Введение в анализ данных

Лекция 12

Рекомендательные системы

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2017

Опрос

- Какие рекомендательные системы вы знаете?

Опрос

- Какие рекомендательные системы вы знаете?
- Рекомендации чего вы хотели бы получать?

Рекомендательные системы

- Фильмы, видео
- Музыка
- Книги
- Приложения
- Товары
- Посты в социальных сетях
- Баннерные системы
- Люди (социальные сети, сервисы знакомств)
- Услуги (рестораны, отели, …)
- Научные публикации

Рекомендательные системы

- Рекомендательные системы сокращают объём информации, необходимый для принятия решения
- Не нужно читать отзывы на 1000 фильмов — модель сама выберет лучший

- Netflix: 2/3 просмотренных фильмов найдены через рекомендательную систему
- Amazon: 35% продаж через полки рекомендаций

На основе чего можно строить рекомендации?

На основе чего можно строить рекомендации?

- Профиль пользователя — «что посмотреть, если я люблю комедии с известными актёрами?»
- Данные по другим пользователям — «что смотрят люди с похожими на мои интересами?»
- Данные по объектам (фильмам) — «какие фильмы похожи на те, которые мне понравились?»

Цели с точки зрения продавца

- ?

Цели с точки зрения продавца

- Продать больше товаров
- Продать больше редких товаров
- Повысить лояльность пользователя
- Лучше понять покупателей

Цели с точки зрения покупателя

Цели с точки зрения покупателя

- Купить то, что нужно
- Понять, что покупать вместе с данным товаром
- Понять, что интересно (если нет задачи купить что-то конкретное)

Краткая история

- Начало 90-х: одна из первых рекомендательных систем (GroupLens, рекомендации записей в Usenet)
- Начало 2000-х: активные исследования, коммерциализация
- 2006: Netflix Prize
- 2007: первая конференция RecSys

Netflix Prize

- Предсказываем, какую оценку пользователь поставит фильму
- Метрика: RMSE
- Задача: улучшить на 10% качество предсказания
- Конкурс шёл с 02.10.2006 по 21.09.2009
- Главный приз: $1,000,000
- Размеры:
  - 500 тысяч пользователей
  - 17 тысяч фильмов
  - 10^8 рейтингов

Netflix Prize

- Одно из первых крупных соревнований по анализу данных (предшественник kaggle и т.д.)
- Первый большой открытый набор данных для тестирования алгоритмов рекомендаций
- Алгоритмы, разработанные участниками конкурса, до сих пор популярны в индустрии
- Netflix Prize привёл к большой популярности RMSE как метрики качества рекомендаций (не самый лучший результат)

Netflix Prize

Методы

- Коллаборативная фильтрация
  - Рекомендации на основе сходства действий пользователей
- Контентные рекомендации
- Гибридные методы

Memory-based models

Обозначения

- Множество товаров: I
- Множество пользователей: U
- Множество пар «пользователь-товар», для которых известны оценки: R
- Если для пары (u, i) известен рейтинг, то будем писать $\exists r_{ui}$
- Оценки — рейтинги фильмов, индикаторы покупки товара и т.д.

Оценки

- Оценки (или фидбэк) бывают явные и неявные
- Явные оценки
  - Пользователь поставил оценку фильму/товару
  - Пользователь написал отзыв
  - Пользователь поставил лайк
- Неявные оценки
  - Пользователь посмотрел фильм
  - Пользователь добавил товар в корзину
  - Пользователь долго смотрел на запись в социальной сети
- Неявные оценки более шумные, но их больше

Сходство пользователей

- $I_{uv} = \{i \in I \mid \exists r_{ui} \text{ и } \exists r_{vi}\}$ — множество товаров, которые оценили и пользователь u, и пользователь v
- Сходство пользователей:

$$w_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{uv}} (r_{vi} - \bar{r}_v)^2}},$$

где $\bar{r}_u$ и $\bar{r}_v$ — средние рейтинги пользователей

User-based collaborative filtering

- Дан пользователь u_0
- Найдём пользователей, которые похожи на него:

$$U(u_0) = \{v \in U \mid w_{u_0 v} > \alpha\}$$

- Порекомендуем те товары, которые часто покупались пользователями из $U(u_0)$

User-based collaborative filtering

Товары

Пользователи

1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1

User-based collaborative filtering

Товары

Пользователи

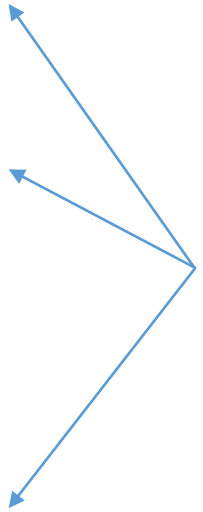
1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1

User-based collaborative filtering

Товары

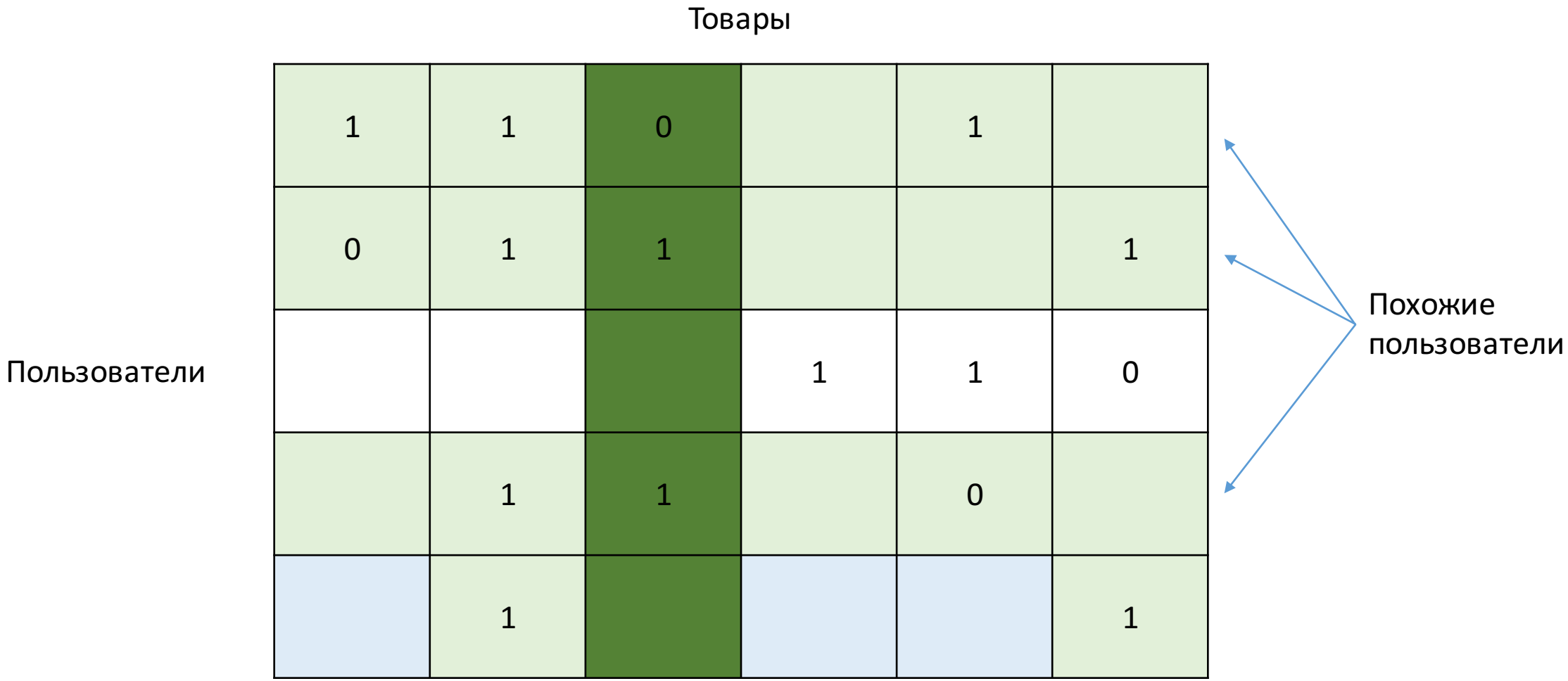
Пользователи

1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1



Похожие пользователи

User-based collaborative filtering



User-based collaborative filtering

Недостатки:

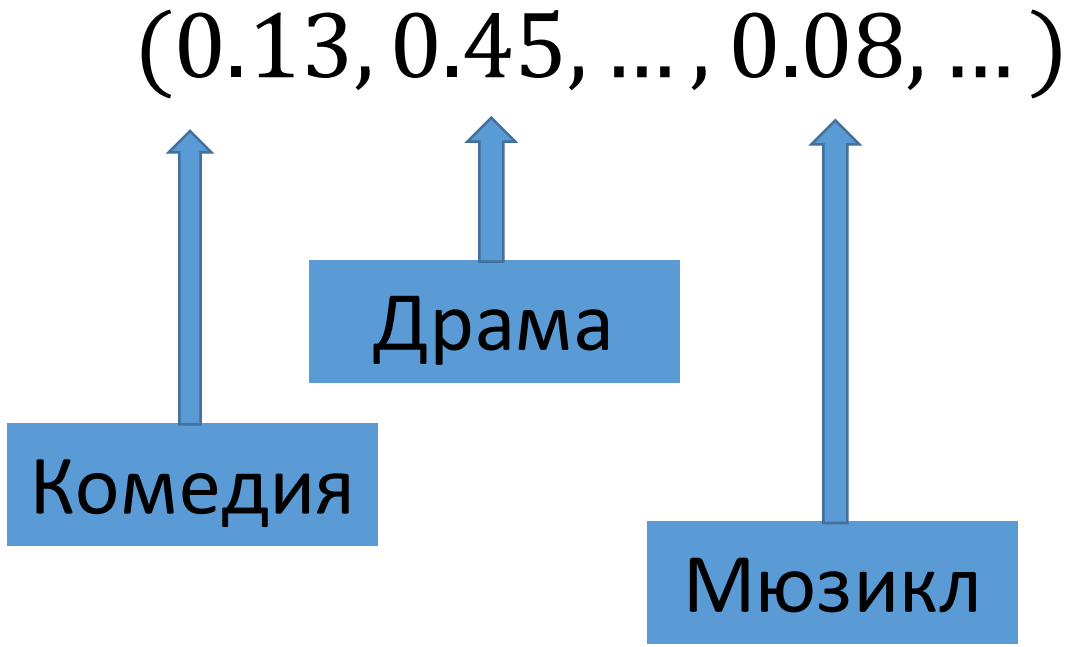
- Много параметров, которые сложно выбирать
  - Какой порог сходства для пользователей?
  - Сколько похожих пользователей должны были купить товар, чтобы мы его порекомендовали?
- Требуется хранить всю матрицу оценок

Есть и другие методы, основанные на сходствах, но все обладают теми же недостатками.

Модели со скрытыми переменными

Векторы интересов

- Для пользователя — насколько он интересуется каждым жанром
- Для фильма — насколько он относится к каждому жанру



Рейтинг

- Предположение: заинтересованность определяется как скалярное произведение векторов пользователя и фильма

$$(0.1, 0.5, 0.01, 0.92) \times (0, 0, 0.1, 0.95) = 0.875$$

$$(0.1, 0.5, 0.01, 0.92) \times (0.9, 0, 0, 0.1) = 0.182$$

Пользователь

Фильм

Модели со скрытыми переменными

- Обучим вектор p_u для каждого пользователя u
- Обучим вектор q_i для каждого товара i
- Оценка приближается их скалярным произведением:

$$r_{ui} \approx \langle p_u, q_i \rangle$$

- Находим векторы только по известным оценкам
- После этого можем предсказать оценку для любой пары «пользователь-товар»

Модели со скрытыми переменными

- Оптимизационная задача:

$$\sum_{(u,i)\in R} \left(r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle\right)^2 \to \min_{P,Q}$$

- Решение: градиентный спуск, Alternating Least Squares (ALS) и другие методы

Модели со скрытыми переменными

2	5	
5		4
	1	
	2	5

Модели со скрытыми переменными

	(0.9, 0.05)	(0.02, 1.1)	(1.05, 0.01)
(2.1, 5)	2	5	
(4.6, 0)	5		4
(0, 1)		1	
(4.9, 0.9)		1	5

Контентные методы

Контентные рекомендации

- Сведём задачу к обычному обучению с учителем
- Объект: пара «пользователь-товар» (u, i)
- Ответ: отклик пользователя
- Факторы: информация про пользователя и про товар
- Обучаем любую модель на этих данных
- Среди факторов могут быть и прогнозы коллаборативных моделей

Метрики качества рекомендаций

Качество предсказаний

- Насколько хорошо мы предсказываем оценки r_{ui}?
- Разделяем сессии пользователей на две части: обучаемся на первой, измеряем качество предсказания второй
- Оцениваем, насколько хорошо предсказываем поведение пользователя — но не факт, что нужно именно это
- Зачем рекомендовать то, что он и так купил бы?

Другие метрики

- Покрытие
  - Какая доля товаров рекомендовалась хотя бы раз?
  - Какой доле пользователей хотя бы раз показаны рекомендации?
- Новизна
  - Как много рекомендованных товаров пользователь встречал раньше?
- Прозорливость (serendipity)
  - Способность предлагать товары, которые отличаются от купленных ранее
- Разнообразие

Резюме

- Рекомендации — широкая задача с большим количеством коммерческих применений
- Модели: коллаборативная фильтрация, контентный подход
- Рекомендации товаров на основе сходства пользователей
- Модели со скрытыми переменными
- Обилие метрик качества