

Машинное обучение

Лекция 12.1. Рекомендательные системы



План рассказа

- Рекомендательные системы: что про них рассказывают и с чем вы столкнетесь
- Как строят рекомендательные системы на практике
- Оценка качества и проблемы из реального мира
- Необычные применения рекомендаций

Рекомендательные системы:
что про них рассказывают и с чем вы столкнетесь

Что хотим научиться делать

- 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

- 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

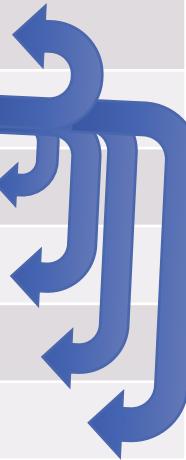
	Пила	Улица Вязов	Ванильное небо	1+1
Маша	5	4	1	2
Юля		5	2	
Вова			3	5
Коля	3		4	5
Петя				4
Ваня		5	3	3

Рекомендации фильмов

	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

	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

	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

	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

	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

j

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

i

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

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

Как измерить качество?

Качество модели = качество прогноза оценок?

- Среднеквадратичное отклонение (RMSE)
- Среднее абсолютное отклонение (MAE)

Правильно ли мы живем?

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

Другой пример: рекомендации товаров

j

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

i

Отличия от прошлой задачи

- Нет негативных примеров
- Понятней связь с прибылью

С чем мы столкнулись?

Товар 1	Товар 2	Товар 3	Товар 4

Нужно заполнить так, чтобы увеличить выручку

В чем проблема с kNN и MF?

В чем проблема с kNN и MF?

- Ничего не обязаны сделать с выручкой, просто что-то рекомендуют

В чем проблема с kNN и MF?

- Ничего не обязаны сделать с выручкой, просто что-то рекомендуют
- Возможно и дадут прирост денег, но наверняка можно больше (оптимизируя именно деньги)

В чем проблема с kNN и MF?

- Ничего не обязаны сделать с выручкой, просто что-то рекомендуют
- Возможно и дадут прирост денег, но наверняка можно больше (оптимизируя именно деньги)
- Ограниченные возможности включать новые признаки и объединять в себе другие модели

Как строят рекомендательные системы на практике

Что можем делать

- Прогнозировать, какие товары будут куплены
- Максимизировать прибыль

Остается вопрос: какие прогнозы нужны и как их использовать, чтобы денег стало больше?

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Максимизация количества покупок

Товар 1	Товар 2	Товар 3	Товар 4

Вероятность:	p_1	p_2	p_3	p_4
--------------	-------	-------	-------	-------

Максимизация дохода

	Товар 1	Товар 2	Товар 3	Товар 4
Вероятность:	p_1	p_2	p_3	p_4
Цена:	c_1	c_2	c_3	c_4

Максимизация дохода



Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

Вероятность:	0.05	0.02	0.015	0.009
Цена:	3490	1990	1590	1970

Максимизация прибыли



Puma
Ветровка
3 490 руб.

Crocs
Сланцы
1 990 руб.

Tony-p
Слипоны
~~1 999 руб.~~ 1 590 руб.

Champion
Брюки спортивные
~~3 599 руб.~~ 1 970 руб.

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

- 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

- Популярные
- Из тех же категорий
- Часто покупаемые с уже просмотренными/понравившимися
- Из заранее подготовленных списков похожих товаров

Генерация негативных примеров

- 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

Мини-задача

Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

Мини-задача

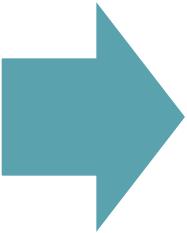
Как изменится построение модели, если нам нужно максимизировать количество просмотренных пользователем товаров?

А если нужно максимизировать количество проданных товаров из категории «аксессуары»?

Дополнительные ограничения

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

User 2 item модели



Item 2 item модели



Вам также могут понравиться



3.5

до -17%



4.5



4.0

Item 2 item по РМI

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

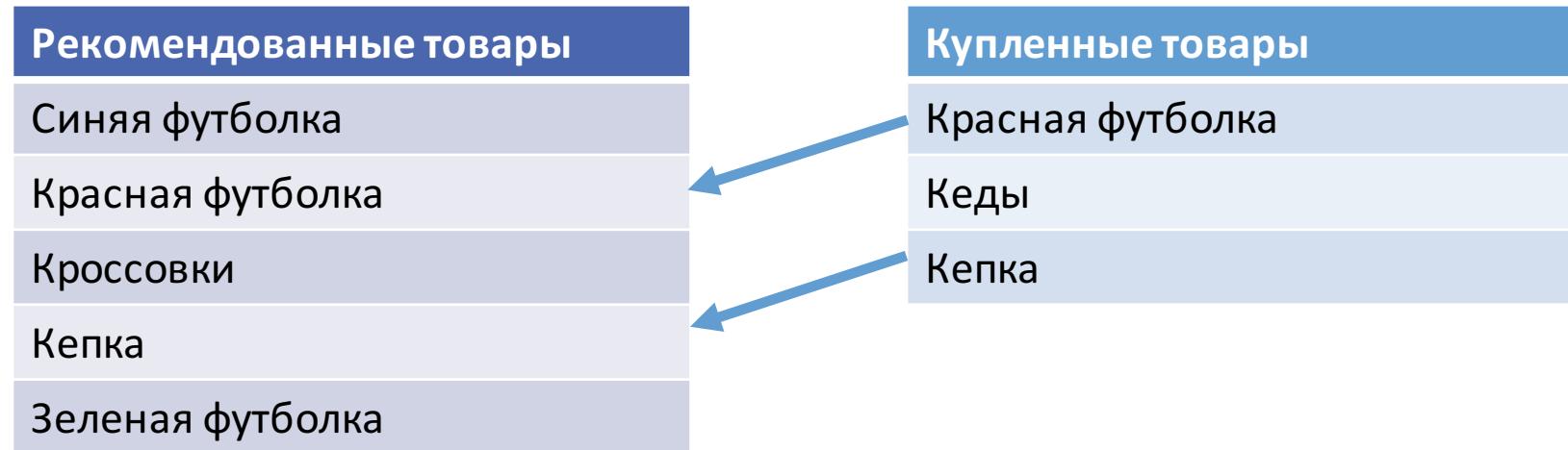
$p(x)$ – вероятность встретить объект x (в пользовательской сессии/среди купленных/среди понравившихся)

$p(y)$ – вероятность встретить объект y

$p(x, y)$ – вероятность встретить оба

Оценка качества и проблемы из реального мира

Точность (Precision@k)



k – количество
рекомендаций

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

AveragePrecision@k - усредненный по сессиям Precision@k

Полнота (Recall@k)

Рекомендованные товары	Купленные товары
Синяя футболка	Красная футболка
Красная футболка	Кеды
Кроссовки	
Кепка	Кепка
Зеленая футболка	

k – количество
рекомендаций

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

AverageRecall@k - усредненный по сессиям Recall@k

Взвешенный ценами recall@k

Рекомендованные товары	Купленные товары
Синяя футболка – 1000р	Красная футболка – 1200р
Красная футболка – 1200р	Кеды – 3000р
Кроссовки – 3500р	Кепка – 900р
Кепка – 900р	
Зеленая футболка – 800р	

Взвешенный ценами Recall@k = $\frac{\text{стоимость купленного из рекомендованного}}{\text{стоимость покупок}}$

AverageRecall@k - усредненный по сессиям Recall@k

Качество классификации против качества рекомендаций

Пример – 2 решения для прогноза купит/не купит товар

	Алгоритм 1	Алгоритм 2
AUC классификатора	0.52	0.85
Recall@5	0.72	0.71

Онлайновая оценка качества

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

Онлайновая оценка качества

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

Идеи:

1. А/В тест
2. Оценка статзначимости результата

A/B тест

1. Случайным образом делим клиентов на равные группы
2. Измеряем целевые метрики (например, доход с каждой группы клиентов) в каждой группе за длительный период времени
3. Получаем какое-то число для каждой группы
4. Что дальше?

Статистическая значимость: пример



Статистическая значимость: пример



Одна кривая отличается от других на 10%
Но разбиение на самом деле – случайное

Истории из практики: разбиение на группы

- Предложено аналитиками:
 - Брать hash от user_id
 - Смотреть на остаток от деления на 2
- Сделано:
 - Брать hash от user_id+user_email
 - Смотреть на остаток от деления на 2

Истории из практики: дизайн

Сопутствующие товары	Похожие товары		
Товар 1	Товар 2	Товар 3	Товар 4

Истории из практики: сравнение методов

- Интегрировали чужое решение, чтобы сравнить качество со своим
- Оценили качество у обоих
- Совпало до тысячных долей
- Не стали использовать чужое решение

Что еще нам может помешать?

1. Проблема холодного старта
2. Плохое логгирование данных
3. Нехватка данных

Про нехватку данных

$$P(\text{покупка} \mid \text{показ}) = P(\text{покупка} \mid \text{клик}) P(\text{клик} \mid \text{показ})$$



мало данных

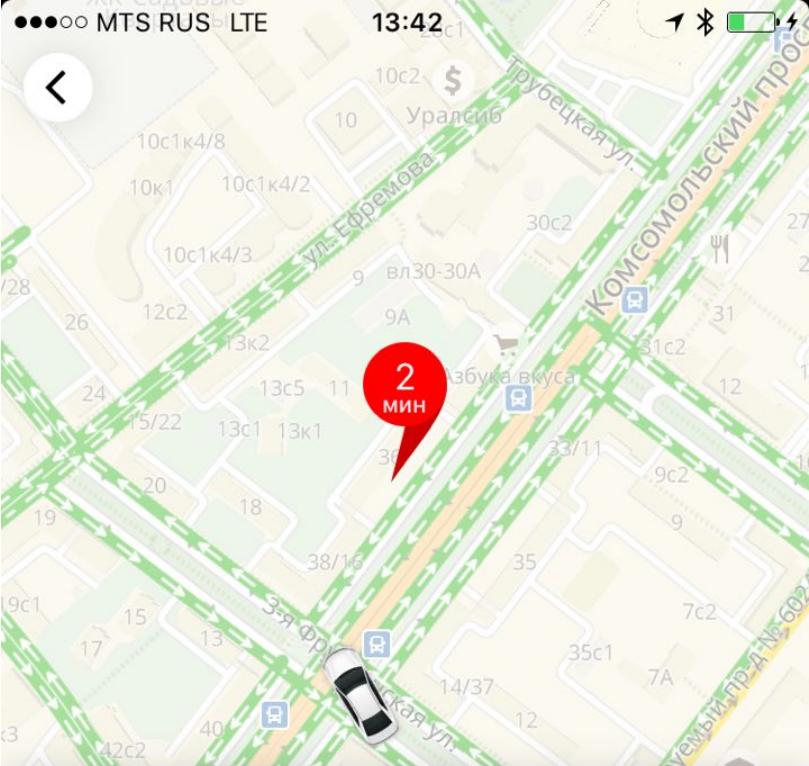
чуть больше данных

На какие метрики смотрят в онлайне

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

«Необычные» рекомендации

Прогнозирование точки Б в Такси



Комсомольский
проспект, 36

Подъезд

куда

Льва Толстого, 14

Бол

Время подачи

Ближайшее

VISA ****

Комментарий, пожелания

Вызвать такси

109 ₽ за первые 4 мин и 2 км

Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В

Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В

Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В
3. Задача классификации: объекты - пары (пользователи, момент времени), классы – адреса мест, куда может поехать

Что делаем

Какая постановка задачи машинного обучения должна быть?

1. Угадать точку В
2. Задача регрессии: объекты – пользователи, целевое значение – координаты точки В
3. Задача классификации: объекты - пары (пользователи, момент времени), классы – адреса мест, куда может поехать
4. Задача классификации: объекты - тройки (пользователь, место, момент времени, классы – 1 и 0 (поехал или нет), 1 при точном попадании в адрес или в окрестность 20 метров

Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)

Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC

Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC
3. Precision@1, Precision@2

Как оцениваем качество

Какая метрика более уместна?

1. Accuracy классификатора (доля угадываний поедет/не поедет)
2. ROC-AUC
3. Precision@1, Precision@2
4. Recall@1, Recall@2

Признаки объекта

- # 0 – общее количество кандидатов (может быть и меньше 20, если поездок мало)
- # 1 - максимальный score (это баллы) кандидатов (доля от суммы score по всем кандидатам)
- # 2 - минимальные score кандидатов (доля от суммы)
- # 3 - отношение максимума к минимуму score кандидатов
- # 4 - количество заказов
- # 5 - время с последнего заказа в секундах
- # 6 - время с первого заказа в секундах
- # 7 - среднее время между заказами
- # 8 - среднее время между заказами
- # 9 - среднее расстояние поездки
- # 10 - количество раз текущего старта
- # 11 - вероятность адреса текущего старта
- # 12 - вероятность старта из 100 метровой окрестности текущей точки старта
- # 13 - вероятность старта из 500 метровой окрестности текущей точки старта
- # 14 - вероятность старта из 100 метровой окрестности текущей точки старта в окрестности 2 часа от текущего времени

Признаки объекта

- # 15 - вероятность старта из 100 метровой окрестности текущей точки старта в окрестности 2 часа от текущего времени при условии старта в окрестности 2 часа от текущего времени
- # 16 - количество заказов в истории в окне 2 часа от текущего времени
- # 17 - вероятность старта из 100 метровой окрестности текущего старта в окрестности 2 часа от текущего времени при условии старта в окрестности 2 часа от текущего времени
- # 18 - количество заказов за последние две недели
- # 19 - количество заказов за последние две недели по отношению ко всем заказам
- # 20 - день недели
- # 21 - расстояние до кандидата
- # 22 - расстояние до кандидата по отношению к среднему расстоянию поездки
- # 23 - ранг (номер после сортировки) кандидата по score
- # 24 - отношение score кандидата к максимальному score среди кандидатов
- # 25 - количество кандидатов в радиусе 500 метров от текущего кандидата

Результаты

- Было до нас: (эвристический алгоритм)

$\text{recall}@1 = 63,7\%$;

$\text{recall}@2 = 78,5\%$;

$\text{recall}@3 = 84,6\%$

- Стало:

$\text{recall}@1 = 72,1\%$;

$\text{recall}@2 = 82,6\%$;

$\text{recall}@3 = 88\%$

Еще примеры

- Рекомендации методов удержания клиента
- Персонализация контента сайта
- Персонализация привлечения клиентов
- ...

Спасибо за внимание



info@applieddatascience.ru



https://t.me/joinchat/B10lThC96v0BQCvs_joNew



https://github.com/vkantor/ml2018jan_feb