Zaawansowane zagadnienia systemów rekomendacyjnych

Michał Maternik, Kacper Rączy, Maksym Telepchuk

Dane

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

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

Mining Frequent Itemsets

1. Podejście z dziedziny Collaborative Filtering (CF)
2. Model-based
3. Działa na zbinaryzowanych macierzach rankingu
4. Idea: Im większe jest wsparcie elementu, tym większe jest prawdopodobieństwo, że ten element pojawi się w jakimś zestawie przedmiotów

Do P., Nguyen V.T., Dung T.N. (2013) A New Approach for Collaborative Filtering Based on Mining Frequent Itemsets.

Binary matrix

Table 1. Rating matrix table

	Item 1	Item 2	Item 3	Item 4
User 1	*3*	*5*	*2*	*1*
User 2	*3*	*5*	*2*	*1*
User 3	*1*	*5*	*4*	

Table 2. Bit rating matrix table

	User 1	User 2	User 3
Item_1_1	*0*	*0*	*1*
Item_1_3	*1*	*1*	*0*
Item_2_5	*1*	*1*	*1*
Item_3_2	*1*	*1*	*0*
Item_3_4	*0*	*0*	*1*
Item_4_1	*1*	*1*	*0*

Łatwe porównanie wektorów

Binary matrix

Table 1. Rating matrix table

	Item 1	Item 2	Item 3	Item 4
User 1	*3*	*5*	*2*	*1*
User 2	*3*	*5*	*2*	*1*
User 3	*1*	*5*	*4*	

Table 2. Bit rating matrix table

	User 1	User 2	User 3
Item_1_1	*0*	*0*	*1*
Item_1_3	*1*	*1*	*0*
Item_2_5	*1*	*1*	*1*
Item_3_2	*1*	*1*	*0*
Item_3_4	*0*	*0*	*1*
Item_4_1	*1*	*1*	*0*

Łatwe porównanie wektorów
Problem: przy dużej ilości wektorów algorytm jest wolny

Reguły asocjacyjne

- Zbinaryzowaną macież nazywamy ścianą.
- Na ścianie wszystkie elementy są pokazane w kolejności malejącej według wartości wsparcia (support)
- Posiadamy wałek, którym jedziemy po ścianie, od pozycji do pozycji, w kolejności malejącej.
- Jeśli zostanie znaleziony przedmiot spełniający minimalne wsparcie (support), jest dodawany do częstego zestawu przedmiotów (frequent itemset), a zadanie miningu jest kontynuowane, aż nie ma elementu, który spełnia minimalne wsparcie.
- Wszystkie przedmioty z tego częstego zestawu przedmiotów zostaną usunięte ze ściany, oraz mining zaczyna się od nowa aż na ścianie nie zostanie się żadnego przedmiotu.

Przykład

	U1	U2	U3
I_{11}	0	0	1
I_{13}	1	1	0
I_{25}	1	1	1
I_{32}	1	1	0
I_{34}	0	0	1
I_{41}	1	1	0

⇨

	U1	U2	U3
I_{25}	1	1	1
I_{13}	1	1	0
I_{41}	1	1	0
I_{32}	1	1	0
I_{34}	0	0	1
I_{11}	0	0	1

$p(s_i)$ = null
$s_i = \{\}$
$S = \{\}$

⇨

	U1	U2	U3
I_{13}	1	1	0
I_{41}	1	1	0
I_{32}	1	1	0

$p(s_i) = 111$
$s_i = \{I_{25}\}$
$S = \{\}$

Przykład

	U1	U2	U3
I_{13}	1	1	0
I_{41}	1	1	0
I_{32}	1	1	0

$p(s_i) = 111$
$s_i = \{I_{25}\}$
$S = \{\}$

	U1	U2	U3
I_{41}	1	1	0
I_{32}	1	1	0

$p(s_i) = 110$
$s_i = \{I_{25}, I_{13}\}$
$S = \{\}$

	U1	U2	U3
I_{32}	1	1	0

$p(s_i) = 110$
$s_i = \{I_{25}, I_{13}, I_{41}\}$
$S = \{\}$

$p(s_i) = null$
$s_i = \{\}$
$S = \{\{I_{25}, I_{13}, I_{41}, I_{32}\}\}$

Wyniki - Frequent Itemset Mining

	MAE CV 5	Fit time	Test time
SVD	0.7392∓0.002	4.54∓0.02	0.22∓0.08
Random	1.2202∓0.004	0.14∓0.01	0.21∓0.09
Frequent Itemset Mining	1.6007∓0.04	27.54∓1.56	0.008∓0.0002
Improved FIM	1.4074	1096	0.01

Pseudokod

```
B = bit_transform(D)
S = mining_frequent_itemset(B)
matched_itemset = null
max_count = -1
For each s ∈ S
   bs = bitset(u) AND bitset(s)
   If bs = bitset(u) && count(bs) > max_count then
      matched_itemset = s
      max_count = count(bs)
   End If
End For
r_item = bitset(matched_itemset) AND (NOT bitset(u))
```

```
O = sort(I)
i = 1
While (true)
   c = first(O)
   s_i = s_i U {c}
   O = O / {c}
   If O = Ø then return S
   While (true)
      If c = last(O) then
         S = S U s_i
         O = O / S
         i = i +1
         break
      Else
         c = next(O, c)
         If support(c) < min_sup continue
         b = bitset(S) AND bitset(c)
         If count(b) ≥ min_sup then
            s_i = s_i U {c}
         End If
      End If
   End While
End While
```

Partitioning clustering

- Podejście typu collaborative filtering
- Użytkownicy i obiekty (filmy) przyporządkowywane do klastrów
- Swoboda w doborze metody klastrowania:
    - **K-means**
    - K-way
    - Bisecting K-means

S. Merugu and T. George, "A Scalable Collaborative Filtering Framework Based on Co-Clustering," in Proceedings. Fifth IEEE International Conference on Data Mining, Houston, TX, 2005 pp. 625-628.

doi: 10.1109/ICDM.2005.14

Co-clustering

- Użytkownicy i obiekty (filmy) przyporządkowywane do klastrów C_u, C_i oraz C_ui
- Predykcja uzyskiwana jako:

$$\hat{r}_{ui} = \overline{C_{ui}} + (\mu_u - \overline{C_u}) + (\mu_i - \overline{C_i}),$$

S. Merugu and T. George, "A Scalable Collaborative Filtering Framework Based on Co-Clustering," in Proceedings. Fifth IEEE International Conference on Data Mining, Houston, TX, 2005 pp. 625-628.

doi: 10.1109/ICDM.2005.14

Co-Clustering - Wyniki

	MAE CV 5	Fit time	Test time
Co-Clustering	**0.7590**∓0.005	**2.26**∓0.06	**0.16**∓0.05
KNN (cosine)	0.7413∓0.005	1.64∓0.03	4.45∓0.06
SVD	0.7392∓0.002	4.54∓0.02	0.22∓0.08
Random	1.2202∓0.004	0.14∓0.01	0.21∓0.09

Co-Clustering - Wpływ zmian hiperparametrów

	Liczba klastrów	MAE CV 5	Fit time	Test time
Co-Clustering	**3**	0.7561∓0.005	**2.26**∓0.06	0.21∓0.07
Co-Clustering	**5**	0.7585∓0.005	**2.85**∓0.06	0.23∓0.06
Co-Clustering	**7**	0.7624∓0.006	**3.18**∓0.04	0.25∓0.08
Co-Clustering	**11**	0.7658∓0.004	**3.85**∓0.06	0.22∓0.09

Co-clustering - Podsumowanie

Zalety podejścia:

- Skuteczność porównywalna z metodami opartymi na faktoryzacji macierzy
- Znacznie niższy koszt obliczeniowy
- Możliwość przeprowadzania przyrostowych aktualizacji modelu

Wady:

- Porównywalna, ale jednak gorsza od SVD skuteczność

S. Merugu and T. George, "A Scalable Collaborative Filtering Framework Based on Co-Clustering," in Proceedings. Fifth IEEE International Conference on Data Mining, Houston, TX, 2005 pp. 625-628.

doi: 10.1109/ICDM.2005.14

SlopeOne

- Predykcja uzyskiwana jako:

$$\hat{r}_{ui} = \mu_u + \frac{1}{|R_i(u)|} \sum_{j \in R_i(u)} \mathrm{dev}(i,j),$$

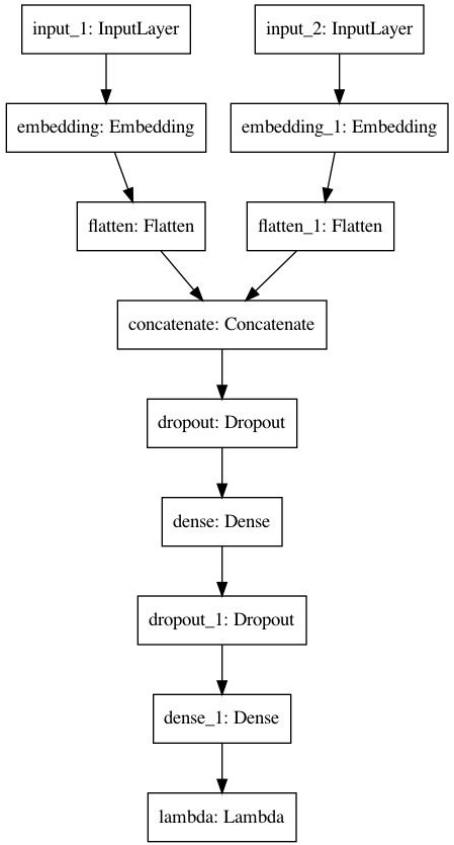
gdzie: R_i(u) ,

$$\mathrm{dev}(i,j) = \frac{1}{|U_{ij}|} \sum_{u \in U_{ij}} r_{ui} - r_{uj}$$

Lemire, D., & Maclachlan, A. (2005). Slope One Predictors for Online Rating-Based Collaborative Filtering. SDM.

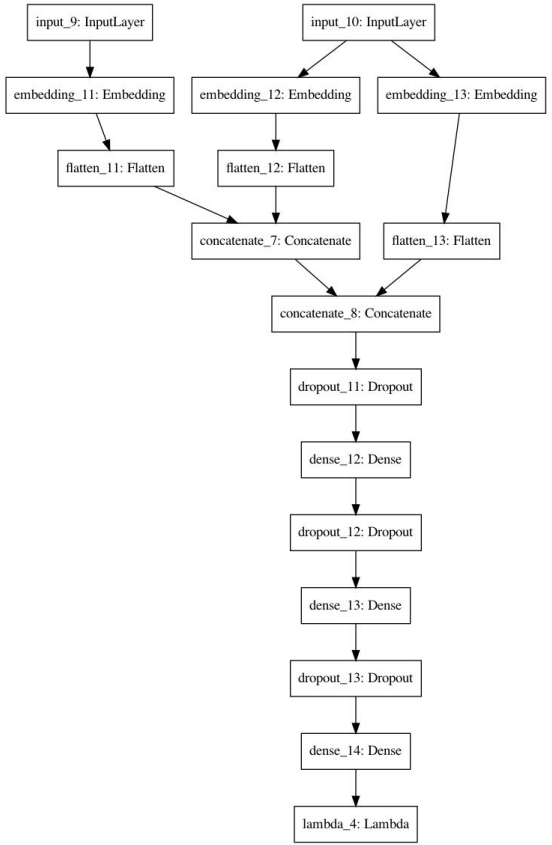
Embedding & NN

- input: (user id, movie id)
- output: rating
- loss: MSE, optimizer: Adam(lr=0.001)
- wymiar osadzeń: 50
- regularyzacja L2 na macierzach osadzeń + dropout na jednostkach FC



Embedding & NN - podejście hybrydowe

- Dla każdego filmu pobrano zarys fabuły z imdb.com
- Teksty osadzono na word2vec
- Konkatenacja wektora osadzenia obiektu z embeddings ID user i obiekt



Embedding & NN - wyniki

Method	MAE Train-Test (30%)
KNN (Pearson)	0.7338
SVD	0.7420
SVD++	0.7239
Embedding + NN	0.6711
Embedding + content + NN	0.6687

Dziękujemy