OTUS

Рекомендательные системы

Сегментация пользователей и задача персонализации

otus.ru

Проверить, идет ли запись

Меня хорошо видно && слышно?

Ставим “+”, если все хорошо
“-”, если есть проблемы

Тема вебинара

Сегментация пользователей и задача персонализации

Елена Позднеева

Team Lead Data Scientist, MegaTex

Как подобрать **персональный** тариф?
Как **удержать/привлечь** клиента?
Как принести компании **финансовый** эффект?
Как повышать **лояльность** клиента?

Контакты:
@frideliya
www.linkedin.com/in/elena-pozdneeva

Правила вебинара

Активно
участвуем

Off-topic обсуждаем
в учебной группе **#RecSys-2024-10**
https://t.me/+tr0h3CYc3xs4OTdi

Задаем вопрос
в чат или голосом

Вопросы вижу в чате,
могу ответить не сразу

Условные обозначения

Индивидуально

Время, необходимое
на активность

Пишем в чат

Говорим голосом

Документ

Ответьте себе или
задайте вопрос

Маршрут вебинара

- Задача персонализации в бизнесе
- Сегментация на основе эвристик
- Сегментация при помощи K-means и DBSCAN
- Сегментация при помощи логистической регрессии
- Практика
- Рефлексия

Цели вебинара

К концу занятия вы сможете

1. Научиться проводить сегментацию пользователей на основе эвристик
2. Научиться применять кластеризацию для задачи сегментации
3. Научиться проводить сегментацию при помощи логистической регрессии

Смысл

Зачем вам это уметь

1. Понимать, что такое сегментация и чем она важна с точки зрения бизнеса
2. Расширить кругозор в области методики сегментации
3. Научиться подбирать подходящий под задачу способ сегментации

Оцените, пожалуйста, насколько вы знакомы с методами сегментации пользователей?

1 - не знаком, не занимаюсь такой задачей
2 - слышал, что-то знаю
3 - я эксперт!

Задача персонализации в бизнесе

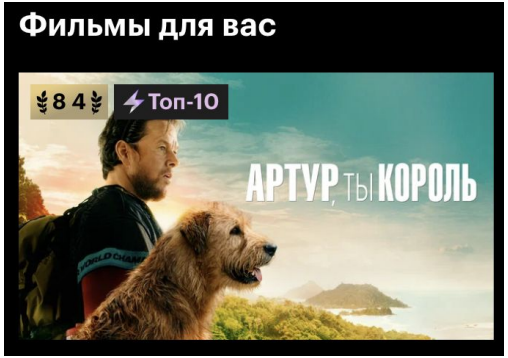
Определение

Персонализация - это процесс адаптации продуктов, услуг или контента для удовлетворения индивидуальных потребностей и предпочтений конкретных пользователей.

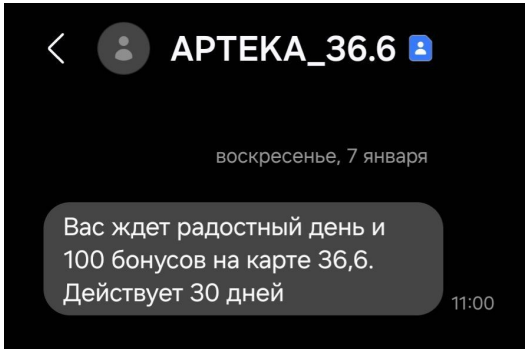
- Улучшение пользовательского опыта
- Повышение вовлеченности пользователей
- Увечение конверсий и продаж
- Повышение лояльности клиентов
- Снижение оттока пользователей

Примеры

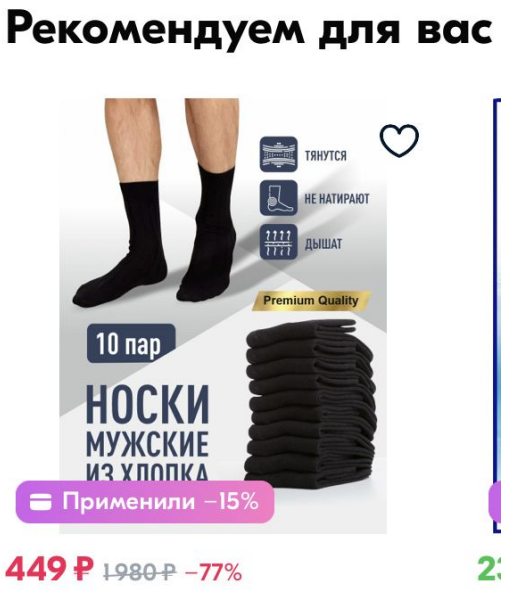
Рекомендации фильмов



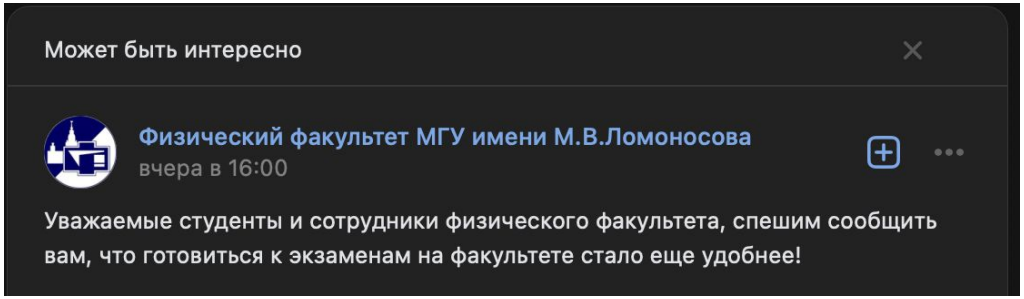
Программы лояльности



Предложения и скидки в интернет-магазинах



Персонализированные новостные ленты в соцсетях



Данные

Источники данных:

- веб-аналитика
- социальные сети
- транзакционные данные
- клиентские опросы и отзывы
- CRM системы

Методы анализа:

- descriptive analytics
- predictive analytics

Типы данных:

- демографические данные
- поведенческие данные
- психографические данные
- контекстуальные данные
- исторические данные

Проблемы и вызовы:

- качество данных
- приватность и безопасность данных
- интеграция данных из различных источников

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Сегментация на основе эвристик

Определение

Эвристическая сегментация - это метод разделения пользователей на группы на основе эмпирических правил и упрощенных моделей, основанных на опыте и наблюдениях.

Плюсы:

- простота и скорость
- интерпретируемость
- практическая применимость

Минусы:

- ограниченная точность
- субъективность
- невозможно учесть сложные зависимости

Методы

демографическая сегментация

психографическая сегментация

поведенческая сегментация

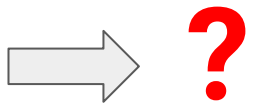
географическая сегментация

RFM-анализ:

- **Recency** (от англ. давность)
  когда пользователь последний раз совершал покупку
- **Frequency** (от англ. частота)
  как часто пользователь совершает покупки
- **Monetary** (от англ. деньги)
  сколько денег пользователь потратил за определенный период

Методика RFM-анализа

Давность покупки	Частота покупки	Сумма покупок
1 — недавние	1 — часто	1 — большая
2 — средняя давность	2 — средняя частота	2 — средняя
3 — давние	3 — низкая частота	3 — маленькая



Методика RFM-анализа

Давность покупки	Частота покупки	Сумма покупок
1 — недавние	1 — часто	1 — большая
2 — средняя давность	2 — средняя частота	2 — средняя
3 — давние	3 — низкая частота	3 — маленькая

27 сегментов

111	121	131
112	122	132
113	123	133
211	221	231
212	222	232
213	223	233
311	321	331
312	322	332
313	323	333

Активные: постоянные клиенты, готовые регулярно покупать
клиенты, которые покупают время от времени на разную сумму
новые клиенты, которые недавно совершили 1-2 покупки

Пассивные: недавние клиенты, которые часто покупали, но потом перестали
недавние клиенты, которые покупают с разной частотой и на разную сумму
недавние клиенты, которые редко покупают на разную сумму

Отток: ранее бывшие активными клиенты, почему снизили активность
клиенты, покупавшие время от времени, но нерегулярно
разовые клиенты, которые не продолжили взаимодействие после первой покупки.

Алгоритм сегментации

1. Собрать данные о транзакциях клиентов на выбранный период (год, квартал)
2. По каждому клиенту рассчитать Recency, Frequency, Monetary
3. Разделить клиентов на квантильные группы по каждому из показателей R, F и M. Определите количество групп в зависимости от потребностей бизнеса. Групп может быть 3, 4, 5 и т.д.
4. Создайте сегменты на основе RFM-баллов (например, 111 или 312)
5. Определите маркетинговые стратегии для каждого сегмента
6. Регулярно обновляйте RFM-анализ, чтобы учитывать изменения в поведении клиентов и реагировать на динамику рынка

https://mindbox.ru/journal/education/chto-takoe-rfm-analiz/

LIVE

Вопросы?

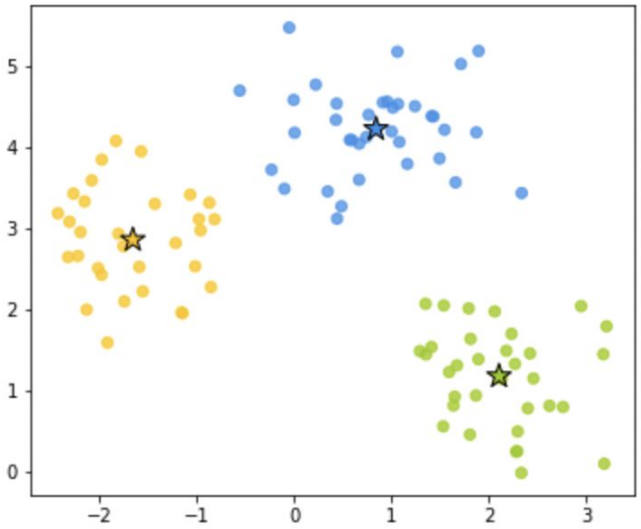
Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Сегментация при помощи K-means

Определение

k-means - это метод кластеризации, который делит данные на K кластеров на основе схожести, минимизируя среднеквадратическую **ошибку разбиения**.

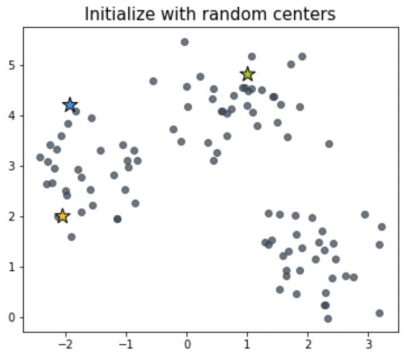


$$e^2 = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|^2 \rightarrow min$$

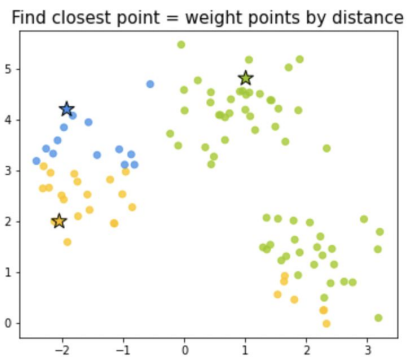
K — количество кластеров, K = 3
n_j — количество объектов в кластере j
c_j — «центр масс» кластера j (точка со средними значениями характеристик для данного кластера)

https://education.yandex.ru/handbook/ml/article/klasterizaciya

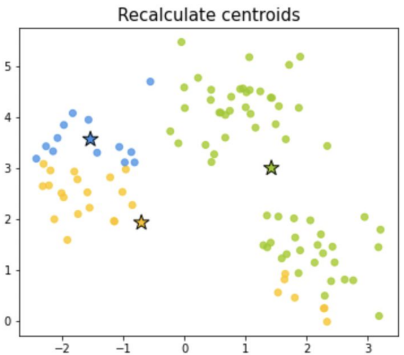
Алгоритм k-means



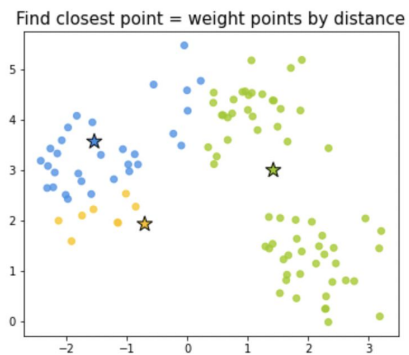
1



2

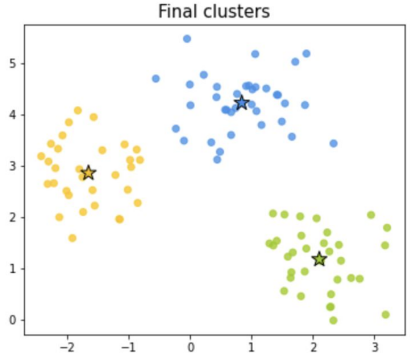


3



4

1. Выбор K начальных случайных центров
2. Для каждого объекта **рассчитываем расстояние** до каждого из K кластеров и относим его к ближайшему
3. **Пересчитываем центры** кластеров как среднее значение точек, принадлежащих каждому кластеру
4. и далее… повторяем расчет расстояний и пересчет центров до тех пор, пока центры не перестанут значительно меняться



https://education.yandex.ru/handbook/ml/article/klasterizaciya

Плюсы и минусы

Плюсы:

- простота и скорость
- интерпретируемость

Минусы:

- выбор K
- чувствительность к начальным центроидам
- форма кластеров

Методы улучшения:

- **метод локтя:** определение оптимального количества кластеров на основе графика суммы квадратов ошибок
- **K-means++:** улучшение инициализации центроидов для достижения более стабильных результатов
- **силуэтный анализ:** оценка качества кластеров на основе внутрикластерного расстояния и расстояния до ближайшего кластера

LIVE

Вопросы?

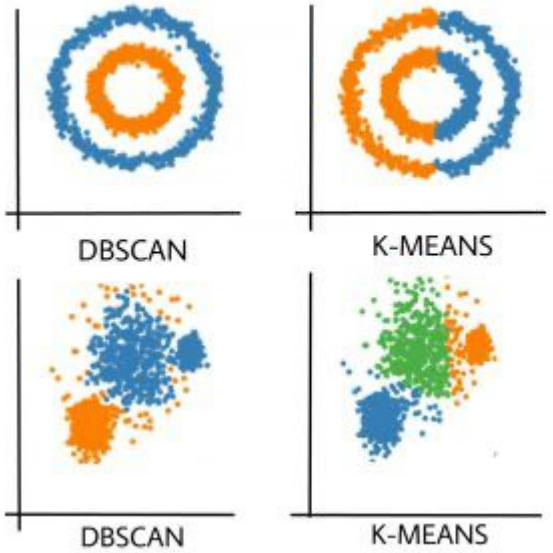
Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Сегментация при помощи DBSCAN

Определение

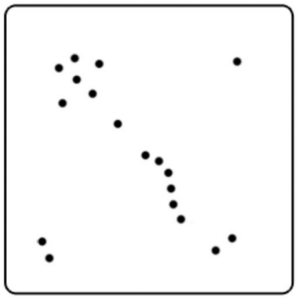
DBSCAN (Density-based spatial clustering of applications with noise) - плотностной алгоритм пространственной кластеризации с присутствием шума.



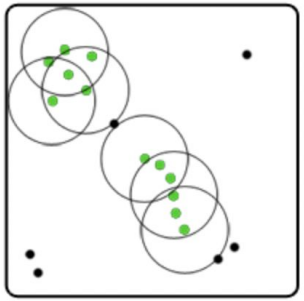
умеет работать с кластерами сложной формы

умеет находить шумы

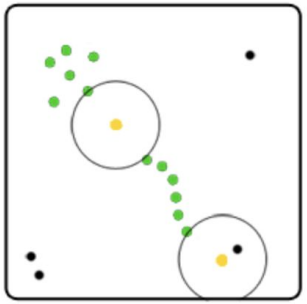
Алгоритм DBSCAN



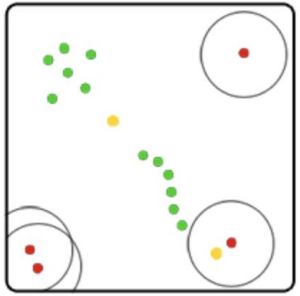
Задача: разбить толпу на кластеры



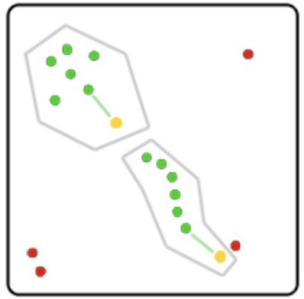
Все, у кого есть хотя бы три соседа на расстоянии метра, берут в руки зелёные флажки



Если меньше трёх соседей? Если хотя бы один сосед держит зеленый флаг, вручим жёлтые флаги



Красные: меньше трех соседей и нет соседей с зелеными флагами



Присвоить желтых в один из зеленых кластеров

Можно не помечать всех зеленых, а начать с одного рандомного:

- соседей меньше трех? кандидат в “отшельники”
- соседей 3 и больше? красим его зеленым и обходим соседей

- Распределяем “отшельников”

Плюсы и минусы

Плюсы:

- обнаружение кластеров произвольной формы
- работа с шумом и выбросами
- не требует предварительного задания количества кластеров
- эффективен для больших наборов данных

Минусы:

- чувствительность к параметрам
- трудности с кластеризацией данных с переменной плотностью
- проблемы с масштабируемостью для очень больших данных
- неопределенность при выборе параметров

Сравнение DBSCAN и K-means

Критерий	DBSCAN	K-means
Основные параметры	epsilon, min_samples	кол-во кластеров
Инициализация	не требует инициализации центроидов	требует случайной инициализации центроидов
Форма кластеров	произвольная	сферическая (круговая)
Работа с шумом и выбросами	эффективно выявляет шум и выбросы	плохо работает с шумом и выбросами
Кол-во кластеров	определяется автоматически	задается заранее
Подходящие данные	кластеры произвольной формы и плотности	кластеры одинаковой плотности и формы
Примеры применения	геопространственная аналитика, анализ изображений, выявление аномалий	маркетинговая сегментация, анализ потребителей

LIVE

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Сегментация при помощи логистической регрессии

Определение

Логистическая регрессия - метод статистического анализа, используемый для предсказания бинарного исхода (класса) на основе одной или нескольких независимых переменных.

Математическая модель: $P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}}$

$P(Y = 1 \mid X)$ - вероятность положительного исхода

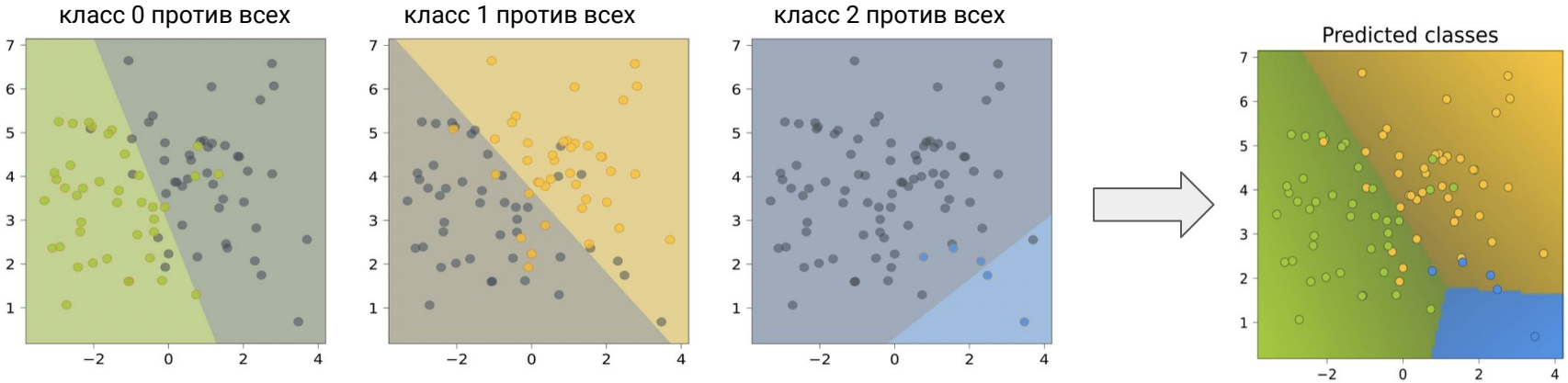
X_i - независимые переменные

β_i - коэффициенты модели показывают, как изменение независимой переменной влияет на вероятность положительного исхода. Положительное значение коэф-та указывает на увеличение вероятности события с увеличением значения переменной

Мультиномиальная логистическая регрессия - метод статистического анализа, используемый для предсказания одного из K классов на основе одной или нескольких независимых переменных.

Строим K линейных моделей и используем нормировку **softmax**.

Логрег для многих классов



Алгоритм разделения на 3 сегмента:

1. обучаем 3 линейный классификатора для выявления одного класса среди всех остальных
2. каждому классу присваивается самый “уверенный” классификатор

Проблема?

- выходы классификаторов могут иметь разные масштабы
- нормировка искажает результат

https://education.yandex.ru/handbook/ml/article/klasterizaciya

LIVE

Вопросы?

Ставим “+”,
если вопросы есть

Ставим “–”,
если вопросов нет

Цели вебинара

Проверка достижения целей

1. Научились проводить сегментацию пользователей на основе эвристик
2. Научились применять K-means и DBSCAN для задачи сегментации
3. Научились проводить сегментацию при помощи логистической регрессии

Вопросы для проверки

По пройденному материалу всего вебинара

1. Зачем бизнесу сегментация базы?
2. Какие эвристические методы сегментации вы знаете?
3. Какие методы сегментации на основе близости вы знаете?

Ключевые тезисы занятия

Подведем итоги

1. Сегментация помогает бизнесу найти персональный подход к категориям своих клиентов
2. Очень часто задачи сегментации можно решить простыми эвристическими методами, например с помощью RFM-анализа
3. Для кластеризации клиентов в маркетинговых задачах можно использовать K-means
4. Для решения задач с геоданными или данными с выбросами можно использовать DBSCAN
5. Если требуется построить линейный классификатор, можно использовать линейную логистическую регрессию

Рефлексия

Рефлексия

Будете ли применять на практике то,
что узнали на вебинаре?

Следующий вебинар

12 ноября 2024

Look-a-like выделение сегментов пользователей

Ссылка на вебинар
будет в ЛК за 15 минут

Заполните, пожалуйста, опрос о занятии по ссылке в чате

Спасибо за внимание!

Приходите на следующие вебинары

Елена Позднеева

Team Lead Data Scientist, МегаТех

@frideliya + чат группы https://t.me/+tr0h3CYc3xs4OTdi