Algorytmy Numeryczne – Zadanie 3

12 listopada 2019

Metoda ALS w systemach rekomendacji

Systemy rekomendacji są stosowane między innymi w sklepach internetowych. Na podstawie profilu użytkownika, dotychczasowej historii zakupów, zadanych zapytań w wyszukiwarce, wystawionych ocen itp. można próbować przewidzieć jakie inne produkty mogłyby być przez niego dobrze ocenione. Rekomendować możemy na przykład filmy, książki, utwory muzyczne, artykuły do przeczytania, inne produkty do kupienia.

Metoda faktoryzacji macierzy

W metodzie faktoryzacji macierzy zakładamy, że ocena produktu może być wyliczona na podstawie współczynników dwóch macierzy: U – opisującej współczynniki użytkowników i P – opisującej współczynniki produktów. Przybliżenie kompletnej macierzy rekomendacji R otrzymujemy jako iloczyn macierzowy: $R \approx X^T Y$.

Jeśli użytkownik u ocenił produkt p, to ocenę tę oznaczamy jako $\overline{r_{up}}$ i przyjmujemy za docelową wartość r_{up} – odpowiedniego elementu macierzy R. Z drugiej strony $\overline{r_{up}} \approx U_u^T P_p$, gdzie U_u i P_p są wektorami kolumnowymi w macierzach U i P odpowiednio.

Wszystkie znane wartości $\overline{r_{up}}$ oznaczmy jako $\overline{R}$. Na podstawie tych znanych wartości staramy się wyliczyć możliwie najlepsze macierze U i P, których iloczyn da macierz R – przybliżenie dla wszystkich ocen których nie znamy.

Jako kryterium optymalizacji przyjmujemy następującą funkcję celu:

$$f(U,P) = \sum_{\overline{r_{up}} \in \overline{R}} (\overline{r_{up}} - U_u^T P_p)^2 + \lambda \left(\sum_u \|U_u\|^2 + \sum_p \|P_p\|^2 \right), \tag{1}$$

którą staramy się zminimalizować. Pierwszy składnik powyższej sumy pokazuje jak rozwiązanie różni się dla znanych wartości rekomendacji, natomiast drugi składnik to tak zwana regularyzacja, która ma przeciwdziałać nadmiernemu dopasowaniu.

Jeśli liczba wszystkich użytkowników wynosi n a liczba wszystkich produktów wynosi m, to macierz R jest typu $n \times m$. Natomiast typy macierzy U i P to odpowiednio $d \times n$ i $d \times m$, gdzie d jest parametrem rozwiązania.

Metoda ALS (ang. Alternating Least Square)

Ponieważ trudno jest znaleźć minimum funkcji 1 bezpośrednio, to stosujemy metodę najmniejszych kwadratów na przemian raz ustalając współczynniki macierzy P raz macierzy U. W kolejnych krokach iteracyjnego algorytmu ALS powinniśmy otrzymać coraz "lepsze" macierze U i P.

Algorytm ALS[2]

1. Nadaj wartości początkowe macierzom P i U.
2. Powtarzaj do osiągnięcia zadowalającego rezultatu:
3. **for** $u = 1, 2, \ldots n$
4. **Pod** U_u **podstaw rozwiązanie układu równań:**

$$A_u X = V_u$$

5. **for** $p = 1, 2, \ldots m$
6. **Pod** P_p **podstaw rozwiązanie układu równań:**

$$B_p X = W_p$$

Oznaczamy: $A_u = P_{I_u} P_{I_u}^T + \lambda E$ jest macierzą typu $d \times d$, zbiór I_u zawiera te indeksy p dla których użytkownik u ocenił produkt p ($I_u = \{p : \overline{r_{up}} \in \overline{R}\}$), P_{I_u} to podmacierz macierzy P zawierająca kolumny o numerach w I_u, a $V_u = \sum_{i \in I_u} r_{ui} P_i$.

I analogicznie: $B_u = U_{I_p} U_{I_p}^T + \lambda E$ jest macierzą typu $d \times d$, zbiór I_p zawiera te indeksy u dla których użytkownik u ocenił produkt p ($I_p = \{u : \overline{r_{up}} \in \overline{R}\}$), a $W_p = \sum_{i \in I_p} r_{ip} U_i$.

E jest macierzą jednostkową (na przekątnej same jedynki poza przekątną zera).

Zadanie

W wybranym przez siebie języku programowania zaimplementuj metodę ALS korzystając z gotowej metody Gaussa z drugiego zadania.

Dane testowe

Do testów proszę wykorzystać prawdziwe, archiwalne, zanonimizowane dane o produktach i rekomendacjach ze sklepu Amazon udostępnione w ramach projektu SNAP[1]. Plik zawiera rekordy dotyczące ponad 500 000 produktów i łącznie ponad 7 mln ocen.

Testy proszę przeprowadzić na 3 wybranych samodzielnie podzbiorach danych:

S: małym, zawierającym od 10 do 100 produktów,

M: średnim, około 10 razy większym niż mały, zawierającym od 100 do 1000 produktów,

*B: dużym, około 10 razy większym niż średni zawierającym więcej niż 1000 produktów (im więcej tym lepiej).

Dla każdego zbioru produktów należy dobrać użytkowników (być może nie wszystkich), którzy oceniali produkty z danej grupy.

Dobór parametrów

Parametrami metody ALS są:

λ – współczynnik λ.

d – liczba wierszy w macierzach U i P.

Optymalny dobór parametrów zależy od konkretnego zagadnienia. Pierwszym podejściem dla małego zbioru danych może być $\lambda = 0.1$, $d = 3$. Dalej proszę spróbować dobrać optymalne parametry dla wybranych zbiorów danych.

Sprawozdanie

Proszę przeprowadzić testy używając typu podwójnej precyzji: `double` (lub odpowiednika w wybranym języku programowania). Korzystając z metody Gaussa zaimplementowanej w pierwszym zadaniu w wariancie z częściowym wyborem elementu podstawowego.

W pierwszej części sprawozdania proszę przedstawić sposób doboru danych oraz argumenty za prawidłową implementacją.

W drugiej części sprawozdania proszę przedyskutować uzyskane wyniki, w tym:

- przydatność implementowanej metody dla testowanych danych i obranych parametrów,
- tempo zbieżności metody ALS w zależności od obranych parametrów,
- wpływ parametru d na jakość stworzonych rekomendacji i czas obliczeń.

Uwagi i wskazówki

S1: Proszę zwrócić uwagę, że w stosunku do opisu z pracy [2] w funkcji celu (1) nie stosujemy wag.

S2: Spodziewamy się, że każdy prawidłowo zbudowany układ równań w tym zdaniu ma jedno rozwiązanie.

S3: Spodziewamy się, że każda kolejna iteracja prawidłowo zaimplementowanej metody ALS zmniejsza funkcję celu.

S4: Aby sprawdzić czy wyliczone przez nas macierze U i P dobrze przybliżają oceny użytkowników postępujemy w następujący sposób: dzielimy interesujące nas oceny na dwa podzbiory: uczący i testowy. Na podstawie podzbioru uczącego obliczamy macierze U i P. Na podstawie podzbioru testowego sprawdzamy uzyskane rozwiązanie.

S5: Jeśli znane rekomendacje są wartościami liczbowymi z pewnego zakresu (np. 1–10) to pożądanym jest by wyliczone przez nas nieznane rekomendacje były również liczbami z tego zakresu.

S6: Docelowo cała macierz R zawiera tyle elementów, że ich jednoczesne przechowywanie w pamięci nie jest możliwe.

S7: W charakterze wartości początkowych dla U i P można użyć niewielkich liczb pseudolosowych.

S8: Wyliczone rekomendacje powinny być zgodne z intuicją, np: spodziewamy się, że produkt oceniany źle lub bardzo źle we wszystkich ocenach nie będzie rekomendowany innym użytkownikom; pierwszym zarekomendowanym produktem z kategorii X użytkownikowi u który do tej pory oceniał produkty z tej kategorii wyłącznie negatywnie nie będzie inny produkt z kategorii X; jeśli użytkownik dobrze oceniał produkty z kategorii X to z wysokim prawdopodobieństwem zostanie mu zarekomendowany inny produkt z tej kategorii; itp.

Ocena i elementy nieobowiązkowe

- Pominięcie prób doboru optymalnego współczynnika λ : -10 pkt. (w tym wariancie proszę przyjąć ustaloną wartość np. $\lambda = 0.1$).
- Pominięcie testów na dużym zbiorze danych (B): -20 pkt.

Praca zespołowa

Zadanie można wykonać w zespole nie więcej niż trzyosobowym. W takim przypadku proszę dokładnie oznaczyć jaki był zakres pracy członków zespołu. W oddaniu projektu musi uczestniczyć cały zespół.

Literatura

[1] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[2] Yunhong Zhou, Dennis Wilkinson, Robert Schreiber, and Rong Pan. Large-scale parallel collaborative filtering for the netflix prize. In *Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management*, AAIM '08, pages 337–348, Berlin, Heidelberg, 2008. Springer-Verlag.