

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

1	Wstęp	3
2	Preliminaria	4
2.1	Elementy rachunku prawdopodobieństwa i statystyki	4
2.2	Elementy algebry liniowej	6
3	Elementy eksploracji danych wykorzystywane w systemach rekomen-	
	dujących	10
3.1	Preprocessing danych	10
3.1.1	Miary podobieństwa	11
3.1.2	Redukcja wymiaru	12
3.2	Metody eksploracji danych	19
3.2.1	k - Najbliższych Sąsiadów	20
3.2.2	k -Średnich	21
3.2.3	Maszyna Wektorów Nośnych	23
3.3	Szacowanie błędów obliczeń	27
3.3.1	Ocena dokładności metody	27
3.3.2	Ocena jakości modelu	28
3.3.3	Ocena rankingów	30
4	Modele tworzenia rekomendacji	32
4.1	Systemy rekomendujące oparte na treści(Content-based recommender systems):	34
4.1.1	Wygenerowanie profilu produktu - algorytm TF-IDF	35
4.1.2	Wygenerowanie profilu użytkownika	37
4.1.3	Rozpoznanie cech produktu odpowiednich dla użytkownika . . .	37
4.2	Filtrowanie kolaboratywne (Collaborative filtering)	38
4.2.1	Filtrowanie kolaboratywne oparte na użytkowniku	38
4.2.2	Filtrowanie kolaboratywne oparte na elementach	40
4.3	Hybrydowe systemy rekomendacji	42

4.4	Systemy rekomendujące kontekstowe (Context – aware recommender systems):	43
5	Eksperymenty / część praktyczne	45
5.1	Faktoryzacja macierzy w regułach rekomendujących	45
5.2	ALS z Apache Spark i MLlib	46
5.2.1	Apache Spark	46
5.2.2	ALS	46
5.2.3	Implementacja algorytmu	47
6	Podsumowanie	48

Rozdział 1

Wstęp

Rozdział 2

Preliminaria

2.1 Elementy rachunku prawdopodobieństwa i statystyki

Niech F oznacza σ - algebrę podzbiorów z przestrzeni Ω oraz niech X oznacza funkcję rzeczywistą określoną na przestrzeni Ω , to znaczy:

$$X : \Omega \longrightarrow \mathbb{R}.$$

Definicja 2.1 (Zmienna losowa [2, Sec 2.2 Def.2.2]). *Zmienną losową nazywamy funkcję X , która jest F - mierzalna, to znaczy jeżeli dla każdego $a \in \mathbb{R}$ zachodzi:*

$$\{\omega : X(\omega) < a\} = X^{-1}((-\infty, a)) \in F,$$

gdzie X^{-1} jest operacją przeciwobrazu zbioru przez funkcję X .

Niech (Ω, F, P) będzie przestrzenią probabilistyczną skończoną. X niech będzie zmienną losową prostą, która przyjmuje wartości x_1, \dots, x_n .

Przez A_k , $k = \{1, \dots, n\}$ oznaczmy: $\{\omega : X(\omega) = x_k\}$.

Wtedy zmienną losową prostą X możemy zapisać w postaci $X(\omega) = \sum_{k=1}^n x_k I_{A_k}(\omega)$, gdzie A_1, \dots, A_n tworzą mierzalny podział Ω , $A_k \in F$, $A_k \neq \emptyset$, $A_i \cap A_j = \emptyset$, dla $i \neq j$ oraz $\cup_{k=1}^n A_k = \Omega$.

Definicja 2.2 (Wartość oczekiwana [2, Sec 2.7 Def.2.21]). *Wartością oczekiwaną zmiennej losowej prostej X po zbiorze Ω względem miary probabilistycznej P nazywamy liczbę*

$$E(X) = \sum_{k=1}^n x_k P(A_k)$$

Skoro $A_k = \{\omega : X(\omega) = x_k\}$ i $P_X(x_k) = P(A_k)$, więc:

$$E(X) = \sum_{k=1}^n x_k P_X(x_k).$$

Definicja 2.3 (Kowariancja [2, Sec 2.8 Def.2.32]). *Kowariancją zmiennych losowych X, Y nazywamy liczbę:*

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

gdzie $E(X)$ oznacza wartość oczekiwaną zmiennej losowej X .

Definicja 2.4 (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 2.5 (Odchylenie standardowe [2, Sec 2.8 Def.2.28]). *Odchyleniem standardowym zmiennej losowej X nazywamy liczbę:*

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

Definicja 2.6 (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 2.7 (Macierz kowariancji). *Macierzą kowariancji nazywamy macierz*

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

gdzie:

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

Definicja 2.8 (Macierz korelacji). *Macierzą korelacji nazywamy*

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

gdzie:

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

2.2 Elementy algebry liniowej

Definicja 2.9 (Iloczyn skalarny [3, Sec 14.3 Def. 14.7]). Niech U oznacza przestrzeń liniową nad ciałem \mathbb{R} . Iloczynem skalarnym nazywamy formę dwuliniową

$$d : U \times U \longrightarrow \mathbb{R},$$

gdy:

- dla każdego $u \in U$ zachodzi:

$$d(u, u) \geq 0$$

- jest symetryczna, oznacza to, że dla dowolnych $u, v \in U$ zachodzi:

$$d(u, v) = d(v, u)$$

- $d(u, u) = 0 \Leftrightarrow u = \Theta_U$

Przestrzeń liniową U nad ciałem liczb rzeczywistych z iloczynem skalarnym

$$d : U \times U \longrightarrow \mathbb{R}$$

nazywamy przestrzenią euklidesową.

Definicja 2.10 (Macierz [5, Sec 8.1 Def. 8.1]). Macierzą o m wierszach i n kolumnach (o wymiarach $m \times n$) i wyrazach w ciele \mathbb{K} nazywamy funkcję określoną w zbiorze

$$\{1, 2, \dots, m\} \times \{1, 2, \dots, n\},$$

gdzie $n \in \mathbb{N}, m \in \mathbb{N}$ i przyjmującą wartości w ciele \mathbb{K} . Zapisujemy ją w postaci tabeli

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

a to oznacza, że wartością naszej funkcji dla argumentu (i, j) jest element a_{ij} , czasami oznaczany też jako $a_{i,j}$, należący do ciała \mathbb{K} .

Uwaga 2.11 ([5, Sec 8.1]). Macierz zapisujemy na następujące sposoby:

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

przy czym ostatnich dwóch sposobów używamy wtedy, gdy liczba kolumn i wierszy danej macierzy są ustalone.

Definicja 2.12 (Element macierzy [5, Sec 8.1]). *Elementem macierzy nazywamy a_{ij} , który stoi w i - tym wierszu i j - tej kolumnie macierzy $[a_{ij}]$. Element macierzy nazywany jest też wyrazem macierzy lub współczynnikiem macierzy.*

Definicja 2.13 (Macierz transponowana [5, Sec 8.1]). *Przez $M_{m \times n}(\mathbb{K})$ oznaczmy zbiór wszystkich macierzy o wymiarach $m \times n$ i elementach z ciała \mathbb{K} . Niech A , gdzie $A = [a_{ij}]_{j=1, \dots, n}^{i=1, \dots, m}$, będzie macierzą ze zbioru $M_{m \times n}(\mathbb{K})$. Macierz B , gdzie $B = [b_{ij}]_{j=1, \dots, m}^{i=1, \dots, n}$ ze zbioru $M_{n \times m}(\mathbb{K})$ nazywamy macierzą transponowaną macierzy A , jeśli $b_{ji} = a_{ij}$ dla każdego i ze zbioru $\{1, \dots, n\}$ oraz j ze zbioru $\{1, \dots, m\}$. Macierz transponowaną macierzy A oznaczamy jako A^t , A^T lub A^* .*

Twierdzenie 2.14 (Własności transpozycji macierzy [3, Sec 5.1 Tw. 5.1]). *Niech A i B będą macierzami o współczynnikach z K oraz niech posiadają tyle kolumn i wierszy, że operacje występujące powyżej są określone. Niech $\lambda \in K$. Zachodzą następujące równości:*

- $(A^T)^T = A$,
- $(A + B)^T = A^T + B^T$,
- $(\lambda A)^T = \lambda A^T$,
- $(AB)^T = B^T A^T$.

Definicja 2.15 (Macierz kwadratowa [5, Sec 8.1]). *Macierzą kwadratową nazywamy macierz, w której liczba wierszy i liczba kolumn są takie same. Liczbę tę nazywamy stopniem macierzy kwadratowej.*

Definicja 2.16 (Macierz diagonalna [5, Sec 8.1]). *Macierzą diagonalną nazywamy macierz kwadratową $[a_{ij}]$, gdzie wszystkie elementy poza główną przekątną są równe zero. Macierz diagonalną oznaczamy $\text{diag}(a_{11}, a_{22}, \dots, a_{nn})$.*

Definicja 2.17 (Macierz jednostkowa [5, Sec 8.1]). *Macierzą jednostkową stopnia n nazywamy taką macierz diagonalną, w której na głównej przekątnej wszystkie elementy są równe 1.*

Definicja 2.18 (Macierz ortogonalna [5, Sec 14.3 Def. 14.26]). *Macierz kwadratową C , gdzie $C = [c_{ij}]$, nazywamy macierzą ortogonalną, jeżeli spełniony jest warunek*

$$C^t \cdot C = C \cdot C^t = E,$$

gdzie E oznacza macierz jednostkową stopnia n .

Definicja 2.19 (Macierz nieosobliwa [5, Sec 10.4]). *Macierz nieosobliwą nazywamy macierz kwadratową, której wyznacznik jest różny od zera.*

Definicja 2.20 (Macierz osobliwa [5, Sec 10.4]). *Macierz osobliwą nazywamy macierz kwadratową, której wyznacznik jest równy zeru.*

Definicja 2.21 (Ślad macierzy). *Śladem macierzy $A \in M_{n,n}(K)$ nazywamy wielkość:*

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

Uwaga 2.22 ([5, 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 2.23 (Rząd kolumnowy i wierszowy macierzy [5, Sec 8.1]). *Niech $A \in M_{m \times n}(\mathbb{K})$. Rzędem kolumnowym macierzy A nazywamy wymiar podprzestrzeni przestrzeni \mathbb{K}^n generowanej przez kolumny macierzy A . Rząd ten oznaczamy symbolem $r_k(A)$. Rzędem wierszowym macierzy A nazywamy wymiar podprzestrzeni generowany przez wiersze macierzy A i oznaczamy go $r_w(A)$.*

Uwaga 2.24. *Rząd macierzy wyznacza maksymalną liczbę wektorów niezależnych, które są kolumnami lub wierszami tej macierzy.*

Twierdzenie 2.25 ([5, Sec 8.1 Tw. 8.10]). *Dla dowolnej macierzy rząd wierszowy jest równy rządowi wierszowemu.*

Definicja 2.26 (Rząd macierzy [5, Sec 8.1]). *Rzędem macierzy A nazywamy wspólną wartość rzędu kolumnowego i wierszowego macierzy A . Rząd macierzy oznaczamy symbolem $\text{rank}(A)$ lub $\text{rz}(A)$.*

Definicja 2.27 (Mnożenie macierzy [5, Sec 9.3 Def 9.13]). *Niech $A \in M_{m \times n}(\mathbb{K})$ i $B \in M_{k \times m}(\mathbb{K})$. Jeśli*

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

oraz

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k1} & b_{k2} & \cdots & b_{km} \end{bmatrix},$$

to iloczynem macierzy B i A nazywamy macierz C taką, że

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

gdzie

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

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

Definicja 2.28 (Przestrzeń liniowa [5, Sec 7.1]). ?

Definicja 2.29 (Przestrzeń generowana przez zbiór [5, Sec 7.1 Def 7.13]). Niech X będzie dowolnym i niepustym podzbiorem przestrzeni liniowej V . Podprzestrzenią generowaną przez zbiór X nazywamy zbiór wszystkich skończonych kombinacji liniowych wektorów ze zbioru X . Zbiór ten oznaczamy symbolem $\text{span}(X)$ lub $\text{lin}(X)$.

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

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

gdzie V jest przestrzenią liniową nad ciałem liczb rzeczywistych lub ciałem liczb zespolonych K .

Rozdział 3

Elementy eksploracji danych wykorzystywane w systemach rekomendujących

Większość systemów rekomendujących opiera swój rdzeń na algorytmach, które możemy rozumieć jako konkretne przypadki technik eksploracji danych. Proces eksploracji danych składa się z trzech kroków:

1. preprocesing 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 Preprocesing 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, jak filtrowanie kolaboratywne 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 [6]). *Niech $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$. Normą x nazywamy:*

$$\|x\| = \sqrt{x_1^2 + \dots + x_n^2}.$$

Jeśli $x, y \in \mathbb{R}^n$ to liczbę: $\|x - y\|$ nazywamy odległością euklidesową punktów x i y , gdzie:

$$(x - y) = (x_1 - y_1, \dots, x_n - y_n).$$

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

Uwaga 3.2. *Odległość Minkowskiego wyrażamy wzorem:*

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

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 [6]). *Niech $x, y \in \mathbb{R}^n$, $n \in \mathbb{N}$. Odległością kosinusową nazywamy:*

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

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

$$\text{sim}(x, y) = \frac{x \cdot y}{\|x\| \|y\|},$$

zatem

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

Innym podejściem jest korelacja Pearsona, którą definiujemy następująco:

Definicja 3.4 (współczynnik korelacji Pearsona [6]). *Niech $X, Y \in \mathbb{R}^n$ 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:*

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k,$$

$$\bar{y} = \frac{1}{n} \sum_{k=1}^n y_k.$$

Wówczas współczynnikiem korelacji Pearsona nazywamy:

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

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. Podobne zjawiska możemy dostrzec także w regułach rekomendujących. Proces redukcji wymiaru pozwala przezwyciężyć ten problem poprzez transformację przestrzeni danych do przestrzeni o mniejszej liczbie wymiarów. W poniższym rozdziale przyjrzymy się najczęściej wybieranym algorytmom redukcji wymiarów w kontekście reguł rekomendujących. Są to analiza głównych składowych oraz rozkład według wartości osobliwych.

Analiza Czynnika

[4, Sec 11.1]

Głównym celem analizy czynnikowej jest odnalezienie nowego zbioru zmiennych, który jest mniej liczny niż zbiór zmiennych wejściowych a jednocześnie wyraża zależności między zmiennymi obserwowalnymi.

Definicja 3.6 (Matematyczny model analizy czynnikowej). *Matematycznym modelem analizy czynnikowej nazywamy zapis postaci:*

$$Y = AX + BZ, \quad (3.1)$$

gdzie:

- $Y = [Y_1, Y_2, \dots, Y_m]^T$, $m \in \mathbb{N}$ to macierz standaryzowanych zmiennych, $Y_j = (y_{1j}, y_{2j}, \dots, y_{Tj})$, $j \in \{1, \dots, m\}$ - numer zmiennej, T - liczba obserwacji,
- $A = [a_{jl}]_{m \times k}$, gdzie $k \leq m$ - macierz ładunków czynnikowych czynników wspólnych, $l \in \{1, 2, \dots, k\}$ - numer czynnika wspólnego,
- $X = [X_1, X_2, \dots, X_k]^T$ - macierz czynników wspólnych, gdzie $X_l = (x_{l1}, x_{l2}, \dots, x_{lT})$,
- $B = [\text{diag}(b_j)]_{m \times m}$ - macierz ładunków czynnikowych czynników specyficznych,
- $Z = [Z_1, Z_2, \dots, Z_m]^T$ - macierz czynników specyficznych, gdzie $Z_j = (z_{1j}, z_{2j}, \dots, z_{Tj})$.

Czynniki wspólne i specyficzne są standaryzowane i wzajemnie ortogonalne. Zmienne obserwowalne poddaje się procesowi standaryzacji.

Definicja 3.7 (Normalizacja). *Normalizacją nazywamy procedurę wstępnej obróbki danych umożliwiającą ich dalsze porównanie i analizę.*

Definicja 3.8 (Standaryzacja). *Standaryzacją nazywamy rodzaj normalizacji zmiennej losowej, w którym wartość oczekiwana zmiennej losowej otrzyma wartość 0, natomiast odchylenie standardowe wartość 1, opisany za pomocą wzoru:*

$$y_{ij} = \frac{v_{ij} - \bar{v}_j}{s_j},$$

gdzie:

- $i \in \{1, 2, \dots, n\}$ - numer obiektu,
- $j \in \{1, 2, \dots, m\}$ - numer zmiennej,
- v_{ij} - wartość j - tej zmiennej pierwotnej dla i - tego obiektu
- \bar{v}_j - średnia arytmetyczna j - tej zmiennej pierwotnej,
- s_j - odchylenie standardowe dla j - tej zmiennej pierwotnej.

Równanie (3.1) można zatem zapisać w następujący sposób:

$$y_{ij} = a_{j1}x_{1i} + a_{j2}x_{2i} + \dots + a_{jk}x_{ki} + b_j z_{ij} = \sum_{t=1}^k a_{jt}x_{ti} + b_j z_{ij}.$$

Metoda analizy czynnikowej ma na celu wyznaczyć ładunki czynnikowe a_{jl} oraz b_j na podstawie zmiennych wejściowych v_{ij} . Między zmiennymi zachodzą związki, które są określane za pomocą współczynników korelacji Pearsona i określone w macierzy:

$$R = [r_{jh}]_{m \times m}.$$

Wyrazy macierzy obliczane są za pomocą wzoru:

$$r_{jh} = \frac{\frac{1}{T} \sum_{i=1}^T (y_{ij} - \bar{y}_j)(y_{ih} - \bar{y}_h)}{s_j s_h},$$

gdzie:

- \bar{y}_j, \bar{y}_h - średnie arytmetyczne odpowiednio j - tej i h - tej zmiennej,
- s_j, s_h - odchylenie standardowe j - tej i h - tej zmiennej.

Na głównej przekątnej macierzy korelacji znajdują się współczynniki korelacji zmiennej z nią samą. Zatem:

$$r_{jj} = \frac{1}{T} \sum_{i=1}^T y_{ij} y_{ij} = \frac{1}{T} \sum_{i=1}^T y_{ij}^2 = \frac{1}{T} \sum_{i=1}^T (y_{ij} - \bar{y}_j)^2 = s_{ij}^2 = 1.$$

Pozostałe elementy macierzy R są współczynnikami korelacji między różnymi zmiennymi i przyjmują postać:

$$r_{jh} = \frac{1}{T} \sum_{i=1}^T y_{ij} y_{ih}. \quad (3.2)$$

Stąd:

$$r_{jh} = \frac{1}{T} \sum_{i=1}^T [(a_{j1}x_{1i} + a_{j2}x_{2i} + \dots + a_{jk}x_{ki} + b_j z_{ij}) \cdot (a_{h1}x_{1i} + a_{h2}x_{2i} + \dots + a_{hk}x_{ki} + b_h z_{ih})],$$

a zatem:

$$r_{jh} = \frac{1}{T} \sum_{i=1}^T [(a_{j1}a_{h1}x_{1i}^2 + a_{j2}a_{h2}x_{2i}^2 + \dots + a_{jk}a_{hk}x_{ki}^2 + b_j b_h z_{ij} z_{ih}) + a_{j1}a_{h2}x_{1i}x_{2i} + \dots + a_{j1}b_h x_{1i}z_{ih} + \dots + b_j a_{hk} z_{ij} x_{ki}].$$

Czynniki są standaryzowane i nieskorelowane stąd otrzymujemy

$$\frac{1}{T} \sum_{i=1}^T x_{li}^2,$$

takie, że $\frac{1}{T} \sum_{i=1}^T x_{li}^2 = 1$ oraz wyrażenia

$$\frac{1}{T} \sum_{i=1}^T x_{li} x_{oi},$$

$$\frac{1}{T} \sum_{i=1}^T z_{ij} z_{ih},$$

$$\frac{1}{T} \sum_{i=1}^T x_{oi} z_{ij},$$

gdzie $o \in \{1, 2, \dots, k\}$ będące współczynnikami korelacji czynników swoistych i wspólnych. Są one równe 0. Wzór (3.2) przyjmuje zatem postać:

$$r_{jh} = a_{j1}a_{h1} + a_{j2}a_{h2} + \dots + a_{jk}a_{hk}.$$

Elementem odgrywającym dużą rolę w analizie czynnikowej jest podział wariancji całkowitej j - tej zmiennej na wariancję wspólną i specyficzną. Nich więc:

$$s_j^2 = h_j^2 + d_j^2.$$

Zakładamy przy tym, że $h_j^2 = \sum_{l=1}^k a_{jl}^2$ i oznacza zmienność wspólną j - tej zmiennej objaśnianą przez czynniki główne oraz $d_j^2 = 1 - h_j^2$ i oznacza zmienność specyficzną j - tej zmiennej.

Zależności między składowymi wariancji całkowitej zostały wyprowadzone ze wzoru:

$$s_j^2 = \frac{1}{T} \sum_{i=1}^T (y_{ij} - \bar{y}_j)^2.$$

Zauważmy, że $\bar{y}_j = 0$. Otrzymujemy zatem zależność:

$$\begin{aligned} s_j^2 &= \frac{1}{T} \sum_{i=1}^T (a_{i1}x_{1i} + a_{j2}x_{2i} + \dots + a_{jk}x_{ki} + b_j z_{ij})^2 \\ &= \frac{1}{T} [a_{j1}^2 \sum_{i=1}^T x_{1i}^2 + a_{j2}^2 \sum_{i=1}^T x_{2i}^2 + \dots + a_{jk}^2 \sum_{i=1}^T x_{ki}^2 + b_j^2 \sum_{i=1}^T z_{ij}^2 \\ &\quad + 2(a_{j1}a_{j2} \sum_{i=1}^T x_{1i}x_{2i} + \dots + a_{j1}a_{jk} \sum_{i=1}^T x_{1i}x_{ki} \\ &\quad + a_{j1}b_j \sum_{i=1}^T x_{1i}z_{ij} + \dots + a_{jk}b_j \sum_{i=1}^T x_{ki}z_{ij})]. \end{aligned}$$

Ze standaryzacji czynników i zmiennych wynika, że wariancja l - tego czynnika ma następującą postać:

$$\frac{1}{T} \sum_{i=1}^T x_{li}^2,$$

natomiast współczynnik korelacji między o -tym i l -tym czynnikiem:

$$\frac{1}{T} \sum_{i=1}^T x_{oi} x_{li} = \frac{\sum_{i=1}^T (x_{oi} - \bar{x}_o)(x_{li} - \bar{x}_l)}{s_{X_o} s_{X_l}} = r_{X_o, X_l}.$$

Wariancja ma zatem postać:

$$s_j^2 = 1 = a_{j1}^2 + a_{j2}^2 + \dots + a_{jk}^2 + b_j^2 + 2(a_{j1}a_{j2}r_{X_1, X_2} + \dots + a_{jk}b_j r_{X_k, Z_j}),$$

ponieważ czynniki są nieskorelowane to wzór ma postać:

$$s_j^2 = 1 = a_{j1}^2 + a_{j2}^2 + \dots + a_{jk}^2 + b_j^2.$$

Analiza Głównych Składowych

[4, Sec 11.1]

Jedną z najczęściej stosowanych metod wyodrębniania czynników głównych jest analiza głównych składowych. Celem tej metody nie jest wyjaśnienie korelacji między zmiennymi ale objaśnienie wariancji danych.

Definicja 3.9 (Matematyczny model analizy głównych składowych). *Matematycznym modelem analizy głównych składowych nazywamy model bazujący na równaniu charakterystycznym*

$$RV = \lambda V,$$

gdzie:

- R - macierz korelacji zmiennych,
- V - wektor własny,
- λ - wartość własna.

Rozwiązanie można przedstawić w postaci:

$$\det(R - I\lambda) = 0.$$

Zauważmy, że

$$\det \begin{bmatrix} 1 - \lambda & r_{12} \\ r_{12} & 1 - \lambda \end{bmatrix} = 0 \Leftrightarrow (1 - \lambda)(1 - \lambda) - r_{12}r_{12} = 0 \Leftrightarrow \lambda^2 - 2\lambda + (1 - r_{12}^2) = 0.$$

Zatem dla macierzy kowariancji stopnia drugiego wartości własne są następujące:

$$\lambda_1 = 1 + r_{12},$$

$$\lambda_2 = 1 - r_{12}.$$

Warto zauważyć, że jeżeli wartość korelacji między dwiema zmiennymi jest równa 1 to wartości własne przyjmują wartości 2 i 0. W przypadku, gdy korelacja wynosi 0 wartości własne przyjmują wartość 1. Suma:

$$\lambda_1 + \lambda_2 = (1 + r_{12}) + (1 - r_{12}) = 2$$

i jest równa liczbie zmiennych. Iloczyn:

$$\lambda_1 \lambda_2 = (1 - r_{12})^2$$

jest natomiast równy wyznacznikowi macierzy kowariancji. W macierzach korelacji każdego wymiaru powyższe wartości są spełnione.

Dodatkowo warto zauważyć, że największa wartość własna odzwierciedla liczbę wariancji wyjaśnioną przez pierwszą oś główną, druga co do wielkości wartość odzwierciedla liczbę wariancji wyjaśnioną przez drugą oś i kolejne wartości analogicznie.

Kolejną rzeczą wartą uwagi jest fakt, że skoro suma wartości własnych jest równa liczbie zmiennych to dzieląc i - tą wartość własną przez ilość zmiennych otrzymamy proporcję wariancji wyjaśnioną przez i - tą oś składową. Zatem:

$$p_i = \frac{\lambda_i}{m},$$

gdzie:

- p_i - proporcja wariancji,
- λ_i - i - ta wartość własna,
- m - liczb zmiennych.

Dla wartości własnych wyznacza się odpowiadające wektory własne, przy czym długość każdego z nich wynosi 1. Kolejnym krokiem jest wyznaczenie ładunków głównych składowych przez iloczyn macierzy, której wiersze składają się z wektorów własnych i macierzy zawierającej względne stosunki liczb wariancji opisane przez odpowiedni zmienne.

Analiza głównych składowych często jest błędnie postrzegana jako odmiana analizy czynnikowej. Analiza czynnikowa dąży do wyjaśnienia wariancji wspólnej pozostającej pod wpływem wspólnego czynnika. Analiza głównych składowych ma natomiast na celu wyjaśnienie całkowitej wariancji wszystkich zmiennych.

Rozkład Według Wartości Osobliwych (ang. Singular Value Decomposition (SVD))

Definicja 3.10 (Rozkład Według Wartości Osobliwych [10]). *Rozkładem według wartości osobliwych $m \times n$ - wymiarowej macierzy \mathbb{X} , gdzie $m \geq n$ oraz $r \in \mathbb{N}$ jest rzędem*

macierzy A nazywamy rozkład:

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

gdzie:

- \mathbb{U} jest macierzą ortogonalną $m \times m$ - wymiarową,
- Σ jest macierzą diagonalną, $m \times n$ - wymiarową o nieujemnych wartościach, $\Sigma = \text{diag}(d_1, d_2, \dots, d_n)$, $n \in \mathbb{N}$ tak, że $d_i > 0$ dla $1 \leq i \leq r$ i $d_i = 0$ dla $i \geq r + 1$,
- \mathbb{V} są macierzą ortogonalną $n \times n$ - wymiarową.

Definicja 3.11 (Norma Frobeniusa [10]). Normą Frobeniusa nazywamy:

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

gdzie A jest macierzą $m \times n$ - wymiarową. A^T jest sprzężeniem macierzy, a $\text{tr}(A)$ śladem macierzy A .

Twierdzenie 3.12 (Warunki równoważne SVD [10]). Niech rozkład według wartości osobliwych macierzy A będzie dany wzorem

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

gdzie $U = [u_1, u_2, \dots, u_m]$, $V = [v_1, v_2, \dots, v_n]$, $\Sigma = \text{diag}(d_1, d_2, \dots, d_n)$ oraz $d_1 \geq d_2 \geq \dots \geq d_r > d_{r+1} = \dots = d_n = 0$. $R(A)$ i $N(A)$ oznaczają zakres i jądro macierzy. Wtedy:

1. właściwości rzędu macierzy: $\text{rank}(A) = r$, $N(A) = \text{span}\{v_{r+1}, \dots, v_n\}$, $R(A) = \text{span}\{u_1, u_2, \dots, u_r\}$,
2. $A = \sum_{i=1}^r u_i \cdot d_i \cdot v_i^T$,
3. $\|A\|_F^2 = d_1^2 + \dots + d_r^2$ i $\|A\|_2^2 = d_1^2$.

Twierdzenie 3.13 (Twierdzenie Eckart - Younga [10]). Niech X będzie macierzą $m \times n$ - wymiarową z rozkładem według wartości osobliwych $\mathbb{X} = \mathbb{U} \Sigma \mathbb{V}^T$, $r \in \mathbb{N} = \text{rank}(A)$ niech będzie rzędem macierzy i $r \leq p = \min(m, n)$. Zdefiniujemy:

$$A_k = \sum_{i=1}^k u_i \cdot d_i \cdot v_i^T,$$

wtedy

$$\min_{\text{rank}(B)=k} \|A - B\|_F^2 = \|A - A_k\|_F^2 = d_{k+1}^2 + \dots + d_p^2.$$

Dowód. Niech $A \in \mathbb{R}^{m \times n}$ będzie macierzą o wartościach rzeczywistych, gdzie $m \geq n$. Załóżmy, że

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

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

$$A_k = \sum_{i=1}^k u_i \cdot d_i \cdot v_i^T,$$

gdzie u_i i v_i oznaczają odpowiednio i -te kolumny macierzy U i V . Zauważmy, że

$$\|A - A_k\|_F^2 = \left\| \sum_{i=k+1}^n u_i \cdot d_i \cdot v_i^T \right\|_F^2 = \sum_{i=k+1}^n d_i^2.$$

Stąd należy udowodnić, że $B_k = XY^T$, gdzie X i Y są macierzami oraz

$$\|A - A_k\|_F^2 = \sum_{i=k+1}^n d_i^2 = \|A - B_k\|_F^2.$$

Z nierówności trójkąta, jeżeli $A = A' + A''$ wtedy $d_1(A) \leq d_1(A') + d_1(A'')$. Przez A'_k i A''_k oznaczmy przybliżenia SVD odpowiednio macierzy A' i A'' . Stąd dla każdego $i, j \geq 1$

$$\begin{aligned} d_i(A') + d_j(A'') &= d_1(A' - A'_{i-1}) + d_1(A'' - A''_{j-1}) \geq d_1(A - A'_{i-1} - A''_{j-1}) \geq \\ &= d_1(A - A_{i+j-2}) (\text{rank}(A'_{i-1} + A''_{j-1})) \geq d_{i+j-1}(A), \end{aligned}$$

gdy $\text{rank}(A'_{i-1} + A''_{j-1}) \leq \text{rank}(A_{i+j-2})$. Jeżeli

$$d_{k+1}(B_k) = 0,$$

kiedy $A' = A - B_k$ i $A'' = B_k$ wnioskujemy, że dla $i \geq 1, j = k + 1$

$$d_i(A - B_k) \geq d_{k+1}(A).$$

Stąd:

$$\|A - B_k\|_F^2 = \sum_{i=1}^n d_i(A - B_k)^2 \geq \sum_{k+1}^n d_i(A)^2 = \|A - A_k\|_F^2.$$

□

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

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 w jednym ze swoich artykułów Tadeusz Morzy [13] metody eksploracji można podzielić na sześć klas:

- 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 k - Najbliższych Sąsiadów

[6, 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 idea 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

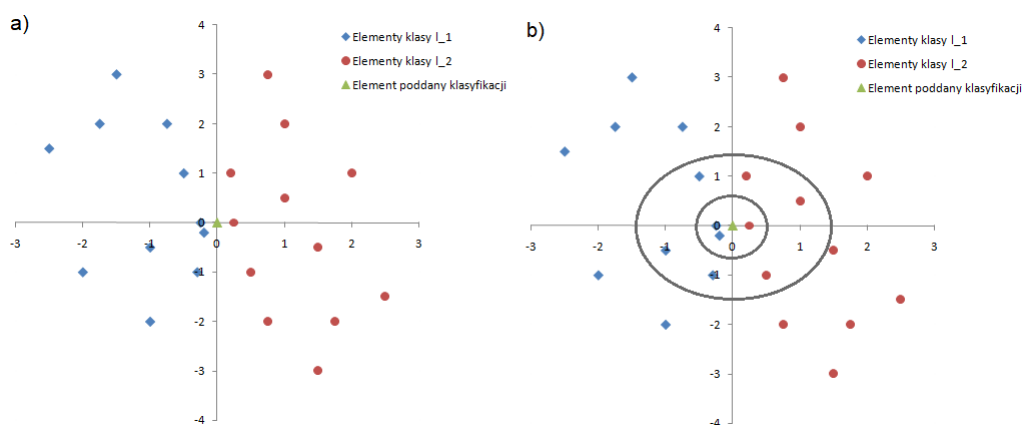
$$Y = \{\{y_1, l_1\}, \dots, \{y_k, l_k\}\},$$

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

$$\sum_{i=1}^k d(q, y_i)$$

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 k -Średnich

[9, Sec 2.3.1]

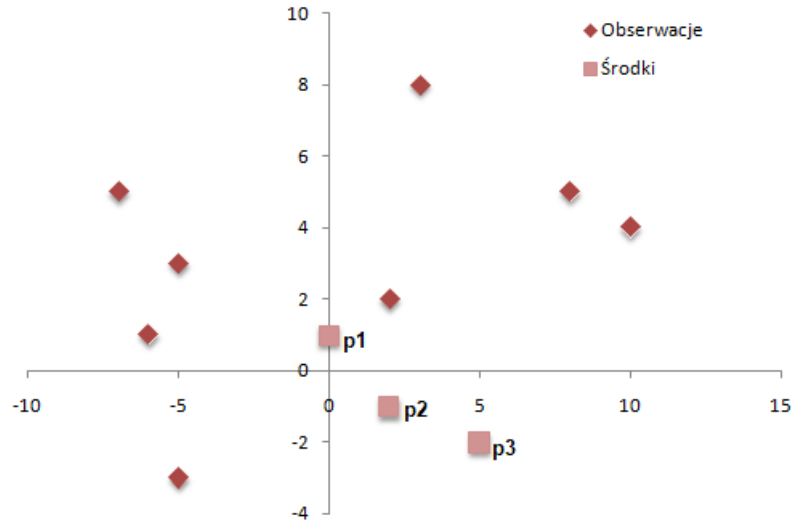
Algorytm k -średnich jest prostym i zarazem efektywnym algorytmem grupowania. Głównym celem algorytmu jest podział pewnego zbioru X :

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

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

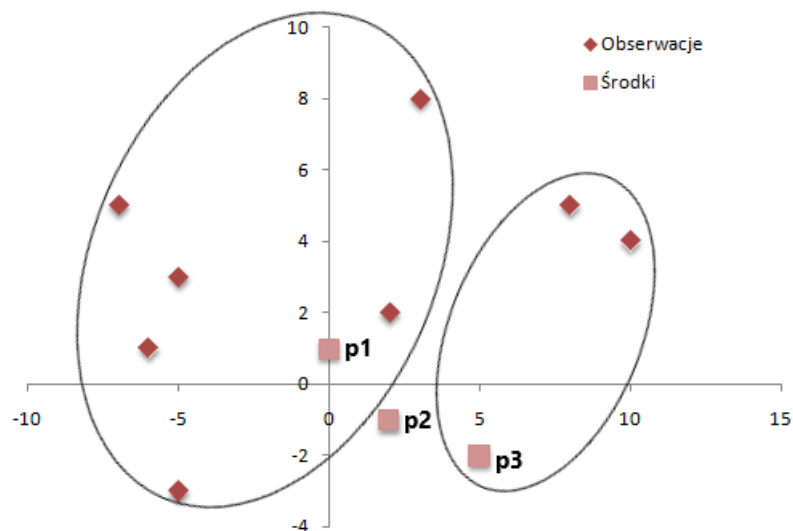
W wyniku grupowania n - elementowego zbioru X na k podgrup jest macierz podziału A o wymiarach $k \times n$. Każdy z elementów tej macierzy a_{ik} oznacza stopień w

jakim wektor x_k przynależy do grupy. Na wstępie algorytmu ustalamy wartość parametru k jako liczbę grup, które zostaną wydzielone. Wybieramy k reprezentantów, które stanowią prototypy grup.



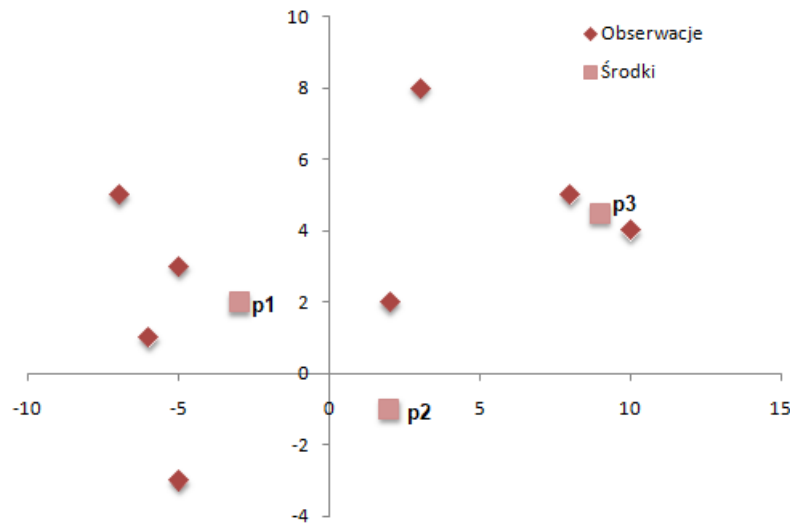
Rysunek 3.2: Metoda k -średnich. Wybór początkowych środków.

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:

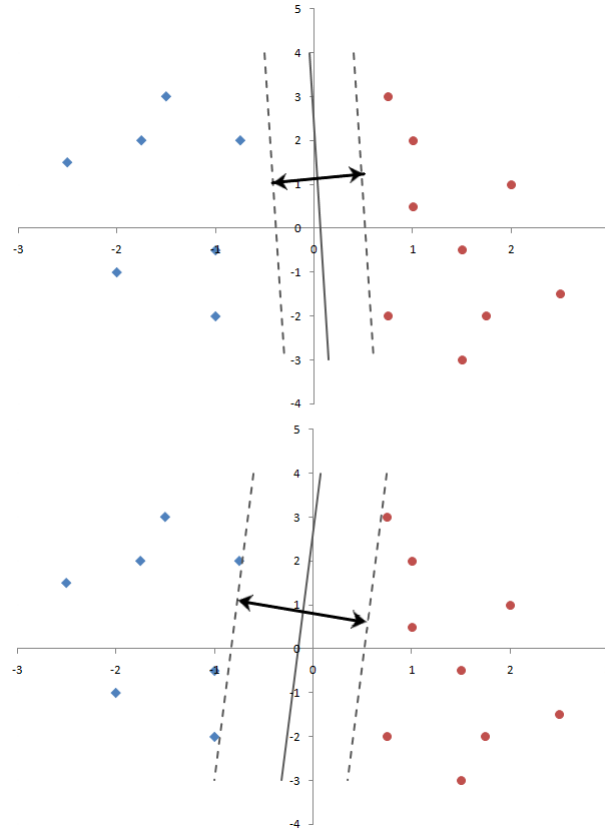
$$J(u, B) = \sum_{i=1}^k \sum_{k=1}^N b_{ki} d(u_i, x_k)^2,$$

gdzie

- $d(u, x)$ oznacza odległość elementu x od grupy wyznaczonej przez środek u ,
- N to liczebność zbioru X ,
- B oznacza macierz podziału.

3.2.3 Maszyna Wektorów Nośnych

Maszyny wektorów nośnych (ang. Support Vector Machines (SVM)) jest algorytmem uczenia maszynowego używanym zarówno w przypadku zadań klasyfikacji jak i regresji. Głównym celem algorytmu jest znalezienie w przestrzeni n - wymiarowej hiperpłaszczyzny, która wyraźnie sklasyfikuje obserwacje i podzieli na dwie klasy maksymalizując odległość między obserwacjami. Podejście to pozwala uniknąć błędnego przyporządkowania marginalnych obserwacji w przyszłości.



Rysunek 3.5: Maszyna wektorów nośnych - maksymalizacja odległości między klasami.

Aby formalnie zdefiniować maszyny wektorów nośnych zaczniemy od formalnego zdefiniowania problemu klasyfikacji.

Definicja 3.14 (Klasyfikacja [12]). *Niech w przestrzeni Ω znajdują się wektory danych $x_i \in \mathbb{R}^n$, $i \in \{1, \dots, n\}$. P niech będzie próbą uczącą. Wtedy:*

$$P = \{(x_i, c_i) : x_i \in \mathbb{R}^n, c_i \in \{1, -1\}\}.$$

Należy znaleźć klasyfikator dzielący całą przestrzeń Ω na dwie klasy $\{-1, 1\}$, który pozwoli na przyporządkowywać nowego elementu do klasy. Szukamy funkcji $f(x)$ - granicy między klasami.

Definicja 3.15 (Liniowa Separowalność [12]). *Dwie klasy nazywamy liniowo separowalnymi, jeżeli istnieje hiperpłaszczyzna $f(x)$ postaci:*

$$f(x) = w^T x + b$$

taka, że:

$$\begin{cases} f(x_i) > 0, & x_i \in \{1\} \\ f(x_i) < 0, & x_i \in \{-1\}. \end{cases}$$

Liniową separowalność definiujemy więc za pomocą funkcji:

$$\langle w, x \rangle + b = 0,$$

gdzie w, b są parametrami spełniającymi założenia modelu:

$$f(x) = \begin{cases} 1, & \text{gdy } \langle w, x \rangle + b > 0 \\ -1, & \text{gdy } \langle w, x \rangle + b < 0. \end{cases}$$

Parametry w i b wyznaczamy w taki sposób, by odległość między płaszczyznami

$$b_{i1} \langle w, x \rangle + b = 1,$$

$$b_{i2} \langle w, x \rangle + b = -1$$

była maksymalna. Zatem:

$$f(x) = \begin{cases} 1, & \text{gdy } \langle w, x \rangle + b \geq 1 \\ -1, & \text{gdy } \langle w, x \rangle + b \leq -1, \end{cases}$$

oraz szukana odległość wyznaczana jest przez maksymalizowanie funkcji:

$$Margin = \frac{2}{\|w\|^2}.$$

Jest to równoważne minimalizowaniu funkcji odwrotnej:

$$L(W) = \frac{\|w\|^2}{2},$$

przy ograniczeniach

$$y_i(\langle w, x \rangle + b) \geq 1.$$

Problem ten nazywamy problemem optymalizacji kwadratowej i rozważamy przy użyciu mnożników Lagrange’a.

Minimalizujemy więc funkcję Lagrange’a:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i(\langle w, x \rangle + b) - 1),$$

gdzie $\alpha_i \geq 0$ dla $i \in \{1, \dots, n\}$ to mnożniki Lagrange’a spełniające ograniczenia Karusha-Kuhna-Tuckera:

- $\alpha_i \geq 0$,
- $\alpha_i (y_i(\langle w, x \rangle + b) - 1) = 0$.

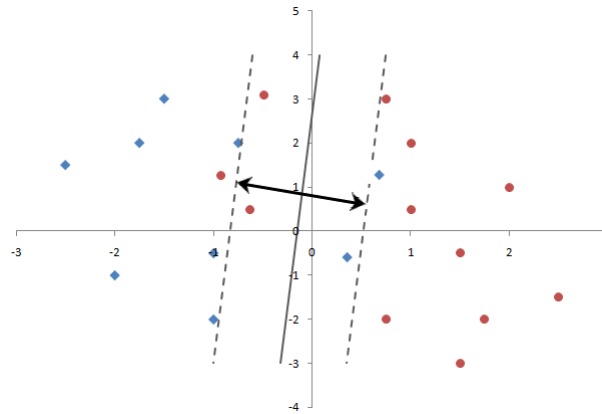
Możemy zauważyć, że w konsekwencji mnożniki Lagrange’a są niezerowe tylko dla wektorów nośnych. Liczba parametrów do oszacowania pozostaje jednak nadal zbyt duża. Częstym podejściem w tym przypadku jest przejście na postać dualną zadania optymalizacyjnego. Maksymalizujemy tu funkcję:

$$\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j),$$

przy ograniczeniach:

- $\alpha_i \geq 0$,
- $\forall_i \sum_{i=1}^n \alpha_i y_i = 0$.

Uwaga 3.16 ([12]). *Jeżeli elementy nie są jednak separowalne liniowo:*



Rysunek 3.6: Maszyna wektorów nośnych - elementy nieseparowalne liniowo.

wprowadzamy klasyfikator *Soft Margin SVM* oraz zmienne osłabiające (pomocnicze). W tym przypadku należy poddać minimalizacji funkcję:

$$L(W) = \frac{\|w\|^2}{2} + C \sum_{i=1}^n \varepsilon_i,$$

gdzie parametr C ocenia starty związane z każdym błędnie zaklasyfikowanym punktem. Funkcja $f(x)$ jest natomiast dana wzorem:

$$f(x) = \begin{cases} 1, & \text{gdy } \langle w, x \rangle + b \geq 1 - \varepsilon \\ -1, & \text{gdy } \langle w, x \rangle + b \leq -1 + \varepsilon. \end{cases}$$

Interpretacja rozwiązań:

- $\varepsilon_i = 0$ i $\alpha_i = 0$ - poprawne zaklasyfikowanie obserwacji, obserwacja leży poza marginesem,

- $\varepsilon_n = 0$ i $0 < \alpha_i < C$ - poprawne zaklasyfikowanie obserwacji, obserwacja leży na jednej z płaszczyzn ograniczających margines,
- $0 < \varepsilon_n < 1$ i $\alpha_i = C$ - poprawne zaklasyfikowanie obserwacji, obserwacja leży wewnątrz marginesu,
- $\varepsilon_n \geq 1$ i $\alpha_i = C$ - niepoprawne zaklasyfikowanie obserwacji.

3.3 Szacowanie błędów obliczeń

3.3.1 Ocena dokładności metody

Kolejna kwestia często poruszana w regułach rekomendujących to dokładność przewidywań. 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 3.17 (błąd średni [6, Sec 4.1.1]). *Średnim błędem nazywamy wartość wyrażenia:*

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

Definicja 3.18 (średni błąd bezwzględny [6, Sec 4.1.1]). *Średnim błędem bezwzględnym nazywamy wartość wyrażenia:*

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

Definicja 3.19 (średni błąd kwadratowy [6, 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 3.20. [6, 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 3.21 (średni błąd RMSE [6, 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 3.22 (macierz pomyłek [7, Sec 4.8.1]). Macierzą 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 - f_{ij} oznacza liczbę rekordów z etykietą E_i , którym błędnie została przypisana etykieta E_j .

Etykieta przewidziana	E_1	E_2
Etykieta rzeczywista		
E_1	f_{11}	f_{12}
E_2	f_{21}	f_{22}

Tabela 3.1: Macierz pomyłek.

Uwaga 3.23. [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:

Etykieta przewidziana	+	−
Etykieta rzeczywista		
+	<i>TP</i>	<i>FN</i>
−	<i>FP</i>	<i>TN</i>

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 3.24 (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 $f_b = f_t - f_p$.*

Kosztom 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 rate).

Definicja 3.25 (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 3.26 (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 3.27. [7, Sec 4.8.1] Innymi miarami, które można wywnioskować bezpośrednio z macierzy pomyłek dla klasyfikacji binarnej (Tabla 3.2) są:

- *WTP - współczynnik TP (czułość):*

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

- *WFP - współczynnik FP:*

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

- *WTN - współczynnik TN (specyficzność):*

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

- *precyzja:*

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

- *zwrot:*

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

- *F-miara:*

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

3.3.3 Ocena rankingów

W poprzedniej sekcji została omówiona metoda wyboru odpowiedniego modelu klasyfikującego. Każdy z takich modeli kończy się przedstawieniem pewnej listy wyników.

W tej części zostaną zaprezentowane metody służące do oceny otrzymanych na podstawie modelu rankingów i pomagające zapewnić odpowiedni porządek wyników. Jeżeli elementy posiadają oceny (np. użytkowników) intuicyjnym jest stworzyć ranking przez uporządkowanie tych ocen w malejący sposób. W przypadku gdy jednak nie mamy takich danych lub nie jest odpowiednie tworzenie takiego rankingu użyjemy znormalizowanej miary opartej na odległości (ang. Normalized Distance-based Performance Measure (NDPM)):

Definicja 3.28 (miara znormalizowana oparta na odległości [14, Sec 2.2.1]). Niech r_{ui} będzie rankingiem odniesienia i \hat{r}_{ui} rankingiem stworzonym przez wybrany system rekomendujący n_u elementów i użytkownikowi u . Ponadto niech:

$$C^+ = \sum_{i < j} \text{sgn}(r_{ui} - r_{uj}) \text{sgn}(\hat{r}_{ui} - \hat{r}_{uj})$$

$$C^- = \sum_{i < j} \text{sgn}(r_{ui} - r_{uj})^2 \text{sgn}(\hat{r}_{uj} - \hat{r}_{ui})$$

$$C^u = \sum_{i < j} \text{sgn}(r_{ui} - r_{uj})^2$$

$$C^s = \sum_{i < j} \text{sgn}(\hat{r}_{ui} - \hat{r}_{uj})$$

$$C^0 = C^u - (C^+ - C^-),$$

gdzie C^u jest liczbą par dla których ranking referencyjny okazuje się lepszą możliwością, C^+ i C^- to liczba par z poprawną i niepoprawną kolejnością, C^0 jest liczbą par dla których ranking referencyjny nie jest wiążący, kiedy ranking systemu rekomendującego jest.

Miarą NDPM nazywamy:

$$NDPM = \frac{C^- + 0.5C^0}{C^u}$$

Uwaga 3.29. [14, Sec 2.2.1] Powiązania w rankingu referencyjnym pojawiają się naturalnie, kiedy znamy preferencje użytkownika (np. przydziela on oceny). Czasami jednak rankingi są bardziej specyficzne (np. kiedy dajemy użytkownikowi wybór między dwoma elementami). Wtedy też system rekomendujący nie powinien tworzyć rankingu klasyfikując jedno elementy wyżej niż inne. W takich przypadkach przychodzą nam z pomocą

- miara Kendall's τ :

$$\tau = \frac{C^+ - C^-}{\sqrt{C^u} \sqrt{C^s}},$$

- miara Spearman's ρ :

$$\rho = \frac{1}{n_u} \frac{\sum_i (r_{iu} - \bar{r})(\hat{r}_{iu} - \bar{\hat{r}})}{\sigma(r)\sigma(\hat{r})},$$

gdzie \bar{r} i $\bar{\hat{r}}$ oznaczają średnie, natomiast $\sigma(r)$ i $\sigma(\hat{r})$ odchylenia standardowe.

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 [14, Sec 1.3]). *Przedmiotem nazwiemy klasę obiektów tego samego typu, nierozróżnialnych dla obserwatora, reprezentowanych przez co najmniej jeden element.*

Przedmioty stanowią podstawową grupę elementów w rozważaniach systemów rekomendujących.

Definicja 4.2 (Użytkownik [14, Sec 1.3]). *Użytkownikiem nazywamy osobę zdolną do przedstawienia własnej oceny wybranego przedmiotu.*

W pracy [14] użyty jest zawsze ten sam zbiór ocen, jednak łatwo możemy pokusić się o jego uogólnienie.

Definicja 4.3 (Zbiór ocen [14, Sec 1.3]). *Podzbiór skończony zbioru \mathbb{N} lub $\mathbb{N} \cup \{0\}$ nazywamy zbiorem ocen dla przedmiotów.*

Definicja 4.4 (Macierz preferencji [14, Sec 1.3]). *Rozważmy zbiór przedmiotów o liczności n oraz grupę użytkowników o liczności m . Macierzą preferencji M nazwiemy macierz o wymiarach $n \times m$ i wartościach w ustalonym zbiorze ocen.*

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

Definicja 4.5 (Własność [14, 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 P . Dla każdej $p \in P$ poprzez V_p rozumiemy zbiór wszystkich dopuszczalnych wartości własności p .

Definicja 4.6 (Funkcja anotująca [14, Sec 1.3]). Funkcją anotującą własność $p \in P$ nazwiemy funkcję

$$a_p: S \rightarrow V_p,$$

gdzie S oznacza zbiór wszystkich przedmiotów.

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

Definicja 4.7 (Problem tworzenia rekomendacji [14, Sec 1.3]). Rozważmy pewien zbiór przedmiotów S oraz pewien zbiór użytkowników U oraz pewien zbiór ocen Z . Niech ponadto R będzie funkcją działającą $R: S \times U \rightarrow Z$. Załóżmy, że dla funkcji R znane są wartości dla pewnych par przedmiotów i użytkowników. Zadaniem naszym jest zaproponowanie sposobu predykcji wartości tej funkcji R dla brakujących par w sposób minimalizujący wybrany funkcjonal błędu.

Przyjrzyjmy się następującemu przykładowi, który wykorzystamy do ilustracji tej sytuacji.

Przykład 4.1. Załóżmy, że mamy zbiór sześciu książek $\{s_1, s_2, s_3, s_4, s_5, s_6\}$, oraz zbiór sześciu czytelników $\{u_1, u_2, u_3, u_4, u_5, u_6\}$. Poniższa tabela przedstawia oceny jakie czytelnicy przyporządkowali po przeczytaniu poszczególnym pozycjom. Znak '?' oznacza, że czytelnik danej książki nie czytał lub czytał lecz jej nie ocenił.

Czytelnicy		u_1	u_2	u_3	u_4	u_5	u_6
Książki	s_1	6	3	?	6	4	?
	s_2	?	6	6	5	6	?
	s_3	7	7	8	7	8	9
	s_4	8	10	10	7	6	8
	s_5	9	6	6	6	6	?
	s_6	5	7	7	5	4	2

Celem jest przewidzieć brakujące oceny w sposób minimalizujący błąd popełniany przez nasz model.

Uwaga 4.8. Systemy rekomendujące mogą zostać podzielone na systemy oparte na kontekście i systemy oparte na samych ocenach. W pierwszym przypadku predykcja jest dokonywana na podstawie przydzielonych ocen oraz pod względem wektorów cech rozważanych elementów. Główne założenie wymaga, że jeżeli użytkownik ocenił w pewien

sposób przedmiot A w przeszłości oraz przedmiot B jest podobna do A , to użytkownik będzie skłonny w podobny sposób ocenić przedmiot B . W systemach opartych na ocenach, badaniu podlegają natomiast zależności które występują między produktami, a użytkownikami. Zakłada się, że jeżeli użytkownicy A i B wykazują podobieństwo oraz użytkownik A oceni pewien przedmiot, którego użytkownik B jeszcze nie ocenił, to prawdopodobnie ocena użytkownika B będzie podobna do oceny użytkownika A .

4.1 Systemy rekomendujące oparte na treści(Content-based recommender systems):

Systemy oparte na treści wyróżnia ukierunkowanie na spersonalizowany poziom użytkownika oraz treść produktu.

Problem 4.1. Głównym zadaniem w tej metodzie jest scharakteryzowanie wektora cech G , które użytkownik ceni, a następnie zasugerowanie mu produktów, które te cechy posiadają.

Wspomniana metoda opiera się na obliczaniu podobieństw oraz wykorzystuje zadania uczenia maszynowego, takie jak klasyfikacja.

Algorytm 4.9. W typie rekomendacji opartym na treści stworzenie rekomendacji i wygenerowanie listy elementów, które mogą być odpowiednie użytkownikowi możemy przedstawić w trzech krokach:

1. wygenerowanie profilu produktu,
2. wygenerowanie profilu użytkownika,
3. rozpoznanie cech produktu odpowiednich dla użytkownika.

Przykład 4.2. Niech poniższa tabela określa w jakim stopniu (skala 0-1) każda z książek reprezentuje cechy gatunków g_1, g_2 .

Gatunki		g_1	g_2
Książki	s_1	0,9	0
	s_2	1	0,01
	s_3	0,99	0
	s_4	0,1	1
	s_5	0	0,9
	s_6	0,8	0,3

Dodatkowo zakładamy, że istnieje gatunek g_0 , którego cechy reprezentują wszystkie książki oraz dla każdej z książek $g_0 = 1$. Zatem $g^{(1)} = \begin{bmatrix} 1 \\ 0,9 \\ 0 \end{bmatrix}$ jest wektorem cech

odpowiadającym książce s_1 , $g^{(2)} = \begin{bmatrix} 1 \\ 1 \\ 0,01 \end{bmatrix}$ jest wektorem cech odpowiadającym książce s_2 . Analogicznie możemy wyznaczyć podobne wektory dla każdej z książek. Należy jednak zaznaczyć, że liczba wybranych przez nas cech to $n = 2$.

Dla każdego użytkownika j wyznaczamy wektor parametrów $\Theta^{(j)} \in \mathbb{R}^3$ (w ogólnej postaci $\Theta^{(j)} \in \mathbb{R}^{n+1}$), który przedstawia (w skali 0-5) preferencje użytkownika dotyczące poszczególnych gatunków. I tak odpowiednio $\Theta^{(1)}$ to wektor preferencji czytelnika u_1 , $\Theta^{(2)}$ wektor preferencji użytkownika u_2 , $\Theta^{(3)}$ wektor preferencji użytkownika u_3 i $\Theta^{(4)}$ wektor preferencji użytkownika u_4 , $\Theta^{(5)}$ wektor preferencji użytkownika u_5 i $\Theta^{(6)}$ wektor preferencji użytkownika u_6 .

Aby przewidzieć ocenę jaką użytkownik j wystawiłby książce i po jej przeczytaniu użyjemy równania:

$$(\Theta^{(j)})^T g^{(i)}.$$

Obliczmy ocenę jaką książce k_3 wystawiłby użytkownik u_1 przy założeniu, że wektor preferencji użytkownika u_1 jest postaci $\Theta^{(1)} = \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$. Użytkownik ten preferuje więc książki gatunku g_1 , gdy książki gatunków g_0 i g_2 są dla niego nieatrakcyjne. Zatem:

$$(\Theta^{(1)})^T g^{(3)} = [0, 5, 0] \begin{bmatrix} 1 \\ 0,99 \\ 0 \end{bmatrix} = 0 \cdot 1 + 5 \cdot 0,99 + 0 \cdot 0 = 4,95.$$

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 produktu - algorytm TF-IDF

W większość systemów rekomendacji opartych na treści można zauważyć użycie prostych modeli wyszukiwujących. Jednym z najbardziej popularnych jest model przestrzeni wektorowej (*ang. Vector Space Model*) z algorytmem TF-IDF (*ang. TF – term frequency, IDF – inverse document frequency*).

Definicja 4.10 (Model przestrzeni wektorowej [6, Sec 3.3.1.1]). *Modelem przestrzeni wektorowej nazywamy formę reprezentacji dokumentów, gdzie dokument i jest reprezentowany przez wektor x_i w przestrzeni n -wymiarowej, a każdy z n wymiarów stanowi rozważaną cechę produktu.*

Uwaga 4.11. *Dla danych w formie dokumentów tekstowych cechami, które pomagają scharakteryzować temat dokumentu, np. artykułu na stronie internetowej są słowa. Dla każdego ze słów zostaje obliczona wartość funkcji TFIDF (została ona dokładnie omówiona poniżej). W końcowym etapie słowa, które otrzymały najwyższe wyniki zostają uznane za charakteryzujące rozważany dokument. Formalnie rzecz ujmując każdy z dokumentów jest przedstawiony za pomocą wektora wag, gdzie waga w odpowiedni sposób wyraża zależność między dokumentem, a badanym terminem.*

Definicja 4.12 (TFIDF [6, Sec 3.3.1.1]). *Niech $S = \{s_1, s_2, \dots, s_n\}$, $n \in \mathbb{N}$ będzie zestawem dokumentów / analizowanych przedmiotów. Natomiast $G = \{g_1, g_2, \dots, g_n\}$, $n \in \mathbb{N}$ zestawem cech rozważanych w przedmiotach. Każdy z dokumentów s_j , $j \in \{1, \dots, n\}$ jest reprezentowany jako wektor w przestrzeni wektorowej n -wymiarowej. Zatem $s_j = [w_{1j}, w_{2j}, \dots, w_{nj}]$, gdzie w_{kj} jest wagą dla cechy g_k w dokumencie s_j .*

TFIDF nazywamy funkcję:

$$TFIDF(t_k, d_j) = TF(g_k, k_j) \cdot IDF,$$

gdzie:

- $TF(g_k, k_j)$ (macierz term frequency) przedstawia odniesienie każdego z podanych terminów do każdego z badanych elementów:

$$TF(g_k, k_j) = \frac{f_{k,j}}{\max_z f_{z,j}},$$

$\max_z f_{z,j}$ - maksymalna w odniesieniu do wszystkich cech c_z , które pojawiły się w dokumencie p_j częstotliwość wystąpień ($f_{z,j}$)

- IDF (inverse dokument frequency) wyraża się formułą:

$$IDF = \log \frac{N}{n_k}$$

N - całkowita liczba dokumentów w zbiorze, n_k - liczba dokumentów w których cecha c_k wystąpiła przynajmniej raz.

Ponadto w związku z założeniem o normalizacji wagi, które zostały uzyskane w wyniku $TFIDF(t_k, d_j)$ poddane zostaną metodzie transformacji kosinusowej:

$$w_{k,j} = \frac{TFIDF(t_k, d_j)}{\sqrt{\sum_{i=1}^{|T|} TFIDF(t_i, d_j)^2}}.$$

Tak wygenerowany został zbiór słów, które są reprezentacją danego dokumentu oraz jego tematu.

Przy stosowaniu algorytmu *TFIDF* zakładamy, że:

- rzadkie cechy są równie istotne jak częste (założenie IDF),
- kilkukrotne wystąpienie terminu w rozważanym dokumencie jest równie istotne jak pojedyncze (założenie TF),
- długość dokumentu (filmu, książki) nie ma znaczenia (założenie normalizacji).

Zatem jeżeli termin występuje często w konkretnym przedmiocie rozważań (TF) i równocześnie rzadko w pozostałych elementach zbioru (IDF) ma większe prawdopodobieństwo stać się jedną z istotnych cech rozważanych w temacie. Ponadto normalizacja wektorów wag pozwala zrównoważyć wartość wyników i umożliwia ich porównywanie w dalszej analizie.

Uwaga 4.13. *[6, Sec 3.3.1.1] Ostatnim krokiem przed zarekomendowaniem pozycji czytelnikom jest wyestymowanie podobieństw występujących między poszczególnymi dokumentami. Naturalnymi miarami odległości, które można tu wykorzystać są:*

- *współczynnik Jaccarda mierzący odległość między zbiorami słów,*
- *podobieństwa kosinusów mierzące odległość między wektorami. Przy estymowaniu odległości kosinusów między dwoma dokumentami elementami wektora są słowa, które zostały wybrane na podstawie wcześniej przeprowadzonego algorytmu. W wektorze tym wartość 1 oznacza, że słowo pojawiło się w zbiorze opisującym dokument, natomiast 0 oznacza sytuację przeciwną.*

4.1.2 Wygenerowanie profilu użytkownika

Przy generowaniu wektora opisującego rozważane dokumenty należy uwzględnić komponenty opisujące preferencje użytkowników. Dla każdego użytkownika u przedstawiamy jego preferencje w postaci wektora x_u , gdzie użytkownik pozycjonuje element i poprzez wektor cech x_i .

4.1.3 Rozpoznanie cech produktu odpowiednich dla użytkownika

Proces rekomendacji bazuje na dopasowaniu cech profilu użytkownika i wartości opisujących treść obiektu. Rezultatem jest stwierdzenie czy rozważany kandydat jest zainteresowany analizowanym przedmiotem. Zainteresowanie użytkownika danym przedmiotem można wyestymować używając podobieństwa kosinusów. Rozważmy dokument,

gdzie użytkownik wyraża zainteresowanie większością cech w wektorze x_i opisującym dokument i . Otrzymamy cosinus kąta między użytkownikiem i dokumentem będący dodatnim ułamkiem. Oznacza to, że kąt będzie bliski 0° a tym samym odległość między wektorami będzie mała. W przypadku przeciwnym otrzymamy cosinus bliski zeru lub mniejszy od zera, oznacza to kąt należący do przedziału $(90^\circ; 180^\circ)$ oraz małe podobieństwo między rozważanymi wektorami.

4.2 Filtrowanie kolaboratywne (Collaborative filtering)

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

Wyróżniamy dwa podstawowe typy filtrowania kolaboratywnego:

- filtrowanie kolaboratywne oparte na użytkowniku (ang. user-based),
- filtrowanie kolaboratywne oparte na elementach (ang. item-based).

4.2.1 Filtrowanie kolaboratywne oparte na użytkowniku

Algorytm 4.14. *Stworzenie rekomendacji filtrowania kolaboratywnego opartej na użytkowniku wykonamy w dwóch krokach:*

1. *Znalezienie podobieństwa między czytelnikami opierającego się na znanych informacjach. Najczęstszymi stosowanymi podejściami do obliczania szukanego podobieństwa są Metryka Euklidesowa i Współczynnik Korelacji Pearsona.*
2. *Wystymowanie ocen, które użytkownicy (w szczególności osoba dla której tworzymy rekomendację) mogliby wystawić dla nieocenionych przez nich elementów z rozważanego zbioru.*

W celu dokładniejszego zrozumienia wróćmy do przykładu.

Przykład 4.3. *Chcąc obliczyć podobieństwo między użytkownikiem u_2 i u_3 wybierzmy książki, które przeczytali obaj użytkownicy. W tym przypadku są to: s_2, s_3, s_4, s_5, s_6 .*

Na tej podstawie tworzymy odpowiednio wektory ocen użytkowników $o_{2,3}^{(2)}$ i $o_{2,3}^{(3)}$. Zatem

$$o_{2,3}^{(2)} = \begin{bmatrix} 6 \\ 7 \\ 10 \\ 6 \\ 7 \end{bmatrix} \wedge o_{2,3}^{(3)} = \begin{bmatrix} 6 \\ 8 \\ 10 \\ 6 \\ 7 \end{bmatrix}.$$

Posługując się odległością euklidesową

$$d_e(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$$

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:

Czytelnicy	u₁	u₂	u₃	u₄	u₅	u₆
u₁	0	5,099	4,243	3	4,359	3,606
u₂	5,099	0	1	4,796	5,196	5,745
u₃	4,243	1	0	3,873	5	5,477
u₄	3	4,796	3,873	0	2,828	3,742
u₅	4,359	5,196	5	2,828	0	3
u₆	3,606	5,745	5,477	3,742	3	0

W celu uzyskania wartości z przedziału $[0, 1]$ poddajemy macierz normalizacji ($\max\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} - \min\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$):

Czytelnicy	u₁	u₂	u₃	u₄	u₅	u₆
u₁	0	0,5099	0,4243	0,3	0,4359	0,3606
u₂	0,5099	0	0,1	0,4796	0,5196	0,5745
u₃	0,4243	0,1	0	0,3873	0,5	0,5477
u₄	0,3	0,4796	0,3873	0	0,2828	0,3742
u₅	0,4359	0,5196	0,5	0,2828	0	0,3
u₆	0,3606	0,5745	0,5477	0,3742	0,3	0

Im większa odległość między czytelnikami tym mniejsze jest podobieństwo jakie między nimi występuje. Zakładając, że największa wartość prawdopodobieństwa to 1 macierz podobieństwa przyjmuje wartości:

<i>Czytelnicy</i>	u_1	u_2	u_3	u_4	u_5	u_6
u_1	1	0,4901	0,5757	0,7	0,5641	0,6394
u_2	0,4901	1	0,9	0,5204	0,4804	0,4255
u_3	0,5757	0,9	1	0,6127	0,5	0,4523
u_4	0,7	0,5204	0,6127	1	0,7172	0,6258
u_5	0,5641	0,4804	0,5	0,7172	1	0,7
u_6	0,6394	0,4255	0,4523	0,6258	0,7	1

Bazując na przedstawionych podobieństwach wyestymujemy ocenę jaką użytkownik u_1 zaproponuje dla książki s_2 . W poniższym równaniu obliczymy średnią ważoną ocen wykorzystując wartość podobieństwa między użytkownikiem u_1 , a pozostałymi użytkownikami, którzy wcześniej ocenili książkę:

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

Na podstawie metody filtrowania kolaboratywnego opartej na użytkownikach wnioskujemy, że użytkownik u_1 wystawiłby książce s_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

na których podstawie możemy dokonać rekomendacji dla naszych użytkowników. Możemy wnioskować, że dla użytkownika u_3 książka s_1 prawdopodobnie nie będzie zbyt atrakcyjna. Użytkownik u_6 , natomiast, z chęcią przeczyta książkę s_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 4.15. Podobnie jak w przypadku opartym na użytkowniku, w tym również należy wykonać dwa kroki.

1. Pierwszym etapem jest znalezienie podobieństw występujących między elementami. Najczęstszą miarą podobieństwa w tym przypadku jest podobieństwo kosinusów. Miara ta wyraża podobieństwo między n -wymiarowymi wektorami poprzez kąt między nimi w przestrzeni wektorowej. Wraz ze wzrostem wartości kąta rośnie podobieństwo.
2. Następnie, na podstawie wydanych przez użytkownika ocen, należy wyestymować noty dla elementów przez niego nieocenionych.

Przykład 4.4. W tym przypadku wektory odzwierciedlają zbiory ocen wystawione poszczególnym książkom przez użytkowników. Wektor mają zatem postać:

$$o^{(s_1)} = \begin{bmatrix} 6 \\ 3 \\ 6 \\ 4 \end{bmatrix}, o^{(s_2)} = \begin{bmatrix} 6 \\ 6 \\ 5 \\ 6 \end{bmatrix}, o^{(s_3)} = \begin{bmatrix} 7 \\ 7 \\ 8 \\ 7 \\ 8 \\ 9 \end{bmatrix}, o^{(s_4)} = \begin{bmatrix} 8 \\ 10 \\ 10 \\ 7 \\ 6 \\ 8 \end{bmatrix}, o^{(s_5)} = \begin{bmatrix} 9 \\ 6 \\ 6 \\ 6 \\ 6 \end{bmatrix}, o^{(s_6)} = \begin{bmatrix} 5 \\ 7 \\ 7 \\ 5 \\ 4 \\ 2 \end{bmatrix}$$

Warto zauważyć, że wektory są różnej długości.

Aby obliczyć podobieństwo między książkami s_1 i s_2 wyznaczmy wektory ocen w których uwzględnimy przypadki, gdzie jeden użytkownik ocenił obie pozycje. Zatem:

$$\overline{o^{(s_1)}} = \begin{bmatrix} 3 \\ 6 \\ 4 \end{bmatrix}, \overline{o^{(s_2)}} = \begin{bmatrix} 6 \\ 5 \\ 6 \end{bmatrix}.$$

Następnie używając wzoru na podobieństwo kosinusowe

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

obliczamy podobieństwo między wybranymi książkami

$$\text{sim}(s_1, s_2) = \frac{\overline{o^{(s_1)}} \cdot \overline{o^{(s_2)}}}{|\overline{o^{(s_1)}}| |\overline{o^{(s_2)}}|} = \frac{3 \cdot 6 + 6 \cdot 5 + 4 \cdot 6}{\sqrt{6^2 + 3^2 + 6^2 + 4^2} \sqrt{6^2 + 6^2 + 5^2 + 6^2}} = 0,6339.$$

Postępując w analogiczny sposób otrzymamy macierz podobieństwa elementów:

Czytelnicy	s₁	s₂	s₃	s₄	s₅	s₆
s₁	1	0,6339	0,7372	0,7195	0,8935	0,7599
s₂	0,6339	1	0,7951	0,8150	0,7977	0,8898
s₃	0,7372	0,7951	1	0,9780	0,8586	0,9200
s₄	0,7195	0,8150	0,9780	1	0,8860	0,9681
s₅	0,8935	0,7977	0,8586	0,8860	1	0,9413
s₆	0,7599	0,8898	0,9200	0,9681	0,9413	1

Wystymujmy teraz ocenę jaką użytkownik u_6 zaproponuje dla książki s_2 . Ponownie obliczymy średnią ważoną ocen, tym razem, wykorzystując wartość podobieństwa między książką k_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 s_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 Hybrydowe systemy rekomendacji

Aby uzyskać lepsze wyniki możemy wykorzystać podczas tworzenia rekomendacji zarówno filtrowanie kolaboratywne jak i metody oparte na treści.

Niech:

- n_u oznacza liczbę użytkowników,
- n_m oznacza liczbę elementów.

Dodatkowo wprowadźmy parametry:

- $r(i, j)$, gdzie $r(i, j) = 1$, jeżeli użytkownik j ocenił element i ,
- $y(i, j)$ - ocena jaką użytkownik j przydzielił elementowi i (ocenę definiujemy tylko wtedy gdy $r(i, j) = 1$).

Znając wektory preferencji użytkowników $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(n_u)}$ odnajdźmy wektor cech konkretnego elementu $g^{(i)}$, gdzie $i \in \{1, \dots, n_m\}$? Minimalizując następujące równanie:

$$\min_{g^{(i)}} \frac{1}{2} \sum_{j:r(i,j)=1} ((\Theta^{(j)})^T g^{(i)} - y(i, j))^2 + \frac{\lambda}{2} \sum_{k=1}^n (g_k^{(i)})^2.$$

Aby odnaleźć wektory cech $g^{(1)}, g^{(2)}, \dots, g^{(n_m)}$ wszystkich elementów równanie, które minimalizujemy przyjmuje następującą postać:

$$\min_{g^{(1)}, \dots, g^{(n_m)}} \frac{1}{2} \sum_{i=1}^{n_m} \sum_{j:r(i,j)=1} ((\Theta^{(j)})^T g^{(i)} - y(i, j))^2 + \frac{\lambda}{2} \sum_{i=1}^{n_m} \sum_{k=1}^n (g_k^{(i)})^2.$$

Znając wektory cech elementów i oceny wystawione przez użytkowników możemy przewidzieć wektory preferencji użytkowników $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(n_u)}$ poprzez:

$$\min_{\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(n_u)}} \frac{1}{2} \sum_{i=1}^{n_u} \sum_{j:r(i,j)=1} ((\Theta^{(j)})^T g^{(i)} - y(i, j))^2 + \frac{\lambda}{2} \sum_{i=1}^{n_u} \sum_{k=1}^n (\Theta^{(j)})^2.$$

Ponownie znając $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(n_u)}$ możemy wyestymować $g^{(1)}, g^{(2)}, \dots, g^{(n_m)}$.

Powtarzając powyższe kroki odnajdowania wektorów cech elementów i wektorów preferencji użytkowników stworzymy algorytm, który z każdą iteracją będzie dokładniej estymował szukane wartości.

4.4 Systemy rekomendujące kontekstowe (Context – aware recommender systems):

Systemy rekomendujące kontekstowe są systemami rekomendującymi opartymi na treści w których zostaje uwzględniony dodatkowy wymiar zwany kontekstem. Upřednio 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:

$$R : U \times S \times K \rightarrow Z$$

Definicja 4.16 (Kontekst [6, Sec 3.3.1.1]). *Kontekstem nazywamy wektor preferencji.*

Algorytm 4.17. *W modelu kontekstowym rekomendacje są generowane w dwóch krokach:*

1. *Metody systemów rekomendujących opartych na treści służące do wygenerowania listy rekomendacji bazującej na preferencjach użytkownika.*
2. *Odfiltrowanie rekomendacji, które odpowiadają przyjętemu kontekstowi. Wyróżniamy tutaj dwa warianty. W pierwszym etap filtrowania zostaje dokonany na końcu, natomiast w drugim filtrowanie jest etapem wstępnym do tworzenia rekomendacji.*

Filtrowanie jako etap wstępny (ang. Pre-Filtering) W tym podejściu informacje kontekstowe używane są do odfiltrowania najbardziej istotnych informacji i konstruowania dwuwymiarowego zbioru danych. Bardzo dużą zaletą tego podejścia jest możliwość implementacji w kolejnym kroku wcześniej opisanych metod rekomendacji.

Filtrowanie jako etap końcowy (ang. Post-Filtering) Informacje o kontekście są ignorowane w wejściowych danych, a rekomendacja dokonywana jest na całym

zbiorze. To w następnym kroku lista rekomendacji stworzona dla użytkownika jest zawężana przez uwzględnienie kontekstu.

Rozdział 5

Eksperymenty / część praktyczne

5.1 Faktoryzacja macierzy w regułach rekomendujących

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 elementów. Otrzymujemy $n \times m$ - wymiarową macierz R , w której wyraz r_{ui} jest oceną wystawioną przez użytkownika u elementowi i , $i \in \{1, \dots, m\}$. Naszym celem jest wypełnić macierz R brakującymi ocenami.

W przypadku filtrowania kolaboratywnego znaną metodą rozwiązania takiego problemu jest faktoryzacja macierzy.

Definicja 5.1 (Faktoryzacja macierzy [11]). *Faktoryzacją macierzy nazywamy rozkład macierzy na produkt macierzy.*

W kontekście rekomendacji algorytm faktoryzacji macierzy działa przez dekompozycję macierzy R w produkt dwóch macierzy X i Y . X jest macierzą użytkowników (wiersze reprezentują poszczególnych użytkowników), natomiast macierz Y to macierz elementów (kolumny oznaczają elementy).

Definicja 5.2 (Postać macierzowa faktoryzacji macierzy [11]). *Niech $x_1, \dots, x_n \in \mathbb{R}^k$ będą faktorem użytkowników oraz $y_1, \dots, y_m \in \mathbb{R}^k$ będą faktorem elementów. Wtedy $k \times n$ - wymiarowa macierz użytkowników X oraz $k \times m$ - wymiarowa macierz elementów*

Y mają następujące postaci:

$$X = \begin{bmatrix} | & & | \\ x_1 & \dots & x_n \\ | & & | \end{bmatrix},$$
$$Y = \begin{bmatrix} | & & | \\ y_1 & \dots & y_m \\ | & & | \end{bmatrix}.$$

5.2 ALS z Apache Spark i MLlib

5.2.1 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.

5.2.2 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 X i Y .

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

$$\min_{X,Y} \sum_{r_{ui}} (r_{ui} - x_u^T y_i)^2 + \lambda (\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2),$$

gdzie

- r_{ui} są znanymi ocenami wystawionymi przez użytkowników,
- λ jest czynnikiem regulującym.

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

Algorytm 5.3 (ALS [11]). *Zainicjowanie X , Y .*

Powtarzamy:

- *for $u = 1, \dots, n$ wykonujemy:*

$$x_u = (\sum_{r_{ui} \in r_{u*}} y_i y_i^T + \lambda I_k)^{(-1)} \sum_{r_{ui} \in r_{u*}} r_{ui} y_i$$

koniec for

- *for $i = 1, \dots, m$ wykonujemy:*

$$y_i = (\sum_{r_{ui} \in r_{*i}} x_u x_u^T + \lambda I_k)^{(-1)} \sum_{r_{ui} \in r_{*i}} r_{ui} x_u$$

koniec for

do momentu zbieżności.

5.2.3 Implementacja algorytmu

Rozdział 6

Podsumowanie

Bibliografia

- [1] *Wstęp do analizy matematycznej funkcje jednej zmiennej.*
- [2] *Wykład z teorii prawdopodobieństwa.* Wydawnictwo Naukowe - Techniczne, 2000.
- [3] *Elementy algebry liniowej część I i II.* Wydawnictwo Naukowe - Techniczne, 2002.
- [4] *Statystyczna analiza danych z wykorzystaniem programu R.* Wydawnictwo Naukowe PWN, 2009.
- [5] *Algebra liniowa z elementami geometrii analitycznej.* Politechnika Łódzka, 2011.
- [6] *Recommender Systems Handbook.* Springer, 2011.
- [7] *Eksploracja danych. Metody i algorytmy.* Wydawnictwo Naukowe PWN, 2013.
- [8] *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] Stefanowski J. Svm - support vector machines. metoda wektorów nośnych.
- [13] Morzy T. Eksploracja danych: problemy i rozwiązania. Instytut Informatyki Politechniki Poznańskiej, 1999.
- [14] Kidziński Ł. Statistical foundation of recommender systems. Master's thesis, University of Warsaw, 2011.