Рекомендательная система

Дипломный проект

Ипатов Александр, февраль 2023

Постановка задачи

- Задача рекомендаций товаров пользователям
- Выводить для конкретного пользователя 3 рекомендуемых товара
- Бизнес метрика: повысить продажи (оборот)
- Техническая метрика: Precision@3, MAP@3

Описание данных

На входе имеется 4 csv-файла, из которых собраны 3 датафрейма:

- events.csv
- category_tree.csv
- item_properties_part1.csv, item_properties_part2.csv

```
1 events.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2756101 entries, 0 to 2756100
Data columns (total 5 columns):
 #   Column         Dtype  
---  ------         -----  
 0   timestamp      int64  
 1   visitorid      int64  
 2   event          object 
 3   itemid         int64  
 4   transactionid  float64
dtypes: float64(1), int64(3), object(1)
memory usage: 105.1+ MB
```

```
1 categories.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1669 entries, 0 to 1668
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   categoryid  1669 non-null   int64  
 1   parentid    1644 non-null   float64
dtypes: float64(1), int64(1)
memory usage: 26.2 KB
```

```
1 properties.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 20275902 entries, 0 to 9275902
Data columns (total 4 columns):
 #   Column     Dtype  
---  ------     -----  
 0   timestamp  int64  
 1   itemid     int64  
 2   property   object 
 3   value      object 
dtypes: int64(2), object(2)
memory usage: 773.5+ MB
```

В результате преобразований для работы с коллаборативной фильтрацией подготовлен следующий датафрейм

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1363315 entries, 0 to 1929269
Data columns (total 3 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   visitorid    1363315 non-null  int64
 1   itemid       1363315 non-null  int64
 2   transaction  1363315 non-null  int64
dtypes: int64(3)
memory usage: 41.6 MB
```

	visitorid	itemid	transaction
0	257597	355908	0
1	992329	248676	0
3	483717	253185	0
4	951259	367447	0
5	972639	22556	0

Модель SVD

- Для построения модели используем алгоритм Коллаборативной фильтрации, основная идея которого состоит в том, что похожим пользователям нравятся похожие товары.
- Для расчета потребовалась библиотека **surprise**, выбрана модель **SVD** с гиперпараметрами: n_factors=5, n_epochs=200, biased=True, lr_all=0.002, reg_all=0.05, init_mean=0, init_std_dev=0.01, verbose=False).
- Суть SVD в том, что в разрезе коллаборативной фильтрации данные (таблица) обычно является разреженной (много 0, мало 1), а значит что при произведении матричных операций произведения можно пренебрегать отдельными членами, в связи с чем матричные операции выполняются быстрее при должном доверии к результатам.

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

Предсказание

$$\sum_{r_{ui} \in R_{train}} \left(r_{ui} - \hat{r}_{ui}\right)^2 + \lambda \left(b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2\right)$$

Функция ошибок (которую нужно минимизировать)

$$\begin{aligned} b_u &\leftarrow b_u &+ \gamma(e_{ui} - \lambda b_u) \\ b_i &\leftarrow b_i &+ \gamma(e_{ui} - \lambda b_i) \\ p_u &\leftarrow p_u + \gamma(e_{ui} \cdot q_i - \lambda p_u) \\ q_i &\leftarrow q_i + \gamma(e_{ui} \cdot p_u - \lambda q_i) \end{aligned}$$

Минимизируем стохастическим градиентным спуском (правила)

Значение метрики

В результате подбора гиперпараметров с помощью GridSearchCV с сеткой *{'n_factors': [5, 10, 20], 'n_epochs': [50, 100, 200], 'lr_all': [0.002, 0.005, 0.01], 'reg_all': [0.02, 0.05, 0.1]}* были выбраны оптимальные.

Значение метрики Precision@3 = 0.8694 (train), 0.8447 (test)