

Введение в анализ данных

Лекция 17

Рекомендательные системы

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2020

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

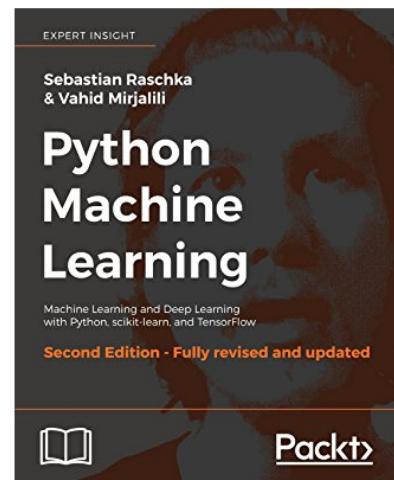
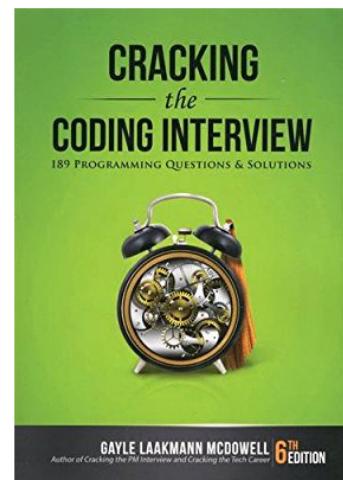
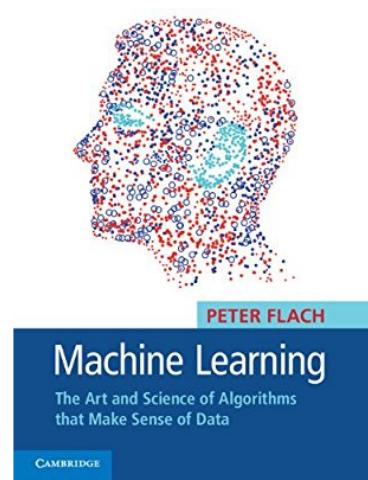
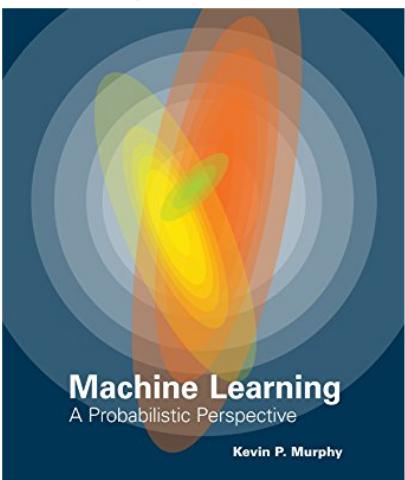
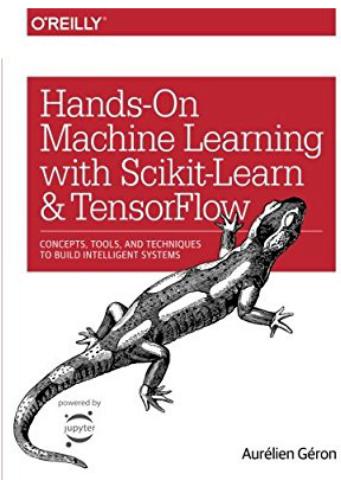


Рекомендательные системы

- 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
- Youtube: 60% просмотров благодаря рекомендациям

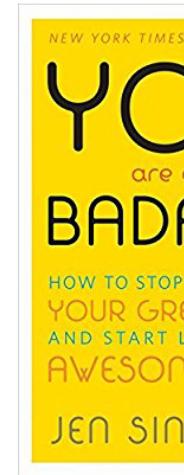
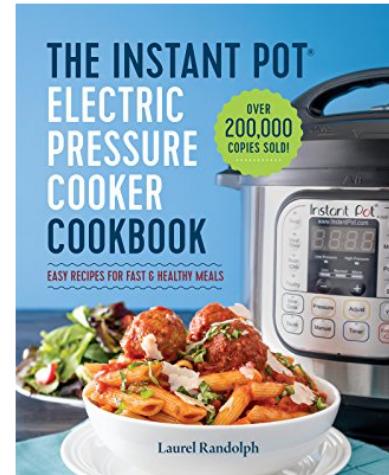
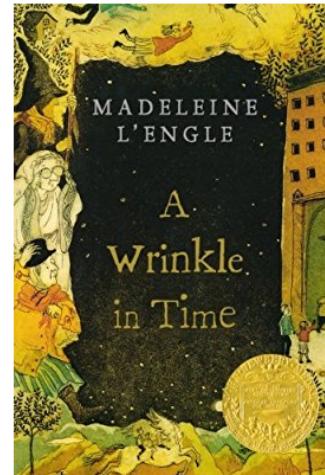
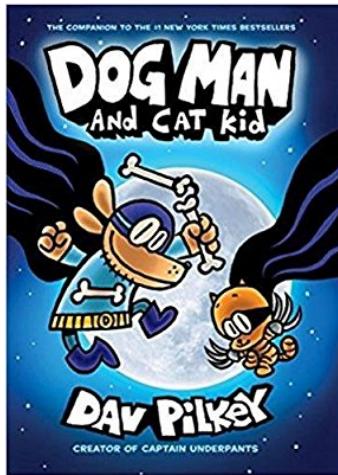
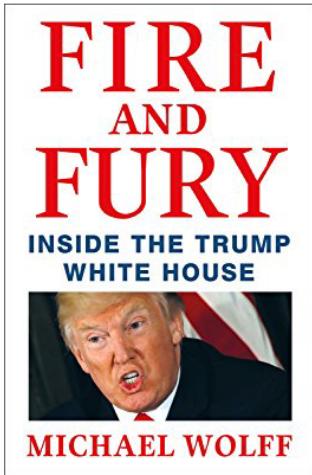
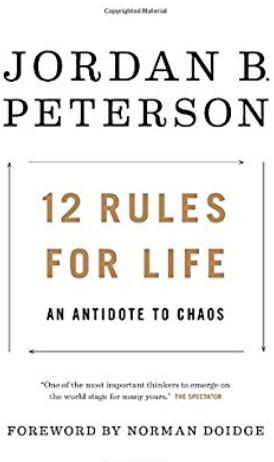
Amazon

Recommendations for you in Books



Amazon

Books best sellers [See more](#)



Netflix

House of Cards

★★★★★ 2013 TV-MA 1 Season HD 51

Sharks gliding ominously beneath the surface of the water? They're a lot less menacing than this Congressman.

This winner of three Emmys, including Outstanding Directing for David Fincher, stars Kevin Spacey and Robin Wright.

Because you watched Orange Is the New Black

Because you watched Red Lights

Netflix

Profile Type	Score Image A	Score Image B
Comedy	5.7	6.3
Romance	7.2	6.5



Image A



Image B

Рекомендации контента

- Медийный бум приводит к взрывному росту объёмов информации в сети
- Рекомендательные системы помогают ориентироваться
- Для авторов — поиск целевой аудитории
- Пионеры в Китае — Toutiao (более 100 миллионов активных пользователей) и другие платформы

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



На основе чего можно строить
рекомендации?

На основе чего можно строить рекомендации?

- Данные по другим пользователям — «что смотрят люди с похожими на мои интересами?»
- Данные по объектам (фильмам) — «какие фильмы похожи на те, которые мне понравились?»

Типичная рекомендательная система

- Объект: пара «user-item»
- Целевая переменная: клики, длинные клики, просмотры, покупки, прослушивания, лайки и т.д.
- Решаем задачу классификации/регрессии/ранжирования

Типичная рекомендательная система

- Объект: пара «user-item»
- Целевая переменная: клики, длинные клики, просмотры, покупки, прослушивания, лайки и т.д.
- Решаем задачу классификации/регрессии/ранжирования

Особенности:

- Выбор целевой переменной
- Выбор метрики качества
- Факторы для модели
- Слишком много товаров/видео/песен/...

Отбор кандидатов

- Простая и быстрая модель, которая отбирает тысячи товаров для данного пользователя
- Сложная модель применяется только к отобранным кандидатам

Основные подходы

- Есть методы, разработанные напрямую для рекомендаций
- Коллаборативная фильтрация
 - Рекомендации на основе сходства действий пользователей
- Контентные рекомендации

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			
		1	1	0	
		1	0	1	
Пользователи					1
				1	1
				0	
		1	1		
	1				1

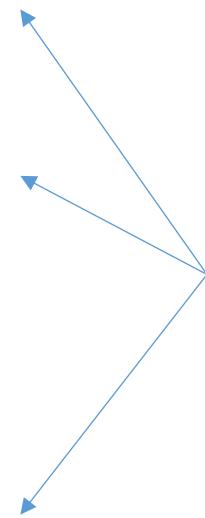
User-based collaborative filtering

		Товары					
		1	1	0		1	
		0	1	1			1
Пользователи					1	1	0
			1	1		0	
			1				1

User-based collaborative filtering

		Товары					
		1	1	0		1	
		0	1	1			1
Пользователи					1	1	0
			1	1		0	
		1					1

Похожие пользователи



User-based collaborative filtering

		Товары			
		1	1	0	
		0	1	1	
Пользователи					
				1	1
			1	1	0
		1	1		
	1			0	
					1

Похожие пользователи

The diagram illustrates a user-based collaborative filtering matrix. The matrix has 'Пользователи' (Users) on the rows and 'Товары' (Products) on the columns. The matrix entries are binary values (0 or 1). A blue arrow points from the bottom row (User 5) to the right, labeled 'Похожие пользователи' (Similar users), indicating that User 5 is being compared against other users to find similarities.

User-based collaborative filtering

Недостатки:

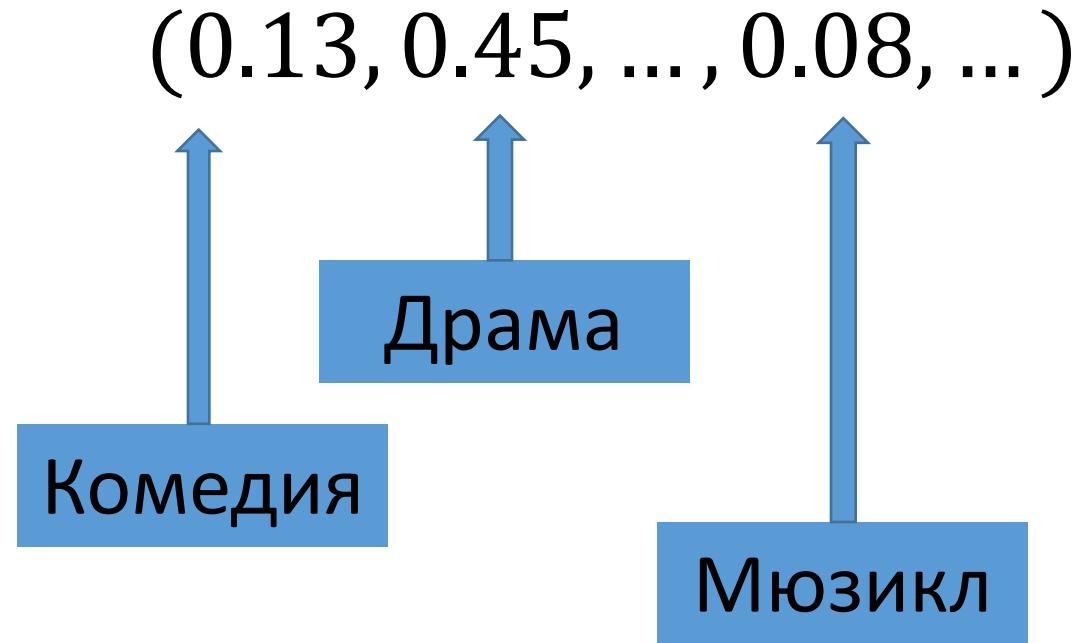
- Много параметров, которые сложно выбирать
 - Какой порог сходства для пользователей?
 - Сколько похожих пользователей должны были купить товар, чтобы мы его порекомендовали?
- Требуется хранить всю матрицу оценок

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:

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

- 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

В зависимости от целевой переменной:

- MSE, MAE, R²
- Accuracy, HitRate, precision/recall, AUC-ROC
- Метрики качества ранжирования (далее в курсе)

Качество предсказаний

Позиция	Релевантность
1	1
2	0
3	1
4	1
5	0
6	0
7	1
...	...

Качество предсказаний

Позиция	Релевантность
1	1
2	0
3	1
4	1
5	0
6	0
7	1
...	...

$$\text{precision}@5 = \frac{3}{5}$$

$$\text{recall}@5 = \frac{3}{4} \text{ (если всего 4 релевантных айтема)}$$

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