

Sisteme de recomandare

Sistemele de recomandare (*Recommender Systems* - RS) sunt algoritmi care realizează predicții pentru luarea de decizii prin prelucrări statistice ale caracteristicilor elementelor dintr-un sistem de tip utilizator-produs sau stare-acțiune. RS se utilizează cu precădere pentru identificarea celor mai bune sugestii privind produsele și serviciile oferite de o platformă (ex. magazin online), target-ul fiind comunitatea de utilizatori ai acelei platforme.

RS sunt de mai multe tipuri, funcție de prelucrările pe care le realizează. Cele mai multe astfel de sisteme sunt:

- Bazate pe conținut - RS realizează legături între o baza de date de utilizatori și o baza de date cu produse prin cuvinte cheie deduse pe baza caracteristicilor utilizatorilor și produselor. În acest sens, se iau în calcul atât specificațiile produselor, cât și detalii ce provin din profilurile utilizatorilor
- Colaborative – RS care realizează asocierea produs-utilizator pe baza popularității produselor. Aceste sisteme iau în calcul măsura în care produsele au fost selectate/cumpărate/evaluate de către utilizatorii unei comunități, pentru a face recomandări utilizatorilor care încă nu au întâlnit acele produse.

RS utilizate de către platformele de mari dimensiuni sunt sisteme hibride, constând într-un pipeline complex ce implică și prelucrare de conținut și procesare de tip colaborativ. În continuare vom studia RS ce presupun prelucrare de tip colaborativ, întrucât acestea apar cel mai frecvent în sistemele din lumea reală.

Adesea, RS colaborative primesc datele de intrare sub forma unei matrice de rating-uri, care are următoarele caracteristici:

- matricea are dimensiunile $N \times M$
 - o N utilizatori
 - o M produse
- doar o submulțime a utilizatorilor au dat rating-uri unei submulțimi de produse. Perechile utilizator-produs pentru care există rating al produsului din partea utilizatorilor sunt în general mult mai mici decât $M \times N$
- prin urmare, matricea de rating-uri este *rară* (engl. *sparse matrix*) (Fig 1.)
 - o elementele definite sunt în număr restrâns
 - o majoritatea elementelor sunt nedefinite



					
	5	3		3	
		1			5
	5	4	3		
	1	2			4
		1			
			4	5	

Fig. 1. Exemplu de matrice de rating-uri

Scopul unui RS colaborativ este ca, pornind de la valorile disponibile din matrice, să se determine estimări ale valorilor lipsă. Un exemplu în acest sens este ilustrat în Fig. 2, unde problema care se pune este: care dintre produsele pe care al doilea utilizator nu le-a evaluat încă i se pot recomanda acestuia? Decizia se ia prin realizarea de predicții (estimări) ale valorilor lipsă din linia corespunzătoare utilizatorului. Funcție de rating-urile estimate, se decide dacă un anumit produs i se poate recomanda acestuia. De exemplu, dacă ratingurile $\in [1, 5]$ și valoarea estimată pe poziția (2, 1) este 4.5, atunci se poate decide faptul că primul produs constituie o recomandare adecvată pentru al doilea utilizator.

		Produse				
						
Utilizatori		5	3		3	
		?	1	?	?	5
		5	4	3		
		1	2			4
			1			
				4	5	

Fig. 2. Matrice de rating-uri unde problema care se pune este în ce măsură sunt recomandabile utilizatorului al doilea produsele pentru care acesta nu a dat rating

Sarcini:**1. Fie următoarea matrice de ratinguri:**

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

Să se determine rating-ul pe care utilizatorul **Alice** l-ar da produsului **Item5** prin următoarele metode:

1.1. Colaborare bazată pe similaritatea dintre utilizatori

Se estimează valoarea rating-ului lipsă pe baza similarității dintre Alice și ceilalți utilizatori, care au dat rating pentru Item5. Vom calcula similaritatea dintre doi utilizatori folosind **corelația Pearson**:

$$sim(a, b) = \frac{\sum_{p \in P} (r_{a,p} - \bar{r}_a)(r_{b,p} - \bar{r}_b)}{\sqrt{\sum_{p \in P} (r_{a,p} - \bar{r}_a)^2} \sqrt{\sum_{p \in P} (r_{b,p} - \bar{r}_b)^2}}$$

unde:

- a, b – doi utilizatori
- $r_{a,p}$ – rating-ul dat de utilizatorul a obiectului p
- \bar{r}_a = rating-ul mediu dat de utilizatorul a
- P – mulțimea de obiecte care au primit rating-uri de la a și b

Rating-ul lipsă se poate estima astfel:

$$pred(a, p) = \bar{r}_a + \frac{\sum_{b \in N} sim(a, b) * (r_{b,p} - \bar{r}_b)}{\sum_{b \in N} sim(a, b)}$$

Formula realizează o estimare a rating-ului pe care utilizatorul a l-ar da produsului p :

- a – utilizatorul care nu a dat rating produsului p
- \bar{r}_a = rating-ul mediu dat de utilizatorul a
- b - unul dintre utilizatorii care au dat rating produsului p
- N – mulțimea utilizatorilor care au dat rating obiectului p
- $r_{b,p}$ = rating-ul dat de utilizatorul b produsului p
- \bar{r}_b = rating-ul mediu dat de utilizatorul b

1.2. Colaborare bazată pe similaritatea dintre produse

Se estimează valoarea rating-ului lipsă pe baza similarității dintre Item5 și celelalte produse, care au primit rating de la Alice. Vom calcula similaritatea dintre două produse folosind **similaritatea cosinus**:

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|}$$

a, b sunt două produse. Se tratează a, b ca fiind vectori și se determină cosinusul unghiului format de cei doi. Cosinusul se calculează ca fiind raportul dintre produsul scalar al vectorilor și produsul lungimilor lor.

De exemplu, presupunem că a și b au primit ratingurile r_{a1} , r_{a2} , r_{b1} , r_{b2} de la doi utilizatori. Atunci:

$$\vec{a} = [r_{a1} \ r_{a2}]$$

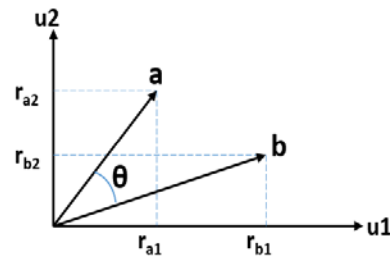
$$\vec{b} = [r_{b1} \ r_{b2}]$$

$$\text{sim}(\vec{a}, \vec{b}) = \cos(\theta)$$

$$\text{sim}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}$$

$$\vec{a} \cdot \vec{b} = r_{a1} * r_{b1} + r_{a2} * r_{b2}$$

$$|\vec{a}| = \sqrt{r_{a1}^2 + r_{a2}^2}$$



Calcululele sunt similare în cazul în care produsele au rating-uri de la mai mulți utilizatori, singura diferență fiind faptul că vectorii au mai multe coordonate (r_{a1} , r_{a2} , r_{a3} , ...)

Rating-ul lipsă se poate estima astfel:

$$\text{pred}(u, p) = \frac{\sum_{i \in \text{ratedItem}(u)} \text{sim}(i, p) * r_{u,i}}{\sum_{i \in \text{ratedItem}(u)} \text{sim}(i, p)}$$

u – utilizatorul care nu a dat rating produsului p

i – produsele care au primit rating de la u

r_{ui} – ratingul dat de u produsului i

2. Fie următoarea matrice de ratinguri:

		Item			
		W	X	Y	Z
User	A		4.5	2.0	
	B	4.0		3.5	
	C		5.0		2.0
	D		3.5	4.0	1.0

Rating Matrix

Să se determine elementele lipsă (estimări ale rating-urilor utilizatorilor A, B, C, D pentru produsele W, X, Y, Z pe care nu le-au evaluat încă) prin **factorizarea matricei**.

Factorizarea presupune descompunerea matricei într-un produs de două matrice de dimensiuni mai mici. În cazul nostru, vom descompune matricea rating-urilor (4x4) într-un produs de două matrice de dimensiuni (4x2) și (2x4):

		Item			
		W	X	Y	Z
User	A		4.5	2.0	
	B	4.0		3.5	
	C		5.0		2.0
	D		3.5	4.0	1.0

Rating Matrix

$$=$$

User	A		
	B		
	C		
	D		

User Matrix

$$\times$$

		W	X	Y	Z

Item Matrix

Trebuie să determinăm elementele matricelor User și Item matrix. Odată determinate, pentru a realiza estimarea elementului BX din matricea de ratinguri se face produsul liniei B din matricea User și coloanei X din matricea Item. Identificarea matricelor factori se realizează prin metode numerice. Vom folosi în acest scop metoda gradientului descendent.

Notăm

- N = numărul de utilizatori
- M = numărul de produse
- n = indexul unui utilizator, $n \in [0..N-1]$
- m = indexul unui produs, $m \in [0..M-1]$
- K = numărul liniilor, respectiv al coloanelor din cele două matrice factor
- k = indexul liniilor, respectiv coloanelor din cele două matrice factor, $k \in [0..K-1]$
- R = matricea ratingurilor, de dimensiuni NxM

- U = matricea utilizatorilor, de dimensiuni $N \times K$
- V = matricea produselor, de dimensiuni $K \times M$

Atunci descompunerea matricei arată astfel:

$$R_{N \times M} = U_{N \times K} \times V_{K \times M}$$

Se cunosc o parte din elementele din R (celelalte trebuie estimate). Scopul este determinarea elementelor din U și V , pe baza cărora se va face estimarea.

- se inițializează u_{nk} , v_{km} cu valori aleatoare
- dorim să determinăm valorile u_{nk} , v_{km} pentru care este minimă eroarea:

$$MSE = \frac{1}{2} (r_{nm} - \sum_k u_{nk} v_{km})^2$$

Pe parcursul mai multor iterații:

- se determină eroarea
- se determină gradientii erorii în raport cu u_{nk} , v_{km} , care au următoarea expresie:

$$\frac{\partial MSE}{\partial u_{nk}} = (\sum_k u_{nk} v_{km} - r_{nm}) v_{km}$$

$$\frac{\partial MSE}{\partial v_{km}} = (\sum_k u_{nk} v_{km} - r_{nm}) u_{nk}$$

- se ajustează valorile u_{nk} , v_{km} funcție de valorile gradientilor și ai ratei de învățare α :

$$u_{nk} = u_{nk} - \alpha \frac{\partial MSE}{\partial u_{nk}}$$

$$v_{km} = v_{km} - \alpha \frac{\partial MSE}{\partial v_{km}}$$

- algoritmul se oprește fie după un număr limită de iterații, fie atunci când eroarea nu se mai modifică semnificativ de la o iterație la alta.