Podstawowe zagadnienia systemów rekomendacyjnych

Michał Maternik, Kacper Rączy, Maksym Telechuk

Dane

MovieLens 100K - zbiór danych z ocenami filmów (od 1 do 5)

- Liczba użytkowników: 943
- Liczba filmów: 1682
- Liczba opinii: 100000

Collaborative filtering

- Korzysta z danych historycznych preferencji dla zbioru itemów
- Zakłada, że osoby mające podobne preferencje będą mieli takie preferencje w przyszłości
- 2 rodzaje rankingu zwykle są uwzględniane: jawny (explicit) in niejawny (implicit).

Zalety:

- nie wymaga szczegółowych danych o produktach
- przypisywanie wag do użytkowników które są bardziej podobne

Wady:

- wymaga danych o feedback'u użytkownika
- wysoki narzut obliczeniowy, trzeba wszystko obliczyć na nowo po dodaniu nowego użytkownika/produktu
- nowi użytkownicy/produkty nie są brani pod uwage, dopóki nie wygenerują ruchu

Wykorzystanie KNN classification

Algorytmy oparte o algorytm najbliższych sąsiadów są podstawowymi metodami podejścia collaborative Filtering. Rozróżniamy metody oparte o podobieństwo użytkowników lub obiektów. Jako rekomendacje zwracana jest zadana ilość najbardziej podobnych użytkowników lub obiektów, a rating jest średnią ważoną ratingów składowych, z wagami pochodzącymi z zastosowanej miary podobieństwa.

Metoda umożliwia stosowanie różnych miar podobieństwa:

- Miara cosinusowa
- Korelacja Pearsona

Uzyskiwane wyniki są lepsze (wg metryki MAE) przy zastosowaniu korelacji Pearsona.

Wady i zalety KNN classification

Główną zaletą podejścia jest stabilność, rozumiana w ten sposób że gdy metoda oparta jest podobieństwo obiektów, to ratingi nie zmieniają się gwałtownie wraz ze zmianą preferencji użytkowników.

Z ograniczeń metody można wskazać słabe radzenie sobie z danymi rzadkimi, a także słabe skalowanie przy wzroście liczby użytkowników i produktów.

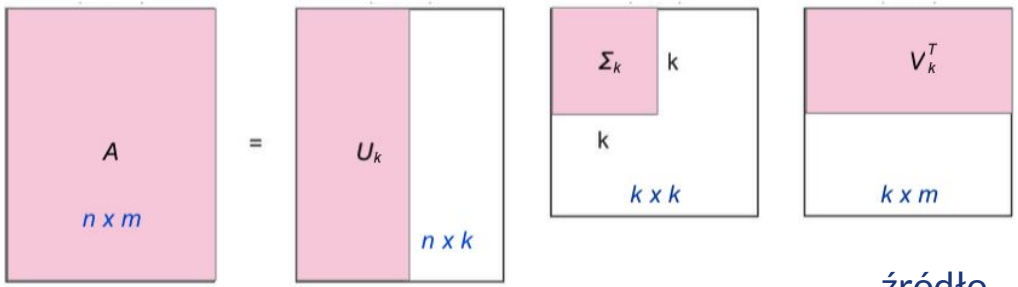
Wykorzystanie faktoryzacji macierzy

Macierz rankingu jest to macierz NxM, gdzie N - liczba użytkowników, M - liczba itemów.

Macierze rankingu w rzeczywistości są rzadkie, co prowadzi czasami do konieczności przeprowadzenia faktoryzacji na tej macierzy, czyli rozłożenia oryginalną macierz na macierze niskowymiarowe z ukrytymi cechami i mniejszą rzadkością.

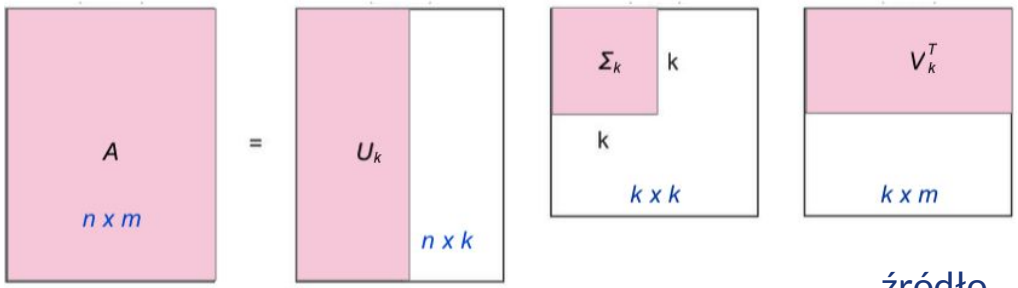
Metody faktoryzacji wprowadzają cechy ukryte, co pozwala porównywać dwóch użytkowników, nawet jeśli oni nie oceniali tych samych itemów.

Singular Value Decomposition(SVD)



źródło

Singular Value Decomposition(SVD)



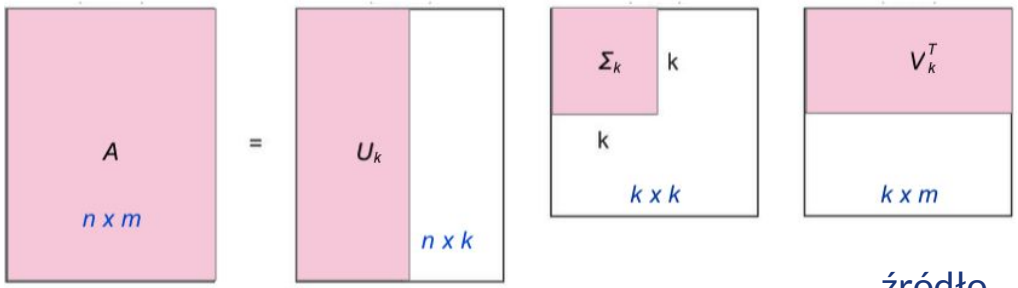
źródło

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

r_ui - znany ranking itemu i od użytkownika u
r^_ui - przewidwalny ranking itemu i od użytkownika u liczony
na podstawie macierzy U i V

<- koszt, minimalizowany metodą SGD

Singular Value Decomposition(SVD)



źródło

r_ui - znany ranking itemu i od użytkownika u
r^_ui - przewidwalny ranking itemu i od użytkownika u liczony
na podstawie macierzy U i V

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

<- koszt, minimalizowany metodą SGD

SVD++ - rozszerzenie o niejawny ranking (w tym przypadku jest to fakt, że użytkownik ocenił item, niezależnie od wartości oceny)

Wyniki

	MAE Train-Test (30%)	MAE CV 5	Fit time	Test time
KNN (cosine)	0.7382	0.7413∓0.005	1.64∓0.03	4.45∓0.06
KNN (Pearson)	0.7338	0.7297∓0.004	1.49∓0.03	4.35∓0.12
SVD	0.7420	0.7392∓0.002	4.54∓0.02	0.22∓0.08
SVD++	0.7239	0.7216∓0.003	164.94∓1.11	4.23∓0.13
Random	1.2116	1.2202∓0.004	0.14∓0.01	0.21∓0.09

Content-based filtering

W przeciwieństwie do CF nie wykorzystuje danych historycznych, zamiast tego skupia się na podobieństwu cech rekomendowanych obiektów.

Zalety:

- jako że nie korzysta z feedback'u innych użytkowników, może być zintegrowany i działać od razu dla nowych obiektów (brak problemu cold start)
- można rekomendować nowe lub mniej popularne obiekty

Wady:

- wymaga szczegółowych informacji na temat obiektów, a nie wszystkie cechy się do czegoś nadają

Rekomendacja filmów na podstawie gatunków

Miara porównawcza: TF-IDF na gatunkach filmów z profilu użytkownika i porównanie z nowymi wykorzystując odległość consinusową.

1. TF-IDF dla każdego filmu -> macierz [N, 19] (19 gatunków)
2. Podobieństwo cosinusowe -> macierz C o wymiarach [N, N]
3. Dla filmu m weź wiersz C[m], następnie posortuj -> lista rekomendacji posortowana według trafności

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

$\mathrm{tf}_{i,j}$ = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

Lion King, The (1994)

98	Snow White and the Seven Dwarfs (1937)
102	All Dogs Go to Heaven 2 (1996)
417	Cinderella (1950)
419	Alice in Wonderland (1951)
431	Fantasia (1940)
472	James and the Giant Peach (1996)
500	Dumbo (1941)
537	Anastasia (1997)
587	Beauty and the Beast (1991)
595	Hunchback of Notre Dame, The (1996)
623	Three Caballeros, The (1945)
988	Cats Don't Dance (1997)
94	Aladdin (1992)
541	Pocahontas (1995)
1090	Pete's Dragon (1977)
992	Hercules (1997)
101	Aristocats, The (1970)
403	Pinocchio (1940)
624	Sword in the Stone, The (1963)
945	Fox and the Hound, The (1981)

Pulp Fiction (1994)

75	Carlito's Way (1993)
181	GoodFellas (1990)
292	Donnie Brasco (1997)
345	Jackie Brown (1997)
503	Bonnie and Clyde (1967)
627	Sleepers (1996)
910	Twilight (1998)
1105	Newton Boys, The (1998)
1121	They Made Me a Criminal (1939)
1155	Cyclo (1995)
1190	Letter From Death Row, A (1998)
1193	Once Were Warriors (1994)
1225	Night Falls on Manhattan (1997)
1438	Jason's Lyric (1994)
1452	Angel on My Shoulder (1946)
1504	Killer: A Journal of Murder (1995)
1518	New Jersey Drive (1995)
1637	Normal Life (1996)
129	Kansas City (1996)
308	Deceiver (1997)