

# POLITECHNIKA ŁÓDZKA

WYDZIAŁ FIZYKI TECHNICZNEJ, INFORMATYKI  
I MATEMATYKI STOSOWANEJ

Kierunek: Matematyka Stosowana

Specjalność: Analiza Danych w Biznesie i Logistyce

---

## **Matematyczne modele wykorzystywane w systemach rekomendacji.**

Anita Kudaj  
Nr albumu: 220020

---

Praca magisterska napisana w Instytucie Matematyki  
Politechniki Łódzkiej

Promotor: dr, mgr inż. Piotr Kowalski

ŁÓDŹ, 07.2019

# Spis treści

<b>1</b>	<b>Wstęp</b>	<b>3</b>
<b>2</b>	<b>Preliminaria</b>	<b>4</b>
2.1	Oznaczenia używane w pracy . . . . .	4
2.2	Elementy algebry liniowej . . . . .	5
2.3	Elementy rachunku prawdopodobieństwa i statystyki . . . . .	8
<b>3</b>	<b>Elementy eksploracji danych wykorzystywane w systemach rekomendujących</b>	<b>11</b>
3.1	Wstępne przetwarzanie danych . . . . .	11
3.1.1	Miary podobieństwa . . . . .	12
3.1.2	Redukcja wymiaru . . . . .	13
3.2	Metody eksploracji danych . . . . .	16
3.2.1	Algorytm k - najbliższych sąsiadów . . . . .	16
3.2.2	Algorytm k - średnich . . . . .	17
3.3	Szacowanie błędów obliczeń . . . . .	19
3.3.1	Ocena dokładności metody . . . . .	19
3.3.2	Ocena jakości modelu . . . . .	21
<b>4</b>	<b>Modele tworzenia rekomendacji</b>	<b>24</b>
4.1	Systemy rekomendujące oparte na treści - Content-based recommender systems . . . . .	26
4.1.1	Wygenerowanie profilu dokumentu tekstowego - algorytm TFIDF . . . . .	28
4.2	Filtrowanie kolaboratywne - Collaborative filtering . . . . .	29
4.2.1	Filtrowanie kolaboratywne oparte na użytkowniku . . . . .	29
4.2.2	Filtrowanie kolaboratywne oparte na elementach . . . . .	32
4.3	Systemy rekomendujące kontekstowe - Context-aware recommender systems . . . . .	33
4.4	Dekompozycja macierzy ocen - SVD . . . . .	34

<b>5</b>	<b>Eksperymenty / część praktyczne</b>	<b>37</b>
5.1	ALS z Apache Spark i MLlib . . . . .	37
5.1.1	Apache Spark . . . . .	37
5.1.2	ALS i MLlib . . . . .	38
5.1.3	Implementacja algorytmu . . . . .	39
<b>6</b>	<b>Podsumowanie</b>	<b>40</b>

# Rozdział 1

## Wstęp

# Rozdział 2

## Preliminaria

### 2.1 Oznaczenia używane w pracy

$\mathbb{N}$  - zbiór liczb naturalnych,

$\mathbb{R}$  - zbiór liczb rzeczywistych,

$\mathbb{K}$  - ciało liczb rzeczywistych lub zespolonych,

$X$  - zbiór,

$\mathbf{x}$  - wektor,

$X$  - zmienna losowa,

$\mathbb{X}$  - macierz,

$[a_{ij}]_{j=1, \dots, n}^{i=1, \dots, m}$  - macierz o  $m$  wierszach i  $n$  kolumnach,

$[a_{ij}]$  - macierz kwadratowa,

$\mathbb{M}_{m \times n}(\mathbb{K})$  - zbiór wszystkich macierzy o wymiarach  $m \times n$  i elementach z ciała  $\mathbb{K}$ ,

$\mathcal{V}$  - przestrzeń liniowa,

$\text{span}(X)$  - przestrzeń generowana przez zbiór  $X$ ,

$(\Omega, \mathcal{F}, P)$  - przestrzeń probabilistyczna,

$\Omega$  - zbiór zdarzeń elementarnych,

$\mathcal{F}$  - rodzina podzbiorów zbioru  $\Omega$ ,

$P$  - funkcja prawdopodobieństwa,

$B(\mathbb{R}^n)$  -  $\sigma$ -algebra borelowska zbiorów w  $\mathbb{R}^n$ ,

$E(X)$  - wartość oczekiwana,

$\text{Cov}(X; Y)$  - kowariancja zmiennych losowych  $X, Y$ ,

$\text{Var}(X)$  - wariancję zmiennej losowej  $X$ ,

$\sigma(X)$  - odchylenie standardowe zmiennej losowej  $X$ ,

$\rho(X, Y)$  - współczynnik korelacji zmiennych losowych  $X, Y$ ,

$\Sigma$  - macierz kowariancji,

$\mathcal{P}$  - macierz korelacji,

$d(x, y)$  - odległość punktów  $x$  i  $y$ ,  
 $d_e(x, y)$  - odległość euklidesowa punktów  $x$  i  $y$ ,  
 $d_r(x, y)$  - odległość Minkowskiego punktów  $x$  i  $y$ ,  
 $\text{sim}(X, Y)$  - współczynnik podobieństwa wektorów  $X$  i  $Y$ ,  
 $\rho^p(X, Y)$  - współczynnik korelacji Pearsona,  
 $J(A, B)$  - indeks Jaccarda,  
 $\|\cdot\|_F$  - norma Frobeniusa.

## 2.2 Elementy algebry liniowej

W definicjach poniżej korzystamy z pojęcia ciała, którego wyjaśnienie odnajdziemy w książce Tadeusza Poredy i Jacka Jędrzejewskiego *Algebra liniowa z elementami geometrii analitycznej* [6, Sec 4.4].

**Definicja 2.1** (Macierz [6, Sec 8.1 Def. 8.1]). *Niech  $n, m \in \mathbb{N}$ . Macierzą o  $m$  wierszach,  $n$  kolumnach (o wymiarach  $m \times n$ ) i wyrazach w ciele  $\mathbb{K}$  nazywamy funkcję*

$$\mathbb{A} : \{1, 2, \dots, m\} \times \{1, 2, \dots, n\} \rightarrow \mathbb{K}.$$

*Wartością funkcji dla argumentu  $(i, j)$  jest element  $a_{ij}$  należący do ciała  $\mathbb{K}$ . Macierz zapisujemy w postaci tabeli*

$$\mathbb{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

Przez  $\mathbb{M}_{m \times n}(\mathbb{K})$  oznaczmy zbiór wszystkich macierzy o wymiarach  $m \times n$  i elementach z ciała  $\mathbb{K}$ .

**Uwaga 2.2** ([6, Sec 8.1]). *Przykładowe sposoby zapisu macierzy:*

$$[a_{ij}]_{j=1, \dots, n}^{i=1, \dots, m}, (a_{ij})_{j=1, \dots, n}^{i=1, \dots, m}, [a_{ij}]_{j \leq n}^{i=1 \leq m}, (a_{ij})_{j \leq n}^{i=1 \leq m}, [a_{ij}], (a_{ij}).$$

*Sposobów  $[a_{ij}]$ ,  $(a_{ij})$  używamy, gdy liczba kolumn i wierszy danej macierzy jest ustalona.*

*W tej pracy używać będziemy zapisu  $[a_{ij}]_{j=1, \dots, n}^{i=1, \dots, m}$  oraz zapisu  $[a_{ij}]$  w przypadku macierzy kwadratowych.*

**Definicja 2.3** (Macierz transponowana [6, Sec 8.1]). *Niech  $\mathbb{A}$ , gdzie  $\mathbb{A} = [a_{ij}]_{j=1, \dots, n}^{i=1, \dots, m}$ , będzie macierzą ze zbioru  $\mathbb{M}_{m \times n}(\mathbb{K})$ . Macierz  $\mathbb{B} = [b_{ij}]_{j=1, \dots, m}^{i=1, \dots, n}$  nazywamy macierzą transponowaną macierzy  $\mathbb{A}$ , jeśli*

$$b_{ji} = a_{ij}$$

dla każdego  $i \in \{1, \dots, n\}$  oraz  $j \in \{1, \dots, m\}$  i oznaczamy jako  $\mathbb{A}^T$ .

**Uwaga 2.4** (Rodzaje macierzy [6, Sec 8.1, Sec 10.4]). Poniżej zostały zdefiniowane niektóre rodzaje macierzy.

- Macierz kwadratową nazywamy macierz, w której liczb wierszy i liczba kolumn są równe. Liczbę tę nazywamy stopniem macierzy kwadratowej.
- Macierz diagonalną nazywamy macierz kwadratową  $[a_{ij}]$ , gdzie wszystkie elementy poza główną przekątną są równe 0. Macierz diagonalną oznaczamy  $\text{diag}(a_{11}, a_{22}, \dots, a_{nn})$ .
- Macierz jednostkową stopnia  $n$  nazywamy macierz diagonalną, w której na głównej przekątnej wszystkie elementy są równe 1. Macierz jednostkową będziemy oznaczać  $\mathbb{I}$ .
- Macierz kwadratową  $\mathbb{C}$ , gdzie  $\mathbb{C} = [c_{ij}]$ , nazywamy macierzą ortogonalną, jeżeli spełniony jest warunek

$$\mathbb{C}^T \cdot \mathbb{C} = \mathbb{C} \cdot \mathbb{C}^T = \mathbb{I}.$$

- Macierz nieosobliwą nazywamy macierz kwadratową, której wyznacznik jest różny od 0.
- Macierz osobliwą nazywamy macierz kwadratową, której wyznacznik jest równy 0.

**Definicja 2.5** (Mnożenie macierzy [6, Sec 9.3 Def 9.13]). Niech  $\mathbb{A} \in \mathbb{M}_{m \times n}(\mathbb{K})$  i  $\mathbb{B} \in \mathbb{M}_{k \times m}(\mathbb{K})$ . Jeśli

$$\mathbb{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbb{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{km} \end{bmatrix}$$

to iloczynem macierzy  $\mathbb{B}$  i  $\mathbb{A}$  nazywamy macierz  $\mathbb{C}$  taką, że

$$\mathbb{C} = [c_{lj}]_{j=1, \dots, n}^{l=1, \dots, k},$$

gdzie

$$c_{lj} = \sum_{i=1}^m b_{li} \cdot a_{ij}.$$

Element  $c_{lj}$  tego iloczynu to iloczyn  $l$ -tego wiersza macierzy  $\mathbb{B}$  przez  $j$ -tą kolumnę macierzy  $\mathbb{A}$ .

**Definicja 2.6** (Dodawanie macierzy [6, Sec 8.1]). Niech  $\mathbb{A}, \mathbb{B} \in \mathbb{M}_{m \times n}(\mathbb{K})$ . Dodawaniem macierzy  $(\mathbb{B} + \mathbb{A})$  nazywamy macierz  $\mathbb{C} \in \mathbb{M}_{m \times n}(\mathbb{K})$  taką, że

$$c_{ij} = b_{ij} + a_{ij}.$$

**Definicja 2.7** (Mnożenie macierzy przez element ciała [6, Sec 8.1]). Niech  $\mathbb{A} \in \mathbb{M}_{m \times n}(\mathbb{K})$  oraz  $\lambda \in \mathbb{K}$ . Mnożeniem macierzy przez element z ciała  $(\lambda \cdot \mathbb{A})$  nazywamy macierz  $\mathbb{C} \in \mathbb{M}_{m \times n}(\mathbb{K})$  taką, że

$$c_{ij} = \lambda \cdot a_{ij}.$$

**Twierdzenie 2.8** (Własności transpozycji macierzy [4, Sec 5.1 Tw. 5.1]). Niech  $\mathbb{A}$  i  $\mathbb{B}$  będą macierzami o współczynnikach z  $\mathbb{K}$  oraz niech posiadają tyle kolumn i wierszy, że operacje występujące powyżej są określone. Niech  $\lambda \in \mathbb{K}$ . Zachodzą następujące równości:

- $(\mathbb{A}^T)^T = \mathbb{A}$ ,
- $(\mathbb{A} + \mathbb{B})^T = \mathbb{A}^T + \mathbb{B}^T$ ,
- $(\lambda \mathbb{A})^T = \lambda \mathbb{A}^T$ ,
- $(\mathbb{A}\mathbb{B})^T = \mathbb{B}^T \mathbb{A}^T$ .

**Definicja 2.9** (Ślad macierzy). Śladem macierzy  $\mathbb{A} = [a_{ij}]$  nazywamy wielkość:

$$\text{tr}(\mathbb{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}.$$

**Uwaga 2.10** ([6, Sec 8.1]). Wiersz macierzy o wymiarach  $m \times n$  traktować możemy jako wektor z przestrzeni  $\mathbb{K}^n$ , natomiast kolumnę jako wektor przestrzeni  $\mathbb{K}^m$ .

**Definicja 2.11** (Rząd kolumnowy i wierszowy macierzy [6, Sec 8.1]). Niech  $\mathbb{A} \in \mathbb{M}_{m \times n}(\mathbb{K})$ . Rzędem kolumnowym macierzy  $\mathbb{A}$  nazywamy wymiar przestrzeni  $\mathbb{K}^n$  generowanej przez kolumny macierzy  $\mathbb{A}$ . Rząd ten oznaczamy symbolem  $r_k(\mathbb{A})$ . Rzędem wierszowym macierzy  $\mathbb{A}$  nazywamy wymiar podprzestrzeni generowanej przez wiersze macierzy  $\mathbb{A}$  i oznaczamy go  $r_w(\mathbb{A})$ .

**Definicja 2.12** (Rząd macierzy [6, Sec 8.1]). Rzędem macierzy  $\mathbb{A}$  nazywamy wspólną wartość rzędu kolumnowego i wierszowego macierzy  $\mathbb{A}$ . Rząd macierzy oznaczamy symbolem  $\text{rz}(\mathbb{A})$ .

W definicji poniżej korzystamy z pojęcia przestrzeni liniowej. Wyjaśnienie tego pojęcia możemy odnaleźć w książce *Algebra liniowa z elementami geometrii analitycznej* [6, Sec 7.1].



**Definicja 2.13** (Przestrzeń generowana przez zbiór [6, Sec 7.1 Def 7.13]). *Niech  $X$  będzie dowolnym i niepustym podzbiorem przestrzeni liniowej  $\mathcal{V}$ . Podprzestrzenią generowaną przez zbiór  $X$  nazywamy zbiór wszystkich skończonych kombinacji liniowych wektorów ze zbioru  $X$ . Zbiór ten oznaczamy symbolem  $\text{span}(X)$ .*

*Symbolicznie zapisujemy zbiór  $\text{span}(X)$  jako:*

$$\left\{x \in \mathcal{V} : \exists_{n \in \mathbb{N}} \exists_{(\alpha_1, \dots, \alpha_n) \in \mathbb{K}^n} \exists_{(a_1, \dots, a_n) \in X^n} (x = \alpha_1 \cdot a_1 + \dots + \alpha_n \cdot a_n)\right\},$$

gdzie  $\mathcal{V}$  jest przestrzenią liniową nad ciałem liczb rzeczywistych lub ciałem liczb zespolonych  $\mathbb{K}$ .

**Definicja 2.14** (Wartość własna macierzy kwadratowej [6, Sec 12.2]). *Liczbę  $\lambda$  nazywamy wartością własną macierzy kwadratowej  $\mathbb{A}$ , jeżeli istnieje niezerowy wektor  $\mathbf{x}$  taki, że*

$$\mathbb{A}\mathbf{x} = \lambda\mathbf{x}.$$

*Każdy niezerowy wektor  $\mathbf{x}$  spełniający powyższe równania nazywamy wektorem własnym macierzy  $\mathbb{A}$  odpowiadającym wartości własnej  $\lambda$ .*

## 2.3 Elementy rachunku prawdopodobieństwa i statystyki

**Definicja 2.15** (Przestrzeń probabilistyczna [2, Sec 1.2, Sec 1.4]). *Przestrzenią probabilistyczną nazywamy uporządkowaną trójkę  $(\Omega, \mathcal{F}, P)$ , gdzie:*

- $\Omega$  to zbiór wszystkich zdarzeń elementarnych i  $\Omega \neq \emptyset$ ,
- $\mathcal{F}$  rodzina podzbiorów zbioru  $\Omega$  taka, że:
  - $\emptyset \in \mathcal{F}$ ,
  - jeżeli  $A \in \mathcal{F}$ , to  $\bar{A} = \Omega \setminus A \in \mathcal{F}$ ,
  - jeżeli  $A_n \in \mathcal{F}$  dla  $n = 1, 2, \dots$ , to  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ ,
- $P : \mathcal{F} \rightarrow [0, 1]$  taka, że:
  - $\forall_{A \in \mathcal{F}} P(A) \geq 0$ ,
  - $P(\Omega) = 1$ ,
  - jeżeli  $A_n \in \mathcal{F}$ ,  $n = 1, 2, \dots$  są takie, że  $A_i \cap A_j = \emptyset$  dla  $i \neq j$  to

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

W poniższej definicji zostało użyte pojęcie  $\sigma$ -algebry, którego wyjaśnienie można odnaleźć w [2, Sec 1.2 Def. 1.2].

**Definicja 2.16** ( $B(\mathbb{R}^n)$  [2, Sec 1.12]). *Niech  $\mathcal{I}$  oznacza klasę wszystkich zbiorów, składających się ze skończonych sum rozłącznych zbiorów  $I = I_1 \times I_2 \times \dots \times I_k$ , gdzie  $I_k = [a_k, b_k]$ . Najmniejszą  $\sigma$ -algebrę  $\sigma(\mathcal{I})$  generowaną przez klasę zbiorów  $\mathcal{I}$  nazywa się  $\sigma$ -algebrą borelowską zbiorów w  $\mathbb{R}^n$  i oznacza się  $B(\mathbb{R}^n)$ .*

**Definicja 2.17** (Zmienna losowa [3, Sec 5.1 Def. 1]). *Odwzorowanie  $X : \Omega \rightarrow \mathbb{R}^n$  nazywamy zmienną losową o wartościach w  $\mathbb{R}^n$ , jeśli dla każdego  $A \in B(\mathbb{R}^n)$  zbiór  $X^{-1}(A) \in \mathcal{F}$ .*

**Definicja 2.18** (Wartość oczekiwana [3, Sec 5.6 Def. 2]). *Wartością oczekiwaną zmiennej losowej  $X$  o wartościach w  $\mathbb{R}$  nazywamy liczbę:*

$$E(X) = \int_{\Omega} X dP,$$

*jeżeli  $X$  jest  $P$ -całkowalna, tzn. jeżeli zachodzi:*

$$E(X) = \int_{\Omega} |X| dP < \infty.$$

**Definicja 2.19** (Kowariancja [2, Sec 2.8 Def.2.32]). *Kowariancją zmiennych losowych  $X, Y$  nazywamy liczbę:*

$$\text{Cov}(X; Y) = E((X - E(X))(Y - E(Y))).$$

**Definicja 2.20** (Wariancja zmiennej losowej [2, Sec 2.8 Def.2.28]). *Wariancją zmiennej losowej  $X$  nazywamy liczbę:*

$$\text{Var}(X) = E((X - E(X))^2),$$

*jeżeli po prawej stronie równości wartość oczekiwana istnieje.*

**Definicja 2.21** (Odchylenie standardowe [2, Sec 2.8 Def.2.28]). *Odchyleniem standardowym zmiennej losowej  $X$  nazywamy liczbę:*

$$\sigma(X) = \sqrt{\text{Var}(X)}.$$

**Definicja 2.22** (Współczynnik korelacji [2]). *Współczynnikiem korelacji nazywamy charakterystykę ilościową stopnia zależności dwóch zmiennych losowych  $X$  i  $Y$  zdefiniowaną następująco:*

$$\rho(X, Y) = \frac{\text{Cov}(X; Y)}{\sigma(X) \sigma(Y)}.$$

**Definicja 2.23** (Macierz kowariancji). *Macierz kowariancji nazywamy macierz*

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix},$$

gdzie:

- $\sigma_{ii}$  - wariancja zmiennej  $X_i$ ,
- $\sigma_{ij} = \text{Cov}(X_i; X_j)$  - kowariancja między zmiennymi  $X_i$  i  $X_j$ .

**Definicja 2.24** (Macierz korelacji). *Macierz korelacji nazywamy*

$$\mathcal{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix},$$

gdzie:

- $\rho_{ij} = \frac{\text{Cov}(X_i; X_j)}{\sigma_i \sigma_j}$  - współczynnik korelacji  $X_i$  i  $X_j$ .

## Rozdział 3

# Elementy eksploracji danych wykorzystywane w systemach rekomendujących

Większość systemów rekomendujących opiera się na algorytmach, które możemy rozumieć jako różne technik eksploracji danych. Zazwyczaj proces eksploracji danych składa się z trzech kroków:

1. wstępne przetwarzanie danych,
2. analiza danych,
3. interpretacja wyników.

W tym rozdziale zostaną przeanalizowane najważniejsze i najczęściej używane w regułach rekomendujących metody. Zaczniemy od miar podobieństw i redukcji wymiaru. W kolejnym etapie spojrzymy na metody klasyfikacji, grupowania i regresji, aby zakończyć interpretacją wyników i oceną błędów obliczeń.

### 3.1 Wstępne przetwarzanie danych

Przed przystąpieniem do kroku analizy dane wymagają przygotowania: wyczyszczenia, przefiltrowania, transformacji. Dopiero tak przygotowane dane mogą zostać poddane zadaniom uczenia maszynowego. W tej sekcji zostaną przedstawione problemy, które spotykamy przy tworzeniu reguł rekomendujących.

### 3.1.1 Miary podobieństwa

W systemach rekomendujących bardzo częstym podejściem jest używanie metod klasyfikacji i grupowania. Metody te opierają się na obliczaniu podobieństw i odległości. Najprostszym i jednocześnie najczęściej używanym podejściem jest odległość euklidesowa.

**Definicja 3.1** (Odległość euklidesowa [1]). *Niech  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ,  $n \in \mathbb{N}$ . Odległością euklidesową  $\mathbf{x}$  i  $\mathbf{y}$  nazywamy:*

$$d_e(x, y) = \sqrt{\sum_{k=1}^n (\mathbf{x}_k - \mathbf{y}_k)^2}.$$

Warto również wspomnieć o uogólnionej wersji odległości euklidesowej - odległości Minkowskiego.

**Uwaga 3.2.** *Odległość Minkowskiego wyrażamy wzorem:*

$$d_r(x, y) = \left( \sum_{k=1}^n |x_k - y_k|^r \right)^{\frac{1}{r}}.$$

*W zależności od wartości stopnia odległości  $r$  odległość Minkowskiego przyjmuje konkretne nazwy:*

- $r = 1$  - odległość manhatan,
- $r = 2$  - wspomniana wcześniej odległość euklidesowa,
- $r \rightarrow \infty$  - supremum.

Kolejnym podejściem, gdzie poszczególne elementy są postrzegane jako  $n$  - wymiarowe wektory, a podobieństwo między nimi jest obliczane na podstawie kąta, który tworzą jest odległość kosinusowa.

**Definicja 3.3** (Odległość kosinusowa [1]). *Niech  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ,  $n \in \mathbb{N}$ . Odległością kosinusową nazywamy:*

$$d(\mathbf{x}, \mathbf{y}) = 1 - \text{sim}(\mathbf{x}, \mathbf{y}),$$

*gdzie  $\text{sim}(\mathbf{x}, \mathbf{y})$  to współczynnik podobieństwa wektorów  $\mathbf{x}$  i  $\mathbf{y}$ :*

$$\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n x_k y_k}{\sqrt{\sum_{k=1}^n x_k^2} \sqrt{\sum_{k=1}^n y_k^2}}.$$

Innym podejściem jest korelacja Pearsona, którą definiujemy następująco:

**Definicja 3.4** (Współczynnik korelacji Pearsona [1]). *Niech  $X, Y$  będą zmiennymi losowymi o rozkładach ciągłych oraz niech  $x_k, y_k$ , gdzie  $k \in \{1, \dots, n\}$  oznaczają wartości prób losowych tych zmiennych. Przez  $\bar{x}$  i  $\bar{y}$  oznaczmy:*

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad \bar{y} = \frac{1}{n} \sum_{k=1}^n y_k.$$

*Wówczas współczynnikiem korelacji Pearsona nazywamy:*

$$\rho^p(X, Y) = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^n (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^n (y_k - \bar{y})^2}}.$$

Indeks Jaccarda (współczynnik podobieństwa Jaccarda) to kolejny wskaźnik opisujący podobieństwo.

**Definicja 3.5** (Indeks Jaccarda [8]). *Niech  $A$  i  $B$  oznaczają zbiory. Indeks Jaccarda (podobieństwem Jaccarda) nazywamy funkcję:*

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

### 3.1.2 Redukcja wymiaru

Wraz ze wzrostem ilości obserwacji rośnie ich dokładność. Warto zauważyć, że tym samym rośnie stopień komplikacji w interpretacji otrzymanych wyników. Zbyt duża ilość zmiennych, które opisują obserwacje powoduje wzrost prawdopodobieństwa, że zmienne te są ze sobą skorelowane, a informacje wnoszone przez część zmiennych są redundantne. Proces redukcji wymiaru pozwala przezwyciężyć ten problem poprzez transformację przestrzeni danych do przestrzeni o mniejszej liczbie wymiarów. W poniższym rozdziale przyjrzymy się najczęściej wybieranemu algorytmom redukcji wymiarów w kontekście reguł rekomendujących. Jest to Rozkład Według Wartości Osobliwych (ang. Singular Value Decomposition (SVD)).

**Definicja 3.6** (Rozkład Według Wartości Osobliwych [10]). *Rozkładem według wartości osobliwych  $m \times n$  - wymiarowej macierzy  $\mathbb{X}$ , gdzie  $m \geq n$  nazywamy rozkład:*

$$\mathbb{X} = \mathbb{U} \Sigma \mathbb{V}^T,$$

gdzie:

- $\mathbb{U}^T \mathbb{U} = \mathbb{V}^T \mathbb{V} = \mathbb{I}$ ,  $\mathbb{U}$  jest wymiaru  $m \times m$  oraz  $\mathbb{V}$  wymiaru  $n \times n$ ,
- $\Sigma$  jest macierzą diagonalną o nieujemnych wartościach,  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$ ,  $n \in \mathbb{N}$  taką, że  $\sigma_1 > 0$  dla  $1 \leq i \leq \text{rz}(\mathbb{X})$  i  $\sigma_i = 0$  dla  $i \geq \text{rz}(\mathbb{X}) + 1$  (macierz  $\Sigma$  jest macierzą wymiaru  $m \times n$ , gdzie  $\sigma_{ij} = 0$ , gdy  $i \neq j$ ).

**Uwaga 3.7.** Niezerowe wyrazy macierzy  $\Sigma$  nazywamy wartościami osobliwymi macierzy  $\mathbb{X}$ . Kolumny macierzy  $\mathbb{U}$  i  $\mathbb{V}$  nazywamy odpowiednio lewymi i prawymi wektorami szczególnymi macierzy  $\mathbb{X}$ .

**Definicja 3.8** (Wartość osobliwa macierzy [13]). Wartością osobliwą  $\sigma_k$  macierzy  $\mathbb{X}$  nazywamy

$$\sigma_k = \sqrt{\lambda_k},$$

gdzie  $\lambda_k$ ,  $k \in \mathbb{N}$  jest wartością własną macierzy  $\mathbb{X}\mathbb{X}^T$ .

**Definicja 3.9** (Norma Frobeniusa[10]). Normą Frobeniusa nazywamy:

$$\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\text{tr}(\overline{A}^T A)},$$

gdzie  $A$  jest macierzą  $m \times n$  - wymiarową.  $\overline{A}^T$  nazywamy transpozycją sprzężenia macierzy.

**Twierdzenie 3.10** (Warunki równoważne SVD [10]). Niech rozkład według wartości osobliwych macierzy  $\mathbb{X}$  będzie dany wzorem

$$\mathbb{X} = \mathbb{U}\Sigma\mathbb{V}^T$$

gdzie  $\mathbb{U} = [u_1, u_2, \dots, u_m]$ ,  $\mathbb{V} = [v_1, v_2, \dots, v_n]$ ,  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$  oraz  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$ .  $R(\mathbb{X})$  i  $N(\mathbb{X})$  oznaczają zakres i jądro macierzy. Wtedy:

1. właściwości rzędu macierzy:  $\text{rz}(\mathbb{X}) = r$ ,  $N(\mathbb{X}) = \text{span}(v_{r+1}, \dots, v_n)$ ,  $R(\mathbb{X}) = \text{span}(u_1, u_2, \dots, u_r)$ ,
2.  $\mathbb{X} = \sum_{i=1}^r u_i \cdot \sigma_i \cdot v_i^T$ ,
3.  $\|\mathbb{X}\|_F^2 = \sigma_1^2 + \dots + \sigma_r^2$  i  $\|\mathbb{X}\|_2^2 = \sigma_1^2$ .

**Twierdzenie 3.11** (Twierdzenie Eckart - Younga [10]). Niech  $\mathbb{X}$  będzie macierzą  $m \times n$  - wymiarową z rozkładem według wartości osobliwych  $\mathbb{X} = \mathbb{U}\Sigma\mathbb{V}^T$ ,  $\text{rz}(\mathbb{X}) \in \mathbb{N}$  niech będzie rzędem macierzy i  $\text{rz}(\mathbb{X}) \leq p$ . Zdefiniujmy:

$$\mathbb{X}_k = \sum_{i=1}^k u_i \cdot \sigma_i \cdot v_i^T,$$

wtedy

$$\min_{\text{rz}(\mathbb{B})=k} \|\mathbb{X} - \mathbb{B}\|_F^2 = \|\mathbb{X} - \mathbb{X}_k\|_F^2 = \sigma_{k+1}^2 + \dots + \sigma_p^2.$$

Aby udowodnić twierdzenie Eckart - Younga wprowadźmy twierdzenie Weylsa:

**Twierdzenie 3.12** (Twierdzenie Weylsa [13]). *Niech  $\mathbb{X}, \mathbb{Y} \in \mathbb{M}_{m \times n}(\mathbb{K})$  oraz  $r = \min\{m, n\}$ . Dodatkowo niech odpowiednio  $\sigma_1(\mathbb{X}) \geq \sigma_2(\mathbb{X}) \geq \dots \geq \sigma_r(\mathbb{X}) \geq 0$ ,  $\sigma_1(\mathbb{Y}) \geq \sigma_2(\mathbb{Y}) \geq \dots \geq \sigma_r(\mathbb{Y}) \geq 0$  i  $\sigma_1(\mathbb{Z}) \geq \sigma_2(\mathbb{Z}) \geq \dots \geq \sigma_r(\mathbb{Z}) \geq 0$  będą wartościami osobliwymi macierzy  $\mathbb{X}, \mathbb{Y}, \mathbb{Z} = \mathbb{X} + \mathbb{Y}$ . Wtedy:*

$$\sigma_{i+j-1}(\mathbb{Z}) \leq \sigma_i(\mathbb{X}) + \sigma_j(\mathbb{Y}),$$

gdzie  $1 \leq i, j \leq r$ ,  $i + j \leq r + 1$ .

*Dowód.* Niech  $\mathbb{X} \in \mathbb{M}_{m \times n}(\mathbb{R})$  będzie macierzą o wartościach rzeczywistych, gdzie  $m \geq n$ . Załóżmy, że

$$\mathbb{X} = \mathbb{U}\Sigma\mathbb{V}^T$$

jest rozkładem według wartości osobliwych macierzy  $\mathbb{X}$ . Chcemy pokazać, że najlepszym przybliżeniem macierzy  $\mathbb{X}$  w normie Frobeniusa (oznaczamy  $\|\cdot\|_F$ ) jest

$$\mathbb{X}_k = \sum_{i=1}^k u_i \cdot \sigma_i \cdot v_i^T,$$

gdzie  $u_i$  i  $v_i$  oznaczają odpowiednio  $i$ -te kolumny macierzy  $\mathbb{U}$  i  $\mathbb{V}$ . Zauważmy, że z własności 3. twierdzenia o warunkach równoważnych SVD mamy:

$$\|\mathbb{X} - \mathbb{X}_k\|_F^2 = \left\| \sum_{i=k+1}^n u_i \cdot \sigma_i \cdot v_i^T \right\|_F^2 = \sum_{i=k+1}^n \sigma_i^2.$$

Stąd należy udowodnić, że  $\mathbb{B}_k = \mathbb{X}\mathbb{Y}^T$ , gdzie  $\mathbb{X}$  i  $\mathbb{Y}$  są macierzami oraz

$$\|\mathbb{X} - \mathbb{X}_k\|_F^2 = \sum_{i=k+1}^n \sigma_i^2 = \|\mathbb{X} - \mathbb{B}_k\|_F^2.$$

Niech  $\mathbb{X} = \mathbb{X}' + \mathbb{X}''$ . Korzystając z Twierdzenia Weylsa mamy:

$$\sigma_{i+j-1}(\mathbb{X}) \leq \sigma_i(\mathbb{X}') + \sigma_j(\mathbb{X}'').$$

Jeżeli

$$\sigma_{k+1}(\mathbb{B}_k) = 0,$$

kiedy  $\mathbb{X}' = \mathbb{X} - \mathbb{B}_k$  i  $\mathbb{X}'' = \mathbb{B}_k$  wnioskujemy, że dla  $i \geq 1, j = k + 1$

$$\sigma_i(\mathbb{X} - \mathbb{B}_k) \geq \sigma_{k+1}(\mathbb{X}).$$

Stąd:

$$\|\mathbb{X} - \mathbb{B}_k\|_F^2 = \sum_{i=1}^n \sigma_i(\mathbb{X} - \mathbb{B}_k)^2 \geq \sum_{k+1}^n \sigma_i(\mathbb{X})^2.$$

Z własności 3. twierdzenia o warunkach równoważnych SVD mamy:

$$\sum_{k+1}^n \sigma_i(\mathbb{X})^2 = \|\mathbb{X} - \mathbb{X}_k\|_F^2.$$

□

Zawsze jest możliwe dokonać dekompozycji macierzy  $\mathbb{X}$  do postaci  $\mathbb{X} = \mathbb{U}\Sigma\mathbb{V}^T$ .



## 3.2 Metody eksploracji danych

Termin eksploracja danych jest często używany jako określenie procesu odkrywania wiedzy z danych. Często jednak terminem "proces odkrywania wiedzy" określamy cały proces pracy z danymi, natomiast termin "eksploracja danych" odnosi się do etapu odkrywania pewnego rodzaju reguł.

Jak podaje Tadeusz Morzy [7] w metodach eksploracji można wyróżnić:

- odkrywanie asocjacji,
- klastrowanie,
- odkrywanie wzorców sekwencji,
- odkrywanie klasyfikacji,
- odkrywanie podobieństw w przebiegach czasowych,
- wykrywanie zmian i odchyłeń.

W tej sekcji zostaną przedstawione te metody, które stosowane są najczęściej w regułach rekomendujących.

### 3.2.1 Algorytm $k$ - najbliższych sąsiadów

Opis poniższego algorytmu oparty został na [1, Sec 2.3.1].

Algorytm  $k$ -najbliższych sąsiadów ( $k$ -NN) jest najczęściej używanym algorytmem klasyfikacji.

Przyporządkowanie nowych elementów zostaje przeprowadzone na podstawie porównania obserwacji z  $k$  najbardziej podobnymi jej obiektami ze zbioru treningowego. Podstawowa założenie algorytmu mówi, że jeżeli nowy rekord znajduje się w pewnym otoczeniu, to na podstawie  $k$  - najbliższych mu obserwacji zostanie przyporządkowana do niego klasa, której pojawienie się w rozważanym zbiorze jest najliczniejsze.

Niech  $q$  będzie punktem dla którego chcemy odnaleźć jego klasę  $l$ .

$X = \{\{x_1, l_1\}, \dots, \{x_n, l_n\}\}$  niech będzie zbiorem treningowym, gdzie  $x_j$  jest  $j$ -tym elementem zbioru, natomiast  $l_j$  etykietką klasy do której zbiór należy,  $j \in \{1, \dots, n\}$ .

Przeprowadzając algorytm  $k$ -NN zaczynamy od wyboru podzbioru

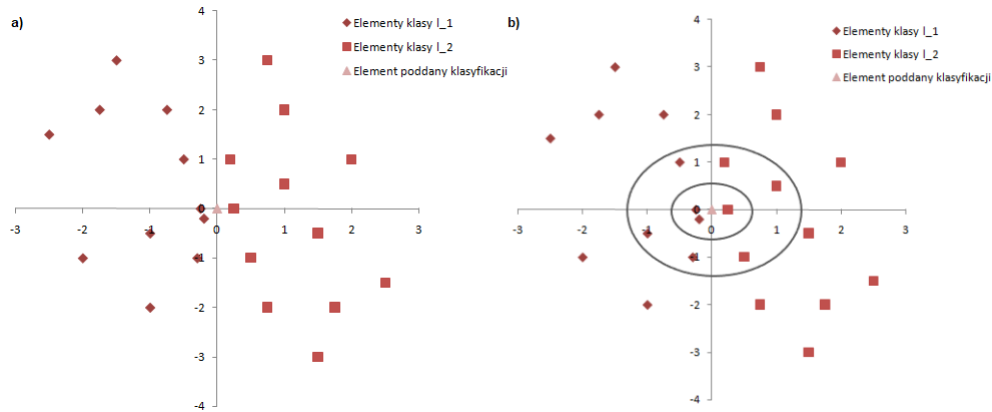
$$Y = \{\{y_1, l_1\}, \dots, \{y_k, l_k\}\},$$

$k \in \{1, \dots, n\}$  takiego, że  $Y \in X$  oraz

$$\sum_1^k d(q, y_k)$$

jest minimalna.  $Y$  zawiera więc  $k$  punktów z  $X$ , które leżą najbliżej rozważanego punktu  $q$ . Następnie do punktu  $q$  zostaje przyporządkowana klasa taka, że

$$l = f(\{l_1, \dots, l_k\}).$$



Rysunek 3.1: Metoda  $k$  - najbliższych sąsiadów.

Na powyższym rysunku widzimy przykładowe zastosowanie algorytmu  $k$ -NN. W części a) przedstawiony został zbiór treningowy z podziałem na dwie klasy (rąby, koła) oraz punkt, który będziemy chcieli przyporządkować do jednej z nich (trójkąt). W części b) przedstawiono dwa koła, jedno prezentujące najbliższe sąsiedztwo dla  $k = 3$ , drugie dla  $k = 9$ . W obu przypadkach nowy punkt (trójkąt) zostanie przyporządkowany do klasy  $l_1$ . Warto jednak zauważyć, że znajduje się on na granicy dwóch klastrów przez co przy innym wyborze  $k$  może zostać przyporządkowany do klasy  $l_2$ .

Opisana metoda przyporządkowuje wybranemu rekordowi najbardziej mu podobne wykorzystując miary odległości.

Najtrudniejszym zadaniem przy przeprowadzaniu algorytmu  $k$ -NN jest wybór  $k$ . Jeżeli  $k$  będzie zbyt małe - klasyfikator stanie się bardzo wrażliwy, jeżeli jednak  $k$  będzie zbyt duże sąsiedztwo może zawierać zbyt dużo punktów z innych klas. Rozważając przypadek z rysunku 3.1 łatwo zauważyć, że nawet mała zmiana w obserwacjach zbioru treningowego może doprowadzić do zmiany wyniku.

### 3.2.2 Algorytm k - średnich

Opis poniższego algorytmu oparty został na [9, Sec 2.3.1].

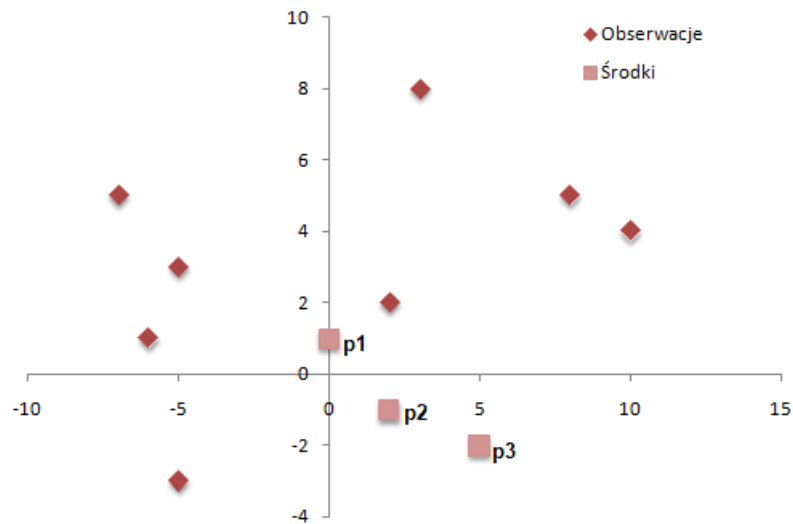
Algorytm k-średnich jest prostym i zarazem efektywnym algorytmem grupowania.

Głównym celem algorytmu jest podział pewnego zbioru  $X$ :

$$X = \{x_i = (x_{i1}, \dots, x_{id}) : i \in \{1, \dots, N\}\},$$

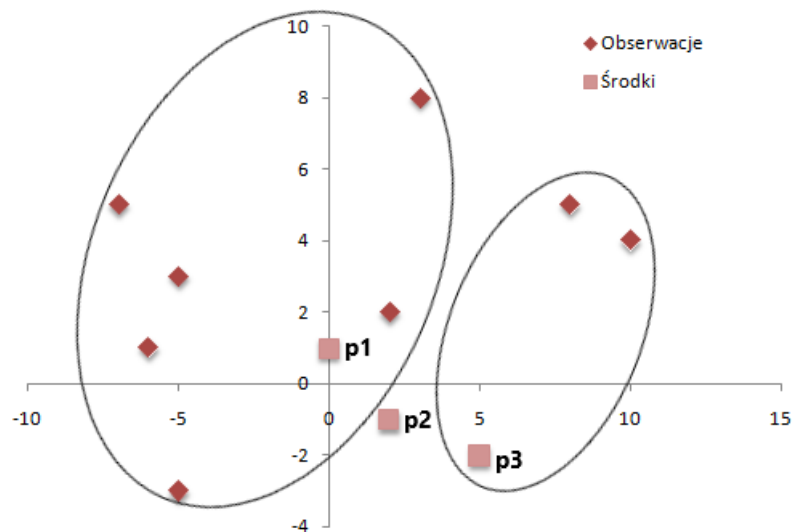
gdzie  $x_i$  jest  $d$  - wymiarowym wektorem cech opisującym obiekt na podzbiory.

W wyniku grupowania  $n$  - elementowego zbioru  $X$  na  $k$  podgrup jest macierz podziału  $\mathbb{A}$  o wymiarach  $k \times n$ . Każdy z elementów tej macierzy  $a_{ik}$  oznacza stopień w jakim wektor  $x_k$  przynależy do grupy. Na wstępie algorytmu ustalamy wartość parametru  $k$  jako liczbę grup, które zostaną wyodrębnione. Wybieramy  $k$  reprezentantów, które stanowią prototypy grup.



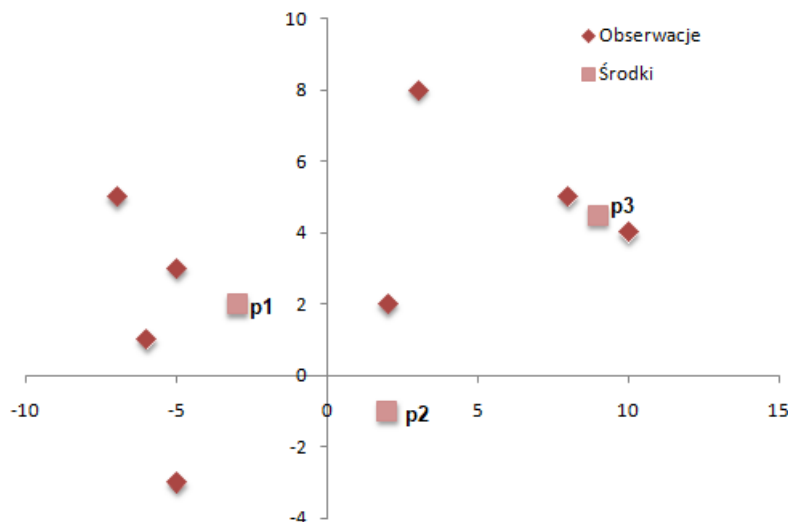
Rysunek 3.2: Metoda  $k$ -średnich. Wybór początkowych środków.

W powyższym przykładzie (rysunek 3.2) wybranymi środkami są punkty  $p1$ ,  $p2$ ,  $p3$ . Kolejnym krokiem jest przypisanie każdego z elementów do najbliższej mu grupy.



Rysunek 3.3: Metoda  $k$ -średnich. Przypisanie elementów do grup.

Dla każdej z tak ustalonych grup obliczamy średnią arytmetyczną współrzędnych, które staną się kolejnymi środkami.



Rysunek 3.4: Metoda  $k$ -średnich. Wybór nowych środków.

Kroki te są wykonywane do momentu występowania migracji między obiektami. W algorytmie  $k$ -średnich liczba grup pozostaje więc niezmienną, zmienna jest tylko przynależność do grup. W metodzie tej poszukiwanie optymalnego podziału odpowiada wyznaczaniu takich grup, które minimalizują następującą funkcję kryterialną:

$$J(u, \mathbb{A}) = \sum_{i=1}^k \sum_{k=1}^N a_{ki} d(p_i, x_k)^2,$$

gdzie

- $d(p_i, x_k)$  oznacza odległość elementu  $x_k$  od grupy wyznaczonej przez środek  $p_i$ ,
- $N$  to liczebność zbioru  $X$ ,
- $\mathbb{A}$  oznacza macierz podziału.

## 3.3 Szacowanie błędów obliczeń

### 3.3.1 Ocena dokładności metody

Najczęściej używanymi miarami dokładności modelu są:

- błąd średni (Mean Error),
- średni błąd bezwzględny (Mean Absolute Error),
- średni błąd kwadratowy (Mean Squared Error).

Niech dla elementu  $i$  ze zbioru testowego  $P$  będzie dostarczona predykcja  $\hat{r}_i$ . Aby ocenić jakość jej wyniku należy porównać ją ze znaną wartością  $r_i$ .

**Definicja 3.13** (Błąd średni [1, Sec 4.1.1]). *Średnim błędem nazywamy wartość wyrażenia:*

$$ME = \frac{1}{|P|} \sum_{i \in P} (\hat{r}_i - r_i).$$

**Definicja 3.14** (Średni błąd bezwzględny [1, Sec 4.1.1]). *Średnim błędem bezwzględnym nazywamy wartość wyrażenia:*

$$MAD = \frac{1}{|P|} \sum_{i \in P} |\hat{r}_i - r_i|.$$

**Definicja 3.15** (Średni błąd kwadratowy [1, Sec 4.1.1]). *Średnim błędem kwadratowym nazywamy wartość wyrażenia:*

$$MSE = \frac{1}{|P|} \sum_{i \in P} (\hat{r}_i - r_i)^2.$$

**Uwaga 3.16.** [1, Sec 4.1.1] Funkcja kwadratowa jest funkcją monotoniczną co pozwala na dość częste zastępowanie średniego błędu kwadratowego przez średnią kwadratową błędów (Root Mean Squared Error (RMSE)):

$$RMSE = \sqrt{MSE}$$

Normalized RMSE (NRMSE) oraz Normalized MAE (NMAE) są znormalizowanymi, przez użycie zakresu wartości  $r_{max} - r_{min}$ , wersjami błędów RMSE i MAE.

Kolejnym rodzajem powszechnie używanego błędu, który pozwala na użycie sum ważonych jest średni błąd RMSE (Average RMSE).

**Definicja 3.17** (Średni błąd RMSE [1, Sec 4.1.1]). *Niech  $w_i > 0$  będzie wagę dla elementu  $i$  oraz niech  $\sum w_i = 1$ .*

*Średnim błędem RMSE nazywamy wartość wyrażenia:*

$$ARMSE = \sqrt{\sum_{i \in P} w_i (\hat{r}_i - r_i)^2}.$$

### 3.3.2 Ocena jakości modelu

Ocenę jakości modelu przeprowadza się na zbiorze testowym. Dla każdego z rekordów jest znana jego etykieta. Rekordy te są poddawane działaniu modelu, a następnie etykiety przypisane rekordom przez model są porównywalne z rzeczywistymi wartościami etykiet.

W następnym kroku zliczana jest liczba rekordów poprawnie i niepoprawnie zaklasyfikowanych przez model, a wynik testu zostaje przedstawiony w postaci macierzy pomyłek.

**Definicja 3.18** (Macierz pomyłek [7, Sec 4.8.1]). *Macierz pomyłek nazywamy macierz kwadratową  $m \times m$  ( $m$  oznacza liczbę etykiet), gdzie wiersze reprezentują etykiety początkowe, natomiast kolumny etykiety przyporządkowane rekordom przez model. Element macierzy  $\mathbb{F}$  oznacza liczbę rekordów z etykietą  $E_i$ , którym błędnie została przypisana etykieta  $E_j$ .*

$\mathbb{F}$	$E_1$	$E_2$
$E_1$	$f_{11}$	$f_{12}$
$E_2$	$f_{21}$	$f_{22}$

Tabela 3.1: Macierz pomyłek.

**Uwaga 3.19.** [7, Sec 4.8.1] Często elementy macierzy pomyłek dla problemów klasyfikacji binarnej oznacza się symbolami :  $TP$ ,  $TN$ ,  $FN$ ,  $FP$ . Oznaczenia te symbolizują cztery możliwe przypadki występujące w klasyfikacji binarnej. Załóżmy, że wyróżniamy klasę pozytywną (+) i negatywną (-). Wtedy :

- $TP$  (ang. *true - positive*) - liczba pozytywnych rekordów testowych zaklasyfikowanych do klasy pozytywnej,
- $FN$  (ang. *false - negative*) - liczba pozytywnych rekordów testowych zaklasyfikowanych do klasy negatywnej,
- $FP$  (ang. *false - positive*) - liczba negatywnych rekordów testowych zaklasyfikowanych do klasy pozytywnej,
- $TN$  (ang. *true - negative*) - liczba negatywnych rekordów testowych zaklasyfikowanych do klasy negatywnej.

Macierz pomyłek przyjmuje wtedy postać:

$\mathbb{F}$	+	-
+	$TP$	$FN$
-	$FP$	$TN$

Tabela 3.2: Macierz pomyłek - przypadek klasyfikacji binarnej.

Poprzez analizę macierzy pomyłek bez problemu obliczymy łączną liczbę rekordów zaklasyfikowanych poprawnie oraz rekordów przypisanych błędnie przez klasyfikator.

Analizę zawartości macierzy można rozszerzyć o dodatkową informację - koszt błędnej klasyfikacji (ang. misclassification cost).

**Definicja 3.20** (Koszt błędnej klasyfikacji [7, Sec 4.8.1]). *Oznaczmy przez  $e_{ij}$  koszt błędnego zaklasyfikowania do klasy  $E_j$  rekordu, który w rzeczywistości należy do klasy  $E_i$ . Koszt poprawnej klasyfikacji oznaczmy przez  $e_{ii}$  oraz załóżmy, że  $\forall_i e_{ii} = 0$ . Dodatkowo niech  $f_t$  oznacza liczbę wszystkich przykładów testowych,  $f_p$  liczbę poprawnie zaklasyfikowanych rekordów testowych oraz  $f_p = \sum_{i=1}^m f_{ii}$ ,  $f_b$  niech natomiast oznacza liczbę błędnych klasyfikacji i  $f_b = f_t - f_p$ .*

*Kosztem błędnej klasyfikacji  $E(f_b)$  nazywamy sumę:*

$$E(f_b) = \sum_{i=1}^m \sum_{j=1}^m f_{ij} \cdot e_{ij}.$$

W przypadkach, gdy błędne zaklasyfikowania rekordów nie różnią się kosztami, do oceny jakości klasyfikatora można wykorzystać miary takie jak trafność klasyfikacji (ang. accuracy) oraz błąd klasyfikacji (ang. error rat).

**Definicja 3.21** (Trafność klasyfikacji [7, Sec 4.8.1]). *Trafnością klasyfikacji nazywamy stosunek liczby poprawnie zaklasyfikowanych rekordów testowych do łącznej liczby rekordów testowych:*

$$TR = \frac{f_p}{f_t} = \frac{\sum_{i=1}^m f_{ii}}{f_t}.$$

**Definicja 3.22** (Błąd klasyfikacji [7, Sec 4.8.1]). *Błędem klasyfikacji nazywamy stosunek liczby błędnie zaklasyfikowanych rekordów testowych do łącznej liczby rekordów testowych:*

$$BK = \frac{f_b}{f_t} = \frac{\sum_{i=1}^m \sum_{j=1}^m f_{ij}}{f_t} = 1 - \frac{f_p}{f_t}.$$

**Uwaga 3.23.** [7, Sec 4.8.1] *Innymi miarami, które można wywnioskować bezpośrednio z macierzy pomyłek dla klasyfikacji binarnej (tabla 3.2) są:*

- współczynnik  $TP$  (czułość):

$$WTP = \frac{TP}{TP + FN},$$

- współczynnik  $FP$ :

$$WFP = \frac{FP}{FP + TN},$$

- współczynnik  $TN$  (specyficzność):

$$WTN = \frac{TN}{FP + TN},$$

- precyzja:

$$precyzja = \frac{TP}{TP + FP},$$

- zwrot:

$$zwrot = \frac{TP}{TP + FN},$$

- $F$ -miara:

$$F - miara = \frac{2 \cdot precyzja \cdot zwrot}{precyzja + zwrot}.$$



# Rozdział 4

## Modele tworzenia rekomendacji

W niniejszym rozdziale zajmiemy się formalnym zdefiniowaniem zadania, które ukrywa się pod nazwą tworzenia rekomendacji. Do jego poprawnego określenia będą przydatne następujące pojęcia.

**Definicja 4.1** (Przedmiot [12, Sec 1.3]). *Przedmiotem nazwiemy klasę obiektów tego samego typu, nierozróżnialnych dla obserwatora i reprezentowanych przez co najmniej jeden element. W dalszej części pracy zbiór przedmiotów będziemy oznaczać przez  $P$ .*

Przedmioty stanowią podstawową grupę elementów w rozważaniach systemach rekomendujących.

**Definicja 4.2** (Użytkownik [12, Sec 1.3]). *Użytkownikiem nazywamy osobę zdolną do przedstawienia własnej oceny wybranego przedmiotu. W dalszej części pracy zbiór użytkowników będziemy oznaczać przez  $U$ .*

W pracy [12] użyty jest zawsze ten sam zbiór ocen, jednak łatwo możemy pokusić się o jego uogólnioną definicję.

**Definicja 4.3** (Zbiór ocen [12, Sec 1.3]). *Podzbiór skończony zbioru  $\mathbb{N}$  lub  $\mathbb{N} \cup \{0\}$  nazywamy zbiorem ocen dla przedmiotów. W dalszej części pracy zbiór ocen będziemy oznaczać przez  $O$ .*

**Definicja 4.4** (Macierz preferencji [12, Sec 1.3]). *Rozważmy zbiór przedmiotów o liczności  $n$  oraz grupę użytkowników o liczności  $m$ . Macierz preferencji  $M$  nazwiemy macierz o wymiarach  $n \times m$  i wartościami w ustalonym zbiorze ocen.*

Z uwagi na to, że przedmioty jako wytwory świata rzeczywistego są niemożliwe do opisanego za pomocą skończonej liczby cech rozważa się ich skończoną reprezentację nazywaną wektorem własności.

**Definicja 4.5** (Własność [12, Sec 1.3]). *Własnością nazwiemy cechę wyrażoną za pomocą wartości liczbowej lub pewnej zmiennej kategorycznej, która reprezentuje cechę przedmiotu istotną dla użytkownika w procesie tworzenia oceny. Zbiór wszystkich własności w rozważanym modelu oznaczamy  $W$ . Dla każdej  $w \in W$  poprzez  $V_w$  rozumiemy zbiór wszystkich dopuszczalnych wartości własności  $w$ .*

**Definicja 4.6** (Funkcja anotująca [12, Sec 1.3]). *Funkcją anotującą własność  $w \in W$  nazwiemy funkcję*

$$a_w: P \rightarrow V_w.$$

Mając na uwadze, że zbiór  $W$  jest skończony (jak również zbiór  $P$ ) można utożsamiać funkcję anotującą z wektorem o długości  $|W|$  nazywanym wektorem własności.

**Definicja 4.7** (Funkcja przynależności). *Funkcją przynależności nazywamy funkcję*

$$\mu_W(p_i): P \rightarrow [0, 1]$$

określa stopień przynależności przedmiotu  $p_i \in P$  do zbioru  $W = \{w_k\}$ .

**Definicja 4.8** (Problem tworzenia rekomendacji [12, Sec 1.3]). *Rozważmy pewien zbiór przedmiotów  $P$ , pewien zbiór użytkowników  $U$  oraz pewien zbiór ocen  $O$ . Niech ponadto  $R$  będzie funkcją taką, że:*

$$R: P \times U \rightarrow O.$$

Załóżmy, że dla funkcji  $R$  znane są wartości dla pewnych par przedmiotów i użytkowników. Naszym zadaniem jest zaproponowanie sposobu predykcji brakujących wartości funkcji  $R$  w sposób minimalizujący wybrany funkcjonal błędu.

Przyjrzyjmy się następującemu przykładowi, który ilustruje istotę problemu.

**Przykład 4.1.** *Niech zbiór przedmiotów będzie w tym przypadku zbiorem sześciu książek. Zatem  $P = \{p_1, p_2, p_3, p_4, p_5, p_6\}$ , gdzie  $p_i$  dla  $i \in \{1, 2, 3, 4, 5, 6\}$  oznacza  $i$  - tą książkę. Zbiór użytkowników niech będzie zbiorem czytelników. Zatem  $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$ , gdzie  $u_i$  dla  $i \in \{1, 2, 3, 4, 5, 6\}$  oznacza  $i$  - tego czytelnika. Poniższa tabela to macierz preferencji dla ustalonego zbioru  $P$  i ustalonego zbioru  $U$ . Znak '?' oznacza brakujące wartości funkcji  $R$ , zatem czytelnik danej książki nie czytał lub czytał lecz jego ocena jest nieznana.*

<i>Czytelnicy</i>		$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$
<i>Książki</i>	$p_1$	6	3	?	6	4	?
	$p_2$	?	6	6	5	6	?
	$p_3$	7	7	8	7	8	9
	$p_4$	8	10	10	7	6	8
	$p_5$	9	6	6	6	6	?
	$p_6$	5	7	7	5	4	2

*Zadaniem jest przewidzieć brakujące wartości funkcji  $R$ , czyli oceny nadane przez użytkowników w sposób minimalizujący błąd popełniany przez model.*

**Uwaga 4.9** (Podział systemów rekomendujących). *Podziału systemów rekomendujących dokonujemy ze względu na zakres wykorzystywanych informacji. Wyróżniamy:*

- *systemy rekomendujące oparte na treści,*
- *filtrowanie kolaboratywne,*
- *systemy rekomendujące kontekstowe.*

*W przypadku systemów rekomendujących opartych na treści predykcja jest dokonywana na podstawie ocen wystawionych przedmiotom przez użytkowników oraz wektorów własności rozważanych przedmiotów.*

*W filtrowaniu kolaboratywnym wektory cech zostają pominięte, a predykcja dokonywana jest na podstawie ocen. Wyróżniamy dwa typy filtrowania kolaboratywnego:*

- *filtrowanie oparte na użytkownikach, gdzie zakładamy, że jeżeli użytkownicy  $u_1$  i  $u_2$  wykazują podobieństwo oraz użytkownik  $u_1$  oceni pewien przedmiot, którego użytkownik  $u_2$  jeszcze nie ocenił, to prawdopodobnie ocena użytkownika  $u_2$  będzie podobna do oceny użytkownika  $u_1$ ,*
- *filtrowanie oparte na elementach, gdzie zakładamy, że jeżeli użytkownik ocenił w pewien sposób przedmiot  $p_1$  w przeszłości oraz przedmiot  $p_2$  jest podobna do  $p_1$ , to użytkownik będzie skłonny w podobny sposób ocenić przedmiot  $p_2$ .*

*Systemy rekomendujące kontekstowe są natomiast systemami rekomendującymi opartymi na treści w których zostaje uwzględniony dodatkowy wymiar - kontekst.*

## **4.1 Systemy rekomendujące oparte na treści - Content-based recommender systems**

Systemy oparte na treści wyróżnia ukierunkowanie na spersonalizowany poziom użytkownika oraz treść produktu. Metoda ta opiera się na obliczaniu podobieństw oraz wykorzystuje techniki uczenia maszynowego, takie jak klasyfikacja.

**Algorytm 4.10.** *W metodzie celem stworzenia rekomendacji i wygenerowania listy przedmiotów, które mogą być odpowiednie użytkownikowi opieramy się na treści rozważanych elementów. Algorytm tego rodzaju rekomendacji możemy przedstawić w następujących krokach:*

1. stworzenie wektora własności  $\mathbf{w} = [w_1, \dots, w_n]$ ,  $n \in \mathbb{N}$ , gdzie  $\forall_{i \in \{1, \dots, n\}} w_i \in W$ ,
2. wygenerowanie profili produktów - stworzenie wektorów własności  $\mathbf{w}_{p_i}$  gdzie poszczególne elementy wektora określają przynależność przedmiotu  $p_i$ ,  $i \in \mathbb{N}$  do odpowiednich elementów wektora własności  $\mathbf{w}$  określonego w kroku 1.,
3. wygenerowanie profili użytkowników - stworzenie wektorów własności  $\mathbf{w}_{u_j}$  przypisanych użytkownikom, gdzie poszczególne elementy wektora określają przynależność opinii użytkownika  $u_j$ ,  $j \in \mathbb{N}$  do elementów wektora własności  $\mathbf{w}$  określonego w kroku 1.,
4. obliczymy ocenę  $\hat{o}_{j,i}$  jaką użytkownik  $j$  wystawiłby dla przedmiotu  $i$ , którego wcześniej nie oceniał za pomocą funkcji

$$\hat{o}_{j,i} = \mathbf{w}_{u_j}^T \mathbf{w}_{p_i},$$

5. porównując otrzymane w kroku 3. oceny dokonujemy rekomendacji nowego przedmiotu.

**Przykład 4.2.** Niech wektor własności będzie określony następująco  $\mathbf{w} = [w_1, w_2]$ , a każdy z elementów wektora  $\mathbf{w}$  niech reprezentuje inny gatunek. Poniższa tabela określa przynależność (w przedziale  $[0, 1]$ ) każdej z książek do elementów wektora własności  $w$ .

<i>Gatunki</i>		$w_1$	$w_2$
<i>Książki</i>	$p_1$	0.9	0
	$p_2$	1	0.01
	$p_3$	0.99	0
	$p_4$	0.1	1
	$p_5$	0	0.9
	$p_6$	0.8	0.3

Dodatkowo zakładamy, że istnieje gatunek  $w_0$ , którego cechy reprezentują wszystkie książki oraz dla każdej z książek  $w_0 = 1$ .

Zatem wektor własności odpowiadające poszczególnym książkom mają postać

$$\mathbf{w}_{p_1} = [1, 0.9, 0]^T, \mathbf{w}_{p_2} = [1, 1, 0.01]^T, \mathbf{w}_{p_3} = [1, 0.99, 0]^T,$$

$$\mathbf{w}_{p_4} = [1, 0.1, 1]^T, \mathbf{w}_{p_5} = [1, 0, 0.9]^T, \mathbf{w}_{p_6} = [1, 0.8, 0.3]^T.$$

Dla każdego użytkownika  $j$  wyznaczamy wektor parametrów  $\mathbf{w}_{u_j} \in \mathbb{R}^3$ , który przedstawia przynależność (w przedziale  $[0, 1]$ ) opinii użytkownika do elementów wektora własności.

Preferencje czytelników zostaną więc opisane za pomocą wektorów:

$$\mathbf{w}_{u_1}, \mathbf{w}_{u_2}, \mathbf{w}_{u_3}, \mathbf{w}_{u_4}, \mathbf{w}_{u_5}, \mathbf{w}_{u_6}.$$

Obliczmy ocenę jaką książce  $p_3$  wystawiłby użytkownik  $u_1$  przy założeniu, że wektor preferencji użytkownika  $u_1$  jest postaci  $\mathbf{w}_{u_1} = [0, 5, 0]^T$ . Użytkownik ten preferuje więc książki gatunku  $w_1$ , gdy książki gatunków  $w_0$  i  $w_2$  są dla niego nieatrakcyjne. Zatem:

$$\hat{o}_{1,3} = \mathbf{w}_{u_1}^T \mathbf{w}_{p_3} = [0, 5, 0] \cdot [1, 0.99, 0]^T = 0 \cdot 1 + 5 \cdot 0.99 + 0 \cdot 0 = 4.95.$$

Przewidywaną oceną jest zatem 4.95.

Po przeprowadzeniu podobnych obliczeń dla wszystkich wcześniej nieznanymi ocen możemy zarekomendować naszemu użytkownikowi nową lekturę.

#### 4.1.1 Wygenerowanie profilu dokumentu tekstowego - algorytm TFIDF

W większość systemów rekomendacji opartych na treści używamy gotowych modeli wyszukiwujących.

W przypadku rozważań przeprowadzanych na dokumentach tekstowych jednym z najbardziej popularnych jest model przestrzeni wektorowej (*ang. Vector Space Model*) z algorytmem TFIDF.

Niech  $P = \{p_1, p_2, \dots, p_n\}, n \in \mathbb{N}$  będzie zestawem analizowanych przedmiotów.  $W = \{w_1, w_2, \dots, w_n\}, n \in \mathbb{N}$  niech będzie zbiorem rozważanych własności.

**Definicja 4.11** (Model przestrzeni wektorowej [1, Sec 3.3.1.1]). *Modelem przestrzeni wektorowej nazywamy formę reprezentacji przedmiotów, w której przedmiot  $p_i$  jest reprezentowany przez wektor z przestrzeni  $n$ -wymiarowej, a każdy z  $n$  wymiarów reprezentuje jedną z rozważanych wartości przedmiotu.*

**Definicja 4.12** (Liczność [1, Sec 3.3.1.1]). *Licznością  $f_{k,j}$  nazywamy liczbę wystąpień własności  $w_k$  w przedmiocie  $p_j$ .*

**Definicja 4.13** (TF [1, Sec 3.3.1.1]). *TF (*ang. term frequency*) nazywamy funkcję przedstawiającą zależność wartości  $w_k$  od przedmiotu  $p_j$ :*

$$TF(w_k, p_j) = \frac{f_{k,j}}{\max_z f_{z,j}},$$

gdzie:

- $\max_z f_{z,i}$  - maksymalna w odniesieniu do wszystkich wartości  $w_z \in W, z \in \{1, \dots, n\}$ , które pojawiły się w przedmiocie  $p_i$ , licznosc wystąpień własności.

**Definicja 4.14** (IDF [1, Sec 3.3.1.1]). *IDF (*ang. inverse document frequency*) nazywamy funkcję:*

$$IDF(w_k) = \log \frac{N}{n_k},$$

gdzie:

- $N$  - całkowita liczba przedmiotów w zbiorze  $P$ ,
- $n_k$  - liczba przedmiotów w których własność  $w_k$ ,  $k \in \{1, \dots, n\}$  wystąpiła przynajmniej raz.

**Definicja 4.15** (TFIDF [1, Sec 3.3.1.1]). *TFIDF* (ang. *TF* – term frequency, *IDF* – inverse document frequency) nazywamy funkcję:

$$TFIDF(w_k, p_i) = TF(w_k, p_i) \cdot IDF(w_k).$$

**Definicja 4.16** (Waga własności w przedmiocie). Wagę własności  $w_k$  w przedmiocie  $p_i$  nazywamy wartością:

$$w_{k,i} = \frac{TFIDF(w_k, p_i)}{\sqrt{\sum_{j=1}^{|W|} TFIDF(w_j, p_i)^2}}.$$

**Uwaga 4.17.** Każdy z dokumentów  $p_i$ ,  $i \in \{1, \dots, n\}$  przedstawiamy jako wektor liczb rzeczywistych w przestrzeni  $n$ -wymiarowej. Zatem  $p_i = [w_{1i}, w_{2i}, \dots, w_{ni}]$ , gdzie  $w_{ki}$ ,  $k \in \{1, \dots, n\}$  jest wagą własności  $w_k$  w przedmiocie  $p_i$ .

**Uwaga 4.18.** Dla danych w formie dokumentów tekstowych własnościami, które charakteryzują temat dokumentu są słowa. Dla każdego ze słów zostaje obliczona wartość funkcji *TFIDF*, a otrzymany następnie wektor wag charakteryzuje rozważany dokument.

## 4.2 Filtrowanie kolaboratywne - Collaborative filtering

Podejście kolaboratywne omija niektóre ograniczenia występujące w metodach kontekstowych. Dzięki temu systemowi możemy dokonywać rekomendacji na przestrzeni kilku kontekstów, co pozwala uniknąć nam konkretne wyspecjalizowanej przestrzeni obecnej w metodach kontekstowych.

### 4.2.1 Filtrowanie kolaboratywne oparte na użytkowniku

**Algorytm 4.19.** Stworzenie rekomendacji filtrowania kolaboratywnego opartej na użytkownikach wykonamy w następujących krokach:

1. wybór użytkowników  $u_j, u_k \in U$ ,  $j, k \in \mathbb{N}$ , między którymi chcemy obliczyć podobieństwo,
2. wybór przedmiotów  $p_i \in P$ ,  $i \in \mathbb{N}$ , dla których znane wartości funkcji  $R(p_i, u_j)$  i  $R(p_i, u_k)$ ,

3. stworzenie wektorów ocen  $o_{j,k}^{(j)}$  i  $o_{j,k}^{(k)}$  dla użytkowników  $u_j$  i  $u_k$  wybranych w kroku 1., których elementy stanowią wartości  $R(p_i, u_j)$  oraz  $R(p_i, u_k)$ , gdzie  $p_i$  to przedmioty wybrane w kroku 2.,
4. wyznaczenie odległości między czytelnikami  $u_j$  i  $u_k$  - najczęstszymi stosowanymi podejściami do obliczania odległości są metryka euklidesowa i współczynnik korelacji Pearsona,
5. wyznaczenie macierzy odległości  $\mathbb{U}_1$  między wszystkimi czytelnikami ze zbioru  $U$ ,
6. wyznaczenie macierzy odległości  $\mathbb{U}_2$  między czytelnikami poprzez normalizację danych w celu uzyskania wartości z przedziału  $[0, 1]$ , wyrazy macierzy przyjmują wartości:

$$u_{ij}^{(2)} = \frac{u_{ij}^{(1)}}{\max_{o_i} \{o_i : o_i \in O, i \in \mathbb{N}\} - \min_{o_i} \{o_i : o_i \in O, i \in \mathbb{N}\}},$$

gdzie  $u_{ij}^{(1)}$  i  $u_{ij}^{(2)}$  są odpowiednio elementami macierzy  $\mathbb{U}_1$  i  $\mathbb{U}_2$ ,  $i, j \in \mathbb{N}$ ,

7. wyznaczenie macierzy podobieństwa  $\mathbb{U}_3$  między użytkownikami - zakładając, że największa wartość prawdopodobieństwa to 1 macierz podobieństwa przyjmuje wartości:

$$u_{ij}^{(3)} = 1 - u_{ij}^{(2)},$$

gdzie  $u_{ij}^{(2)}$  i  $u_{ij}^{(3)}$  są odpowiednio elementami macierzy  $\mathbb{U}_2$  i  $\mathbb{U}_3$ ,

8. wyestymowanie nieznanych wartości funkcji  $R$  dla  $u_j \in U$ ,  $j \in \mathbb{N}$  oraz  $p_i \in P$ ,  $i \in \mathbb{N}$  - niech  $u_j$  będzie konkretnie ustalonym użytkownikiem, w celu obliczenia brakujących wartości funkcji  $R$  dla użytkownika  $u_j$  obliczmy średnią ważoną wykorzystując oceny i przyjmując wartości podobieństwa między  $u_j$  i innymi użytkownikami jako wagi.

W celu dokładniejszego zrozumienia rozważmy ponownie przykład 4.1.

**Przykład 4.3.** Chcąc obliczyć podobieństwo między użytkownikiem  $u_2$  i  $u_3$  wybierzmy książki, które zostały przeczytane przez obu użytkowników. W tym przypadku są to:  $p_2$ ,  $p_3$ ,  $p_4$ ,  $p_5$ ,  $p_6$ . Wektorami ocen uwzględniającymi książki ocenione przez obu użytkowników są więc odpowiednio dla użytkownika  $u_2$  wektor  $o_{2,3}^{(2)} = [6, 7, 10, 6, 7]^T$  oraz dla użytkownika  $u_3$  wektor  $o_{2,3}^{(3)} = [6, 8, 10, 6, 7]^T$ .

Posługując się odległością euklidesową obliczamy odległość między użytkownikami  $u_2$  i  $u_3$ :

$$d_e(o_{2,3}^{(2)}, o_{2,3}^{(3)}) = \sqrt{(6-6)^2 + (7-8)^2 + (10-10)^2 + (6-6)^2 + (7-7)^2} = \sqrt{1} = 1.$$

Postępując w podobny sposób dla każdej z par użytkowników otrzymamy następującą macierz odległości  $\mathbb{U}_1$ :

$\mathbb{U}_1$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
$\mathbf{u}_1$	0	5,099	4,243	3	4,359	3,606
$\mathbf{u}_2$	5,099	0	1	4,796	5,196	5,745
$\mathbf{u}_3$	4,243	1	0	3,873	5	5,477
$\mathbf{u}_4$	3	4,796	3,873	0	2,828	3,742
$\mathbf{u}_5$	4,359	5,196	5	2,828	0	3
$\mathbf{u}_6$	3,606	5,745	5,477	3,742	3	0

W procesie normalizacji danych dzielimy elementy macierzy przez

$$(\max\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} - \min\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}) = 10$$

i otrzymujemy macierz  $\mathbb{U}_2$  postaci:

$\mathbb{U}_2$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
$\mathbf{u}_1$	0	0,5099	0,4243	0,3	0,4359	0,3606
$\mathbf{u}_2$	0,5099	0	0,1	0,4796	0,5196	0,5745
$\mathbf{u}_3$	0,4243	0,1	0	0,3873	0,5	0,5477
$\mathbf{u}_4$	0,3	0,4796	0,3873	0	0,2828	0,3742
$\mathbf{u}_5$	0,4359	0,5196	0,5	0,2828	0	0,3
$\mathbf{u}_6$	0,3606	0,5745	0,5477	0,3742	0,3	0

Macierz podobieństwa  $\mathbb{U}_3$  przyjmuje więc wartości:

$\mathbb{U}_3$	$\mathbf{u}_1$	$\mathbf{u}_2$	$\mathbf{u}_3$	$\mathbf{u}_4$	$\mathbf{u}_5$	$\mathbf{u}_6$
$\mathbf{u}_1$	1	0,4901	0,5757	0,7	0,5641	0,6394
$\mathbf{u}_2$	0,4901	1	0,9	0,5204	0,4804	0,4255
$\mathbf{u}_3$	0,5757	0,9	1	0,6127	0,5	0,4523
$\mathbf{u}_4$	0,7	0,5204	0,6127	1	0,7172	0,6258
$\mathbf{u}_5$	0,5641	0,4804	0,5	0,7172	1	0,7
$\mathbf{u}_6$	0,6394	0,4255	0,4523	0,6258	0,7	1

Obliczamy ocenę jaką użytkownik  $u_1$  zaproponuje dla książki  $p_2$ :

$$\frac{0,4901 \cdot 6 + 0,5757 \cdot 6 + 0,7 \cdot 5 + 0,5641 \cdot 6}{0,4901 + 0,5757 + 0,7 + 0,5641} = 5,7$$

Na podstawie metody filtrowania kolaboratywnego opartej na użytkownikach wnioskujemy, że użytkownik  $u_1$  wystawiłby książce  $p_2$  ocenę 5,7.

Postępując w analogiczny sposób przewidzimy wszystkie brakujące oceny :



<i>Czytelnicy</i>		$u_1$	$u_2$	$u_3$	$u_4$	$u_5$	$u_6$
<i>Książki</i>	$s_1$	6	3	4,57	6	4	4,88
	$s_2$	5,7	6	6	5	6	5,72
	$s_3$	7	7	8	7	8	9
	$s_4$	8	10	10	7	6	8
	$s_5$	9	6	6	6	6	6,67
	$s_6$	5	7	7	5	4	2

Możemy więc wnioskować, że w tym przypadku dla użytkownika  $u_3$  książka  $p_1$  prawdopodobnie nie będzie zbyt atrakcyjna. Użytkownik  $u_6$ , natomiast, z chęcią przeczyta książkę  $p_5$ .

#### 4.2.2 Filtrowanie kolaboratywne oparte na elementach

W przypadku filtrowania kolaboratywnego opartego na elementach wartości podobieństwa między użytkownikami zostaje zastąpiona przez wartości podobieństwa między elementami.

**Algorytm 4.20.** W tym rodzaju rekomendacji należy wykonać następujące kroki:

1. wybór przedmiotów  $p_i, p_k \in P$ ,  $i, k \in \mathbb{N}$ , dla których znamy wartość funkcji  $R(p_i, u_j)$  oraz wartość funkcji  $R(p_k, u_j)$ , gdzie  $u_j$  jest użytkownikiem, który wystawia ocenę w obu przypadkach,
2. stworzenie wektorów ocen  $\overline{o^{(p_i)}}$  i  $\overline{o^{(p_k)}}$  dla przedmiotów  $p_i$  i  $p_k$  wybranych w kroku 2., których elementy stanowią wartości funkcji  $R$  dla użytkownika  $u_j$ ,
3. wyznaczenie odległości między przedmiotami  $p_i$  i  $p_k$  - najczęstszymi stosowanymi podejściami do obliczania odległości jest podobieństwo kosinusów,
4. wyznaczenie macierzy podobieństwa  $\mathbb{P}$  między wszystkimi przedmiotami  $P$ ,
5. wyestymowanie nieznanych wartości funkcji  $R$  dla  $u_j \in U$ ,  $j \in \mathbb{N}$  oraz  $p_i \in P$ ,  $i \in \mathbb{N}$  - niech  $p_i$  będzie konkretnie ustalonym przedmiotem oraz  $u_j$  będzie konkretnie ustalonym użytkownikiem, w celu obliczenia brakujących wartości funkcji  $R$  dla  $u_j$  i  $p_i$  obliczymy średnią ważoną wykorzystując oceny oraz przyjmując wartości podobieństwa między  $p_i$  i innymi przedmiotami ocenionymi przez użytkownika jako wagi.

**Przykład 4.4.** Aby obliczyć podobieństwo między książkami  $p_1$  i  $p_2$  wyznaczmy wektory ocen w których uwzględnimy przypadki, gdzie jeden użytkownik ocenił obie pozycje.

Zatem:  $\overline{o^{(p_1)}} = [3, 6, 4]^T$ ,  $\overline{o^{(p_2)}} = [6, 5, 6]^T$ .

Następnie używając wzoru na podobieństwo kosinusowe obliczamy podobieństwo między wybranymi książkami

$$\text{sim}(p_1, p_2) = \frac{\overline{o(p_1)} \cdot \overline{o(p_2)}}{|\overline{o(p_1)}| |\overline{o(p_2)}|} = \frac{3 \cdot 6 + 6 \cdot 5 + 4 \cdot 6}{\sqrt{6^2 + 3^2 + 6^2 + 4^2} \sqrt{6^2 + 6^2 + 5^2 + 6^2}} = 0,6339.$$

Postępując w analogiczny sposób otrzymamy macierz podobieństwa:

$\mathbb{P}$	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	$\mathbf{p}_4$	$\mathbf{p}_5$	$\mathbf{p}_6$
$\mathbf{p}_1$	1	0,6339	0,7372	0,7195	0,8935	0,7599
$\mathbf{p}_2$	0,6339	1	0,7951	0,8150	0,7977	0,8898
$\mathbf{p}_3$	0,7372	0,7951	1	0,9780	0,8586	0,9200
$\mathbf{p}_4$	0,7195	0,8150	0,9780	1	0,8860	0,9681
$\mathbf{p}_5$	0,8935	0,7977	0,8586	0,8860	1	0,9413
$\mathbf{p}_6$	0,7599	0,8898	0,9200	0,9681	0,9413	1

Wystymujemy teraz ocenę jaką użytkownik  $u_6$  zaproponuje dla książki  $p_2$ . Ponownie obliczymy średnią ważoną ocen, tym razem, wykorzystując wartość podobieństwa między książką  $p_1$ , a książkami ocenionymi wcześniej przez użytkownika oraz oceny jakie nadał on tym pozycjom:

$$\frac{(0,7951 \cdot 9 + 0,8150 \cdot 8 + 0,8898 \cdot 2)}{(0,7951 + 0,8150 + 0,8898)} = 6,16.$$

Na podstawie przeprowadzonych obliczeń zakładamy, że ocena jaką wystawiłby po przeczytaniu użytkownik  $u_6$  książce  $p_2$  to 6,16.

Powtarzając powyższe obliczenia dla każdej z pozycji wcześniej nieocenionej przez wybranego użytkownika otrzymamy wszystkie brakujące opinie. Następnie bazując na zdobytych danych z łatwością odnajdziemy pozycję najbardziej odpowiednią do zarekomendowania użytkownikowi.

### 4.3 Systemy rekomendujące kontekstowe - Context-aware recommender systems

Uprzednio opisane metody opierały się głównie na rozważaniu problemów dwu-wymiarowych. W tym podejściu, przez dodanie nowego wymiaru, jakim jest kontekst ( $K$ ), zaczynamy rozważać problemy trój-wymiarowe:

$$R : U \times P \times K \rightarrow O$$

**Definicja 4.21** (Kontekst [1, Sec 3.3.1.1]). *Kontekstem nazywamy wektor preferencji.*

**Algorytm 4.22.** W modelu kontekstowym rekomendacje są generowane w następujący sposób:

1. za pomocą algorytmu systemów rekomendujących opartych na treści zostają wygenerowana lista rekomendacji bazująca na preferencjach użytkownika,
2. odfiltrowanie rekomendacji, które odpowiadają przyjętemu kontekstowi - wyróżniamy tutaj dwa podejścia:
  - filtrowanie jako etap wstępny (ang. *Pre-Filtering*) - informacje kontekstowe są tu używane do odfiltrowania najbardziej istotnych informacji i skonstruowania dwuwymiarowego zbioru danych,
  - filtrowanie jako etap końcowy (ang. *Post-Filtering*) - informacje o kontekście są ignorowane w wejściowych danych, rekomendacja dokonywana jest na całym zbiorze, a w następnym kroku lista rekomendacji stworzona dla użytkownika jest zawężana przez uwzględnienie kontekstu.

## 4.4 Dekompozycja macierzy ocen - SVD

Poniższe rozważania zostały przeprowadzone na podstawie publikacji Desrosiers K. i Karypis G. *A comprehensive survey of neighborhood-based recommendation methods* [14, Sec 4.1.1].

Popularnym podejściem redukcji wymiaru w regułach rekomendujących jest Ukryte Indeksowanie Semantyczne (ang. *Latent Semantic Indexing (LSI)*). LSI to matematyczna metoda opracowana w celu dokładniejszego wyszukiwania informacji.

W podejściu tym macierz ocen  $\mathbb{O}$  o wymiarach  $|U| \times |P|$  i  $\text{rz}(\mathbb{O}) = n$  jest aproksymowana przez macierz  $\hat{\mathbb{O}}$  taką, że  $\text{rz}(\hat{\mathbb{O}}) = k$ ,  $k < n$ .

Zatem

$$\hat{\mathbb{O}} = \mathbb{P}\mathbb{Q}^T,$$

gdzie:

- $\mathbb{P}$  jest  $(|U| \times k)$ -wymiarową macierzą zawierającą koordynaty użytkowników,
- $\mathbb{Q}$  jest  $(|P| \times k)$ -wymiarową macierzą zawierającą koordynaty przedmiotów.

Intuicyjnie,  $u$ -ty rząd macierzy  $\mathbb{P}$ ,  $\mathbf{p}_u \in \mathbb{R}^k$ , reprezentuje współrzędne użytkownika  $u$  rzutowane w  $k$ -wymiarowej przestrzeni. Podobnie  $i$ -ty wiersz macierzy  $\mathbb{Q}$ ,  $\mathbf{q}_i \in \mathbb{R}^k$ , reprezentuje współrzędne przedmiotu  $i$  w tej przestrzeni.

Macierze  $\mathbb{P}$  i  $\mathbb{Q}$  są odnajdowane przez minimalizowanie błędu przybliżenia zdefiniowanego przez kwadrat normy Frobeniusa:

$$E(\mathbb{P}, \mathbb{Q}) = \|\mathbb{O} - \mathbb{P}\mathbb{Q}^T\|_F^2 = \sum_{u,i} (o_{u,i} - \mathbf{p}_u \mathbf{q}_i^T)^2.$$

Minimalizacja tego błędu jest równoważna wyprowadzeniu SVD macierzy  $\mathbb{O}$ :

$$\mathbb{O} = \mathbb{U}\Sigma\mathbb{V}^T,$$

gdzie:

- $\mathbb{U}$  jest  $(|U| \times n)$ -wymiarową lewą macierzą wektorów szczególnych,
- $\mathbb{V}$  jest  $(|P| \times n)$ -wymiarową prawą macierzą wektorów szczególnych,
- $\Sigma$  jest  $(n \times n)$ -wymiarową macierzą wartości osobliwych.

Przez  $\Sigma_k, \mathbb{U}_k, \mathbb{V}_k$  oznaczmy macierze uzyskane w wyniku wyboru  $k$  największych wartości osobliwych oraz ich odpowiednich wektorów.

Macierze  $\mathbb{P}$  i  $\mathbb{Q}$  odpowiadają więc postaciom:

$$\mathbb{P} = \mathbb{U}_k \Sigma_k^{\frac{1}{2}}, \text{ oraz } \mathbb{Q} = \mathbb{V}_k \Sigma_k^{\frac{1}{2}}.$$

Predykcja oceny zostaje dokonana na podstawie równania:

$$o_{u,i} = \mathbf{p}_u \mathbf{q}_i^T.$$

Główny problem jaki pojawia się przy implementacji metody SVD jest bark dużej liczby wyrazów macierzy  $\mathbb{O}$ . Rozwiązaniem tego problemu jest odnalezienie brakujących elementów macierzy  $\mathbb{P}$  i  $\mathbb{Q}$  używając znanych ocen oraz równania:

$$E(\mathbb{P}, \mathbb{Q}) = \sum_{o_{u,i} \in \mathbb{O}} (o_{u,i} - \mathbf{p}_u \mathbf{q}_i^T)^2 + \lambda(\|\mathbf{p}_u\|^2 + \|\mathbf{q}_i\|^2),$$

gdzie  $\lambda$  jest parametrem kontroli poziomu regularyzacji.

To samo podejście może zostać zastosowane w przypadku obliczania podobieństwa między użytkownikami lub przedmiotami w metodzie filtrowania opartego na treści.

Rozwiązujemy tutaj następujący problem:

$$E(\mathbb{P}, \mathbb{Q}) = \sum_{z_{u,i} \in \mathbb{O}} (o_{u,i} - \mathbf{p}_u \mathbf{q}_i^T)^2,$$

gdzie:

- $\forall_{u \in U} \|\mathbf{p}_u\| = 1,$

- $\forall_{i \in P} \|\mathbf{q}_i\| = 1$ ,
- $z_{ui}$  jest średnią ocen  $o_{ui}$  znormalizowaną do zakresu  $[-1, 1]$ .

Powyższy problem odnosi się do znalezienia koordynat dla każdego użytkownika  $u$  i każdego przedmiotu  $i$  w przestrzeni  $k$ -wymiarowej. Zakładamy, że użytkownik  $u$  nada wysoką ocenę przedmiotowi  $i$  jeżeli ich koordynaty są blisko siebie.

Jeżeli dwaj użytkownicy  $u$  i  $v$  są blisko siebie w przestrzeni, to nadadzą oni podobne oceny dla przedmiotów. Podobieństwo między tymi użytkownikami można obliczyć za pomocą równania:

$$w_{uv} = \mathbf{p}_u \mathbf{p}_v^T.$$

Podobnie, podobieństwo między przedmiotami obliczymy następująco:

$$w_{ij} = \mathbf{q}_i \mathbf{q}_j^T.$$

# Rozdział 5

## Eksperymenty / część praktyczne

Rozważany problem rekomendacji może być sformułowany jako problem uczenia maszynowego w którym znane są oceny jakie użytkownicy wystawili pewnym przedmiotom i którego zadaniem jest predykcja ocen użytkowników dla elementów przez nich nieocenionych.

Założmy, że mamy  $n$  użytkowników i  $m$  przedmiotów. Otrzymujemy  $n \times m$  - wymiarową macierz  $\mathbb{O}$ , w której wyrazy  $o_{u_i, p_j}$  są wartościami funkcji  $R(u_i, p_j)$  wystawioną przez użytkownika  $u_i$  elementowi  $p_j$ ,  $j \in \{1, \dots, m\}$ ,  $i \in \{1, \dots, n\}$ . Naszym celem jest wypełnić macierz  $\mathbb{O}$  brakującymi ocenami.

### 5.1 ALS z Apache Spark i MLlib

#### 5.1.1 Apache Spark

Apache Spark to ciesząca się ostatnio dużą popularnością platforma obliczeniowa stworzona w celu przetwarzania dużych zbiorów danych (BigData). Powstała ona w odpowiedzi na platformę MapReduce wykorzystywaną przez Apache Hadoop. Wspomniany MapReduce przetwarza dane w trybie wsadowym co oznacza, że podczas każdej operacji są one wczytywane i zapisywane na dysku (HDFS) przez co spada znacznie jego wydajność przy algorytmach iteracyjnych. W przypadku Apache Spark i głównej jego idei jaką jest Resilient Distributed Dataset zbiory danych są wczytywane do pamięci i dzięki temu są wykorzystywane przez kolejne kroki algorytmu bez konieczności ponownego wczytywania ich na dysk. Zwiększa to znacznie wydajność i szybkość wykonywania operacji.

Jedną z głównych bibliotek Apache Spark jest biblioteka MLlib. Jest to biblioteka uczenia maszynowego, której celem jest uczynić je łatwym i skalowalnym. MLlib zapewnia narzędzia do obsługi algorytmów klasyfikacji, regresji, klastrowania, redukcji

wymiaru, narzędzia algebry liniowej, statystyki i wiele innych. Biblioteka ta wspiera również narzędzia do obsługi reguł rekomendujących, a w szczególności filtrowania kolaboratywnego.

### 5.1.2 ALS i MLlib

Alternating Least Square (ALS) jest algorytmem faktoryzacji macierzy, który został zaimplementowany bibliotece uczenia maszynowego MLlib należącej do Apache Spark. Algorytm ten został opracowany z myślą o rozwiązywaniu problemów filtrowania na dużą skalę. Jest prosty a zarazem dobrze skalowalny w stosunku do dużych zbiorów danych.

Problem predykcji brakujących ocen formułujemy jako problem optymalizacyjny, gdzie celem jest zminimalizowanie funkcji celu oraz odnalezienie najbardziej optymalnych macierzy  $\mathbb{X}$  i  $\mathbb{Y}$ .

W szczególności staramy się zminimalizować błąd najmniejszych kwadratów postaci [11]:

$$\min_{\mathbb{X}, \mathbb{Y}} \sum_{o_{u_i, p_j}} (o_{u_i, p_j} - x_{u_i}^T y_{p_j})^2 + \lambda (\sum_{u_i} \|x_{u_i}\|^2 + \sum_{p_j} \|y_{p_j}\|^2),$$

gdzie

- $o_{u_i, p_j}$  są znanymi ocenami wystawionymi przez użytkowników,
- $\lambda$  jest czynnikiem regulującym.

Powyższa funkcja nie jest funkcją wypukłą (ze względu na obiekt  $x_{u_i}^T y_{p_j}$ ). Ustalając jednak jedną z macierzy  $\mathbb{X}$  lub  $\mathbb{Y}$ , otrzymujemy postać kwadratową, którą można rozwiązać. Rozwiązanie zmodyfikowanego problemu gwarantuje monotoniczne obniżenie ogólnej funkcji kosztów. Stosując ten krok naprzemiennie do macierzy  $\mathbb{X}$  i  $\mathbb{Y}$ , możemy iteracyjnie poprawiać faktoryzację macierzy. Podejście to określamy jako algorytm ALS (Alternating Least Squares).

**Algorytm 5.1** (ALS [11]). *Zainicjowanie  $\mathbb{X}$  i  $\mathbb{Y}$ .*

*Powtarzamy:*

- dla  $i = 1, \dots, n$  wykonujemy:

$$x_{u_i} = (\sum_{o_{u_i, p_j} \in o_{u_i, *}} y_{p_j} y_{p_j}^T + \lambda \mathbb{I}_k)^{(-1)} \sum_{o_{u_i, p_j} \in o_{u_i, *}} o_{u_i, p_j} y_{p_j}$$

*koniec*

- dla  $j = 1, \dots, m$  wykonujemy:

$$y_{p_j} = \left( \sum_{o_{u_i, p_j} \in o_{*, p_j}} x_u x_u^T + \lambda \mathbb{I}_k \right)^{(-1)} \sum_{o_{u_i, p_j} \in o_{*, p_j}} o_{u_i, p_j} x_{u_i}$$

*koniec*

*do momentu zbieżności.*

**Uwaga 5.2** (Różnica między SVD i ALS).

### 5.1.3 Implementacja algorytmu



## Rozdział 6

## Podsumowanie

# Bibliografia

- [1] Ricci F. and Rokach L. and Shapira B. and Kantor P.: *Recommender Systems Handbook*. Springer, 2011.
- [2] Krzyśko M.: *Wykład z teorii prawdopodobieństwa*. Wydawnictwo Naukowe - Techniczne, 2000.
- [3] Jakubowski J., Sztencel R.: *Wstęp do teorii prawdopodobieństwa*. SCRIPT, 2001.
- [4] Banaszak G., Gajda W.: *Elementy algebry liniowej część I*. Wydawnictwo Naukowe - Techniczne, 2002.
- [5] Walesiak M., Gantar E.: *Statystyczna analiza danych z wykorzystaniem programu R*. Wydawnictwo Naukowe PWN, 2009.
- [6] Poreda T., Jędrzejewski J.: *Algebra liniowa z elementami geometrii analitycznej*. Politechnika Łódzka, 2011.
- [7] Morzy T.: *Eksploracja danych. Metody i algorytmy*. Wydawnictwo Naukowe PWN, 2013.
- [8] Gorakala S. K.: *Building Recommendation Engines*. Packt, 2016.
- [9] Nowak Brzezińska A.: Analiza skupień. konspekt do zajęć: Statystyczne metody analizy danych. 2012.
- [10] O'Brien G.W. Berry M.W., Dumais S.T.: Using linear algebra for intelligent information retrieval. 1995.
- [11] Lublin M. Pere Y. Haoming L., Bangzheng H.: Matrix completion via alternating least square (als). 2015.
- [12] Kidziński Ł.: Statistical foundation of recommender systems. Master's thesis, University of Warsaw, 2011.
- [13] Bystrov V., COMISEF: <http://comisef.wikidot.com/concept:eigenvalue-and-singular-value-inequalities> (dostęp 07.05.2019).

- [14] Desrosiers K., Karypis G.: A comprehensive survey of neighborhood-based recommendation methods.