- 📌 Это позволяет оценить или сопоставить результат с новой выборкой.
- 📌 У нас есть полный контроль над выбором количества классов, которые мы хотим в обучающих данных.



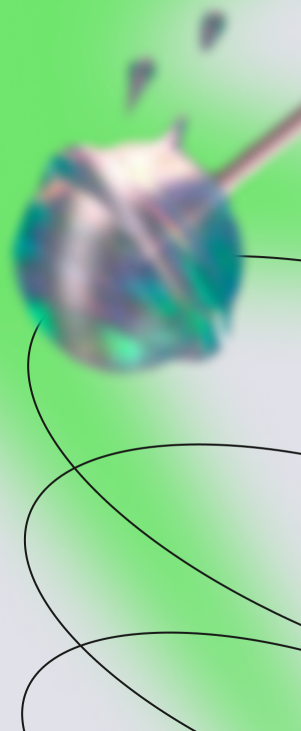
## Недостатки

- 📌 Классификация больших данных может быть сложной задачей.
- 📌 Тренировка моделей обучения с учителем требует много вычислительного времени. Значит, это требует много времени.
- 📌 Обучение с учителем не может справиться со всеми сложными задачами машинного обучения.
- 📌 Время вычислений огромно для обучения с учителем.
- 📌 Для построения моделей такого типа требуется предварительно размеченный набор данных.
- 📌 Это требует тренировочного процесса





# Машинное обучение без учителя





# Обучение без учителя

**Неконтролируемое обучение** — это обучение машины с использованием информации, которая не классифицирована и не помечена, и позволяющее алгоритму действовать на этой информации без руководства

**Неконтролируемое обучение подразделяется на две категории алгоритмов:**

- **Кластеризация** . Проблема кластеризации заключается в том, что вы хотите обнаружить неотъемлемую группу данных, например группу клиентов по покупательскому поведению.
- **Ассоциация** : проблема изучения правил ассоциации заключается в том, что вы хотите обнаружить правила, которые описывают большие части ваших данных, например, люди, которые покупают X, также склонны покупать Y.



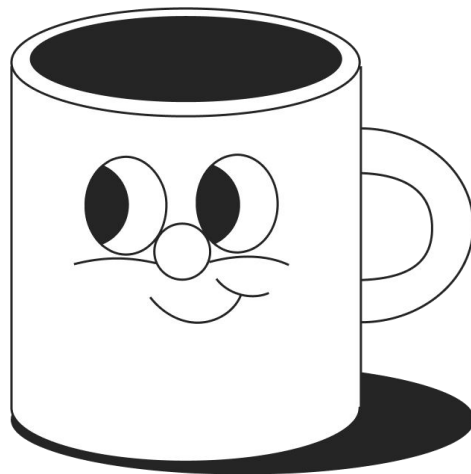
# Типы обучения без учителя

## Типы ассоциаций

1. Эксклюзивный (раздел)
2. Агломеративный
3. Перекрытие
4. Вероятностный

## Типы кластеризации:

1. Иерархическая кластеризация
2. Кластеризация К-средних
3. Анализ главных компонентов
4. Разложение по сингулярным значениям
5. Анализ независимых компонентов











## Преимущества

- 📌 Оно не требует, чтобы обучающие данные были размечены.
- 📌 Уменьшение размерности может быть легко достигнуто с помощью обучения без учителя.
- 📌 Способен находить ранее неизвестные закономерности в данных.
- 📌 **Гибкость:** неконтролируемое обучение является гибким в том смысле, что его можно применять к широкому кругу задач, включая кластеризацию, обнаружение аномалий и анализ правил ассоциации.
- 📌 **Исследование:** неконтролируемое обучение позволяет исследовать данные и обнаруживать новые и потенциально полезные закономерности, которые могут быть не очевидны с самого начала.
- 📌 **Низкая стоимость:** обучение без учителя часто дешевле, чем обучение с учителем, потому что оно не требует размеченных данных, получение которых может занять много времени и средств.



## Недостатки

-  Трудно измерить точность или эффективность из-за отсутствия predetermined ответов во время обучения.
-  Результаты часто имеют меньшую точность.
-  Пользователь должен потратить время на интерпретацию и маркировку классов, которые следуют этой классификации.
-  **Отсутствие руководства** : в неконтролируемом обучении отсутствуют руководство и обратная связь, обеспечиваемые помеченными данными, что может затруднить определение того, являются ли обнаруженные закономерности актуальными или полезными.
-  **Чувствительность к качеству данных** . Неконтролируемое обучение может быть чувствительно к качеству данных, включая пропущенные значения, выбросы и зашумленные данные.
-  **Масштабируемость** : неконтролируемое обучение может быть дорогостоящим в вычислительном отношении, особенно для больших наборов данных или сложных алгоритмов, что может ограничить его масштабируемость.



## Сравнительная таблица машинного обучения с учителем и без

Параметры	Контролируемое машинное обучение	Неконтролируемое машинное обучение
Входные данные	Алгоритмы обучаются на размеченных данных.	Алгоритмы используются для данных, которые не размечены
Вычислительная сложность	Более простой метод	Вычислительно сложный
Точность	Высокая точность	Менее точный
Количество классов	Количество классов известно	Количество классов неизвестно
Анализ данных	Использует автономный анализ	Использует анализ данных в реальном времени
Используемые алгоритмы	Линейная и логистическая регрессия, рандомный лес, Метод опорных векторов, нейронная сеть и т. д.	Кластеризация К-средних, Иерархическая кластеризация, Априорный алгоритм и др.



## Сравнительная таблица машинного обучения с учителем и без

Параметры	Контролируемое машинное обучение	Неконтролируемое машинное обучение
Выход	Дается желаемый результат.	Желаемый результат не указан.
Тренировочные данные	Используйте обучающие данные, чтобы натренировать модель.	Данные обучения не используются.
Сложная модель	Невозможно изучить более крупные и сложные модели, чем при обучении с учителем.	Можно изучать более крупные и сложные модели с помощью обучения без учителя.
Модель	Мы можем протестировать нашу модель.	Мы не можем проверить нашу модель.
Вызывается как	Обучение с учителем также называют классификацией.	Неконтролируемое обучение также называют кластеризацией.
Пример	Пример: Оптическое распознавание символов.	Пример: найти лицо на изображении.



# Классификация

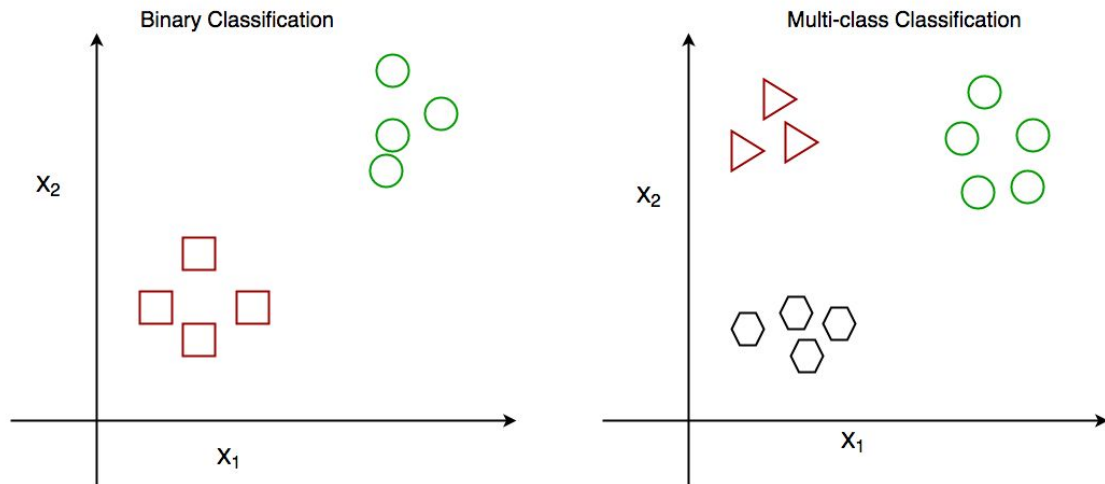






# Классификация

**Классификация** — это процесс поиска или обнаружения модели или функции, которая помогает разделить данные на несколько категориальных классов, т. е. дискретных значений. При классификации данные классифицируются по разным меткам в соответствии с некоторыми параметрами, заданными во входных данных, а затем для данных прогнозируются метки.



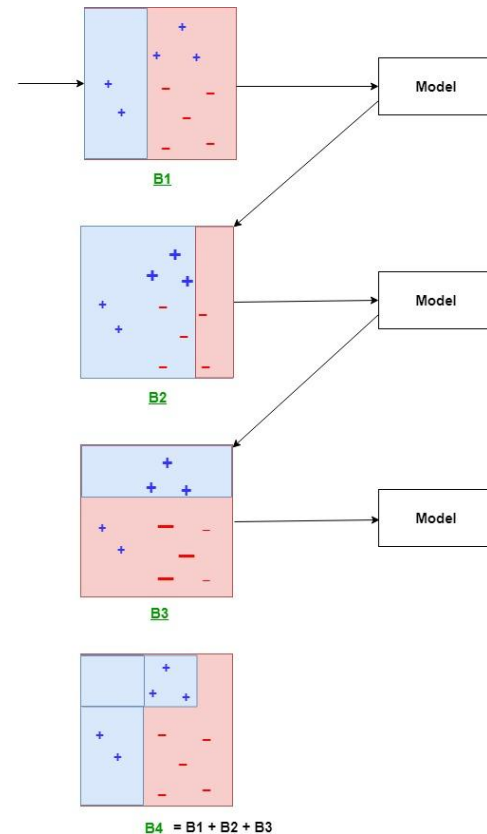


# Бустинг

**Бустинг** – это метод построения модели, который пытается построить сильный классификатор из числа слабых классификаторов.

## Алгоритм:

1. Инициализируйте набор данных и присвойте равный вес каждой точке данных.
2. Предоставьте это в качестве входных данных для модели и определите ошибочно классифицированные точки данных.
3. Увеличьте вес ошибочно классифицированных точек данных.
4. если (получены требуемые результаты)  
    Перейти к шагу 5  
    иначе  
        Перейти к шагу 2
5. Конец





## Бэггинг

**Бэггинг** — это алгоритм метаоценки, которая сопоставляет каждый базовый классификатор со случайными подмножествами исходного набора данных, а затем объединяет их индивидуальные прогнозы (путем голосования или усреднения) для формирования окончательного прогноза.

Такая метаоценка обычно может использоваться как способ уменьшить дисперсию оценки черного ящика (например, дерева решений) путем введения рандомизации в процедуру ее построения и последующего создания из нее ансамбля.



# Сравнение классификации и регрессии

	Классификация	Регрессия
1.	В этой постановке задачи целевые переменные дискретны.	В этой постановке задачи целевые переменные непрерывны.
2.	Такие проблемы, как классификация спама по электронной почте , прогнозирование заболеваний , такие как проблемы, решаются с использованием алгоритмов классификации.	Такие задачи, как прогноз цен на жилье , прогноз осадков, решаются с использованием алгоритмов регрессии.
3.	В этом алгоритме мы пытаемся найти наилучшую возможную границу решения, которая может разделить два класса с максимально возможным разделением.	В этом алгоритме мы пытаемся найти наиболее подходящую линию, которая может представлять общую тенденцию данных.
4.	Показатели оценки, такие как Precision, Recall и F1-Score, используются здесь для оценки производительности алгоритмов классификации.	Показатели оценки, такие как среднеквадратическая ошибка, R2-Score и MAPE , используются здесь для оценки производительности алгоритмов регрессии.
5.	Здесь мы сталкиваемся с такими проблемами, как бинарная классификация или проблемы многоклассовой классификации .	Здесь мы сталкиваемся с такими проблемами, как модели линейной регрессии , а также с нелинейными моделями.
6 .	Входные данные — это независимые переменные и категориальная зависимая переменная.	Входные данные — это независимые переменные и непрерывная зависимая переменная.
7 .	Вывод — категориальные метки.	Выходные данные являются непрерывными числовыми значениями.
8.	Цель состоит в том, чтобы предсказать метки категорий/классов.	Цель состоит в том, чтобы прогнозировать непрерывные числовые значения.
9 .	Примеры использования: обнаружение спама, распознавание изображений, анализ настроений.	Примеры вариантов использования: прогнозирование цен на акции, прогнозирование цен на жилье, прогнозирование спроса.



# Начало работы с классификацией



## Понимание проблемы

Прежде чем приступить к классификации, важно понять проблему, которую вы пытаетесь решить



## Подготовка данных

Как только вы хорошо разберетесь в проблеме, следующим шагом будет подготовка данных .



## Выбор модели

Важно выбрать модель, подходящую для вашей задачи



## Обучение модели

После того, как вы выбрали модель, следующим шагом будет ее обучение на данных обучения



## Оценка модели

После обучения модели важно оценить ее производительность на проверочном наборе



## Тонкая настройка модели

Если производительность модели неудовлетворительна, ее можно настроить, изменив параметры или попробовав другую модель



## Развертывание модели

когда вы удовлетворены производительностью модели, вы можете развернуть ее, чтобы делать прогнозы на основе новых данных



# Типы классификаций



## **Двоичная классификация**

когда нам нужно разделить данные на 2 разных класса

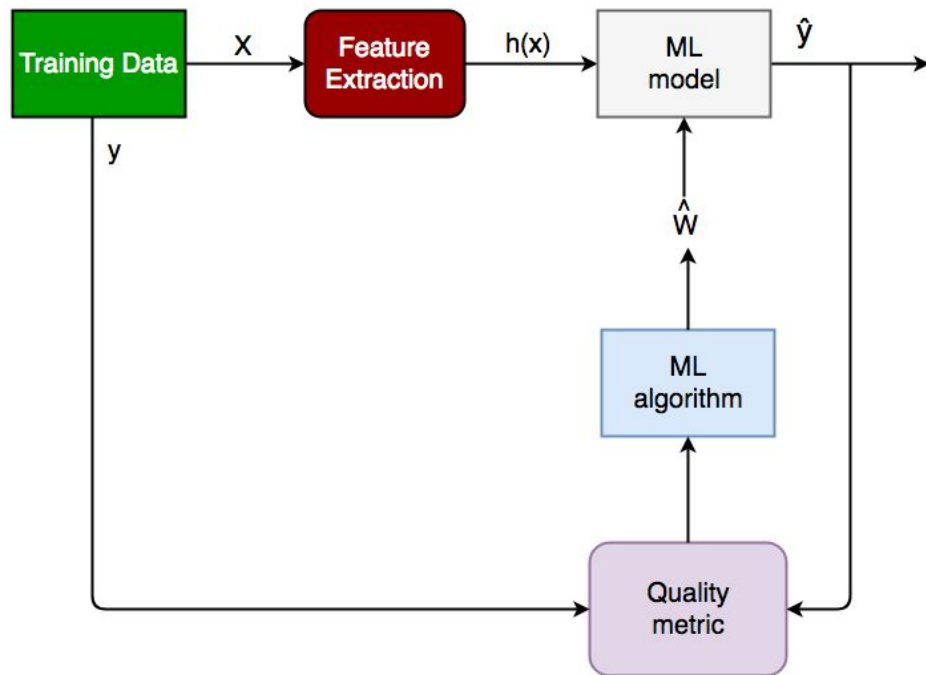


## **Мультиклассовая классификация**

количество классов больше 2











## Обобщенная блок-схема задачи классификации





## Типы классификаторов

-  Линейные классификаторы
-  Древовидные классификаторы
-  Метод опорных векторов
-  Искусственные нейронные сети
-  Байесовая регрессия
-  Гауссовы наивные байесовые классификаторы
-  Классификатор стохастического градиентного спуска
-  Методы ансамбля



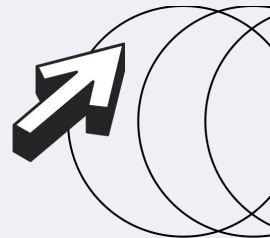


## Практическое применение классификации

1. Беспилотный автомобиль использует методы классификации с поддержкой глубокого обучения, которые позволяют обнаруживать и классифицировать препятствия.
1. Фильтрация спама по электронной почте является одним из наиболее распространенных и общепризнанных способов использования методов классификации.
1. Обнаружение проблем со здоровьем, распознавание лиц, распознавание речи, обнаружение объектов и анализ настроений используют классификацию в своей основе.

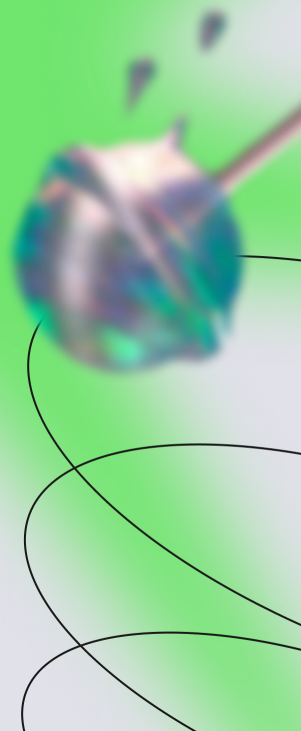
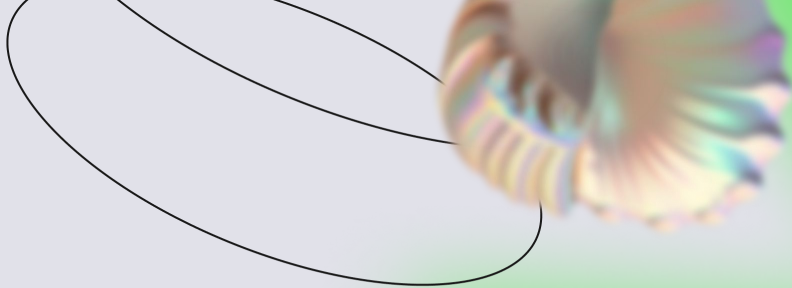


Давайте делать это в коде!





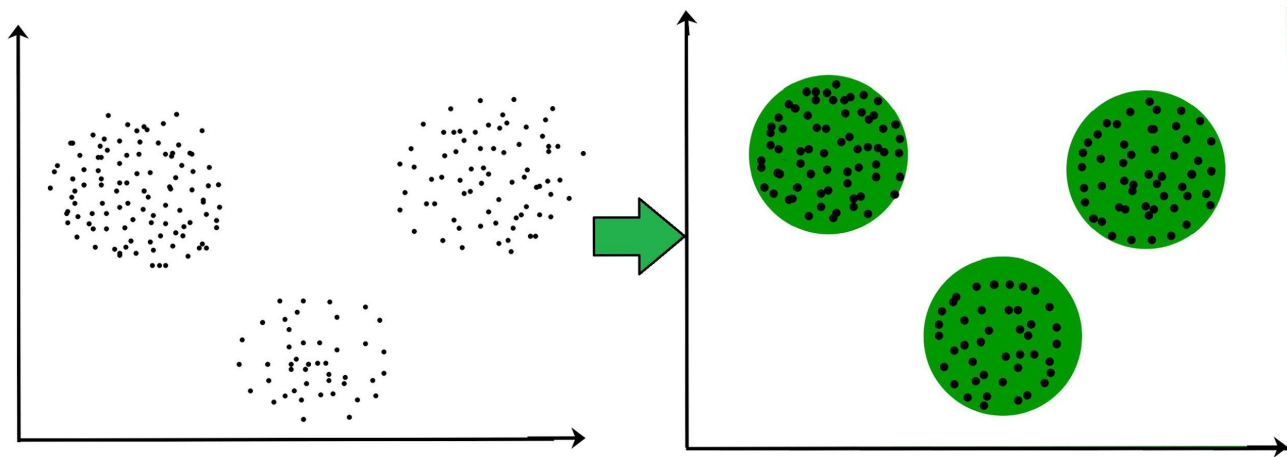
# Кластеризация





# Кластеризация

**Кластеризация** — это задача разделения совокупности или точек данных на несколько групп таким образом, чтобы точки данных в одних и тех же группах были более похожи на другие точки данных в той же группе и отличались от точек данных в других группах. Это в основном совокупность объектов на основе сходства и различия между ними.

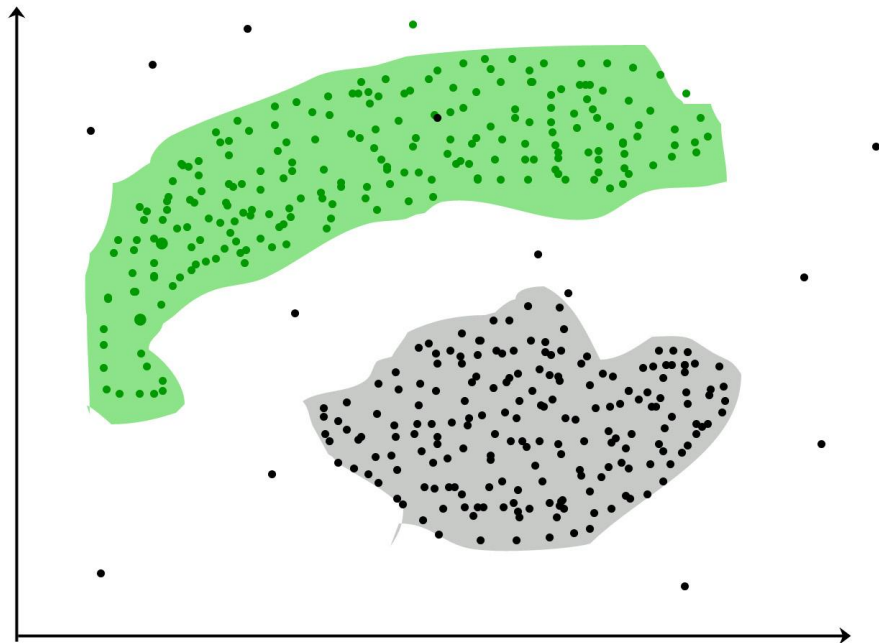




# Кластеризация

## **DBSCAN: пространственная кластеризация приложений с шумом на основе плотности**

Эти точки данных группируются с использованием базовой концепции, согласно которой точка данных находится в пределах заданного ограничения от центра кластера.





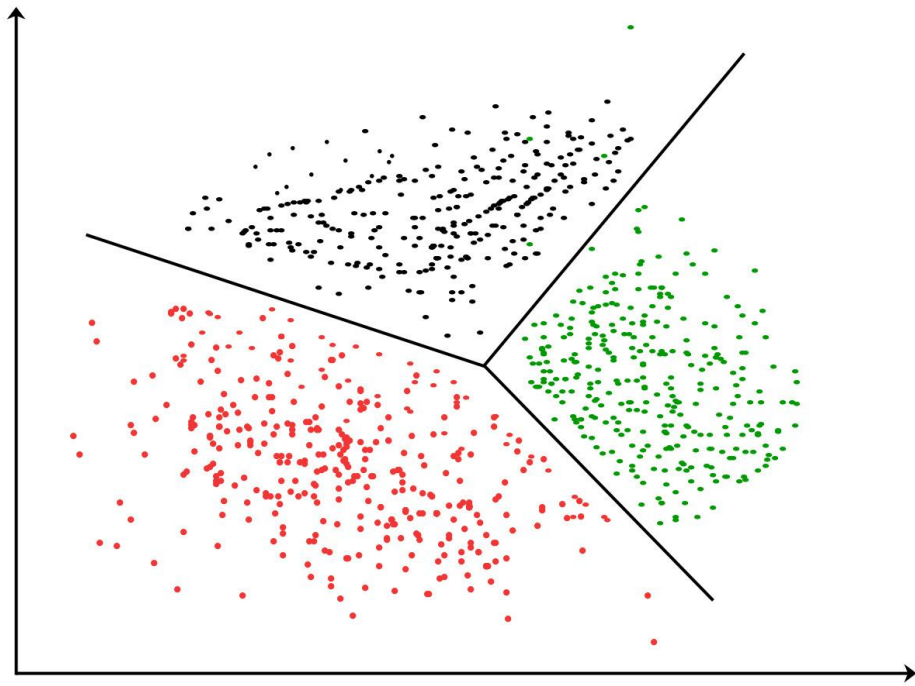
## Методы кластеризации

- **Методы, основанные на плотности:** эти методы рассматривают скопления как плотную область, имеющую некоторые сходства и отличия от более низкой плотной области пространства.
- **Методы, основанные на иерархии:** кластеры, сформированные в этом методе, образуют структуру древовидного типа на основе иерархии. Новые кластеры формируются с использованием ранее сформированного. Делится на две категории
  - 📌 **Агломеративный** (*подход «снизу вверх»*)
  - 📌 **Разделительный** (*подход сверху вниз*)
- **Методы разбиения:** эти методы разбивают объекты на  $k$  кластеров, и каждый раздел образует один кластер.
- **Методы на основе сетки:** в этом методе пространство данных состоит из конечного числа ячеек, которые образуют структуру, подобную сетке.



## Метод к-средних

Это простейший алгоритм обучения без учителя, который решает проблему кластеризации. Алгоритм К-средних разбивает  $n$  наблюдений на  $k$  кластеров, где каждое наблюдение принадлежит кластеру, а ближайшее среднее значение служит прототипом кластера





# Метод к-средних

## Алгоритм работы следующий:

1. Во-первых, мы случайным образом инициализируем  $k$  точек, называемых средними значениями или центроидами кластера.
2. Мы классифицируем каждый элемент по его ближайшему среднему значению и обновляем координаты среднего значения, которые являются средними значениями элементов, классифицированных в этом кластере на данный момент.
3. Мы повторяем процесс для заданного количества итераций, и в конце у нас есть наши кластеры.





## Сферы применения кластеризации



Маркетинг



Обработка изображений



Биология



Генетика



Библиотеки



Финансы



Страхование



Обслуживание клиентов



Городское планирование



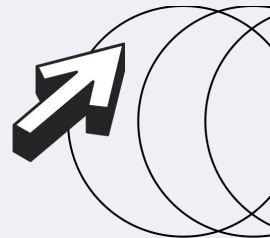
Производство

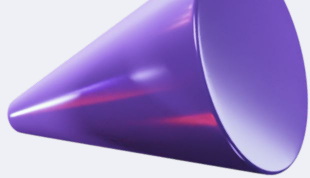


Исследования землетрясений



Давайте делать это в коде!





**Спасибо за внимание**

