Rekrutacja: Intern - Data Science - Zadanie rekrutacyjne

Michał Janik

25 kwietnia 2021

1 Problem

Rekomendacja 10 produktów, którymi może zainteresować się dany użytkownik. Na temat użytkownika, podczas przygotowywania listy produktów, wiemy tylko poniższe rzeczy: 1. trzy ulubione kategorie ($product\_category\_name$) 2. jego geolokacja ($customer\_city, customer\_state$). Gdy użytkownik nie jest zalogowany, nie mamy żadnej informacji na jego temat, należy wymyślić sposób na rekomendację.

2 Dataset

Zbiór danych Brazilian E-Commerce zawiera informacje o ponad 100 tysiącach zamówień z lat 2016 - 2018 dokonanych w Brazylii. Z każdym zamówieniem powiązane są informacje: status i czas zamówienia, cena, rodzaj płatności, lokalizacja klienta, atrybuty produktu i wreszcie recenzje napisane przez klientów.
W tej pracy skorzystamy z poniższych tabel:

- z tabel $olist\_orders\_dataset$ i $olist\_order\_items\_dataset$ aby otrzymać listę zamówionych produktów przez każdego użytkownika
- z tabeli $olist\_order\_reviews\_dataset$ aby otrzymać powiązane oceny
- z tabeli $olist\_products\_dataset$ aby otrzymać kategorie produktów
- z tabeli $olist\_customers\_dataset$ aby otrzymać informacje o geolokacji użytkowników

Dataset dzielimy na train/validation sety, odpowiednio w proporcjach 8:2.

3 Rozwiązanie

3.1 System rekomendacji

Postawiono przed nami problem rekomendacji, mając informacje o preferencjach użytkownika naszym zadaniem jest polecenie mu najbardziej trafnych przedmiotów. Najczęściej stosowanymi metodami są:

- Collaborative filtering - buduje model na podstawie przeszłych zachowań użytkownika (poprzednio zakupionych lub wybranych przedmiotów i/lub ocen wystawionych tym przedmiotom), a także podobnych decyzji podjętych przez innych użytkowników
- Content-based filtering - wykorzystuje szereg wstępnie oznaczonych cech przedmiotu w celu rekomendowania dodatkowych przedmiotów o podobnych cechach

W naszym przypadku brakuje nam cech produktów aby zastosować content-based filtering, mamy jednak dane dotyczące zakupów powiązane z ocenami pozwalające na zastosowanie collaborative filtering. Nasz problem różni się jednak od większości przykładów z literatury przez to, że przy rekomendacji nie posiadamy historycznych transakcji użytkownika, a tylko jego ulubione kategorie.
Ważne punkty, na których będzie się opierać wybór modelu rekomendacyjnego:

- Należy maksymalnie wykorzystać informacje, które mamy dostępne przy rekomendacji, tj:
  - lista 3 ulubionych kategorii produktów użytkownika
  - geolokacja użytkownika - możemy zaobserwować preferencje do kupowania z konkretnych miast
  - aktualny czas - popularność produktów i gust użytkownika mogą zmieniać się w czasie
- Metoda ciesząca się największym sukcesem, Matrix factorization(polegająca na rozłożeniu macierzy interakcji użytkownik-przedmiot na iloczyn dwóch macierzy prostokątnych o mniejszych wymiarach) bezpośrednio nie może zostać użyta w naszym problemie, ponieważ przy rekomendacji nie będziemy w stanie otrzymać wektora użytkownika nie posiadając jego historycznych transakcji.

3.2 Proponowane podejście

Prezentujemy autorskie podejście, który bierze pod uwagę powyższe problemy. Listę będziemy generować w 2 etapach:

1. Model bazujący na faktoryzacji macierzy, który dokonuje predykcji ocen użytkownika u dla każdego przedmiotu $i(\hat{r}_{ui})$. Listę N najwyżej ocenionych przedmiotów przekazujemy do następnego etapu.
2. W tym etapie dokonujemy klasteryzacji przedmiotów (metryka euklidesowa, wektory z poprzedniego modelu, KMeans) na 10 klastrów. W finalnej rekomendacji z każdego klastra zwracamy jeden przedmiot o najwyższej predycji $\hat{r}_{ui}$. W ten sposób otrzymujemy listę 10 przedmiotów, które jednocześnie powinny spodobać się użytkownikowi i nie być zróżnicowane.

W przypadku niezalogowanego użytkownika w 1 etapie wybierzemy N przedmiotów najbardziej popularnych w danym przedziale czasowym, które w 2 etapie zostaną zdywersyfikowane.

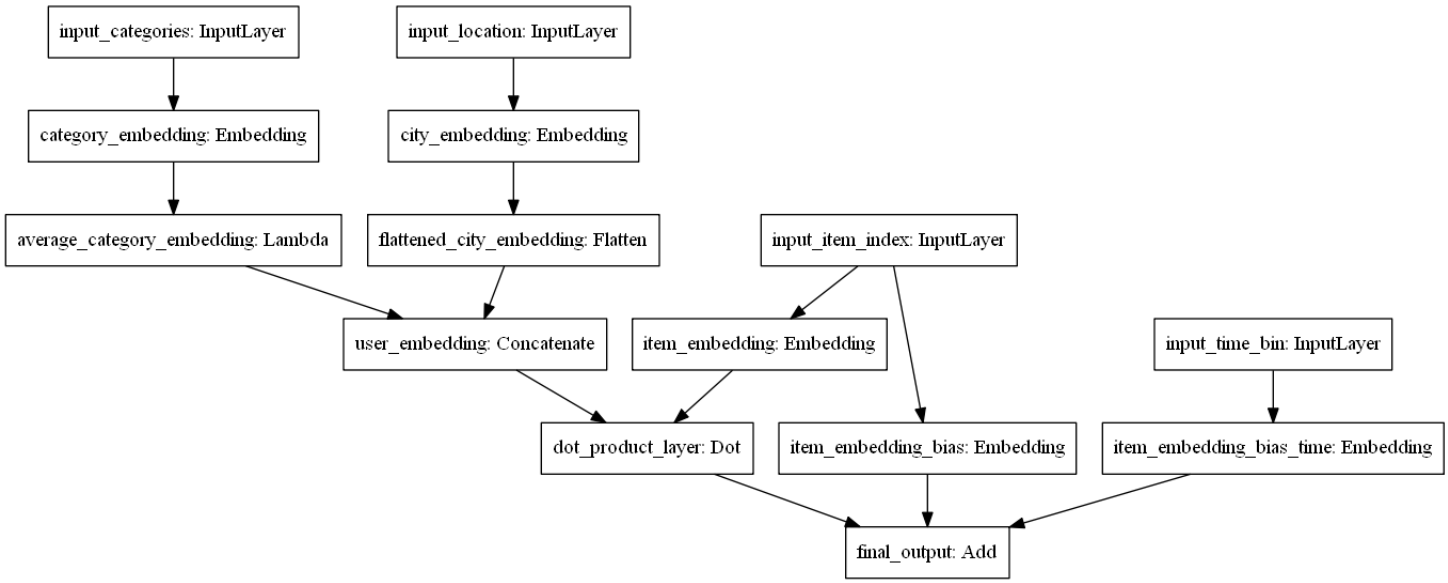
Opis modelu 1: Za pomocą gradient descent minimalizujemy koszt: $||R - \hat{R}||$

r_{ui} - rzeczywista ocena użytkownika u dotycząca przedmiotu i
$\hat{r}_{ui} = q_i^T p_u + b_{ui}(t)$ - przewidywana ocena użytkownika u dotycząca przedmiotu i

q_i - wektor(embedding) przedmiotu i
$p_u = pref_u \frown loc_u$ - wektor użytkownika
$pref_u = \sum_{k \in \{1,2,3\}} category[favourites_u[k]]/3$ - wektor będący reprezentacją upodobań użytkownika u, średnia wektorów jego ulubionych kategorii
$category[k]$ - wektor(embedding) k-tej kategorii
$favourites_u$ - wektor 3-elementowy zawierający ulubione kategorie użytkownika u
loc_u - wektor(embedding) miasta z którego pochodzi użytkownik

$b_{ui}(t) = b_i + b_{i,Bin(t)}$ - Bias przedmiotu podzielony na część stacjonarną i część zmieniającą się w czasie. Dzień t jest związany z liczbą całkowitą Bin(t) (w naszych danych jest to liczba z przedziału od 0 do 25 oznaczająca miesiąc).
b_i - stancjonarny bias przedmiotu
$b_{i,Bin(t)}$ - bias przedmiotu zależny od czasu

Bazując na pracy (1) wykorzystaliśmy modelowanie czasowych dynamik. Wektor 3 ulubionych kategorii użytkownika u ($favourites_u$) składa się z kategorii, z których użytkownik dokonał najliczniejszych zakupów. W przypadku, gdy użytkownik nie kupił produktów z 3 różnych kategorii, duplikujemy najliczniejszą kategorię.
Nasz model jest hybrydą między content-based(dane o preferencjach użytkownika i jego geolokalizacji) oraz collaborative filtering(model uczymy na podstawie preferencji wszystkich użytkowników). Model pierwszego etapu mając tylko te dane o użytkowniku, które są dostępne dla nas przy rekomendacji musi przewidywać oceny przedmiotów. Dodatkowo, wektory przedmiotów i użytkowników nie mają wiele wymiarów, konieczne jest więc modelowanie preferencji. Z tego powodu twierdzimy, że w przypadku większej ilości danych model byłby w stanie rekomendować produkty, które podobałyby się użytkownikom.



4 Trenowanie

- Optimizer: Adam($beta_1 = 0.9$, $beta_2 = 0.999$,),
- Learning rate: 10^{-3},
- Loss: Mean Squared Error

5 Predykcje

Prediction for user ($'cama\_mesa\_banho'$, $'papelaria'$, $'fashion\_calcados'$)(sao paulo):
'629e019a6f298a83aeecc7877964f935', '719d571299707561c34ba04ab867b32a', '2b4609f8948be18874494203496bc318',
'3e4176d545618ed02f382a3057de32b4', '6a8631b72a2f8729b91514db87e771c0', '73326828aa5efe1ba096223de496f596',
'5dee2c14e1989141e15d341d4c62d72a', 'e0cf79767c5b016251fe139915c59a26', 'aa280035c50ba62c746480a59045eec4',
'437c05a395e9e47f9762e677a7068ce7'
Prediction for user ($'esporte\_lazer'$, $'moveis\_decoracao'$, $'telefonia'$)(rio de janeiro):
'6a8631b72a2f8729b91514db87e771c0', 'e0cf79767c5b016251fe139915c59a26', '3e4176d545618ed02f382a3057de32b4',
'629e019a6f298a83aeecc7877964f935', '24c66f106f642621e524291a895c9032', '473795a355d29305c3ea6b156833adf5',
'54d9ac713e253fa1fae9c8003b011c2a', 'f9259c9e7c0f12c70f7a81409680a5ff', 'aadff88486740e0b0ebe2be6c09476ae', '921d31a1daa51

Prediction for user ()():
'6a8631b72a2f8729b91514db87e771c0', 'e0cf79767c5b016251fe139915c59a26', '3e4176d545618ed02f382a3057de32b4', '629e019a6f298a83aeecc7877964f935', '921d31a1daa51460b7a95ea5f3ab64d5', '473795a355d29305c3ea6b156833adf5', '73326828aa5efe1ba096223de496f596', 'aadff88486740e0b0ebe2be6c09476ae', '5dee2c14e1989141e15d341d4c62d72a', '54d9ac713e253fa1fae9c8003b011c2a'

6 Przyszły kierunek prac

Wykorzystaliśmy bardzo prosty model, głębsze modele są obiecujące i wymagają eksploracji. Kolejny kierunek prac to spojrzenie na nasze dane jako 'implicit data' - zignorowanie ocen produktów i skupienie się na interakcji z produktem/jej braku. Dobrym startem jest praca (3)

Literatura

[1] Koren, Yehuda. (2009). Collaborative filtering with temporal dynamics. Proceedings of the 15th. 53. 447-456. 10.1145/1557019.1557072.

[2] Y. Koren, R. Bell and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems,"in Computer, vol. 42, no. 8, pp. 30-37, Aug. 2009, doi: 10.1109/MC.2009.263.

[3] Hu, Yifan Koren, Yehuda Volinsky, Chris. (2008). Collaborative Filtering for Implicit Feedback Datasets. Proceedings - IEEE International Conference on Data Mining, ICDM. 263-272. 10.1109/ICDM.2008.22.