Metody Matematyczne w Uczeniu Maszynowym

Soma Dutta

Uniwersytet Warmińsko-Mazurski w Olsztynie

Semestr Zimowy, 2025

Trzy podstawowe koncepcje uczenia maszynowego

1. Dane (Data): Uczenie maszynowe jest z natury oparte na danych.
2. Model: Model zapewnia formalny sposób opisywania danych, ich wzorców, ich relacji z powiązanymi koncepcjami itp.
3. Uczenie się (Learning): Uczenie się to proces ulepszania wydajności klasyfikowania niewidzianych danych na podstawie wzorców widzianych danych poprzez optymalizację parametrów modelu

Przykład: Z dużego korpusu dokumentów automatyczne wyszukiwanie odpowiednich tematów udostępnianych w dokumentach

Cztery filary uczenia maszynowego

1. Pierwszym krokiem do uczenia maszynowego jest reprezentacja i agregacja danych: Algebra liniowa
2. Drugim krokiem jest zaprojektowanie modelu, który może opisać naturę i zachowanie danych, które często są niepewne, oraz ocenić wydajność modelu w niewidzianych przypadkach testowych: Prawdopodobieństwo i statystyka
3. Trzecim krokiem jest nauczenie się, jak można poprawić wydajność, minimalizując błąd klasyfikacji danych: Rachunek różniczkowy
4. Ponadto niektóre techniki optymalizacyjne i wiedza z zakresu matematyki dyskretnej pomagają dostroić niektóre parametry wykorzystywane w algorytmach uczenia się i zwiększyć wydajność algorytmów.

Ogólny Plan

- Poznamy każdy z istotnych tematów matematycznych
- Dowiemy się, gdzie i jak taka wiedza matematyczna jest wykorzystywana w uczeniu maszynowym
- Podczas sesji ćwiczeniowych będą wykonywane ćwiczenia związane z rozwiązywaniem problemów (na papierze) z zakresu matematyki oraz ćwiczenia odnoszące się do praktyki (pisanie programów).

Bibliografia

- Matematyka w uczeniu maszynowym Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, Helion 2022.
- Matematyczne modele uczenia maszynowego w językach MATLAB i PYTHON. Osowski, Stanisław, Robert Szmurło; Oficyna Wydawnicza Politechniki Warszawskiej; 2023

Algebra Liniowa

- Dane liczbowe są reprezentowane przez wektor, a tabela takich danych jest reprezentowana przez macierz.
- Każdy wektor przedstawia opis obiektu lub sytuacji w świecie rzeczywistym. Oczekuje się, że wektory o podobnej charakterystyce dadzą podobne wyniki, gdy stosuje się algorytmy uczenia maszynowego.
- Potrzebujemy więc operacji przedstawiających podobieństwo wektorów.
- Wektory są szczególnymi przypadkami macierzy.

Treści: Algebra Liniowa

1. Macierze i ich operacje
2. Układ równań liniowych
3. Przestrzeń wektorowa
4. Transformacja liniowa (z jednej przestrzeni wektorowej do innej przestrzeni wektorowej)

Macierze

Biorąc pod uwagę $m, n \in \mathbf{N}$, macierz o wartościach rzeczywistych A formatu $m \times n$ jest $m.n$-krotna i ma elementy a_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, które są uporządkowane zgodnie z prostokątną tablicą składającą się z m wierszy i n kolumn:

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

a_{kl} jest elementem wejściowym k-tego wiersza i l-tej kolumny.

Różne rodzaje macierzy

- Macierz kwadratowa: gdy liczba wierszy i kolumn jest taka sama, powiedżmy n, nazywamy to macierzą kwadratową stopnia n
- Macierz diagonalna: macierz kwadratowa, której jedynymi elementami niezerowymi są elementy na przekątnej
- Macierz jednostka: macierz kwadratowa posiadająca jedynki (1) na głównej przekątnej
- Macierz trójkątna górna: macierz kwadratowa posiadająca zera poniżej głównej przekątnej
- Macierz trójkątna dolna macierz kwadratowa posiadająca zera powyżej głównej przekątnej
- Macierze jednokolumnowe nazywamy wektorami

Operacje na macierzach

- Transpozycja macierzy: Mając daną macierz A formatu $m \times n$, A^t jest macierzą uzyskaną z A przez zamianę wierszy i kolumn macierzy oryginalnej.

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

$$A^t = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}$$

- Macierz kwadratowa A formatu n jest symetryczna, jeśli $A = A^t$.

Dodawanie macierzy

Suma dwóch macierzy A i B formatu $m \times n$ jest definiowana jako suma elementów.

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

$$\mathrm{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mn} \end{pmatrix}$$

$$\mathrm{A} + \mathrm{B} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2n} + b_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \dots & a_{mn} + b_{mn} \end{pmatrix}$$

Właściwości dodawania macierzy

Przypuszczajmy $A = [a_{ij}]$, $B = [b_{ij}]$ i $C = [c_{ij}]$, a każda ma format $m \times n$.

1. (Przemienność) Kolejność dodawania macierzy nie ma znaczenia
   $A + B = B + A$
2. (Łączność) Sposób grupowania macierzy przy dodawaniu nie wpływa na wynik
   $A + (B + C) = (A + B) + C$
3. (Element neutralny): Istnieje unikalna macierz zerowa, która po dodaniu do dowolnej innej macierzy A nie zmienia jej wartości
   $A + \mathbf{0} = A$, gdzie $\mathbf{0}$ to macierz zerowa formatu $m \times n$
4. (Macierz przeciwna) Dla każdej macierzy A można znaleźć macierz przeciwną (-A), która po dodaniu do macierzy A daje macierz zerową
   $A + (-A) = \mathbf{0}$, gdzie $(-A) = [-a_{ij}]$

Mnożenie macierzy

Niech A będzie macierzą formatu $m \times n$, a B macierzą formatu $n \times p$. Iloczyn A i B jest macierzą formatu $m \times p$ zdefiniowaną w następujący sposób.

$$\mathrm{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

$$\mathrm{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{pmatrix}$$

$$\mathrm{AB} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mp} \end{pmatrix}$$

gdzie $c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}$, $i = 1, \dots m$, $j = 1, \dots, p$

Właściwości mnożenia macierzy

1. Nieprzemienny (i) Przypuszczamy $A = [a_{ij}]$ z formatu 2×3 i $B = [b_{ij}]$ z formatu 3×4. Wtedy AB istnieje, ale BA nie istnieje.
   (ii) Nawet jeśli AB i BA istnieją, mogą nie być tym samym.
2. (Łączność) $A(BC) = (AB)C$
   pod warunkiem $A = [a_{ij}]$ formatu $m \times n$, $B = [b_{jk}]$ formatu $n \times p$ i $C = [c_{kl}]$ formatu $p \times s$
3. (Element neutralny) Jeśli $A = [a_{ij}]$ z formatu $m \times n$, wtedy $AI_n = A = I_m A$ gdzie I_m i I_n są to macierze jednostkowe stopnia m i n odpowiednio.
4. Nie dla każdej macierzy A istnieje macierz odwrotna.
5. (Rozdzielność względem dodawania) Dla dwóch macierzy A i B formatu $m \times n$ i dwóch macierzy C i D formatu $n \times p$ $(A + B)C = AC + BC$ i $A(C + D) = AC + AD$

Przykłady mnożenia macierzy w uczeniu maszynowym

- Załóżmy, że prosta sieć neuronowa ma wektor wejściowy $\vec{x} = (1, 0, 0, 5)$ i macierz wagową $W = \begin{pmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{pmatrix}$.

  Do obliczenia wyniku warstwy ukrytej potrzebujemy $\vec{x}.W = (1.0,\ 0.5).\begin{pmatrix} 0.2 & 0.8 \\ 0.6 & 0.4 \end{pmatrix}$.

Wyznacznik macierzy kwadratowej

- Niektóre macierze kwadratowe posiadają macierze odwrotne
- Macierz kwadratową, A stopnia n dla której istnieje macierz odwrotną A^{-1} stopnia n, nazywamy odwracalną, albo nieosobliwą. Macierz, która nie posiada macierzy odwrotnej nazywamy osobliwą.
- Warunek nieosobliwości macierzy
  - Dla macierzy kwadratowej A rzędu n wyznacznik macierzy A, oznaczany jako $|A|$ lub $det(A)$, jest liczbą (rzeczywistą) zdefiniowaną na podstawie różnych możliwych permutacji na $\{1, 2, \ldots, n\}$.
  - Przykład: $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ Możliwe permutacje na $\{1, 2\}$ to $1 \rightarrow 1$, $2 \rightarrow 2$ i $1 \rightarrow 2$, $2 \rightarrow 1$.
  - Każda permutacja ma znak (dodatni lub ujemny) i określa, które elementy należy pomnożyć. Na przykład mnożymy $a_{11}a_{22}$ i $a_{12}a_{21}$. Następnie sumujemy iloczyn z ich znakami.
  - Szczególnie $|A| = a_{11}a_{22}$ - $a_{12}a_{21}$.
  - Macierz A ma macierz odwrotną, jeśli $|A| \neq 0$, tzn., A nie jest osobliwa.

Permutacje na skończonym zbiorze liczb naturalnych i wyznacznika

- Niech A będzie macierzą kwadratową o rozmiarze n.
- Wyznacznik A, oznaczony jako $det(A)$ lub $|A|$, jest zdefiniowany jako: $|A| = \Sigma_{\pi \in S_n} Sign(\pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)}$, gdzie S_n jest zbiorem wszystkich permutacji na $\{1, 2, 3, \ldots, n\}$..
- Przykład: $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ Możliwe permutacje na $\{1, 2\}$: $\pi_1 = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$ i $\pi_2 = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$. π_1 nie ma pary odwrotnej, a π_2 ma jedną parę odwrotną. Więc $Sign(\pi_1) = (-1)^0$ (znak dodatni) i $Sign(\pi_2) = (-1)^1$ (znak ujemny). Wyznacznik A to $|A| = a_{11}a_{22}$ - $a_{12}a_{21}$.

Odwrotność macierzy

- Dla macierzy kwadratowej A istnieje macierz odwrotna, oznaczana A^{-1}, jeśli $|A| \neq 0$.
- Niech A $= \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$ i $|A| \neq 0$.
- $A^{-1} = \frac{1}{|A|} Adj(A)^t$ gdzie $Adj(A)$ (ang. Adjoint of A) to macierz dołączona macierzu A, zdefiniowana jako:

  $Adj(A) = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix}$ i A_{ij} to kofaktor który odnosi się do wyznacznika podmacierzy otrzymanej przez usunięcie i-tego wiersza i j-tej kolumny z A i pomnożenie wartości przez znak $(-1)^{i+j}$.

Właściwości transpozycji i odwrotności

Niech A i B będą macierzami kwadratowymi o rozmiarze n.

- $AA^{-1} = I_n = A^{-1}A$.
- $(AB)^{-1} = B^{-1}A^{-1}$.
- $(A+B)^{-1} \neq A^{-1} + B^{-1}$.

Załóżmy, że A i B są macierzami o rozmiarze $m \times n$.

- $(A^t)^t = A$
- $(A+B)^t = A^t + B^t$.

Załóżmy, że A i B są macierzami kwadratowymi o rozmiarze $m \times n$ i $n \times p$ odpowiednio.

- $(AB)^t = B^t A^t$.

Operacja mnożenia skalarnego na macierzach

Niech $\mathcal{M}_{m\times n}$ będzie zbiorem wszystkich macierzy o rozmiarze $m \times n$ i $\lambda, \delta \in \mathbb{R}$ będą skalarami.

- $\lambda(\delta A) = \lambda\delta(A)$, $A \in \mathcal{M}_{m\times n}$.
- $\lambda(BC) = (\lambda B)C = B(\lambda C) = (BC)\lambda$, $B \in \mathcal{M}_{m\times n}$, $C \in \mathcal{M}_{n\times p}$.
- $\lambda(A + B) = \lambda A + \lambda B$, $A, B \in \mathcal{M}_{m\times n}$.
- $(\lambda + \delta)A = \lambda A + \delta A$, $A \in \mathcal{M}_{m\times n}$.

Taką operację mnożenia liczby rzeczywistej przez macierz można uważać za $\circ : \mathbb{R} \times \mathcal{M} \mapsto \mathcal{M}$ gdzie $\mathcal{M}$ jest zbiorem wszystkich macierzy i jest znana jako operacja zewnętrzna zwracająca macierz.

Układ równań liniowych i reprezentacja macierzowa

- Układ równań liniowych o liczbie równań m i liczbie zmiennych n:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m$$

Liczby a_{ij} nazywamy i współczynnikami układu, a x_j to niewiadome lub zmienne ($i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$).

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$AX = B$$

- Jeśli B to macierz zerowa $AX = B$ to układ jednorodny (ang. homogeneous) równań liniowych, a inaczej to układ niejednorodny (ang. non-homogeneous) równań liniowych.

Rozwiązywanie układu równań liniowych metodą odwracania macierzy

- W przypadku równań liniowych o tej samej liczbie zmiennych i równań (tzn. $m = n$)

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

- Jeśli $|A| \neq 0$, znajdujemy A^{-1}.
- $\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = A^{-1}B$

Elementarne działanie na macierzach

Kluczem do rozwiązania układu równań liniowych są elementarne przekształcenia, które zachowują ten sam zbiór rozwiązań, ale przekształcają układ równań do prostszej postaci

- E1 Wymiana dwóch wierszy w macierzy
- E2 Mnożenie wiersza w macierzy przez stałą (skalar) $\lambda \neq 0$
- E3 Dodawanie dwóch wierszy macierzy

Wierszowa forma schodkowa macierzy (Row Echelon Form)

Niech A będzie macierz wymiaru $m \times n$ gdzie oznaczmy i-ty wiersz A jako $A^{(i,.)}$, a j-ta kolumna A jako $A^{(.,j)}$.

- REF: Macierz A jest w formie schodkowej (REF), jeśli A spełnia następujący warunek.
  - Albo $A^{(1,.)}$ jest wierszem zerowym, albo pierwszy wpis różny od zera (od lewej do prawej) jest 1.
  - Dla dowolnego $i = 1, 2, \ldots \mathrm{m}$, jeśli $A^{(i,.)}$ jest wierszem zerowym, to wszystkie kolejne wiersze będą wierszami zerowymi.
  - Dla dowolnego $i = 1, 2, \ldots, \mathrm{m}$, jeśli $A^{(i,.)}$ nie jest wierszem zerowym, to pierwszym niezerowym wpisem (od lewej do prawej) jest 1 i występuje on na prawo od początkowej 1 w $A^{(i-1,.)}$.
- Początkowy element 1 w każdym wierszu, który nie jest wierszem zerowym, nazywany jest wiodącym elementem (ang. pivot element).
- RREF: Zredukowana forma schodkowa w macierzy: Każda kolumna $A^{(.,j)}$ zawierająca element wiodący ma wszystkie pozostałe elementy równe zero.

Rząd macierzy

- Rząd macierzy to liczba niezerowych wierszy w jej postaci schodkowej (lub schodkowej zredukowanej)
- $A_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$

  A_1 jest w postaci REF i rząd$(A_1) = 3$
- $A_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$

  A_2 jest w postaci REF i rząd$(A_2) = 3$

Uzyskanie RREF macierzy

- $\begin{pmatrix} 2 & 5 & 3 \\ 1 & 2 & 3 \\ 1 & 0 & 8 \end{pmatrix} \longrightarrow^{R^1 \leftrightarrow R^2} \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 3 \\ 1 & 0 & 8 \end{pmatrix} \longrightarrow^{R^2+(-2)R^1,\ R^3+(-1)R^1}$
$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -3 \\ 0 & -2 & 5 \end{pmatrix} \longrightarrow^{R^3+(2)R^2} \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & -1 \end{pmatrix} \longrightarrow^{(-1).R^3}$
$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{pmatrix}$

Rozwiązywanie układu równań liniowych za pomocą elementarnych operacji na wierszach

Układ równań liniowych o liczbie równań m i liczbie zmiennych n:

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m$$

$$\hookrightarrow \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

- Przekształć macierz rozszerzoną $[A|B]$ do postaci schodkowej wierszowej gdzie $[A|B] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{array}\right)$

Przykład

$$2x_1 + 5x_2 + 3x_3 = 5$$
$$x_1 + 2x_2 + 3x_3 = 4$$
$$x_1 + 8x_3 = 9$$

- $[A|B] = \left(\begin{array}{ccc|c} 2 & 5 & 3 & 5 \\ 1 & 2 & 3 & 4 \\ 1 & 0 & 8 & 9 \end{array}\right)$
- Konwertuj do formy schodkowej aby uzyskać rozwiązanie $x_1 = 1$, $x_2 = 0$ i $x_3 = 1$. item Proces ten polegający na przekształceniu układu równań liniowych do postaci schodkowej zredukowanej (RREF) i znalezieniu rozwiązań dla układu równań liniowych nazywa się algorytmem eliminacji Gaussa.

Rozwiązania jednorodnych i niejednorodnych układów równań liniowych

- Jednorodnych układ równań liniowych zawsze ma co najmniej jedno rozwiązanie w którym wszystkie zmienne mają wartość zero. Może mieć nieskończenie wiele rozwiązań, łącznie z rozwiązaniem zawierającym wszystkie wartości zerowe.
- Niejednorodnych układ równań liniowych może mieć jedno rozwiązanie, nieskończenie wiele rozwiązań lub nie mieć rozwiązań.

Warunki istnienia rozwiązań układu równań liniowych

$$a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1$$
$$a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2$$
$$\vdots$$
$$a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m$$

- $[A|B] = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{array}\right)$
- Znajdź rząd A i rząd $[A|B]$.
  - Jeśli rząd$([A|B])$ = rząd(A) wtedy układ równań liniowych ma rozwiązania. Dodatkowo, jeśli $m < n$, to rozwiązań jest nieskończenie wiele. W przeciwnym razie istnieje dokładnie jedno rozwiązanie.
  - Jeśli rząd$([A|B])$ $\neq$ rząd(A) wtedy układ równań liniowych jest niespójne i nie ma żadnego rozwiązania.

Przykłady

1. $x_1 - 2x_2 + x_3 - x_4 = 3$
   $2x_1 - 4x_2 + x_3 = x_4 = 2$
   $x_1 - 2x_2 - 2x_3 + 3x_4 = 1$

- $[A|B]$ (REF) $= \begin{pmatrix} 1 & -2 & 0 & 0 & |3 \\ 0 & 0 & 1 & 0 & |-2 \\ 0 & 0 & 0 & 1 & |-2 \end{pmatrix}$
- Ma nieskończenie wiele rozwiązań, ponieważ x_1, x_3, x_4 są zmiennymi wiodącymi, a x_2 jest zmienną wolną.
- Przestrzeń rozwiązań $= \{(3 + 2x_2, x_2, -2, -2) : x_2 \in \mathbb{R}\}$.

Odwrotność macierzy poprzez zredukowaną formę schodkową wierszową

- Niech A będzie macierzą kwadratową rzędu n i szukamy X takiego, że $AX = I_n$.
- Zaczynamy od kroku: $[A|I_n]$

$$\left(\begin{array}{cccc|cccc} a_{11} & a_{12} & \dots & a_{1n} & 1 & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & a_{2n} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & 0 & 0 & \dots & 1 \end{array}\right)$$

- Konwertuj macierz rozszerzoną w formacie RREF (zredukowana forma wierszowa Eechelon)
- $[I_n|A^{-1}]$

Przykład

- Niech $A = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$
  - $\left(\begin{array}{cccc|cccc} 1 & 0 & 2 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}\right)$

$\downarrow$ RREF

  - $\left(\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -1 & 2 & -2 & 2 \\ 0 & 1 & 0 & 0 & 1 & -1 & 2 & -2 \\ 0 & 0 & 1 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 0 & -1 & 2 \end{array}\right)$
- $A = \begin{pmatrix} -1 & 2 & -2 & 2 \\ 1 & -1 & 2 & -2 \\ 1 & -1 & 1 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix}$

Wektory: specjalne formy macierzy

- Wektor n-wymiarowy można uważać albo za macierz o rozmiarze $1 \times n$, albo o rozmiarze $n \times 1$. Pierwszy typ możemy nazwać wektorami wierszowymi, a drugi typem wektorami kolumnowymi.
- $\mathbb{R}^n = \{(a_1, a_2, \ldots, a_n) : a_i \in \mathbb{R}, 1 \leqslant i \leqslant n\}$ zawiera takie n-wymiarowe (wierszowe) wektory i nazywamy je przestrzenią euklidesową o wymiarze n.
- Tak więc operacje takie jak dodawanie macierzy i mnożenie skalarne są możliwe w $\mathbb{R}^n$ - zwracając wektor z $\mathbb{R}^n$.
- Zanim zaczniemy uczyć się o przestrzeni wektorowej, zobaczmy, jaką strukturę algebraiczną tworzy taka przestrzeń.

Podstawowe struktury algebraiczne

- Grupa: Mając zbiór G i operację binarną $*$ na G, $(G, *)$ jest nazywana grupą, jeśli następujące warunki są spełnione.
  - (i) $x * (y * z) = (x * y) * z$ dla dowolnych $x, y, z \in G$
  - (ii) Isnieje $e \in G$ takie, że dla każdego $x \in G$, $x * e = x = e * x$.
  - (iii) Dla każdego $x \in G$ istnieje $x^{-1} \in G$ takie, że $x * x^{-1} = e = x^{-1} * x$.

  $(G, *)$ nazywana jest grupą abelową (przemienną), jeśli dla dowolnych $x, y \in G$, $x * y = y * x$.
- Przykłady: $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$, $(\mathbb{R} \setminus \{0\}, .)$, $(\mathcal{M}_{m \times n}, +)$

  Ciało: Ciałem nazywamy zbiór F z dwoma dwuargumentowymi działaniami wewnętrznymi, $\oplus$ i $\odot$, oznaczamy przez $(F, \oplus, \odot)$, spełniające następujące warunki.
  - (i) $(F, \oplus)$ to grupa przemienna
  - (ii) $(F \setminus \{0\}, \odot)$ to grupa przemienna gdzie 0 to neutralny element w odniesieniu do operacji $\oplus$.
  - (iii) $x \odot (y \oplus z) = (x \odot y) \oplus (x \odot z)$, dla dowolnych $x, y, z \in F$.
- Przykład: $(\mathbb{R}, +, .)$

Przestrzeń liniowa nad ciałem liczb rzeczywistych

Pojęcie przestrzeni liniowej definiuje się za pomocą dwóch struktur algebraicznych – wektory tworzą grupę przemienną, a skalary pochodzą z ciała K. Tutaj skupimy się na $K = \mathbb{R}$.

Liniowa (wektorowa) przestrzeń $\mathcal{V}$ nad ciałem liczb rzeczywistych, oznaczana przez $(\mathcal{V}, +)$, spełnia następujące własności.

(i) $(\mathcal{V}, +)$ jest grupą przemienną.

(ii) Istnieje operacja zewnętrzna $. : \mathbb{R} \times \mathcal{V} \mapsto \mathbb{R}$ taka, że spełnione są następujące warunki.

- $\lambda(\overrightarrow{x + y}) = \lambda\overrightarrow{x} + \lambda\overrightarrow{y},\ \overrightarrow{x}, \overrightarrow{y} \in \mathcal{V}$ i $\lambda \in \mathbb{R}$.
- $(\lambda + \delta)\overrightarrow{x} = \lambda\overrightarrow{x} + \delta\overrightarrow{x},\ \overrightarrow{x} \in \mathcal{V}$ i $\lambda, \delta \in \mathbb{R}$.
- $\lambda(\delta\overrightarrow{x}) = (\lambda\delta)\overrightarrow{x},\ \overrightarrow{x} \in \mathcal{V}$ i $\lambda, \delta \in \mathbb{R}$.
- $1.\overrightarrow{x} = \overrightarrow{x},\ \overrightarrow{x} \in \mathcal{V}, 1 \in \mathbb{R}$.

Elementy $x \in \mathcal{V}$ nazywane są wektorami. Elementem neutralnym $(\mathcal{V}, +)$ jest wektor zerowy $\overrightarrow{0} = [0, ..., 0]$, a działanie wewnętrzne $+$ nazywane jest dodawaniem wektorów. Elementy $\lambda \in \mathbb{R}$ nazywane są skalarami, a działanie zewnętrzne to mnożenie przez skalary.

Przykłady

- $(\mathcal{M}_{m\times n}, +)$ nad ciałem $\mathbb{R}$ jest przestrzenią liniową.
- $(\mathbb{R}^n, +) = (\mathcal{M}_{1\times n}, +)$ nad ciałem $\mathbb{R}$ jest przestrzenią liniową. Dla dowolnych $\overrightarrow{x}, \overrightarrow{y} \in \mathbb{R}^n$, $\overrightarrow{x} = [x_1, x_2, \ldots, x_n]$, $\overrightarrow{y} = [y_1, y_2, \ldots, y_n]$, mamy
  (i) $\overrightarrow{x} + \overrightarrow{y} = [x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n]$ oraz
  (ii) $\lambda\overrightarrow{x} = [\lambda x_1, \lambda x_2, \ldots, \lambda x_n]$.

Przetwarzanie danych z wykorzystaniem operacji z przestrzeni liniowej

- Wyobraź sobie, że otwierasz zbiór danych: wiersze rekordów klientów, kolumny takie jak wiek, dochód, wydatków.
- Jeśli potraktujesz każdy wiersz jako wektor, od razu wkraczasz na teren algebry liniowej. A kiedy połączysz wszystkie wiersze (wiersze $\times$ kolumny), otrzymasz macierz.
- Aby każda jednostka wartości dla wszystkich wyżej wymienionych cech kolumn była standaryzowana, potrzebna jest operacja mnożenia skalarnego przestrzeni liniowej.
- Aby sprawdzić, czy dwa punkty danych reprezentujące informacje dwóch klientów są zależne czy nie, pomocne może okazać się mnożenie sclara wraz z dodawaniem wektorów.

Podprzestrzenie liniowe

- Definicja: Niech $(\mathcal{V}, +)$ będzie przestrzenią liniową nad $\mathbb{R}$ i $\mathcal{U} \subseteq \mathcal{V}$. $(\mathcal{U}, +)$ nazywa się liniową podprzestrzenią $\mathcal{V}$, jeśli $(\mathcal{U}, +)$ jest przestrzenią liniową nad $\mathbb{R}$.
- Jeśli $\mathcal{U}$ jest podprzestrzenią przestrzeni liniowej $\mathcal{V}$ nad ciałem liczb rzeczywistych, możemy ją oznaczyć przez $\mathcal{U}_{|\mathbb{R}} \subseteq \mathcal{V}_{|\mathbb{R}}$.
- Twierzenie: Niech $(\mathcal{V}, +)$ będzie przestrzenią liniową nad $\mathbb{R}$ i $\mathcal{U} \subseteq \mathcal{V}$. $\mathcal{U}$ jest liniową podprzestrzenią $\mathcal{V}$, jeśli
  (i) $\overrightarrow{0} \in \mathcal{U}$,
  (ii) dla dowolnych $\overrightarrow{x}, \overrightarrow{y} \in \mathcal{U}$, $\overrightarrow{x+y} \in \mathcal{U}$, i
  (iii) dla dowolnej $\lambda \in \mathbb{R}$, $\overrightarrow{x} \in \mathcal{U}$, $\lambda\overrightarrow{x} \in \mathcal{U}$.
- Twierdzenie: Część wspólna dowolnej rodziny podprzestrzeni przestrzeni liniowej $\mathcal{V}_{|\mathbb{R}}$ jest też podprzestrzenią $\mathcal{V}_{|\mathbb{R}}$.

Przykłady podprzestrzeni z macierzy

- Wiemy już, że $(\mathcal{M}_{m\times n}, +)$ jest przestrzenią liniową nad $\mathbb{R}$. Oznaczmy tę szczególną przestrzeń liniową jako $\mathcal{M}_{m\times n|\mathbb{R}}$.
- Rozważmy najpierw następujące podzbiory $\mathcal{M}_{m\times n}$.
  - Zbiór wszystkich macierzy trójkątnych górnych: $UTri^{m,n} = \{A \in \mathcal{M}_{m\times n} : \forall i > j, a_{ij} = 0\}$
  - Zbiór wszystkich dolnych macierzy trójkątnych: $LTri^{m,n} = \{A \in \mathcal{M}_{m\times n} : \forall i < j, a_{ij} = 0\}$
  - Zbiór wszystkich macierzy przekątnych: $Diag^{m,n} = \{A \in \mathcal{M}_{m\times n} : \forall i \neq j, a_{ij} = 0\}$
- $UTri^{m,n}_{|\mathbb{R}} \subseteq \mathcal{M}_{m\times n|\mathbb{R}}$
- $LTri^{m,n}_{|\mathbb{R}} \subseteq \mathcal{M}_{m\times n|\mathbb{R}}$
- $Diag^{m,n}_{|\mathbb{R}} \subseteq \mathcal{M}_{m\times n|\mathbb{R}}$

Rozpiętość przestrzeni liniowej

Niech $\mathcal{V}_{|\mathbb{R}}$ będzie przestrzenią liniową i $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\} \subseteq \mathcal{V}$.

- Kombinacja liniowa: Wektor $\overrightarrow{v} \in \mathcal{V}$ nazywamy kombinacją liniową $\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}$, jeśli istnieją skalary $a_1, a_2, \ldots, a_n \in \mathbb{R}$ takie, że $\overrightarrow{v} = \sum_{i=1}^{n} a_i \overrightarrow{v_i}$.
- Rozpiętość zbioru wektorów $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}$, oznaczona jako $Span(\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\})$, to zbiór wektorów, które można wyrazić jako kombinację liniową $\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}$. Tzn.,
  $Span(\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}) = \{\overrightarrow{v} \in \mathcal{V} : \overrightarrow{v} = \sum_{i=1}^{n} a_i \overrightarrow{v_i}\}$
- Przykład: Rozważmy przestrzeń liniową $\mathbb{R}^2_{|\mathbb{R}}$ i wektory $(0,1), (1,1) \in \mathbb{R}^2_{|\mathbb{R}}$. Niech $(x,y) \in \mathbb{R}^2_{|\mathbb{R}}$.
  - Sprawdzijmy, czy (x,y) jest kombinacją liniową (0, 1) i (1, 1). Tzn., szukamy takie $a_1, a_2 \in \mathbb{R}$, że $(x,y) = a_1(0,1) + a_2(1,1)$
  - Wtedy mamy układ równań liniowych $\begin{cases} a_2 = x \\ a_1 + a_2 = y \end{cases}$
  - Więc, $(x,y) = (y-x)(0,1) + x(1,1)$.
  - Ponieważ (x,y) jest dowolnym wektorem $\mathbb{R}^2_{|\mathbb{R}}$,
    $Span(\{(0,1),(1,1)\}) = \mathbb{R}^2_{|\mathbb{R}}$.

Liniowa (nie)zależność

Niech $\mathcal{V}_{|\mathbb{R}}$ będzie przestrzenią liniową i $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\} \subseteq \mathcal{V}$.

- $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}$ nazywa się liniowo niezależnym, jeśli rozwiązanie zerowe jest jedynym rozwiązaniem dla układu równań otrzymanych z $a_1\overrightarrow{v_1} + a_2\overrightarrow{v_2} + \ldots + a_n\overrightarrow{v_n} = \overrightarrow{0}$. (tzn., jedynym rozwiązaniem jest $a_1 = a_2 = \ldots = a_n = 0$.)
- Jeśli istnieje rozwiązanie niezerowe spełniające $a_1\overrightarrow{v_1} + a_2\overrightarrow{v_2} + \ldots + a_n\overrightarrow{v_n} = \overrightarrow{0}$, wtedy $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}$ nazywa się liniowo zależnym.
- Przykład: Rozważmy przestrzeń liniową $\mathbb{R}^2_{|\mathbb{R}}$ i wektory $(0,1), (1,1) \in \mathbb{R}^2_{|\mathbb{R}}$.
  - Już wiemy, że $(x,y) = (y-x)(0,1) + x(1,1)$.
  - Aby przedstawić $(0,0)$ jako (x,y) mamy tylko jedno rozwiązanie $x = y = 0$.
  - Więc, $\{(0,1),(1,1)\}$ jest liniowo niezależnym.

Baza i wymiar przestrzeni

Definicja: Niech $\mathcal{V}_{|\mathbb{R}}$ będzie przestrzenią liniową i $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\} \subseteq \mathcal{V}$. Dla dowolnej podprzestrzeni $\mathcal{U}_{|\mathbb{R}} \subseteq \mathcal{V}_{|\mathbb{R}}$, $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}$ nazywamy bazą przestrzeni $\mathcal{U}_{|\mathbb{R}}$ jeśli
(i) $\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}$ jest liniowo niezależnym i
(ii) $Span(\{\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n}\}) = \mathcal{U}_{|\mathbb{R}}$.

Twierdzenie: Każda przestrzeń liniowa $\mathcal{U}_{|\mathbb{R}}$ $(\subseteq \mathcal{V}_{|\mathbb{R}})$ ma bazę. Ponadto, wszystkie bazy są równoliczne.

Definicja: Liczbę elementów bazy danej przestrzeni $\mathcal{U}_{|\mathbb{R}}$ nazywamy jej wymiarem i oznaczamy $dim(\mathcal{U}_{|\mathbb{R}})$.

- Przykład: Już zobaczyliśmy $\{(0,1),(1,1)\}$ jest liniowo niezależnym i $Span(\{(0,1),(1,1)\}) = \mathbb{R}^2_{|\mathbb{R}}$. Więc, $\{(0,1),(1,1)\}$ jest bazą przestrzeni $\mathbb{R}^2_{|\mathbb{R}}$ i $dim(\mathbb{R}^2_{|\mathbb{R}}) = 2$.
- Warto zauważyć, że przestrzeń liniowa może mieć różne bazy, ale wszystkie powinny mieć taką samą liczbę elementów.
- Dla przestrzeni $\mathbb{R}^2_{|\mathbb{R}}$, $\{(1,0),(1,0)\}$ jest bazą standardową.

Operacje na przestrzeniach liniowych

- Niech $\mathcal{U}_{|\mathbb{R}}$ i $\mathcal{W}_{|\mathbb{R}}$ będą podprzestrzeniami $\mathcal{V}_{|\mathbb{R}}$. Definiujemy iloczyn i sumę tych podprzestrzeni jako
  - $\mathcal{U}_{|\mathbb{R}} \cap \mathcal{W}_{|\mathbb{R}} = \{\vec{x} \in \mathcal{V}_{|\mathbb{R}} : \vec{x} \in \mathcal{U}_{|\mathbb{R}} \ i \ \vec{x} \in \mathcal{W}_{|\mathbb{R}}\}$
  - $\mathcal{U}_{|\mathbb{R}} \oplus \mathcal{W}_{|\mathbb{R}} = \{\vec{x} \in \mathcal{V}_{|\mathbb{R}} : \vec{x} = \vec{y} + \vec{z}, \vec{y} \in \mathcal{U}_{|\mathbb{R}} \ i \ \vec{z} \in \mathcal{W}_{|\mathbb{R}}\}$.
- Twierdzenie: Zarówno iloczyn $\mathcal{U}_{|\mathbb{R}} \cap \mathcal{W}_{|\mathbb{R}}$ jak i suma $\mathcal{U}_{|\mathbb{R}} \oplus \mathcal{W}_{|\mathbb{R}}$ są podprzestrzeniami $\mathcal{V}_{|\mathbb{R}}$.
- Twierdzenie: Niech $\mathcal{V}_{|\mathbb{R}}$ będzie przestrzenią liniową i $v_1, v_2, \ldots, v_n \in \mathcal{V}_{|\mathbb{R}}$. $Span(\{v_1, v_2, \ldots, v_n\})$ jest podprzestrzenią liniową $\mathcal{V}_{|\mathbb{R}}$.
- Jeśli $\mathcal{V}_{|\mathbb{R}}$ jest skończenie wymiarową przestrzeią liniową, to, dla dowolnej podprzestrzeni $\mathcal{U}_{|\mathbb{R}}$ przestrzeni $\mathcal{V}_{|\mathbb{R}}$, $dim(\mathcal{U}_{|\mathbb{R}}) \leqslant dim(\mathcal{V}_{|\mathbb{R}})$.

Znajdowania bazy $U = Span(\{v_1, v_2, \ldots, v_n\})$

1. Zapisz wektory rozpinające jako kolumny macierzy A.
2. Określ postać schodkową macierzy A.
3. Wektory rozpinające powiązane z wiodącymi kolumnami tworzą bazę U.

   Przykład: Rozważmy przestrzeń liniową $\mathbb{R}^3_{|\mathbb{R}}$. Niech $U = Span(\{(1,0,0),(0,1,0),(1,1,0)\})$. Znajdź bazę i wymiar tej podprzestrzeni.

Znajdowania bazy $U = Span(\{v_1, v_2, \ldots, v_n\})$

1. Zapisz wektory rozpinające jako kolumny macierzy A.
2. Określ postać schodkową macierzy A.
3. Wektory rozpinające powiązane z wiodącymi kolumnami tworzą bazę U.

   Przykład: Rozważmy przestrzeń liniową $\mathbb{R}^3_{|\mathbb{R}}$. Niech $U = Span(\{(1,0,0),(0,1,0),(1,1,0)\})$. Znajdź bazę i wymiar tej podprzestrzeni.

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \hookrightarrow \text{Już jest w formie REF.}$$

   Więc baza U to $\{(1,0,0),(0,1,0)\}$ i ma wymiar 2.

Znajdowania bazy $U = Span(\{v_1, v_2, \ldots, v_n\})$

1. Zapisz wektory rozpinające jako kolumny macierzy A.
2. Określ postać schodkową macierzy A.
3. Wektory rozpinające powiązane z wiodącymi kolumnami tworzą bazę U.

   Przykład: Rozważmy przestrzeń liniową $\mathbb{R}^3_{|\mathbb{R}}$. Niech $U = Span(\{(1,0,0),(0,1,0),(1,1,0)\})$. Znajdź bazę i wymiar tej podprzestrzeni.

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \hookrightarrow \text{Już jest w formie REF.}$$

   Więc baza U to $\{(1,0,0),(0,1,0)\}$ i ma wymiar 2.

$\hookrightarrow$ Warto zauważyć, że liczba liniowo niezależnych kolumn macierzy A jest taka sama jak liczba liniowo niezależnych wierszy macierzy A i jest nazywana rzędem macierzy A.

Właściwości rzędu macierzy

- Dla dowolnej $A = [a_{ij}]_{m \times n}$, $rank(A) = rank(A^t)$.
- Kolumny macierzy A rozpinają podprzestrzeń U przestrzeni liniowej $\mathbb{R}^m_{|\mathbb{R}}$ i $dim(U) = rank(A)$. Bazę U można znaleźć stosując eliminację Gaussa do A w celu identyfikacji kolumn wiodących.
- Wiersze macierzy A rozpinają podprzestrzeń W przestrzeni liniowej $\mathbb{R}^n_{|\mathbb{R}}$ i $dim(W) = rank(A)$. Bazę W można znaleźć stosując eliminację Gaussa do A^t.
- Dla każdej macierzy kwadratowej A rzędu n, A jest regularna (odwracalna) wtedy i tylko wtedy, gdy $rank(A) = n$.

Przestrzeń afiniczna

- Intuicyjnie przestrzeń afiniczną można postrzegać jako przesunięty podzbiór przestrzeni wektorowej, który niekoniecznie zawiera wektor zerowy.
- Matematycznie przestrzeń afiniczną można przedstawić jako: $A = \vec{v} + W$ gdzie
  - $\vec{v}$ jest ustalonym wektorem (używanym do przesunięcia)
  - W jest podprzestrzenią liniową przestrzeni liniowej V,
  - a A jest wynikową przestrzenią afiniczną.
- Definicja: Niech V będzie przestrzenią wektorową. Przestrzeń afiniczna A stowarzyszona z V to niepusty zbiór punktów, w którym dla każdych dwóch punktów $p, q \in A$ istnieje dokładnie jeden wektor $\vec{v} \in V$ taki, że $q = p + \vec{v}$, gdzie $+ : A \times V \to A$ jest działaniem określającym sposób, w jaki przestrzeń wektorowa oddziałuje na przestrzeń afiniczną.
- Formalizuje to koncepcję, zgodnie z którą, wychodząc z punktu p i poruszając się w kierunku wektora $\vec{v}$, docieramy do punktu q.
- Punkty zbioru A mogą nie być wektorami z przestrzeni V, lecz ich różnica stanowi wektor z V.

Podprzestrzeń afiniczna i zastosowania

- Przykład przestrzeni afinicznej Załóżmy, że mamy dwa punkty: $p = (1, 2)$ i $q = (4, 5)$. Wektor określający przesunięcie z punktu p do punktu q wyraża się wzorem $\vec{v} = q - p = (4, 5) - (1, 2) = (3, 3)$. Innymi słowy, zaczynając w punkcie p i stosując wektor przesunięcia $\vec{v}$, docieramy do punktu q.
- Definicja Podprzestrzeń afiniczna to przesunięta podprzestrzeń liniowa. Niech W będzie podprzestrzenią liniową przestrzeni V, a $\vec{v} \in V$. Wówczas zbiór $A = \vec{v} + W = \{\vec{v} + \vec{w} : \vec{w} \in W\}$ stanowi podprzestrzeń afiniczną.

  Wektory z W określają kierunki, natomiast $\vec{v}$ wyznacza orientację.
- Rozważmy przestrzeń wektorową $\mathbb{R}^2$ nad ciałem $\mathbb{R}$ oraz podprzestrzeń $W = \{(x, x) : x \in \mathbb{R}\}$ i wektor $\vec{v} = (2, 1) \in \mathbb{R}^2$. Wówczas zbiór $\vec{v} + W = \{(2 + x, 1 + x) : x \in \mathbb{R}\}$ stanowi w istocie prostą przechodzącą przez punkt $(2, 1)$, której wektorem kierunkowym jest $(1, 1)$.
- Zastosowania przestrzeni afinicznych Grafika komputerowa, Robotyka, Widzenie komputerowe, Modelowanie geometryczne.

Aksjomatyczne definicje iloczynu skalarnego

Iloczyn skalarny w przestrzeni wektorowej V nad ciałem $\mathbb{R}$ to funkcja, która dwóm wektorom $\vec{u}, \vec{v} \in V$ przyporządkowuje liczbę rzeczywistą, oznaczaną jako $\langle \vec{u}, \vec{v} \rangle$. Aby funkcja ta była iloczynem skalarnym, musi spełniać następujące aksjomaty dla wszystkich wektorów $\vec{u}, \vec{v}, \vec{w} \in V$ oraz skalarów $c \in \mathbb{R}$:

(1) Dodatniość (Positivity): $\langle \vec{v}, \vec{v} \rangle \geqslant 0$, przy czym $\langle \vec{v}, \vec{v} \rangle = 0$ wtedy i tylko wtedy, gdy $\vec{v} = \vec{0}$.

(2) Symetria (Symmetry): $\langle \vec{u}, \vec{v} \rangle = \langle \vec{v}, \vec{u} \rangle$.

(3) Addytywność względem pierwszego argumentu (Additivity in first argument): $\langle \vec{u} + \vec{v}, \vec{w} \rangle = \langle \vec{u}, \vec{w} \rangle + \langle \vec{v}, \vec{w} \rangle$.

(4) Jednorodność względem pierwszego argumentu (Homogeneity in first argument): $\langle c\vec{u}, \vec{v} \rangle = c\langle \vec{u}, \vec{v} \rangle$.

Norma Używając tego abstrakcyjnego iloczynu skalarnego możemy zdefiniować normę (lub długość strukturalną) dowolnego wektora jako $||\vec{v}|| = \sqrt{\langle \vec{v}, \vec{v} \rangle}$.

Przestrzeń z iloczynem skalarnym i Dopełnienia ortogonalne

- Przestrzeń z iloczynem skalarnym (inner product space) to przestrzeń wektorowa, w której wprowadzono narzędzie pomiarowe: iloczyn skalarny $\langle \vec{u}, \vec{v} \rangle$.
- Niech $W_{|\mathbb{R}}$ będzie podprzestrzenią przestrzeni z iloczynem skalarnym $V_{|\mathbb{R}}$. Dopełnienie ortogonalne (orthogonal complement) podprzestrzeni $W_{|\mathbb{R}}$ to zbiór wszystkich wektorów w przestrzeni $V_{|\mathbb{R}}$, które są prostopadłe do każdego wektora należącego do $W_{|\mathbb{R}}$: $W_{|\mathbb{R}}^{\perp} = \{\vec{v} \in W_{|\mathbb{R}} : \langle \vec{v}, \vec{w} \rangle = 0, \ \forall \vec{w} \in W_{|\mathbb{R}}\}$
- Korzystając z aksjomatów iloczynu skalarnego, można łatwo wykazać, że $W_{|\mathbb{R}}^{\perp}$ jest poprawną podprzestrzenią przestrzeni $V_{|\mathbb{R}}$.
- Dla dowolnej podprzestrzeni skończenie wymiarowej $V_{|\mathbb{R}} = W_{|\mathbb{R}} \oplus W_{|\mathbb{R}}^{\perp}$. Każdy wektor $\mathbf{v} \in V_{|\mathbb{R}}$ można w jednoznaczny sposób zapisać jako $\vec{v} = \vec{w} + \vec{w}^{\perp}$, gdzie $\vec{w} \in W_{|\mathbb{R}}$ oraz $\vec{w}^{\perp} \in W_{|\mathbb{R}}^{\perp}$. Ponadto, $dim(V_{|\mathbb{R}}) = dim(W_{|\mathbb{R}}) + dim(W_{|\mathbb{R}}^{\perp})$.

Przekształcenia liniowe

- Definicja: Niech $\mathcal{V}_{|K}$ i $\mathcal{W}_{|K}$ będą dwiema przestrzeniami liniowymi nad tym samym ciałem K. Przekształcenie $f : \mathcal{V} \mapsto \mathcal{W}$ nazywamy przekształceniem liniowym $\mathcal{V}$ w $\mathcal{W}$ jeśli (i) dla dowolnych $\overrightarrow{v_1}, \overrightarrow{v_2} \in \mathcal{V}$, $f(\overrightarrow{v_1 + v_2}) = f(\overrightarrow{v_1}) + f(\overrightarrow{v_1})$ i (ii) $f(\lambda\overrightarrow{v_1}) = \lambda f(\overrightarrow{v_1})$ dla dowolnego $\lambda \in K$.
- Warunki (i) i (ii) można również łączyć ze sobą jako $f(\overrightarrow{\lambda_1 v_1 + \lambda_2 v_2}) = \lambda_1 f(\overrightarrow{v_1}) + \lambda_2 f(\overrightarrow{v_1})$ dla dowolnych $\overrightarrow{v_1}, \overrightarrow{v_2} \in \mathcal{V}$ i $\lambda_1, \lambda_2 \in K$.
- Ponieważ tutaj skupiamy się tylko na przestrzeniach liniowych w ciele liczb rzeczywistych, K jest rozpatrywane jako $\mathbb{R}$.

Przykłady

(1) Niech $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}_{|\mathbb{R}}$ będzie funkcją z dwuwymiarowej przestrzeni euklidesowej do jednowymiarowej przestrzeni euklidesowej zdefiniowanej jako $T(x, y) = 2x - y$. Sprawdź czy jest to przekształcenie liniowe.

Przykłady

(1) Niech $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}_{|\mathbb{R}}$ będzie funkcją z dwuwymiarowej przestrzeni euklidesowej do jednowymiarowej przestrzeni euklidesowej zdefiniowanej jako $T(x, y) = 2x - y$. Sprawdź czy jest to przekształcenie liniowe.

- Dla dowolnych $(x_1, y_1), (x_2, y_2) \in \mathbb{R}^2_{|\mathbb{R}}$,
$T(x_1 + x_2, y_1 + y_2) = 2(x_1 + x_2) - (y_1 + y_2)$
$= (2x_1 - y_1) + (2x_2 - y_2) = T(x_1, y_1) + T(x_2, y_2)$
- Dla dowolnego $(x_1, y_1) \in \mathbb{R}^2_{|\mathbb{R}}$ i $\lambda \in \mathbb{R}$,
$T(\lambda x_1, \lambda y_1) = 2\lambda x_1 - \lambda y_1 = \lambda(2x_1 - y_1) = \lambda T(x_1, y_1)$

▶ Więc, T jest przekształceniem liniowym.

(2) Niech $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}^3_{|\mathbb{R}}$ będzie funkcją z dwuwymiarowej przestrzeni euklidesowej do jednowymiarowej przestrzeni euklidesowej zdefiniowanej jako $T(x, y) = (y, x + 2y, x + y)$. Sprawdź czy jest to przekształcenie liniowe.

Przekształcenia liniowe w uczeniu maszynowym

- Dimensionality reduction: Przekształcenia liniowe są wykorzystywane do redukcji wymiarowości
  - Techniki PCA (Principal Component Analysis) (Analiza Głównych Składowych) wykorzystują przekształcenia liniowe w celu ograniczenia liczby cech w zestawie danych, przy jednoczesnym zachowaniu najważniejszych informacji, co przyczynia się do zwiększenia wydajności modeli.
- Feature Engineering: Przekształcenia liniowe pozwalają tworzyć nowe cechy z istniejących. Na przykład łączenie lub manipulowanie cechami za pomocą kombinacji liniowych może zapewnić bardziej użyteczną reprezentację modelu.
- Data preprocessing: Przekształcenia liniowe są używane do przygotowania danych przez normalizację i standaryzację w celu skalowania cech, zapewniając ich zrównoważony wkład do modelu.
- Artificial Neural Network: Mnożenie wektora przez macierz wagową to przekształcenie liniowe, która pomaga sieci neuronowej uczyć się skomplikowanych wzorców.

Obraz, jądro i rząd przekształcenia liniowego

- Definicja: Niech $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{W}_{|\mathbb{R}}$ będzie przekształceniem liniowym.
  (i) Definiujemy obraz przekształcenia liniowego T jako $Im(T) = \{T(\overrightarrow{v}) : \overrightarrow{v} \in \mathcal{V}_{|\mathbb{R}}\}$.
  (ii) Definiujemy jądro przekształcenia liniowego T jako $Ker(T) = \{\overrightarrow{v} \in \mathcal{V}_{|\mathbb{R}} : T(\overrightarrow{v}) = \overrightarrow{0}\}$.
  (iii) Definiujemy rząd przekształcenia liniowego T jako $rank(T) = dim(Im(T))$.
- Twierdzenie: Dla dowolnego przekształcenia liniowego $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{W}_{|\mathbb{R}}$, $dim(\mathcal{V}_{|\mathbb{R}}) = dim(Ker(T)) + dim(Im(T))$.
- Jądro (ang. kernel) przekształcenia liniowego T ($Ker(T)$) często nazywane jest przestrzenią zerową (ang. null space) T, i oznaczone jako $Null(T)$.

Przykład

- Rozważmy przekształcenie liniowe $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}_{|\mathbb{R}}$ zdefiniowane jako $T(x, y) = 2x - y$. Znajdź $Null(T)$.
- $Null(T) = \{(x, y) : T(x, y) = 0\} = \{(x, y) : 2x - y = 0\}$
  $= \{(x, 2x) : x \in \mathbb{R}\}$
- $Null(T) = Span(\{(1, 2)\})$
- $dim(Null(T)) = 1$ bo pojedynczy wektor różny od zera jest zawsze liniowo niezależny.

Reprezentacja macierzowa przekształceni liniowej

- Rozważmy przestrzeń liniową $\mathcal{V}_{|\mathbb{R}}$. Przez uporządkowaną bazę $B = (\overrightarrow{v}_1, \overrightarrow{v}_2, \ldots, \overrightarrow{v}_n)$ tej przestrzeni $\mathcal{V}_{|\mathbb{R}}$ rozumiemy bazę rozpatrywaną w określonym porządku. W przeciwieństwie do bazy, tutaj kolejność wektorów ma znaczenie.
- Rozważmy przestrzeń liniową $\mathcal{V}_{|\mathbb{R}}$ i uporządkowaną bazę $B = (\overrightarrow{v}_1, \overrightarrow{v}_2, \ldots, \overrightarrow{v}_n)$. Dla dowolnego wektora $\overrightarrow{v} \in \mathcal{V}_{\mathbb{R}}$ otrzymujemy unikalną reprezentację (kombinację liniową) $\overrightarrow{v} = a_1\overrightarrow{v_1} + a_2\overrightarrow{v_2} + \ldots + a_n\overrightarrow{v_n}$.
- $a_1, a_2, \ldots, a_n$ to współrzędne $\overrightarrow{v}$ względem B, a $\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$ jest uważane za reprezentację współrzędnych (ang. coordinate representation) $\overrightarrow{v}$ względem B.
- Baza skutecznie definiuje układ współrzędnych. Znamy dwuwymiarowy kartezjański układ współrzędnych, który jest rozpięty przez standardowe wektory bazowe (1, 0), (0, 1).

Reprezentacja współrzędnych zależy od wybranej bazy

- Jednakże każda baza $\mathbb{R}^2_{|\mathbb{R}}$ definiuje poprawny układ współrzędnych, a ten sam wektor (x, y) może mieć różną reprezentację współrzędnych w zależności od innej bazy.
- Przykład:

  $(3, 7) = 3(1, 0) + 7(0, 1)$ względem bazy $\{(1, 0), (0, 1)\}$

  Ale $(3, 7) = 4(0, 1) + (3)(1, 1)$ względem $\{(0, 1), (1, 1)\}$.

Macierz reprezentująca przekształcenią liniową

- Niech $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{W}_{|\mathbb{R}}$ będzie przekształceniem liniowym, a $(\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_n})$ i $(\overrightarrow{w_1}, \overrightarrow{w_2}, \ldots, \overrightarrow{w_m})$ będą odpowiednio uporządkowanymi bazami dla $\mathcal{V}_{|\mathbb{R}}$ i $\mathcal{W}_{|\mathbb{R}}$. Więc, dla każdego $\overrightarrow{v_j}$ $(1 \leqslant j \leqslant n)$, $T(\overrightarrow{v_j})$ ma unikalną reprezentację $T(\overrightarrow{v_j}) = a_{1j}\overrightarrow{w_1} + a_{2j}\overrightarrow{w_2} + \ldots + a_{mj}\overrightarrow{w_m}$ względem $(\overrightarrow{w_1}, \overrightarrow{w_2}, \ldots, \overrightarrow{w_m})$.
- Tworzy to macierz $M(T)$ wymiaru $m \times n$ dla przekształcenia liniowego T, gdzie (i, j)-ty element macierzy to a_{ij}.
- Tzn., z każdej unikalnej reprezentacji $T(\overrightarrow{v_j}) = a_{1j}\overrightarrow{w_1} + a_{2j}\overrightarrow{w_2} + \ldots + a_{mj}\overrightarrow{w_m}$, $1 \leqslant j \leqslant n$, mamy

$$M(T) = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

Przykłady

- Rozważmy wcześniej omówiony przykład przekształceni liniowego $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}^3_{|\mathbb{R}}$ zdefiniowanego jako $T(x, y) = (y, x + 2y, x + y)$.
- Rozważmy również bazy standardowe $\mathbb{R}^2_{|\mathbb{R}}$ i $\mathbb{R}^3_{|\mathbb{R}}$ jako odpowiednio uporządkowane bazy.
- $T(1, 0) = (0, 1, 1) = 0(1, 0, 0) + 1(0, 1, 0) + 1(0, 0, 1)$
  $T(0, 1) = (1, 2, 1) = 1(1, 0, 0) + 2(0, 1, 0) + 1(0, 0, 1)$
- $M(T) = \begin{pmatrix} 0 & 1 \\ 1 & 2 \\ 1 & 1 \end{pmatrix}$
- Teraz, jeśli zmienimy którąkolwiek z uporządkowanych baz rozważanych dla $\mathbb{R}^2_{|\mathbb{R}}$ i $\mathbb{R}^2_{|\mathbb{R}}$, reprezentacja macierzowa ulegnie zmianie. Np. zmieńmy tylko bazę dla $\mathbb{R}^2_{|\mathbb{R}}$ i rozważmy $((0, 1), (1, 1))$ jako uporządkowna nowa baza.
- Wtedy $M(T) = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & 2 \end{pmatrix}$

Przekształcenie liniowe uzyskane z macierzy

- Niech $A = [a_{ij}]$ będzie macierz o wymiarze $m \times n$. Możemy zdefiniować odpowiednio przeksztalcenie liniowe $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{W}_{|\mathbb{R}}$, gdzie $\mathcal{V}_{|\mathbb{R}}$ i $\mathcal{W}_{|\mathbb{R}}$ są przestrzeniami liniowymi o wymiarach odpowiednio n i m, w następujący sposób.
- Najpierw, musimy rozważyć uporządkowane bazy dla $\mathcal{V}_{|\mathbb{R}}$ i $\mathcal{W}_{|\mathbb{R}}$, a niech będą to odpowiednio $(\overrightarrow{v_1}, \ldots, \overrightarrow{v_n})$ i $(\overrightarrow{w_1}, \ldots, \overrightarrow{w_m})$.
- Zdefinujemy, $T(\overrightarrow{v_j}) = \Sigma_{i=1}^{m} a_{ij} \overrightarrow{w_i}$
- Dla dowolnego $\overrightarrow{v} \in \mathcal{V}_{|\mathbb{R}}$, istnieje unikalna reprezentacja, takie jak $\overrightarrow{v} = b_1 \overrightarrow{v_1} + b_2 \overrightarrow{v_2} + \ldots + b_n \overrightarrow{v_n}$

$$
\begin{aligned}
T(\overrightarrow{v}) &= T(b_1 \overrightarrow{v_1} + b_2 \overrightarrow{v_2} + \ldots + b_n \overrightarrow{v_n}) \\
&= b_1 T(\overrightarrow{v_1}) + b_2 T(\overrightarrow{v_2}) + \ldots + b_n T(\overrightarrow{v_n}) \\
&= b_1 \Sigma_{i=1}^{m} a_{i1} \overrightarrow{w_i} + b_2 \Sigma_{i=1}^{m} a_{i2} \overrightarrow{w_i} + \ldots + b_n \Sigma_{i=1}^{m} a_{in} \overrightarrow{w_i}
\end{aligned}
$$

Przykład

- Niech $A = \begin{pmatrix} 2 & 1 \\ -3 & -2 \end{pmatrix}$
- Zdefinujemy przekształcenie liniowe $T : \mathbb{R}^2_{|\mathbb{R}} \mapsto \mathbb{R}^2_{|\mathbb{R}}$ względem bazy $((1,0),(0,1))$.
- $T(1,0) = 2(1,\ 0) + (\text{-}3)(0,\ 1) = (2,\ \text{-}3)$
  $T(0,1) = 1(1,\ 0) + (\text{-}2)(0,\ 1) = (1,\ \text{-}2)$
- $$\begin{aligned} T(x,y) &= T(x(1,0) + y(0,1)) \\ &= xT(1,0) + yT(0,1) \\ &= x(2,-3) + y(1,-2) \\ &= (2x + y, -3x - 2y) \end{aligned}$$

Podprzestrzeń niezmienna

- Przekształcenie liniowe T z przestrzeni liniowej $\mathcal{V}_{|\mathbb{R}}$ do samej siebie nazywa się operatorem liniowym.
- Niech $\mathcal{V}_{|\mathbb{R}}$ będzie skończenie wymiarową przestrzenią liniową o $dim(\mathcal{V}_{|\mathbb{R}}) \geqslant 1$ i $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$ będzie operatorem liniowym. Wówczas podprzestrzeń $\mathcal{U}_{|\mathbb{R}}$ przestrzeni $\mathcal{V}_{|\mathbb{R}}$ nazywana jest podprzestrzenią niezmienną względem T, jeśli $T(u) \in \mathcal{U}_{|\mathbb{R}}$ dla każdego $u \in \mathcal{U}_{|\mathbb{R}}$.
- Przykład: Rozważmy liniowy operator $T : \mathbb{R}^3_{|\mathbb{R}} \mapsto \mathbb{R}^3_{|\mathbb{R}}$ zdefiniowanego jako $T(x, y, z) = (x + 2y, x + y, 2z)$ i podprzestrzeń $\mathcal{U}_{|\mathbb{R}} \subseteq \mathbb{R}^3_{|\mathbb{R}}$ gdzie $\mathcal{U}_{|\mathbb{R}} = Span(\{(0, 0, 1)\}) = \{(0, 0, c) : c \in \mathbb{R}\}$.
  Dla dowolnego $(0, 0, c) \in \mathcal{U}_{|\mathbb{R}}$, $T(0, 0, c) = (0, 0, 2c) \in \mathcal{U}_{|\mathbb{R}}$.

Wartości własne i wektory własne

- Niech $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$ będzie liniowym operatorym. Skalar $\lambda \in \mathbb{R}$ jest nazywane wartością własną (ang. Eigen[1] value) przekształcenia liniowego T jeśli istnieje nie zerowy wektor $\overrightarrow{v} \in \mathcal{V}_{|\mathbb{R}}$ takie, że $T(\overrightarrow{v}) = \lambda \overrightarrow{v}$. Wektor $\overrightarrow{v}$ jest nazywane wektorem własnym (ang. Eigen vector) T odpowiadającym λ.
- Innymi słowy, niech M będzie macierzową reprezentacją przekształcenia liniowego $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$. Więc, M ma wymiar $n \times n$ gdzie $dim(\mathcal{V}_{|\mathbb{R}}) = n$. Skalar $\lambda \in \mathbb{R}$ jest wartością własną macierzy M jeśli istnieje nie zerowy wektor $\overrightarrow{v} \in \mathcal{V}_{|\mathbb{R}}$, takie że $M\overrightarrow{v} = \lambda \overrightarrow{v}$.
- W szczególności, jeśli $\mathcal{V}_{|\mathbb{R}} = \mathbb{R}^n_{|\mathbb{R}}$, wówczas $M \in \mathbb{R}^n \times \mathbb{R}^n$ i $\overrightarrow{v} \in \mathbb{R}^n_{|\mathbb{R}}$.

[1] Eigen to niemieckie słowo oznaczające ‘charakterystyczny’, ‘własny’ lub ‘własny’.

Własności wartości własnych i wektorów własnych

- Poniższe stwierdzenia są równoważne:
  - λ jest wartością własną $T : \mathbb{R}^n_{|\mathbb{R}} \mapsto \mathbb{R}^n_{|\mathbb{R}}$ (lub $M \in \mathbb{R}^n \times \mathbb{R}^n$).
  - Istnieje $\overrightarrow{v} \in \mathbb{R}^n \setminus \{\overrightarrow{0}\}$ takie, że $T(\overrightarrow{v}) = \lambda \overrightarrow{v}$ (lub równoważnie $det(M - \lambda I_n)\overrightarrow{v} = 0$ ma rozwiązanie niezerowe).
  - Rząd$(M - \lambda I_n) < n$.
  - $det(M - \lambda I_n) = 0$.
- Przykład: Rozważmy wcześniej omówiony przykład przekształceni liniowego $T : \mathbb{R}^3_{|\mathbb{R}} \mapsto \mathbb{R}^3_{|\mathbb{R}}$ zdefiniowanego jako $T(x, y, z) = (x + 2y, x + y, 2z)$.
  - Biorąc pod uwagę bazę standardową, mamy macierz $M = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$
  - Niech λ będzie wartość własną T. Więc, $det(M - \lambda I_n) = 0$.
  - $\begin{vmatrix} 1-\lambda & 2 & 0 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} = 0 \hookrightarrow \lambda = 2,\ 1+\sqrt{2},\ 1-\sqrt{2}$.

Przestrzeń własna odpowiadająca wartości własnej

- $T(x, y, z) = (x + 2y, x + y, 2z) = (\lambda x, \lambda y, \lambda z)$

$$(1 - \lambda)x + 2y = 0 \quad (i)$$
$$x + (1 - \lambda)y = 0 \quad (ii)$$
$$(2 - \lambda)z = 0 \quad (iii)$$

- Wektor własny dla $\lambda = 2$:
  - Mamy rozkład równań liniowych: -x + 2y = 0 i x - y = 0
  - Jedynym możliwym rozwiązaniem jest $x = y = 0$ i jest ono całkowicie niezależne od wartości z. Ponieważ jednak wektory własne są wektorami niezerowymi, wektor własny odpowiadający $\lambda = 2$ ma postać $(0, 0, c)$ gdzie $c \in \mathbb{R} \setminus \{0\}$.
  - Przestrzeń własna dla $\lambda = 2$: $V_{\lambda=2} = \{(0, 0, c) : c \in \mathbb{R} \setminus \{0\}\}$.
- Wektor własny dla $\lambda = 1+\sqrt{2}$:
  - Mamy x - $\sqrt{2}$y = 0 i $(1 - \sqrt{2})z = 0$. Więc, $V_{\lambda=1+\sqrt{2}} = \{(\sqrt{2}c, c, 0) : c \in \mathbb{R} \setminus \{0\}\}$.
- Wektor własny dla $\lambda = 1$-$\sqrt{2}$:
  - $V_{\lambda=1-\sqrt{2}} = \{(-\sqrt{2}c, c, 0) : c \in \mathbb{R} \setminus \{0\}\}$.

Związek między wartościami własnymi, wektorami własnymi i bazą

- Twierdzenie: Dla dowolnego operatora liniowego $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$ i wartości własnej λ, V_λ jest niezmienną podprzestrzenią $\mathcal{V}_{|\mathbb{R}}$ względem T.
- Twierdzenie: Dowolny operator liniowy $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$ ma co najwyżej $n = dim(\mathcal{V}_{|\mathbb{R}})$ liczbę wartości własnych.
- Twierdzenie: Niech $T : \mathcal{V}_{|\mathbb{R}} \mapsto \mathcal{V}_{|\mathbb{R}}$ będzie liniowym operatorem i $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}$ będą m różnymi własnymi wartościami T. Wówczas odpowiadające im wektory własne $\overrightarrow{v_1}, \overrightarrow{v_2}, \ldots, \overrightarrow{v_m}$ są liniowo niezależne.
- Przykład: Już zobaczyliśmy $T : \mathbb{R}^3_{|\mathbb{R}} \mapsto \mathbb{R}^3_{|\mathbb{R}}$ zdefiniowanego jako $T(x, y, z) = (x + 2y, x + y, 2z)$ ma $3 = dim(\mathbb{R}^3_{|\mathbb{R}})$ wartości własne. Więc, odpowiedni zbiór wektorów własnych, takich jak $\{(\sqrt{2}, 1, 0), (-\sqrt{2}, 1, 0), (0, 0, 1)\}$ jest liniowym niezależnym.
- Twierdzenie: Wyznacznik macierzy kwadratowej $A = [a_{ij}]_{n \times n}$ jest iloczynem jej wartości własnych (możliwie powtarzalnych).

PageRank Google: Strony internetowe jako wektory własne

- Pomysł na algorytm PageRank, opracowany na Uniwersytecie Stanforda przez Larry'ego Page'a (Larry Page) i Siergieja Brina (Sergey Brin) w 1996 roku, opierał się na założeniu, że znaczenie dowolnej strony internetowej można oszacować na podstawie znaczenia stron, które do niej linkują.
- Wszystkie witryny internetowe są prezentowane w postaci ogromnego, skierowanego grafu, który pokazuje, która strona linkuje do której.
- PageRank oblicza wagę (ważność) $w_i > 0$ witryny internetowej p_i poprzez zliczanie liczby stron wskazujących na p_i i biorąc pod uwagę ważność witryn internetowych, które linkują do p_i.
- Na podstawie grafu tworzona jest macierz przejść A, która szacuje prawdopodobieństwo (kliknięcia), że ktoś rozpocznie przeglądanie jednej strony internetowej i przejdzie na inną stronę.
- PageRank reprezentuje wektor, powiedzmy $\overrightarrow{x}$ rankingu stron internetowych taki, że $A\overrightarrow{x} = \overrightarrow{x}$ - co odpowiada wartości własnej 1.

Rozkład własny i diagonalizacja (ang. Eigen decomposition and diagonalization)

- Przydatność macierzy diagonalnej polega na tym, że pozwala na szybkie obliczenie wyznacznika, potęg i macierzy odwrotnej. Np.,
  - Dla macierzy diagonalnej D wyznacznik jest iloczynem jej elementów diagonalnych.
  - Potęgę macierzy D^k uzyskuje się po prostu podnosząc każdy element przekątnej do potęgi k.
  - Odwrotność D^{-1} jest odwrotnością elementów przekątnej (tzn., $\frac{1}{d_{ii}}$ dla każdego d_{ii} macierzy D), jeśli wszystkie są różne od zera.
- Zatem przekształcenie danej macierzy A (reprezentującej dane) w macierz diagonalną jest jednym z ważniejszych kroków w uczeniu maszynowym.
- Definicja: Macierz kwadratowa $A = [a_{ij}]_{n \times n}$ jest diagonalizowalna (ang. diagonalizable), jeśli istnieje macierz odwracalna P taka, że $D = P^{-1}AP$.

Rola wartości i wektorów własnych w tworzeniu macierzy diagonalizowanej

- Niech A będzie macierzą o wymiaru $n \times n$ i $\lambda_1, \lambda_2, \ldots, \lambda_n$ będą wartościami własnymi macierzy A.
- Skonstruktujemy $D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$
- Stworzymy $P = \begin{pmatrix} \overrightarrow{p_1} & \overrightarrow{p_2} & \ldots & \overrightarrow{p_n} \end{pmatrix}$ gdzie $\overrightarrow{p_i} \in \mathbb{R}^n$ jest wektorym własnym odpowiadającym λ_i.
- Istnieje twierdzenie, które zapewnia, że $AP = PD$
- Warto uwazyć, że macierz $AP = \begin{pmatrix} A\overrightarrow{p_1} & A\overrightarrow{p_2} & \ldots & A\overrightarrow{p_n} \end{pmatrix}$ gdzie dla dowolnego i, $1 \leqslant i \leqslant n$, i-ta kolumna AP jest uzyskiwana przez pomnożenie $A\overrightarrow{p_i}$. Podobnie, $PD = \begin{pmatrix} \lambda_1\overrightarrow{p_1} & \lambda_2\overrightarrow{p_2} & \ldots & \lambda_n\overrightarrow{p_n} \end{pmatrix}$
- Zatem $AP = PD$ oznacza $A\overrightarrow{p_i} = \lambda_i \overrightarrow{p_i}$ dla każdej i, $1 \leqslant i \leqslant n$.

Przykład

- Znajdźmy rozkład własny $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
- Uzyskanie wartości i wektorów własnych: $\begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0$
  - $\lambda_1 = 1$ i $\lambda_2 = 3$
  - $A\overrightarrow{p_1} = \overrightarrow{p_1}$ daje nam $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix} = \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix}$.

    Więc, $\overrightarrow{p_1} = \begin{pmatrix} p_{11} \\ p_{12} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
  - $A\overrightarrow{p_2} = \overrightarrow{p_2}$ daje nam $\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} p_{21} \\ p_{22} \end{pmatrix} = \begin{pmatrix} p_{21} \\ p_{22} \end{pmatrix}$.

    Więc, $\overrightarrow{p_2} = \begin{pmatrix} p_{21} \\ p_{22} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
- Tworzenie P: $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$
- Tworzenie $D = P^{-1}AP = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$

Przydatność macierzy diagonalizowalnych

- Macierze diagonalne D można efektywnie podnieść do potęgi. Dlatego możemy znaleźć potęgę macierzy dla A poprzez rozkład wartości własnych (jeśli istnieje), tak aby $A^k = (PDP^{-1})^k = PD^kP^{-1}$.
  Obliczanie D^k jest efektywne, ponieważ stosujemy tę operację indywidualnie do dowolnego elementu przekątnej.
- Jeśli istnieje rozkład własny $A = PDP^{-1}$, wtedy
  $det(A) = det(PDP^{-1}) = det(P)det(D)det(P^{-1})$
  $= det(D) = \Pi_{i=1}^{n} d_{ii}$

Rozkład wartości osobliwych (ang. Singular Value Decomposition

- Rozkład macierzy na wartości osobliwe (SVD) jest centralną metodą rozkładu macierzy w algebrze liniowej. Nazywa się ją „fundamentalnym twierdzeniem algebry liniowej" (Strang, 1993), ponieważ można ją zastosować do wszystkich macierzy, nie tylko do macierzy kwadratowych, i istnieje ona zawsze.
- Twierdzenie: Niech $A = [a_{ij}]_{m \times n}$ będzie macierzą prostokątną rzędu $r \in [0, min(m, n)]$. SVD macierzy A jest rozkładem postaci $A = U.S.V^t$ gdzie U jest macierzą ortogonalną wymiaru $m \times m$ z wektorami kolumnowymi u_i $(1 \leqslant i \leqslant m)$, V jest macierzą ortogonalną wymiaru $n \times n$ z wektorami kolumnowymi v_j $(1 \leqslant j \leqslant n)$, i S jest macierzą wymiaru $m \times n$, gdzie $s_{ii} = \sigma_i \geqslant 0$ i $s_{ij} = 0$ dla $i \neq j$.

SVD

- Pozycje diagonalne σ_i, $i = 1, \ldots, r$ S nazywane są wartościami osobliwymi (singular values), u_i nazywane są wektorami lewostronnymi osobliwymi (left-singular vector), a v_j nazywane są wektorami prawostronnymi osobliwymi (right-singular vectors). Zgodnie z konwencją wartości osobliwe są uporządkowane, tj. $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_r \geqslant 0$.
- Macierz wartości osobliwych S jest unikalna, ale wymaga pewnej uwagi. S ma taki sam rozmiar jak A. Oznacza to, że S ma diagonalną podmacierz, która zawiera wartości osobliwe i wymaga dodatkowego uzupełnienia zerami.

S z podmacierzą diagonalną

- Jeśli $m > n$, $\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \end{pmatrix}$
- Jeśli $m < n$, $\begin{pmatrix} \sigma_1 & & 0 & 0 & \dots & 0 \\ & \ddots & & \vdots & & \vdots \\ 0 & & \sigma_m & 0 & \dots & 0 \end{pmatrix}$
- SVD istnieje dla dowolnej macierzy A o wymiaru $m \times n$.

Proces rozkładu macierzy według wartości osobliwych

Cel: Wyznaczanie rozkładu SVD dla macierzy $A = [a_{ij}]_{m \times n}$

- Wyznacz AA^t, czyli macierz kwadratową stopnia n.
- Znajdź wartości własne macierzy A^tA - tzn. rozwiązać $det(A^tA - \lambda I) = 0$
- Oblicz wartości osobliwe σ_i, będące pierwiastkami kwadratowymi niezerowych wartości własnych, tj. $\sigma_i = \sqrt{\lambda_i}$ gdzie $\lambda_i \neq 0$
- Uporządkuj wartości osobliwe w kolejności malejącej: $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant 0$
- Utwórz S, wykorzystując wartości osobliwe zgodnie z opisem w definicji
- Znajdź wektor własny $\vec{v_i}$ odpowiadający wartości własnej λ_i macierzy A^tA i znormalizuj tak, żeby $||\vec{v_i}|| = 1$.
- Aby utworzyć V umieścimy te znormalizowane wektory jako kolumny macierzy
- Aby utworzyć macierz U, należy wyznaczyć AA^t, która jest macierzą kwadratową stopnia m i Postępuj zgodnie z krokami podobnymi do tych, które wykonano w celu uzyskania V.

Rzut ortogonalny

- Niech $V_{|\mathbb{R}}$ będzie przestrzenią z iloczynem skalarnym, a $W_{|\mathbb{R}}$ jest podprzestrzenią skończowowymiarową przestrzeni $V_{|\mathbb{R}}$.
- Więc $V_{|\mathbb{R}}$ można rozłożyć na sumę prostą podprzestrzeni $W_{|\mathbb{R}}$ oraz jej dopełnienia ortogonalnego $W_{|\mathbb{R}}^{\perp}$; czyli $V_{|\mathbb{R}} = W_{|\mathbb{R}} \oplus W_{|\mathbb{R}}^{\perp}$
  Gwarantuje to, że każdy wektor $\vec{v} \in V_{|\mathbb{R}}$ można w jednoznaczny sposób zapisać w postaci:
  $\vec{v} = \vec{w} + \vec{w}^{\perp}$ gdzie $\vec{w} \in W_{|\mathbb{R}}$ oraz $\vec{w}^{\perp} \in W_{|\mathbb{R}}^{\perp}$
- Rzut ortogonalny (orthogonal projection) przestrzeni $V_{|\mathbb{R}}$ na podprzestrzeń $W_{|\mathbb{R}}$ to przekształcenie liniowe $P_W : V_{|\mathbb{R}} \to V_{|\mathbb{R}}$ zdefiniowane następująco: $P_W(\vec{v}) = \vec{w}$
- Intuicyjnie, rzut ortogonalny to określone przekształcenie liniowe $P_W : V_{|\mathbb{R}} \to V_{|\mathbb{R}}$, które każdemu wektorowi $\vec{v}$ przyporządkowuje jedyny taki element $\vec{w} \in W$, że pozostały wektor błędu, $(\vec{v} - P_W(\vec{v}))$, w całości należy do $W_{|\mathbb{R}}^{\perp}$. Znalezienie się w $W_{|\mathbb{R}}^{\perp}$ jest istotne, ponieważ poprowadzenie odcinka prostopadłego z punktu danych do tej podprzestrzeni gwarantuje minimalizację odległości $\|\vec{v} - P_W(\vec{v})\|$ (błędu rekonstrukcji lub straty strukturalnej).

Rola rzutu ortogonalnego w uczeniu maszynowym

- W analizie głównych składowych (PCA) dysponujemy wielowymiarowymi wektorami danych, które należą do rozległej przestrzeni wektorowej $V_{|\mathbb{R}}$. Celem jest skompresowanie tych wektorów do podprzestrzeni $W_{|\mathbb{R}}$ o niższym wymiarze, przy jednoczesnym zminimalizowaniu utraty informacji. Dlatego często stosuje się rzut ortogonalny.
- W dziedzinie przetwarzania języka naturalnego (NLP) słowa są odwzorowywane na gęste wektory w przestrzeni z iloczynem skalarnym (np. Word2Vec, LLM embeddings). Rzut ortogonalny wykorzystuje się do wykrywania i usuwania systemowych uprzedzeń (takich jak te dotyczące płci czy rasy) z modeli sztucznej inteligencji.

  Np., można zdefiniować „podprzestrzeń stronniczości" (W), rozpiętą na różnicach wektorów, takich jak $\vec{v}_{\text{mężczyzna}} - \vec{v}_{\text{kobieta}}$. Aby zneutralizować słowo neutralne, na przykład „lekarz" lub „programista" ($\vec{v}$), dokonuje się jego rzutu na dopełnienie ortogonalne podprzestrzeni stronniczości ($W^{\perp}$):

Obrót (Rotation)

- Obrót to przekształcenie liniowe, które obraca wektor lub przestrzeń wokół ustalonego punktu (zazwyczaj początku układu współrzędnych) o określony kąt θ, nie zmieniając przy tym długości żadnych wektorów.
- Zasadniczo, obroty zachowują zarówno długości wektorów, jak i kąty między nimi.
- Wiemy, że każde przekształcenie liniowe można zapisać w odpowiedniej postaci macierzowej. Zatem macierz R reprezentująca obrót jest zawsze macierzą ortogonalną, co oznacza, że $R^T R = I$.
- Wyznacznik macierzy obrotu czystego jest zawsze 1.
- Definicja: Przekształcenie liniowe $T : V_{|\mathbb{R}} \to V_{|\mathbb{R}}$, reprezentowane przez macierz R, jest obrotem, jeśli spełnia następujące warunki: (i) $R^T R = I$ oraz (ii) $det(R) = 1$.

Rozkład Cholesky'ego (Cholesky decomposition)

- Niech A będzie macierzą kwadratową nad ciałem liczb rzeczywistych $\mathbb{R}$. Rozkład Cholesky'ego (lub faktoryzacja Cholesky'ego) macierzy A to przedstawienie tej macierzy w postaci iloczynu: $A = LL^*$ gdzie:
  - L jest macierzą dolnotrójkątną nad ciałem $\mathbb{R}$, której elementy na głównej przekątnej są liczbami rzeczywistymi ściśle dodatnimi, i
  - L^* to sprzężenie hermitowskie macierzy L (conjugate transpose of L) (w przypadku macierzy nad $\mathbb{R}$ zachodzi $L^* = L^T$)
- Macierz A nad $\mathbb{R}$ dopuszcza rozkład Cholesky'ego wtedy i tylko wtedy, gdy spełnia następujące dwie własności: (i) $A = A^T$ i (ii) dla każdego niezerowego wektora $x \in \mathbb{R}^n$, $x^T Ax > 0$.

  Równoważnie, wszystkie wartości własne λ_i macierzy A są liczbami rzeczywistymi ściśle dodatnimi.
- Zastosowanie w uczeniu maszynowym: jest szeroko wykorzystywane w celu zwiększenia szybkości obliczeniowej algorytmów, zapewnienia ich stabilności numerycznej oraz efektywnego wykorzystania pamięci. Ponieważ macierze kowariancji w uczeniu maszynowym są z natury symetryczne i dodatnio określone, rozkład Cholesky'ego stanowi doskonałe narzędzie do pracy z nimi.

Drugi filar uczynia maszynowego: prawdopodobieństwa

- Aby określić niepewność danych, niepewność modelu uczenia maszynowego i niepewność przewidywań wykonywanych przez zaprojektowany model, często stosuje się narzędzia rachunku prawdopodobieństwa.
- Treści
  - Prawdopodobieństwo i istotne pojęcia
  - Zmienna losowa
  - Rozkład prawdopodobieństwa

Pojęcia prawdopodobieństwa według Laplace'a

- Eksperyment to procedura, która przeprowadzana w określonych warunkach prowadzi do uzyskania możliwego zbioru wyników.
- Przykład: Rzucanie kostką jest przykładem eksperymentu.
- Zbiór wszystkich możliwych wyników eksperymentu nazywa się przestrzenią próby (ang. sample or sample space).
- Przykład: Przestrzeń próby rzutu kostką to $\{1, 2, 3, 4, 5, 6\}$.
- Zdarzenie (ang. event) jest identyfikowane z podzbiorem przestrzeni próby eksperymentu. Przestrzeń zdarzeń jest przestrzenią potencjalnych wyników eksperymentu.
- Przykład: Uzyskanie nieparzystej liczby w eksperymencie z rzucaniem kostką odnosi się do zdarzenia $\{1, 3, 5\}$.
- Wyniki przestrzeni próby nazywa się równie prawdopodobnymi, jeśli w trakcie przeprowadzania eksperymentu, w każdym przypadku, wynikiem może być dowolny z elementów przestrzeni próby; to znaczy, że nie ma preferencji jednego wyniku nad innymi.

Prawdopodobieństwo zdarzenia na próbie skończonej

- Prawdopodobieństwo zdarzenia E, które jest podzbiorem skończonej przestrzeni prób S o jednakowo prawdopodobnych wynikach, wynosi $P(E) = \frac{|E|}{|S|}$.
- Przykłady:
  - Prawdopodobieństwo uzyskania nieparzystej liczby w rzucie kostką wynosi $\frac{3}{6} = \frac{1}{2}$.
  - Prawdopodobieństwo uzyskania siódemki w rzucie kostką wynosi 0.
  - Prawdopodobieństwo uzyskania parzystej lub nieparzystej oczki w rzucie kostką wynosi $\frac{6}{6} = 1$.
- Zdarzenie E o prawdopodobieństwie 0 nazywa się zdarzeniem niemożliwym. Zdarzenie polegające na uzyskaniu siódemki w rzucie kostką jest zdarzeniem niemożliwym.
- Zdarzenie E z prawdopodobieństwem 1 nazywa się zdarzeniem pewnym. Zdarzenie polegające na uzyskaniu parzystej lub nieparzystej oczki w rzucie kostką jest zdarzeniem pewnym.
- Dla dowolnego zdarzenia E przestrzeni prób S, $0 \leqslant P(E) \leqslant 1$.

Przykłady

1. Urna zawiera cztery niebieskie i pięć czerwonych kul. Jakie jest prawdopodobieństwo, że wybrana z urny kula będzie niebieska?

   - Istnieje 9 możliwych wyników, jeśli wylosujemy jedną kulę z urny. Spośród nich 4 są korzystne dla danego zdarzenia. Zatem prawdopodobieństwo wylosowania niebieskiej kuli wynosi $\frac{4}{9}$.

2. Jakie jest prawdopodobieństwo, że po rzucie dwiema kostkami suma oczek na obu kostkach wyniesie 7?

   - Jeśli rzucimy dwiema kostkami, liczba wyników w przestrzeni próby wyniesie $6^2 = 36$. Wśród nich wyniki korzystne dla zdarzenia „uzyskania 7 jako sumy oczek dwóch kostek" to (1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3). Zatem zdarzenie ma 6 możliwych przypadków. Stąd prawdopodobieństwo zdarzenia wynosi $\frac{6}{36} = \frac{1}{6}$.

Wybieranie rzeczy z możliwością zastąpienia i bez zastąpienia

- Przykład: Jakie jest prawdopodobieństwo, że liczby 11, 4, 17, 39, 23 zostaną wylosowane w tej kolejności z pojemnika zawierającego 50 kul oznaczonych liczbami od 1 do 50, jeśli (a) kula wybrana w każdej turze nie jest zwracana do pojemnika i (b) kula wybrana w każdej turze jest zwracana do pojemnika?

(a) Wybieranie rzeczy bez zastąpienia: Ponieważ za każdym razem, gdy kula jest wybierana, w koszu jest o jedną kulę mniej, zgodnie z regułą iloczynu (liczenia) istnieje $50,49,48,47,46 = 254251200$ sposobów na wybranie pięciu kul z kosza. Spośród nich jest tylko jeden przypadek, który pasuje do wyboru 11, 4, 17, 39, 23. Zatem prawdopodobieństwo wylosowania tych pięciu liczb w powyższej kolejności wynosi $\frac{1}{254251200}$.

Wybieranie rzeczy z możliwością zastąpienia

- Ponieważ za każdym razem, gdy kulę po wybraniu wraca do kosza, zgodnie z regułą iloczynu istnieje $50^5 = 312500000$ możliwych sposobów wybrania 5 kul. Wśród nich jest tylko jeden przypadek, który pasuje do wyboru 11, 4, 17, 39, 23. Stąd prawdopodobieństwo wylosowania tych pięciu liczb w wyżej wymienionej kolejności wynosi $\frac{1}{312500000}$.

Prawdopodobieństwo kombinacji zdarzeń

- Twierdzenie: Niech E będzie zdarzeniem związanym z przestrzenią prób S. Prawdopodobieństwo zdarzenia E, będącego dopełnieniem E, jest dane wzorem: $P(\overline{E}) = 1\text{-}P(E)$.
- Przykład: Sekwencja dziesięciu bitów jest generowana losowo. Jakie jest prawdopodobieństwo, że przynajmniej jeden z bitów będzie miał wartość 0?
  - Niech E = Bity o długości 10 posiadające co najmniej jedno 0. $\overline{E}$ = Bity o długości dziesięć, niezawierające żadnego 0. Tak więc istnieje tylko jeden możliwy przypadek na korzyść $\overline{E}$. Przestrzeń próby zawiera ciągi $2^{10} = 1024$ bitów. Więc $P(\overline{E}) = \frac{1}{1024}$. Tzn., $P(E) = 1 - \frac{1}{1024} = \frac{1023}{1024}$.
- Twierdzenie: Niech E_1, E_2 będą zdarzeniami przestrzeni prób S. Wtedy $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$.
- Przykład: Jakie jest prawdopodobieństwo, że liczba całkowita dodatnia wybrana losowo z zestawu liczb całkowitych dodatnich nieprzekraczających 100 jest podzielna przez 2 lub 5?

Ograniczenia pojęcia prawdopodobieństwa Laplace'a

- Zakłada dwa warunki: (i) liczba wyników musi być skończona oraz (ii) każdy wynik jest równie prawdopodobny.
- Jednak wiele eksperymentów może nie mieć równie prawdopodobnych wyników i/lub skończenie wielu wyników.

(1) Przestrzeń próby odpowiadająca liczbie rzutów monetą do pierwszego orła to $\{1, 2, 3, \ldots\}$

(2) Moneta może być tak obciążona, że orzeł wypada dwa razy częściej niż reszka.

(3) Rozważmy, że rzucamy dwiema monetami i wygrywamy, jeśli wypadnie przynajmniej jeden orzeł, w przeciwnym razie przegrywamy. Jakie jest prawdopodobieństwo wygranej?

- Jedno rozwiązanie: Jeśli E jest zdarzeniem oznaczającym wygraną w grze, wtedy $E = \{HT, HH, TH\}$ gdzie przestrzeń próbki $S = \{HH, HT, TH, TT\}$. Więc, $P(E) = \frac{3}{4}$.
- Drugie rozwiązanie według D'Alemberta: Jeśli pierwsza moneta wypadnie orłem, drugi rzut nie będzie miał miejsca, ponieważ gra jest już wygrana. Więc, $E = \{H, TH\}$, $S = \{H, TH, TT\}$ i stąd $P(E) = \frac{2}{3}$. (Istnieją kontrowersje dotyczące drugiego rozwiązania.)

Wyzwania uogólniania koncepcji prawdopodobieństwa

- W przypadku nieskończonej przestrzeni prób pojęcie kardynalności zbioru punktów sprzyjających danemu zdarzeniu E zmienia się na pojęcie miary (ang. measure).
  Przykład miary dla zbioru nieskończonego: Praktycznym przykładem miary zbioru jest długość odcinka, gdzie zbiór jest podzbiorem osi liczbowej. Na przykład miarą zbioru punktów z przedziału $[a, b]$ jest jego długość, która wynosi $b - a$.
- Jeśli przestrzeń próby S jest nieprzeliczalnie nieskończona, to nie wszystkie podzbiory S (tj. wszystkie zdarzenia) są mierzalne (zgodnie z pojęciem σ-algebry).
- Czym jest σ-algebra: Składa się z mierzalnych zbiorów, dla których 'miarę' (sposób kwantyfikacji jego wielkości, taki jak długość, powierzchnia lub objętość) można zdefiniować w spójny sposób. Zbiory te są fundamentalne w teorii miary i są zamknięte na operacje takie jak dopełnienie i suma skończona, co oznacza, że jeśli zbiór jest mierzalny, jego dopełnienie również jest mierzalne, a suma skończonej liczby zbiorów mierzalnych również jest mierzalna.

Uogólnione pojęcie prawdopodobieństwa

- Kolmogorov (1933) wprowadził aksjomatyczną definicję miary prawdopodobieństwa dla σ-algebry $F \subseteq P(S)$ dla danej przestrzeni próbek S.
- Niech S będzie przestrzenią prób eksperymentu i $F \subseteq P(S)$ jest σ-algebrą. Prawdopodobieństwo jest funkcją $P : F \mapsto [0, 1]$ taką, że (i) $0 \leqslant P(E)$ dla dowolnego $E \in F$, (ii) $P(S) = 1$, i (iii) Dla dowolnego ciągu wzajemnie rozłącznych zdarzeń $E_1, E_2, \ldots \in F$, $P(\cup_{i=1}^{\infty} E_i) = \Sigma_{i=1}^{\infty} P(E_i)$.
- Z powyższych warunków można udowodnić (iv) $P(\phi) = 0$ gdzie ϕ oznacza niemożliwe zdarzenie i (v) $P(E) \leqslant 1$ dla dowolnego $E \in F$.
- Dla przeliczalnej przestrzeni prób S, $P(S)$ to σ-algebra.
- Definicję prawdopodobieństwa Laplace'a można uzyskać jako szczególny przypadek tej definicji.

Kombinacje zdarzeń

- Przykład: Jakie prawdopodobieństwa powinniśmy przypisać wynikom H (orzeł) i T (reszka) przy rzucie uczciwą monetą? Jakie prawdopodobieństwa powinniśmy przypisać tym zdarzeniom, gdy moneta jest obciążona tak, że orzeł wypada dwa razy częściej niż reszka?
  - W przypadku uczciwej monety prawdopodobieństwo H i T jest jednakowe. To znaczy $P(H) = P(T) = \frac{1}{2}$.
  - Dla stronniczej monety $P(H) = 2P(T)$.
    Więc, $P(H) + P(T) = 1$ daje nam $3P(T) = 1$.
    Stąd, $P(T) = \frac{1}{3}$ i $P(H) = \frac{2}{3}$.
- Każdy wynik $s \in S$ otrzymuje zatem wartość $P(\{s\}) \in [0, 1]$, i możemy to oznaczyć jako $P(s)$.
- Twierdzenie: Dla dowolnych dwóch zdarzeń E_1, E_2,
  (i) jeśli $E_1 \subseteq E_2$, wtedy $P(E_2 \setminus E_1) = P(E_2)$ - $P(E_1)$
  (ii) $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
  (iii) $P(E_1 \cup E_2) \leqslant P(E_1) + P(E_2)$.

Prawdopodobieństwo warunkowe

- Przykład: Załóżmy, że rzucamy trzy razy uczciwą monetą i wiemy, że zdarzenie F oznacza, że pierwszy rzut wypadnie reszką. Biorąc pod uwagę te informacje, jakie jest prawdopodobieństwo zdarzenia E, że wypadnie nieparzysta liczba reszek?
- Rozwiązanie: Widzimy, że $F = \{TTT, TTH, THT, THH\}$ i $E = \{THH, HTH, HHT, TTT\}$. Skoro już określono, że zdarzenie F już wystąpiło, musimy poszukać zdarzenia F i E, tzn. $F \cap E = \{THH, TTT\}$. Ponieważ istnieją 4 możliwe sposoby wystąpienia zdarzenia F, a wśród nich 2 są takie, że zdarzenie E występuje razem z zdarzeniem F, prawdopodobieństwo wystąpienia zdarzenia E, zakładając, że zdarzenie F już wystąpiło, wynosi $\frac{2}{4} = \frac{1}{2}$.
- Definicja: Niech E i F będą dwoma zdarzeniami takimi, że $P(F) > 0$. Prawdopodobieństwo warunkowe E przy danym F, oznaczone jako $P(E|F) = \frac{P(E \cap F)}{P(F)}$.

Niezależne wydarzenia

- Dwa zdarzenia E i F w tej samej przestrzeni próbkowania S są nazywane niezależnymi, jeśli $P(E \cap F) = P(E).P(F)$
- Warto uważyć: Zgodnie z tożsamością wynikającą z definicji prawdopodobieństwa warunkowego, jeśli wiemy, że zdarzenie F już wystąpiło, to $P(E \cap F) = P(E|F).P(F)$. Łącząc to teraz z definicją zdarzenia niezależnego, widzimy, że jeśli wystąpienie zdarzenia F nie ma żadnego wpływu na wystąpienie zdarzenia E, to $P(E \cap F) = P(E).P(F)$ - co oznacza $P(E|F) = P(E)$.
- Przykład: Niech E będzie zdarzeniem polegającym na tym, że losowo wygenerowany ciąg bitów o długości czterech zaczyna się od 1, a F zdarzeniem polegającym na tym, że ciągi bitów zawierają parzystą liczbę jedynek. Czy E i F są niezależne?

  - $E = \{1000, 1001, 1010, 1100, 1011, 1101, 1110, 1111\}$ i $F = \{0000, 0011, 0101, 0110, 1001, 1010, 1100, 1111\}$. Ponieważ w przestrzeni próbek znajduje się 16 ciągów bitowych $P(E) = P(F) = \frac{8}{16} = \frac{1}{2}$. $P(E \cap F) = \frac{4}{16} = \frac{1}{4} = P(E).P(F)$. Zatem E i F są zdarzeniami niezależnymi.

Twierdzenie Bayesa

- Prawdopodobieństwo sumy skończonej liczby zdarzeń: $P(\cup_{i=1}^{n} E_i) = \Sigma_{i=1}^{n} P(E_i)$ - $\Sigma_{i \neq j} P(E_i \cap E_j) +$ $\Sigma_{i \neq j \neq k} P(E_i \cap E_j \cap E_k) + \ldots + (-1)^n P(E_1 \cap E_2 \cap \ldots \cap E_n)$.
- Jeśli zdarzenia $E_1, E_2, \ldots, E_n$ wykluczają się wzajemnie (tj. $E_i \cap E_j = \phi$ dla $i \neq j$), wtedy $P(\cup_{i=1}^{n} E_i) = \Sigma_{i=1}^{n} P(E_i)$.
- Zdarzenia $E_1, E_2, \ldots, E_n$ w przestrzeni próbek S nazywamy wyczerpującym, jeśli za każdym razem, gdy przeprowadzany jest eksperyment, musi nastąpić jedno z tych zdarzeń.
- Niech $E_1, E_2, \ldots, E_n$ będą zbiorem wzajemnie wykluczających się i wyczerpujących zdarzeń w przestrzeni prób S, a B będzie dowolnym zdarzeniem w przestrzeni S. Wtedy $P(B) = \Sigma_{i=1}^{n} P(E_i \cap B) = \Sigma_{i=1}^{n} P(B|E_i)P(E_i)$. Nazywa się to całkowitym prawdopodobieństwem zdarzenia B.
- Twierdzenie Bayesa: Niech $E_1, E_2, \ldots, E_n$ będą zbiorem wzajemnie wykluczających się i wyczerpujących zdarzeń w przestrzeni prób S, a B będzie dowolnym zdarzeniem w przestrzeni S. Wtedy, $P(E_i|B) = \frac{P(B|E_i)P(E_i)}{\Sigma_{i=1}^{n} P(B|E_i).P(E_i)}$.

Przykład

- Załóżmy, że test laboratoryjny mający na celu wykrycie pewnej choroby ma następujące statystyki. Niech A oznacza zdarzenie, że badana osoba jest chora, a B oznacza zdarzenie, że wynik testu jest pozytywny. Wiadomo, że $P(B|A) = 0.99$, $P(B|\overline{A}) = 0.005$, i 0,1 procenta populacji faktycznie cierpi na tę chorobę. Jakie jest prawdopodobieństwo, że dana osoba jest chora, jeśli wynik testu będzie pozytywny?

  Rozwiązanie: $P(A) = 0.001$ i $P(\overline{A}) = 0.999$.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A).P(A)+P(B|\overline{A}).P(\overline{A})} = \frac{0.99\times 0.001}{0.99\times 0.001+0.005\times 0.999} = 0.165$$

Zmienna losowa

- Istnieje wiele problemów, w których musimy radzić sobie z liczbą wystąpień określonego wyniku eksperymentu. Na przykład, możemy potrzebować znać prawdopodobieństwo, że moneta wypadnie z 11 reszkami po rzucie 20 razy. W badaniu tego rodzaju problemów pomocne staje się użycie zmiennej losowej.
- Przykład: Załóżmy, że moneta jest rzucana trzy razy. Niech $X(s)$ reprezentuje liczbę orłów dla każdego $s \in S$, gdzie S jest przestrzenią prób dla eksperymentu polegającego na trzykrotnym rzucie monetą. Wtedy, $X(HHH) = 3$, $X(HHT) = X(HTH) = X(THH) = 2$, $X(TTH) = X(THT) = X(HTT) = 1$, i $X(TTT) = 0$.
- Definicja: Zmienna losowa to funkcja z przestrzeni prób eksperymentu do zbioru liczb rzeczywistych. Oznacza to, że zmienna losowa przypisuje liczbę rzeczywistą każdemu możliwemu wynikowi eksperymentu.

Rozkład prawdopodobieństwa zmiennej losowej

- Przykład: Nawiązując do poprzedniego przykładu, widzimy $P(X = 3) = \frac{1}{8}$, $P(X = 2) = \frac{3}{8} = P(X = 1)$, $P(X = 0) = \frac{1}{8}$, i $P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0) = 1$
- Definicja: Rozkład prawdopodobieństwa zmiennej losowej X w przestrzeni próby S to specyfikacja prawdopodobieństw $P(X = r)$ dla wszystkich możliwych wartości r takich, że $X(s) = r$ dla $s \in S$.
- Przykład funkcji rozkładu prawdopodobieństwa:

X	$P(X)$
0	$\frac{1}{8}$
1	$\frac{3}{8}$
2	$\frac{3}{8}$
3	$\frac{1}{8}$

- Właściwości funkcji rozkładu prawdopodobieństwa:
  Dla zmiennej losowej X w przestrzeni próby S,
  (i) $0 \leqslant P(X = r) \leqslant 1$ dla dowolnej r takiej, że $X(s) = r$ dla $s \in S$, i (ii) $\Sigma_{X=r} P(X = r) = 1$.

Funkcja dystrybucji skumulowanej

- Definicja: Funkcja dystrybucji skumulowanej (ang. cumulative distribution function) (cdf) dla zmiennej losowej X jest zdefiniowany jako $F_X(x) = P(X \leqslant x)$, gdzie -$\infty < x < \infty$
- Przykład: $F_X(0) = P(X \leqslant 0) = \frac{1}{8}$
  $F_X(1) = P(X \leqslant 1) = P(X = 0) + P(X = 1) = \frac{1}{8} + \frac{3}{8} = \frac{1}{2}$.
  $F_X(2) = P(X \leqslant 2) = P(X = 0) + P(X = 1) + P(X = 2)$
  $= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}$.
  $F_X(3) = P(X \leqslant 2) =$
  $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$
  $= \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$.

$$F_X(x) = \begin{cases} 0 & \text{if } -\infty < x < 0 \\ \frac{1}{8} & \text{if } 0 \leqslant x < 1 \\ \frac{1}{2} & \text{if } 1 \leqslant x < 2 \\ \frac{7}{8} & \text{if } 2 \leqslant x < 3 \\ 1 & \text{if } 3 \leqslant x < \infty \end{cases}$$

Dyskretna zmienna losowa i rozkład prawdopodobieństwa

- Właściwości F_X:

(i) $0 \leqslant F_X(x) \leqslant 1$

(ii) Jeśli $x_1 \leqslant x_2$, $F_X(x_1) \leqslant F_X(x_2)$

(iii) $lim_{x \to \infty} F_X(x) = 1$

(iv) $lim_{x \to -\infty} F_X(x) = 0$

- Znajdowanie prawdopodobieństw z F_X:
  - $P(a < X \leqslant b) = F_X(b) - F_X(a)$
  - $P(X > a) = 1 - P(X \leqslant a) = 1 - F_X(a)$
  - $P(X < b) = lim_{x \to b^-} F_X(x)$
- Zmienna losowa jest dyskretna, jeśli przyjmuje wartości z dyskretnego zbioru. Dla zmiennej losowej dyskretnej $F_X(x)$ reprezentuje funkcję schodkową.

Funkcja masy prawdopodobieństwa

- Dla dyskretnej zmiennej losowej prawdopodobieństwo w dowolnym przedziale oblicza się zasadniczo na podstawie prawdopodobieństw w pewnych dyskretnych punktach leżących w tym przedziale. Zatem dla dyskretnej zmiennej losowej X prawdopodobieństwo w dowolnym punkcie x nazywa się funkcją masy prawdopodobieństwa (ang. probability mass function lub pmf) X w x.
- $P(X = x) = p_X(x)$ gdzie p reprezentuje pmf zmiennej X.
- Dla dyskretnej zmienny przyjmującą wartości $x_1, x_2, \ldots,$ $p_X(x_i) = P(X = x_i) = F_X(x_i) - F_X(x_{i-1})$
- Właściwości pmf:

(i) $0 \leqslant p_X(x_i) \leqslant 1$ dla $i = 1, 2, \ldots$

(ii) $p_X(x) = 0$ jeśli $x \neq x_i$ dla dowolnej $i = 1, 2, \ldots$

(iii) $\Sigma_i p_X(x_i) = 1$

(iv) $F_X(x_i) = \Sigma_{j=1}^{i} p_X(x_j)$.

Ciągła zmienna losowa i rozkład prawdopodobieństwa

- Ciągła zmienna losowa X przyjmuje wartości ze zbioru ciągłego.
- W przypadku zbioru ciągłego, takiego jak przedział $[a, b]$, zmienna losowa X ma oczywiście nieprzeliczalnie wiele wartości do przyjęcia, a zatem rozkład prawdopodobieństwa X jest rozumiany jako skumulowany rozkład prawdopodobieństwa w całym przedziale. Zatem w przypadku zbioru ciągłego rozkład prawdopodobieństwa jest znany jako funkcja gęstości prawdopodobieństwa (ang. probability density function lub pdf.
- Funkcja gęstości prawdopodobieństwa dla zmiennej losowej X w zakresie wartości x to $f_X(x) = \frac{d}{dx} F_X(x)$
- Funkcja gęstości prawdopodobieństwa X w $[a, b]$: $\int_a^b f_X(x)dx$
- Właściwości pdf: (i) $f_X(x) \geqslant 0$ (ii) $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  (iii) $P(a \leqslant X \leqslant b) = \int_a^b f_X(x)dx$
  (iv) $P(a < X \leqslant b) = P(a \leqslant X \leqslant b) = P(a < X < b) = P(a \leqslant X < b) = F_X(b) - F_X(a)$.

Niektóre rozkłady dyskretne

1. Załóżmy, że X jest zmienną losową nad S. Dyskretny rozkład prawdopodobieństwa nazywa się rozkładem jednostajnym (ang. uniform distribution), jeśli $P(X = r)$ jest takie samo dla każdej wartości r, tak że $X(s) = r$ dla $s \in S$.
   - Przykład: Rozważmy eksperyment z rzutem kostką. Zatem $S = \{1, 2, 3, 4, 5, 6\}$. Niech X będzie zmienną losową na S taką, że $X(s) = s$. Wówczas $P(X = s) = \frac{1}{6}$ dla każdego $s \in S$.
   - Jeśli eksperyment jest taki, że każda próba może mieć tylko dwa możliwe wyniki, a prawdopodobieństwo każdego wyniku pozostaje takie samo we wszystkich próbach, każda próba takiego eksperymentu nazywana jest próbą Bernoulliego. Dwa możliwe wyniki próby Bernoulliego są zazwyczaj nazywane sukcesem i porażką.
2. Rozkład dwumianowy: Prawdopodobieństwo dokładnie k sukcesów w n niezależnych próbach Bernoulliego z prawdopodobieństwem sukcesu p i prawdopodobieństwem porażki $q = 1 - p$, wynosi $C(n, k)p^k q^{n-k}$

Przykład

- Jakie jest prawdopodobieństwo uzyskania dwóch orłów w 5 rzutach monetą?

  Rozwiązanie: Po próbach Bernoulliego wykonano tutaj 5 prób rzutu monetą i w każdej próbie prawdopodobieństwo sukcesu (czyli wyrzucenia orła) wynosi $\frac{1}{2}$. Więc, prawdopodobieństwo uzyskania 2 orłów 5 rzutach wynosi $C(5,2)(\frac{1}{2})^5 = \frac{10}{32}$.

- Zmienna losowa Bernoulliego: Zmienna losowa X ma dwie wartości: 'sukces' i 'porażkę', które można odpowiednio uznać za 1 i 0. Niech $P(X = 1) = p$ i $P(X = 0) = 1 - p$. pmf jest zdefiniwany jako: $p(X = i) = p^i(1 - p)^{1-i}$ gdzie $i = 0, 1$.

-

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - p & \text{if } 0 \leqslant x < 1 \\ 1 & \text{if } x \geqslant 1 \end{cases}$$

Rozkład Poissona

- To rozkład dwumianowy o dużej liczbie prób. Oznacza to, że $n \to \infty$ i w rezultacie p prawdopodobieństwo sukcesu staje się bardzo małe ($p \to 0$), a relacja między n i p jest taka, że np ($= \lambda$) jest stała.
- Przykład: Liczba połączeń telefonicznych przychodzących w ustalonym odstępie czasu. Niech $\lambda =$ długość odstępu czasu, n liczba przychodzących połączeń, a prawdopodobieństwo konkretnego połączenia będzie równe p. Wiec, $p = \frac{\lambda}{n}$. Teraz, jeśli n wzrasta, p maleje, ponieważ λ jest stała.
- X jest zmienną losową Poissona z parametrem λ (> 0), jeśli jej pmf $p_X(i) = \frac{e^{-\lambda}\lambda^i}{i!}$, $i = 0, 1, 2, \ldots$
- Warto wiedzieć: Prawdopodobieństwo otrzymania i liczby połączeń to $P(X = i) = \frac{n!}{i!(n-i)!}(\frac{\lambda}{n})^i(1 - \frac{\lambda}{n})^{n-i}$. Gdy $n \to \infty$ $P(X = i) \to \frac{e^{-\lambda}\lambda^i}{i!}$.

Rozkład prawdopodobieństwa ciągłego

Rozkład Gaussa (powszechnie znany jako rozkład normalny) jest prawdopodobnie najważniejszym ciągłym rozkładem prawdopodobieństwa w statystyce oraz uczeniu maszynowym. Opisuje on zmienne, których wartości grupują się symetrycznie wokół centralnej średniej.

- Funkcja gęstości prawdopodobieństwa (pdf) dla rozkładu Gaussa: Zmienna losowa ciągła X ma rozkład Gaussa ze średnią μ (parametr położenia) i wariancją σ^2 (parametr skali) - co zapisujemy jako $X \sim \mathcal{N}(\mu, \sigma^2)$ - jeśli jej funkcja gęstości prawdopodobieństwa (PDF) określona jest wzorem: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$.
  - $\frac{1}{\sqrt{2\pi}\sigma}$ jest stałą normalizacyjną. Zapewnia ona, że całkowity obszar pod krzywą jest dokładnie zintegrowany 1 (tzn. $\int_{-\infty}^{\infty} f(x)\, dx = 1$)
  - $e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ wskazuje, że w miarę oddalania się od średniej μ w dowolnym kierunku, wartość ta maleje wykładniczo.
- Krzywa jest idealnie symetryczna względem środka $x = \mu$. W rezultacie średnia, mediana i moda są identyczne i równe μ.
- Ogone rozkładu zbliżają się asymptotycznie do osi poziomej ($f(x) \to 0$, gdy $x \to \pm\infty$), tzn., każda liczba rzeczywista ma niezerową gęstość prawdopodobieństwa.

Właściwości rozkładu Gaussa

- Dla każdej standardowej krzywej Gaussa:
  • ok. 68,2% danych mieści się w granicach 1 odchylenia standardowego ($\mu \pm 1\sigma$).
  • ok. 95,4% danych mieści się w granicach 2 odchyleń standardowych ($\mu \pm 2\sigma$).
  • ok. 99,7% danych mieści się w granicach 3 odchyleń standardowych ($\mu \pm 3\sigma$).
- Standardowy rozkład normalny (Z): Gdy $\mu = 0$ oraz $\sigma = 1$, rozkład ten nazywany jest standardowym rozkładem normalnym, co tradycyjnie oznacza się za pomocą zmiennej losowej $Z \sim \mathcal{N}(0, 1)$. Jego funkcja gęstości prawdopodobieństwa upraszcza się do postaci: $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z)^2}$.
- Standaryzacja (transformacja do wyniku Z) Każdą dowolną zmienną gaussowską $X \sim \mathcal{N}(\mu, \sigma^2)$ można przekształcić w standardową zmienną normalną poprzez przesunięcie i przeskalowanie zgodnie z wzorem: $Z = \frac{X-\mu}{\sigma}$

Zastosowanie rozkładu Gaussa

- Centralne twierdzenie graniczne (Central limit theorem): Niezależnie od kształtu rozkładu populacji wyjściowej, suma (lub średnia) dużej liczby niezależnych zmiennych losowych o tym samym rozkładzie (independent and identically distributed (i.i.d.)) dąży do rozkładu gaussowskiego wraz ze wzrostem liczebności próby ($n \to \infty$).
- Oto dlaczego wzrost ludzi, wyniki testów, błędy pomiarowe oraz szum tła w naturalny sposób tworzą krzywą dzwonową.
- Własność maksymalnej entropii: Spośród wszystkich możliwych rozkładów ciągłych o znanej wartości oczekiwanej i znanej wariancji, rozkład Gaussa charakteryzuje się maksymalną entropią; tzn., zawiera najwięcej losowości, a zarazem najmniej systematycznego ukierunkowania (structural bias).

Wartość Oczekiwana ($\mathbb{E}$)

- Wartość Oczekiwana Funkcji Zmiennej Losowej: $\mathbb{E}[h(X)]$
  - Dla zmiennej dyskretnej (PMF): Przekształcamy każdą możliwą wartość x_i (w zakresie wartości, jakie przyjmuje zmienna losowa X) za pomocą funkcji $h(x_i)$, a następnie mnożymy przez jej oryginalne prawdopodobieństwo $p(x_i)$: $\mathbb{E}[h(X)] = \sum_i h(x_i) \cdot p(x_i)$
  - Dla zmiennej ciągłej (PDF): Podobnie jak wcześniej, całkujemy iloczyn funkcji $h(x)$ oraz gęstość prawdopodobieństwa $f(x)$ na całej przestrzeni Ω: $\mathbb{E}[h(X)] = \int_\Omega h(x) \cdot f(x)\, dx$
- Przykład: Wartość oczekiwana $\mathbb{E}[X]$ (często oznaczana jako μ) to teoretyczna średnia wartość zmiennej losowej przy nieskończonej liczbie powtórzeń eksperymentu.
- Wartość oczekiwana ze stałej: $\mathbb{E}[c] = c$
- Wyciąganie stałej przed operator (Jednorodność): $\mathbb{E}[aX] = a\mathbb{E}[X]$
- Wartość oczekiwana sumy (Adytywność): $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- Kombinacja Liniowa: $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
- Mnożenie zmiennych:
  $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ (gdy X, Y są niezależne)

Rodzina wykładnicza i sprzężenie

- Rodzina wykładnicza (Exponential Family) scala niemal wszystkie powszechnie stosowane rozkłady ciągłe i dyskretne (np. rozkład Gaussa, Bernoulliego, Poissona) w ramach jednego, nadrzędnego formalizmu matematycznego.
- Gęstość kanoniczna rozkładu z rodziny wykładniczej składa się z czterech elementów: miary bazowej $h(x)$, parametru naturalnego η, statystyki dostatecznej[2] $T(x)$ oraz logarytmu funkcji partycji $A(\eta)$. $f(x|\eta) = h(x)exp(\eta T(x) - A(\eta))$.
- Kluczowe spostrzeżenie Postać kanoniczna rozdziela dane od parametru. Funkcja $h(x)$ zależy wyłącznie od danych, $A(\eta)$ zależy wyłącznie od parametru, a iloczyn $\eta T(x)$ stanowi jedyny punkt ich styku. To właśnie pozwala później sprowadzić cały zbiór danych do podsumowania o ustalonym rozmiarze.

[2] Statystyka dostateczna to pojedyncza liczba (lub niewielki zestaw liczb), wyznaczona na podstawie próby, która zawiera pełną informację, jaką dane niosą na temat nieznanego parametru. Gdy znamy już statystykę dostateczną, pozostałe surowe dane nie wnoszą żadnych dodatkowych informacji; w rezultacie estymację, analizę wiarygodności oraz wnioskowanie statystyczne można w całości przeprowadzić w oparciu o to zwięzłe podsumowanie.

Do sprawdzenia

- Proszę identyfikować rozkładu normalnego jako kanonicznej gęstości prawdopodobieństwa z rodziny wykładniczej
- Wskazówka:
  - Rozważ $f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$
  - Rozwiń wyrażenie kwadratowe $-(\frac{x-\mu}{\sigma})^2$ (Sęk w tym, by dostrzeć, że μx staje się $\eta T(x)$, a $\frac{\mu^2}{2}$ staje się $A(\eta)$).
  - Przenieś stałą normalizacyjną $\frac{1}{\sqrt{2\pi}\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}}$ do wnętrza wykładnika
  - $f(x \mid \mu, \sigma^2) =$
    $\exp\left(\left(\frac{\mu}{\sigma^2}\right) x + \left(-\frac{1}{2\sigma^2}\right) x^2 - \left[\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(\sigma^2)\right] - \frac{1}{2}\log(2\pi)\right)$

Zmiana zmiennych i przekształcenie odwrotne

- W przypadku pojedynczej zmiennej losowej Niech X będzie zmienną losową ciągłą o znanej gęstości prawdopodobieństwa $f_X(x)$. Niech $Y = g(X)$ będzie przekształceniem, w którym g jest funkcją gdzie $g(x)$ jest funkcją ściśle rosnącą lub ściśle malejącą (co oznacza, że jest ona różnowartościowa i odwracalna oraz różniczkowalną). Gęstość prawdopodobieństwa zmiennej przekształconej Y wyraża się wzorem: $f_Y(y) = f_X(g^{-1}(y)).|\frac{d}{dy}g^{-1}(y)|$ gdzie $g^{-1}(y)$ Jest to przekształcenie odwrotne. Odwzorowuje ono wartości z nowej przestrzeni (Y) z powrotem do przestrzeni wyjściowej (X) i $|\frac{d}{dy}g^{-1}(y)|$ pełni funkcję współczynnika korygującego (pochodna funkcji odwrotnej).

Przypadek wielowymiarowy (wiele zmiennych losowych)

- W przypadku wektora ciągłych zmiennych losowych (na przykład przy przekształcaniu (X_1, X_2) na (Y_1, Y_2) za pomocą równań $Y_1 = g_1(X_1, X_2)$ oraz $Y_2 = g_2(X_1, X_2)$) współczynnik poprawkowy przyjmuje postać wyznacznika. Jeśli przekształcenie łączne jest wzajemnie jednoznaczne, najpierw wyznaczamy przekształcenia odwrotne: $X_1 = h_1(Y_1, Y_2) \quad \text{oraz} \quad X_2 = h_2(Y_1, Y_2)$
Wielowymiarowy wzór na łączną gęstość prawdopodobieństwa (PDF) ma postać:
$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}\big(h_1(y_1, y_2), h_2(y_1, y_2)\big) \cdot |J|$
gdzie $|J|$ jest wartością bezwzględną wyznacznika jakobianu przekształcenia odwrotnego: $J =$

$$\begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{bmatrix}$$

- Zastosowanie: To właśnie to zachowanie stanowi podstawę metody próbkowania przez transformację odwrotną (ang. *Inverse Transform Sampling*) – fundamentalnego sposobu, z którego korzystają komputery w celu symulowania danych losowych.

Testy hipotez: parametryczne a nieparametryczne

Testowanie hipotez statystycznych dzieli się na dwie główne gałęzie, w zależności od założeń dotyczących struktury populacji.

- Testy parametryczne
  - Zakłada, że dane podlegają określonej rodzinie rozkładów (zazwyczaj rozkładowi Gaussa).
  - Koncentruje się na estymacji określonych parametrów (np. średniej μ, wariancji σ^2).
  - Wymaga danych ciągłych mierzonych w skali przedziałowej lub ilorazowej.
- Testy nieparametryczne
  - Nie przyjmuje żadnych założeń strukturalnych dotyczących populacji.
  - Koncentruje się na rangach, medianach lub rozkładach strukturalnych, a nie na sztywnych wskaźnikach.
  - Bezproblemowo obsługuje dane typu porządkowego, nominalnego, rangowego lub dane o silnie asymetrycznym rozkładzie.

Regresja Liniowa (Linear Regression)

Regresja liniowa służy do modelowania związków pomiędzy ciągłą zmienną zależną (objaśnianą) Y a jedną lub wieloma zmiennymi niezależnymi (objaśniającymi) X.

- Dla zbioru danych zawierającego n obserwacji i p cech, model wielokrotnej regresji liniowej przyjmuje postać: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$ gdzie:
  - β_0 to wyraz wolny (intercept).
  - $\beta_1, \ldots, \beta_p$ to współczynniki regresji (wagi).
  - ϵ to błąd losowy (resztowy), dla którego zakłada się rozkład normalny: $\epsilon \sim \mathcal{N}(0, \sigma^2)$.
- W zapisie macierzowym: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$

Estymacja Parametrów: Metoda Najmniejszych Kwadratów (MNK)

- Celem Klasycznej Metody Najmniejszych Kwadratów (ang. Ordinary Least Squares – OLS) jest zminimalizowanie sumy kwadratów reszt (RSS): $RSS(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x_i^T}\beta)^2 = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$
  - Obliczenie pochodnej względem wektora β i przyrównanie jej do zera prowadzi do tzw. równań normalnych: $\mathbf{X^T X}\beta = \mathbf{X^T Y}$
  - Przy założeniu, że macierz $\mathbf{X}^T\mathbf{X}$ jest odwracalna, jawne rozwiązanie dla parametrów wynosi: $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T Y}$

Regresja Logistyczna (Logistic Regression)

Regresję logistyczną stosuje się, gdy zmienna zależna Y jest kategoryczna (najczęściej binarna, czyli $Y \in \{0, 1\}$). Zamiast przewidywać bezpośrednio wartość Y, modeluje ona prawdopodobieństwo sukcesu $p = P(Y = 1|X)$.

- Funkcja Logistyczna (Sigmoid): Aby przekształcić dowolną wartość liniową na zakres prawdopodobieństwa (), stosujemy funkcję sigmoidalną: $\sigma(z) = \frac{1}{1+e^{-z}}$
  - Modelujemy logarytm szans (ang. log-odds lub logit) jako liniową kombinację cech wejściowych:
    $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p = \mathbf{X}\beta$
    Wyznaczając p, otrzymujemy funkcję prognozy modelu:
    $p = P(Y = 1|\mathbf{X}) = \frac{\mathbf{1}}{\mathbf{1}+\mathbf{e}^{-\mathbf{X}\beta}}$

Estymacja Parametrów: Metoda Największej Wiarygodności (MNW)

- W przeciwieństwie do regresji liniowej, regresja logistyczna nie posiada rozwiązania w postaci jawnego wzoru algebraicznego. Parametry szacuje się za pomocą Metody Największej Wiarygodności (ang. Maximum Likelihood Estimation - MLE), minimalizując funkcję straty logarytmicznej (Binary Cross-Entropy/Log Loss):
  - $L(\beta) = -\frac{1}{n} \sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)]$

    Optymalizację przeprowadza się numerycznie za pomocą algorytmów iteracyjnych, takich jak Metoda Gradientu Prostego (Gradient Descent) lub metoda Newtona-Raphsona.

Metody Monte Carlo

Metody Monte Carlo to szeroka klasa algorytmów obliczeniowych, które wykorzystują wielokrotne losowanie probabilistyczne do uzyskania wyników numerycznych. Są niezastąpione przy rozwiązywaniu problemów deterministycznych, które są trudne lub niemożliwe do rozwiązania analitycznego (np. całki wielowymiarowe).

- Główna Zasada: Prawo Wielkich Liczb

  Załóżmy, że chcemy obliczyć wartość oczekiwaną funkcji h(X), gdzie X ma gęstość prawdopodobieństwa f(X):
  $I = \mathbb{E}[h(X)] = \int_{\Omega} h(x) f(x)\, dx$

- Jeśli analityczne obliczenie całki jest niemożliwe, możemy wylosować N niezależnych próbek $x_1, x_2, \ldots, x_N$ z rozkładu f(x). Na mocy Mocnego Prawa Wielkich Liczb, średnia z próby zbiega się do prawdziwej wartości oczekiwanej wraz z N $\rightarrow$:
  $\hat{I}_N = \frac{1}{N} \sum_{i=1}^{n} h(x_i) \xrightarrow{\text{p.w.}} I$

- Zgodnie z Centralnym Twierdzeniem Granicznym, błąd statystyczny estymatora Monte Carlo maleje w tempie proporcjonalnym do liczby prób: Błąd $\approx \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$. Tzn., aby zmniejszyć błąd dwukrotnie, musimy zwiększyć liczbę losowań N aż czterokrotnie.

Różniczkowanie funkcji jednowymiarowych

- Definicja pochodnej: Dla funkcji $f : \mathbb{R} \to \mathbb{R}$ pochodna mierzy chwilową szybkość zmian wartości funkcji względem zmiany argumentu, co wyraża wzór $f'(x) = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$.
- Interpretacja geometryczna: Wartość $f'(x_0)$ oznacza współczynnik kierunkowy stycznej do wykresu funkcji w punkcie $(x_0, f(x_0))$.
- Reguła łańcuchowa (Chain Rule): Dla funkcji złożonej $h(x) = f(g(x))$ pochodna jest iloczynem pochodnych funkcji zewnętrznej i wewnętrznej: $h'(x) = f'(g(x)) \cdot g'(x)$.
- Przykład: Niech $h(x) = e^{x^2+11}$ gdzie $f(x) = e^x$ i $g(x) = x^2 + 11$. $h'(x) = e^{x^2+11}.(2x)$

Pochodne cząstkowe i gradienty

- Pochodna cząstkowa: Dla funkcji skalarnej wielu zmiennych $f : \mathbb{R}^n \to \mathbb{R}$ pochodna cząstkowa $\frac{\partial f}{\partial x_i}$ to pochodna względem zmiennej x_i przy założeniu, że wszystkie pozostałe zmienne są stałe.
- Wektor gradientu: Gradient $\nabla f(\mathbf{x})$ to wektor kolumnowy zbierający wszystkie pochodne cząstkowe pierwszego rzędu w jednym obiekcie. $\nabla f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \dots \quad \frac{\partial f}{\partial x_n}\right]^T \in \mathbb{R}^n$
- Własności kierunkowe: Wektor gradientu wskazuje kierunek najszybszego wzrostu funkcji w danym punkcie, a jego norma $\|\nabla f(\mathbf{x})\|$ określa tempo tego wzrostu.
- Zastosowanie w optymalizacji: W metodzie spadku gradientowego (Gradient Descent) parametry aktualizuje się w kierunku przeciwnym do gradientu: $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta \nabla f(\mathbf{x}_k)$ gdzie parametr η to współczynnik uczenia (ang. learning rate) lub długość kroku.
- Wektor $-\nabla f(\mathbf{x}_k)$ wskazuje wyłącznie lokalny kierunek najszybszego spadku wartości funkcji, a skalar $\eta > 0$ decyduje o skali przesunięcia wzdłuż tego kierunku. η określa, jak dużą zmianę parametrów $\mathbf{x}$ wykonamy w pojedynczej iteracji w kierunku wyznaczonym przez ujemny gradient.

Przykład analityczny

- Rozważmy prostą funkcję kwadratową $f(x) = x^2$, której minimum znajduje się w punkcie $x^* = 0$. Gradient wynosi $\nabla f(x) = 2x$, więc reguła aktualizacji przyjmuje postać:
  $x_{k+1} = x_k - \eta(2x_k) = (1 - 2\eta)x_k$
- Gdy $\eta = 0.1$: W każdym kroku $x_{k+1} = 0.8x_k$, więc ciąg monotonicznie i płynnie zbiega do zera.
- Gdy $\eta = 0.5$: W pierwszym kroku x = (1 - 1)x = 0, co oznacza trafienie dokładnie w minimum w jednej iteracji.
- Gdy $\eta = 0.9$: W każdym kroku $x_{k+1} = -0.8x_k$, więc punkt przeskakuje na drugą stronę zera (oscyluje), ale amplituda maleje i algorytm nadal zbiega.
- Gdy $\eta > 1.0$: Czynnik $|1 - 2\eta| > 1$, co powoduje coraz większe skoki na boki i ucieczkę do nieskończoności (rozbieżność).

Gradienty funkcji o wartościach wektorowych (Jakobian)

- Funkcja wektorowa: Rozważamy odwzorowanie $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, gdzie każda ze składowych $f_i(\mathbf{x})$ zależy od wektora wejściowego $\mathbf{x} \in \mathbb{R}^n$.
- Macierz Jacobiego (Jakobian): Uogólnieniem gradientu jest macierz $\mathbf{J} \in \mathbb{R}^{m \times n}$, której wiersze to transponowane gradienty kolejnych funkcji składowych. $\mathbf{J_f}(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$
- Uogólniona reguła łańcuchowa: Dla złożenia funkcji wektorowych $\mathbf{h}(\mathbf{x}) = \mathbf{f}(\mathbf{g}(\mathbf{x}))$ jakobian złożenia jest iloczynem macierzowym jakobianów składowych: $\mathbf{J_h}(\mathbf{x}) = \mathbf{J_f}(\mathbf{g}(\mathbf{x})) \cdot \mathbf{J_g}(\mathbf{x})$.

Gradienty macierzy

- Konwencje zapisu: W literaturze stosuje się konwencję licznikową (wymiar wyniku zgodny z licznikiem) lub mianownikową (wymiar zgodny z mianownikiem, częstsza w uczeniu maszynowym).
- Pochodna funkcji skalarnej względem macierzy: Dla $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ gradient względem macierzy $\mathbf{X}$ jest macierzą o identycznych wymiarach $m \times n$ gdzie $(\nabla_{\mathbf{X}} f(\mathbf{X}))_{ij} = \frac{\partial f(\mathbf{X})}{\partial X_{ij}}$

Tożsamości przydatne w obliczeniach gradientów

Zestawienie najważniejszych reguł dla wektorów $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$ oraz macierzy $\mathbf{A}$ (w konwencji mianownikowej):

$f(\mathbf{x})$ lub $f(\mathbf{X})$	Gradient ∇f	Uwagi
$\mathbf{a}^T\mathbf{x}$	$\mathbf{a}$	Iloczyn skalarny (funkcja liniowa)
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$(\mathbf{A}+\mathbf{A}^T)\mathbf{x}$	Forma kwadratowa ogólna
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$2\mathbf{A}\mathbf{x}$	Forma kwadratowa dla symetrycznej $\mathbf{A}$
$\lVert\mathbf{A}\mathbf{x}-\mathbf{b}\rVert_2^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b})$	Klasyczny problem najmniejszych kwadratów
$\mathrm{Tr}(\mathbf{A}\mathbf{X})$	$\mathbf{A}^T$	Pochodna śladu iloczynu macierzy
$\ln\lvert\det(\mathbf{X})\rvert$	$(\mathbf{X}^{-1})^T$	Dla odwracalnej macierzy kwadratowej $\mathbf{X}$

Propagacja wsteczna i różniczkowanie automatyczne

- Graf obliczeniowy: Dowolną złożoną funkcję można przedstawić jako skierowany graf acykliczny (DAG), w którym węzły reprezentują operacje elementarne, a krawędzie przepływ danych.
- Różniczkowanie automatyczne (AD): Nie jest to ani różniczkowanie symboliczne (prowadzące do eksplozji wyrażeń), ani numeryczne (obciążone błędem zaokrągleń), lecz dokładne obliczanie wartości pochodnych na grafie.
- Tryb w przód (Forward-mode AD): Propaguje pochodne równolegle z wartościami funkcji od wejść do wyjść; jest wydajny dla funkcji typu $\mathbb{R} \to \mathbb{R}^m$.
- Tryb wsteczny (Reverse-mode AD / Backpropagation): Najpierw oblicza wartości w przód, a następnie propaguje gradienty od pojedynczego wyjścia (funkcji straty $L \in \mathbb{R}$) do wszystkich wejść w jednym przebiegu; jest fundamentem uczenia głębokich sieci neuronowych ($\mathbb{R}^n \to \mathbb{R}$).

Pochodne wyższych rzędów

- Pochodne mieszane: Drugie pochodne cząstkowe opisują krzywiznę funkcji; zgodnie z twierdzeniem Schwarza dla funkcji klasy C^2 [3]kolejność różniczkowania nie ma znaczenia: $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.
- Macierz Hessego (Hesjan): Dla funkcji $f : \mathbb{R}^n \to \mathbb{R}$ hesjan $\mathbf{H} \in \mathbb{R}^{n \times n}$ to symetryczna macierz wszystkich drugich pochodnych cząstkowych. $\mathbf{H}_f(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$
- Klasyfikacja punktów krytycznych ($\nabla f(\mathbf{x}_0) = \mathbf{0}$):
  • Gdy $\mathbf{H}(\mathbf{x}_0)$ jest dodatnio określona (wszystkie wartości własne > 0), w punkcie $\mathbf{x}_0$ istnieje minimum lokalne.
  • Gdy $\mathbf{H}(\mathbf{x}_0)$ jest ujemnie określona (wszystkie wartości własne < 0), w punkcie $\mathbf{x}_0$ istnieje maksimum lokalne.
  • Gdy $\mathbf{H}(\mathbf{x}_0)$ jest nieokreślona (wartości własne różnych znaków), punkt $\mathbf{x}_0$ jest punktem siodłowym.

[3]Funkcja klasy C^2 to funkcja, która jest dwukrotnie różniczkowalna w sposób ciągły.

Linearyzacja i wielowymiarowe szeregi Taylora

- Linearyzacja (Aproksymacja pierwszego rzędu): W otoczeniu punktu $\mathbf{x}_0$ funkcję różniczkowalną można przybliżyć hiperpłaszczyzną styczną. $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0)$
- Aproksymacja kwadratowa (Szereg Taylora drugiego rzędu): Uwzględnienie hesjanu pozwala uchwycić lokalną krzywiznę funkcji, co wykorzystuje m.in. metoda Newtona w optymalizacji. $f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \mathbf{H}_f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$
- Ogólny wielowymiarowy wzór Taylora: Dla przesunięcia $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ rozwinięcie rzędu k zapisuje się z użyciem notacji wielowskaźnikowej jako sumę pochodnych kolejnych rzędów pomnożonych przez odpowiednie potęgi przyrostów.

Dziękuję za uwagę