РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ

ЧТО МЫ ХОТИМ НАУЧИТЬСЯ ДЕЛАТЬ

МФТИ

- Есть база пользователей и база объектов (фильмов, музыки, товаров в интернет-магазине)
- Есть обратная связь от пользователей: оценки, просмотры, покупки
- Нужно научиться рекомендовать пользователю то, что ему понравится

РЕКОМЕНДАЦИИ ФИЛЬМОВ: ВОЗМОЖНАЯ ПОСТАНОВКА ЗАДАЧИ

МФТИ

- Есть известные оценки, которые пользователи поставили уже просмотренным фильмам
- Нужно:
  - Спрогнозировать оценки, которые поставили бы пользователи другим фильмам
  - Порекомендовать пользователям то, что им больше понравится

РЕКОМЕНДАЦИИ ФИЛЬМОВ

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

РЕКОМЕНДАЦИИ ФИЛЬМОВ

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	?
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

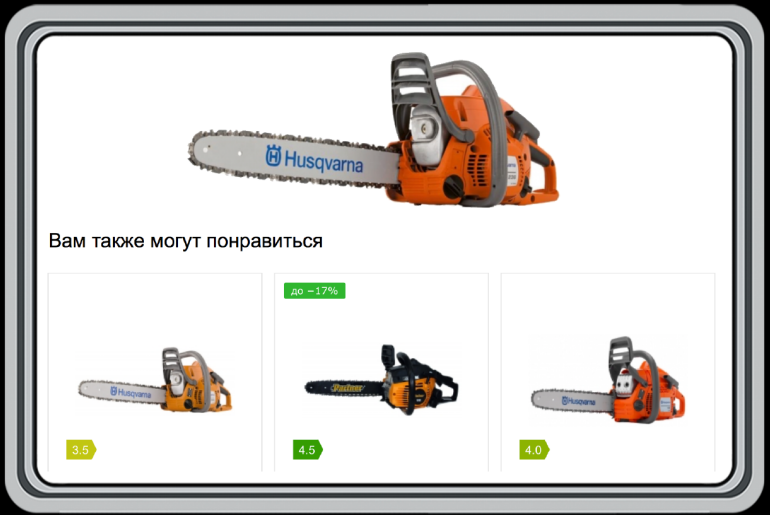
РЕКОМЕНДАЦИИ ТОВАРОВ

МФТИ

	Платье	Туфли	Кожаная куртка	Лабутены
Маша	1	1	1	1
Юля		1	1	?
Вова			1	1
Коля			1	1
Петя				1
Ваня		1	1	1

USER 2 ITEM

ITEM 2 ITEM



ITEM 2 ITEM ПО PMI

МФТИ

$$\text{PMI} = \log \frac{p(x, y)}{p(x)p(y)}$$

- $p(x)$ — вероятность встретить объект (в пользовательской сессии/среди купленных/среди понравившихся)
- $p(y)$ — вероятность встретить объект y

ДОПОЛНИТЕЛЬНЫЕ ОГРАНИЧЕНИЯ

МФТИ

- Хиты
- Из той же категории
- Смотрят/слушают/покупают вместе

РЕЗЮМЕ

МФТИ

- Рекомендации
- User 2 item и item 2 item
- Дополнительные ограничения

kNN
И МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

USER-BASED kNN

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

USER-BASED kNN

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	?
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

ITEM-BASED kNN

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	?
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

ITEM-BASED kNN

МФТИ

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	?
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

МФТИ

j

	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	
Вова			3	5
Коля	3	?	4	5
Петя				4
Ваня		5	3	3

i

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ

МФТИ

- u_i — «интересы пользователей»
- v_j — «параметры фильмов»

$$x_{ij} \approx \langle u_i, v_j \rangle = \sum_{k=1}^{K} u_{ik} v_{jk}$$

МАТРИЧНЫЕ РАЗЛОЖЕНИЯ: НАСТРОЙКА ПРОФИЛЕЙ

МФТИ

$$x_{ij} = \langle u_i, v_j \rangle$$

$$\sum_{i,j} (\langle u_i, v_j \rangle - x_{ij})^2 \to \textbf{min}$$

РЕЗЮМЕ

МФТИ

- Матрица user-item
- kNN
  - User-based
  - Item-based
- Матричные разложения

ПОДХОДЫ К ПОСТРОЕНИЮ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

ПОДХОДЫ К ПОСТРОЕНИЮ РЕКОМЕНДАЦИЙ

МФТИ

- Collaborative filtering
- Content-based
- Demographic
- Utility-based
- Knowledge-based

COLLABORATIVE FILTERING

МФТИ

- Рекомендации для пользователя строятся на основе оценок похожих пользователей

CONTENT-BASED

МФТИ

- Рассчитываются признаки для пользователей и объектов
- Строится модель классификации/регрессии, приближающая оценки пользователей

DEMOGRAPHIC

МФТИ

- Производится сегментация пользователей на группы
- Рекомендации строятся на основе предпочтений группы

UTILITY-BASED

МФТИ

- Для каждого пользователя строится utility function
- Как построить user based utility function?

KNOWLEDGE-BASED

МФТИ

- Строится база знаний о том, как объекты I соотносятся с интересами и предпочтениями пользователя

ПОДХОДЫ К ПОСТРОЕНИЮ РЕКОМЕНДАЦИЙ

МФТИ

- **Collaborative filtering**
- **Content-based**
- Demographic
- Utility-based
- Knowledge-based

COLLABORATIVE VS CONTENT

МФТИ

- Advantages:
  - Cross-genre interests
  - Implicit feedback
  - Quality improving over time
- Problems:
  - “Cold-start” problem for users
  - “Cold-start” problem for items
  - “Gray sheep” problem
  - “Everybody likes bananas”

COLLABORATIVE VS CONTENT

МФТИ

- Advantages:
  - Cross-genre interests
  - Implicit feedback
  - Quality improving over time
- Problems:
  - “Cold-start” problem for users
  - “Cold-start” problem for items
  - “Gray sheep” problem
  - “Everybody likes bananas”

COLLABORATIVE VS CONTENT

МФТИ

- Advantages:
  - Cross-genre interests
  - Implicit feedback
  - Quality improving over time
- Problems:
  - “Cold-start” problem for users
  - “Cold-start” problem for items
  - “Gray sheep” problem
  - “Everybody likes bananas”

WHAT IF...

МФТИ

- Доступно достаточно информации о пользователях и их предпочтениях?
- Доступно достаточно информации об объектах?
- Хочется получить преимущества обоих подходов?
- Давайте их объединять!

МФТИ

ГИБРИДНЫЕ РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ

ЗАЧЕМ СТРОИТЬ ГИБРИДЫ

МФТИ

- Разные имеют свои недостатки и преимущества
- Вместо выбора — гибко используем всё

КАКИЕ БЫВАЮТ ВИДЫ ГИБРИДИЗАЦИИ

МФТИ

- Weighted
- Switching
- Mixed
- Feature combination
- Cascade
- Feature augmentation

WEIGHTED

МФТИ

- Рекомендации строятся на основе комбинирования оценок от разных систем с весами
- Например:
  - Линейная комбинация
  - Голосование

SWITCHING

МФТИ

- Рекомендации строятся путем переключения между системами, работающими независимо, на основании критериев для переключения

MIXED

МФТИ

- Список рекомендаций состоит из «смеси» рекомендаций от разных систем

FEATURE COMBINATION

МФТИ

- Подход основан на content-based
- Признаки от разных систем объединяются в одну выборку для построения единой модели

CASCADE

МФТИ

- Поэтапное применение нескольких моделей для уточнений рекомендаций
- Candidate selection

FEATURE AUGMENTATION

МФТИ

- Выход от одной или нескольких рекомендательных систем используется как входные признаки для другой системы

ГИБРИДИЗАЦИЯ

МФТИ

- Часто улучшает качество рекомендаций
- Иногда положительно сказывается на разнообразии
- Не гарантирует решения всех проблем, связанных с тем или иным подходом

ОФФЛАЙН ОЦЕНКА КАЧЕСТВА

КАК ИЗМЕРИТЬ КАЧЕСТВО?

МФТИ

- Качество модели = качество прогноза оценок?
  - Среднеквадратичное отклонение (RMSE)
  - Среднее абсолютное отклонение (MAE)

ПРАВИЛЬНО ЛИ МЫ ЖИВЁМ?

МФТИ

- Что мы оцениваем: качество прогноза оценок
- Что нужно оценивать: качество рекомендаций

ТОЧНОСТЬ (Precision@k)

МФТИ

- k — количество рекомендаций

$$\text{Precision@}k = \frac{\text{купленное из рекомендованного}}{k}$$

- AveragePrecision@k — усреднённый по сессиям Precision@k

ПОЛНОТА (Recall@k)

МФТИ

- k — количество рекомендаций

$$\text{Recall@}k = \frac{\text{купленное из рекомендованного}}{\text{количество покупок}}$$

- AverageRecall@k — усреднённый по сессиям Recall@k

ВЗВЕШЕННЫЙ ЦЕНАМИ Recall@k

МФТИ

Рекомендованные товары
Синяя футболка — 1000р
Красная футболка — 1200р
Кроссовки — 3500р
Кепка — 900р
Зелёная футболка — 800р

Купленные товары
Красная футболка — 1200р
Кеды — 3000р
Кепка — 900р

Взвешенный ценами Recall@k =

$$= \frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$$

РЕЗЮМЕ

МФТИ

- Проблема оценки качества по MSE и MAE
- Precision@k
- Recall@k
- Учёт цен товаров в Recall@k

ОНЛАЙНОВАЯ ОЦЕНКА КАЧЕСТВА

ОНЛАЙНОВАЯ ОЦЕНКА КАЧЕСТВА

МФТИ

- Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

ОНЛАЙНОВАЯ ОЦЕНКА КАЧЕСТВА

МФТИ

- Допустим, на исторических данных качество алгоритма высокое, а будет ли оно высоким в реальности?

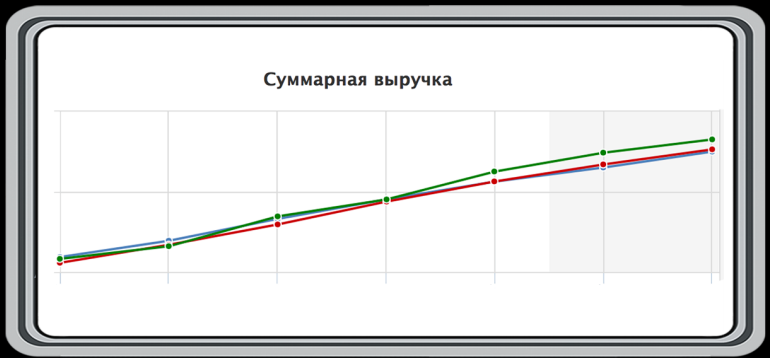
- Идеи:
  - A/B тест
  - Оценка статзначимости результата

A/B ТЕСТ

МФТИ

- Случайным образом делим пользователей на равные группы
- Измеряем целевые метрики (например, количество заказов или доход) в каждой группе за длительный период времени
- Получаем какое-то число для каждой группы
- Что дальше?

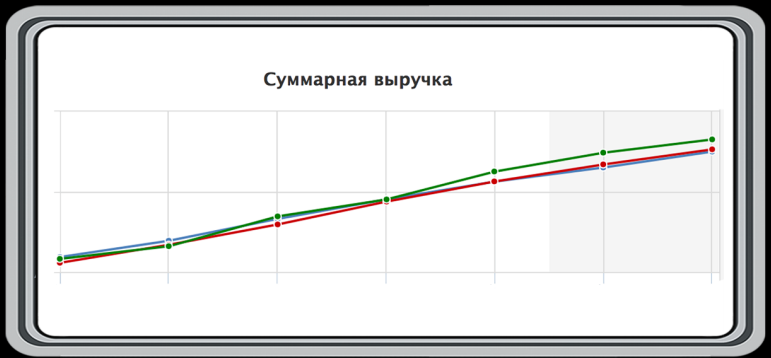
СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ: ПРИМЕР



СТАТИСТИЧЕСКАЯ ЗНАЧИМОСТЬ: ПРИМЕР

МФТИ

- Одна кривая отличается от других на **10%**
  Но разбиение на самом деле — случайное



ЧАСТО ПРИМЕНЯЮТСЯ НА ПРАКТИКЕ

МФТИ

- Приближение нормальным распределением
- Тест Стьюдента
- Бутстреп

НА КАКИЕ МЕТРИКИ СМОТРЯТ В ОНЛАЙНЕ

МФТИ

- Доход в группе
- Доход с пользовательской сессии
- Средняя стоимость купленного товара
- Средний чек
- Конверсия в покупку
- Клики
- Различные модели атрибуции: last click, first click

РЕЗЮМЕ

МФТИ

- A/B тест
- Статзначимость
- Метрики

МАКСИМИЗАЦИЯ ПРИБЫЛИ МАГАЗИНА

РЕКОМЕНДАЦИИ ТОВАРОВ

МФТИ

j

i	Вечернее платье	Кеды	Джинсы	Футболка
Маша	1		1	
Юля	1	1		1
Вова		1	1	
Коля	1	?	1	
Петя		1	1	
Ваня			1	1

ОТЛИЧИЯ ОТ РЕКОМЕНДАЦИЙ ФИЛЬМОВ И МУЗЫКИ

МФТИ

- Нет негативных примеров
- Понятней связь с прибылью

ЧТО МОЖЕМ ДЕЛАТЬ

МФТИ

- Прогнозировать, какие товары будут куплены
- Максимизировать прибыль

МАКСИМИЗАЦИЯ ДОХОДА

МФТИ

МАКСИМИЗАЦИЯ ДОХОДА

МФТИ

Вероятность	p_1	p_2	p_3	p_4
Цена	c_1	c_2	c_3	c_4

Товар 1	Товар 2	Товар 3	Товар 4

МАКСИМИЗАЦИЯ ДОХОДА

МФТИ

Вероятность	0.05	0.02	0.015	0.009
Цена	3490	1990	1590	1970

МАКСИМИЗАЦИЯ ПРИБЫЛИ

МФТИ

Вероятность	0.05	0.02	0.015	0.009
Цена	3490	1990	1590	1970
Маржинальность	0.1	0.4	0.4	0.2

ПРОГНОЗИРОВАНИЕ ВЕРОЯТНОСТИ

МФТИ

- Объекты: тройки (пользователь, товар, момент времени)
- Классы: 1 — товар будет куплен, 0 — товар не будет куплен
- Признаки: параметры пользователя, товара, момента времени и их «взаимодействие»

ОТБОР КАНДИДАТОВ

МФТИ

- Популярные
- Из тех же категорий
- С высоким PMI с уже просмотренными/ понравившимися
- Из заранее подготовленных списков похожих товаров

ГЕНЕРАЦИЯ НЕГАТИВНЫХ ПРИМЕРОВ

МФТИ

- Добавить к каждому позитивному примеру весь каталог как негативный (не реально)
- Случайные с равномерным распределением
- Случайные, с вероятностями, пропорциональными популярности объекта
- Самые популярные примеры
- Те объекты, которые рекомендовал бы какой-то алгоритм, но они не были куплены

РЕЗЮМЕ

МФТИ

- Построение рекомендаций с учетом желаемого экономического эффекта
- Прогнозирование вероятности покупки товара
- Отбор кандидатов
- Генерация негативных примеров