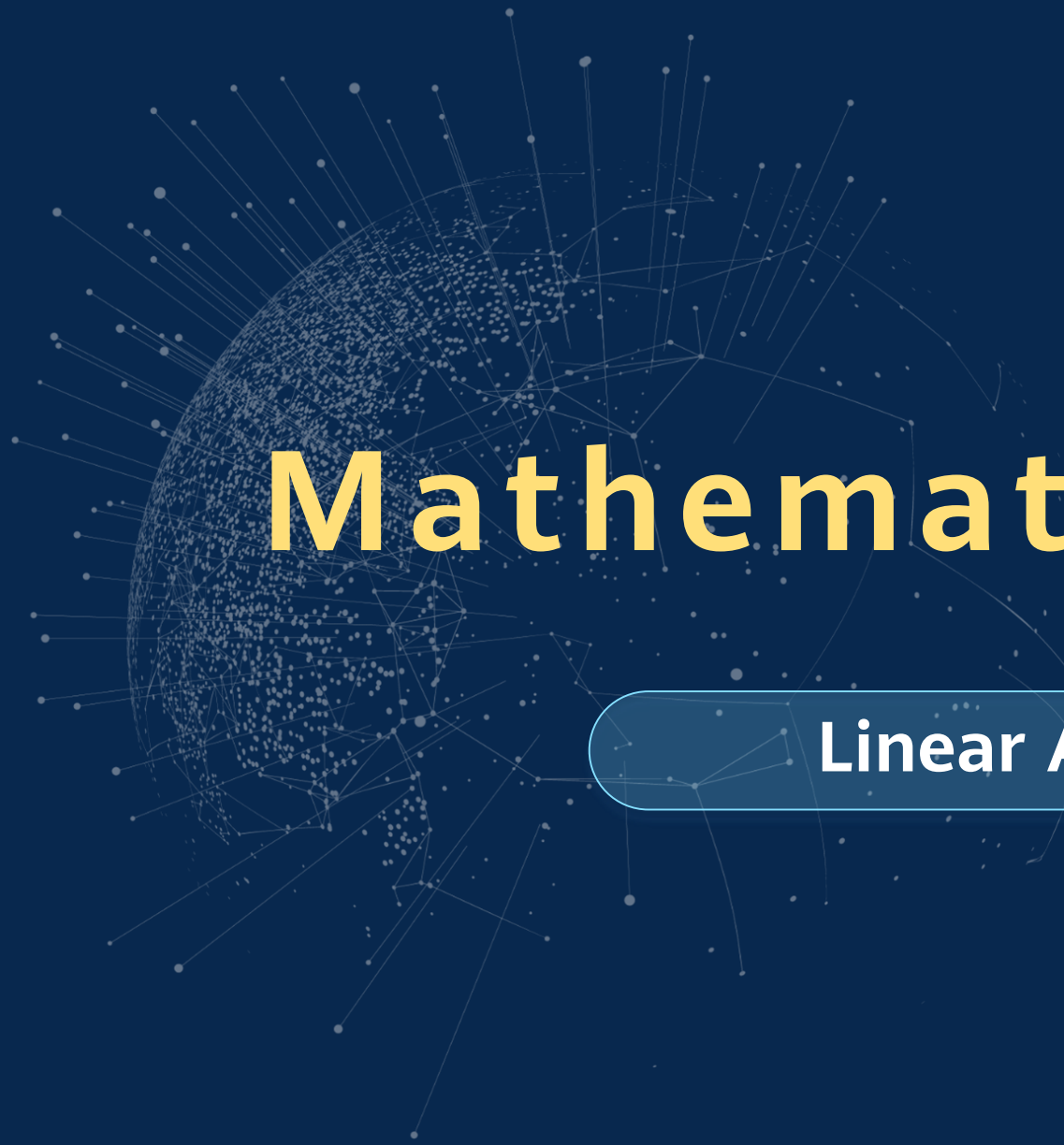


Mathematics Basics

Linear Algebra

HUAWEI

Contents

- **Mathematics and AI**
  - **Linear Algebra**
    - Basics Concepts and operations
    - Linear transformation and matrix
    - Matrix Decompositions
  - **Statistics and Probability**
- **Optimization Problems**

Why Matrix?

- In data management, the fundamental entity is a relation.
  - SQL(relation algebra) is a collection of operations on relations- Select, project, join, group-by.
- In artificial intelligence and data mining, the fundamental entity is a matrix,
  - Therefore we need to know the basic operations on matrices.
  - One important application is to summarise data or extract patterns from data or compress data.

Scalar, Vector and Matrix

- Scalar: a real number, for example: $x = 3$.
- Vector: an array of numbers that are arranged in order, for example: $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
- Matrix: a rectangular array of $m \times n$ numbers arranged in m rows and n columns. The numbers can be represented as $a_{ij}(i = 1, 2, \cdots, m; j = 1, 2, \cdots, n)$, and the matrix can be represented as

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

  - A special case of a matrix is a vector.
  - Square matrix when $m = n$.
  - Each row or column could represent one object. If rows are object then columns are features/ attributes/components.

Numbers vs Matrices

Numbers

- Can add and subtract numbers
- Multiply numbers
- Divide two numbers a/b as long as b is not 0.
- Can factorise positive numbers into product of primes

Matrices

- Can add and substract compatible matrices
- Multiply and divide matrices
- Division of matrices is complicated
- Can factorise any matrix to get data patterns(using a technique called Singular Value Decomposition)

Matrix Operation

- Matrix addition:

  $C_{m\times n} = A_{m\times n} + B_{m\times n}$. Adding matrix A and matrix B together means adding the corresponding entries in these two matrices together: $c_{ij} = a_{ij} + b_{ij}$

  Note: To perform the addition operation on matrix A and matrix B, the number of rows in matrix A must be the same as that in matrix B, and the number of columns in matrix A must be the same as that in matrix B.

- Scalar multiplication:

  If $A = (a_{ij})_{m\times n}$, $k \in K$, the multiplication of scalar k with matrix A is defined as $kA = (ka_{ij})_{m\times n}$.

- Matrix multiplication:

  If matrices $A = (a_{ij})_{m\times n}$, $B = (b_{ij})_{n\times p}$, then

  $$C = AB = (c_{ij})_{m\times p}$$

  Note: AB is meaningful only if the number of columns of matrix A equals the number of rows of matrix B.

$$A_{n\times m} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \quad a_i$$

$$B_{m\times p} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mp} \end{bmatrix} \quad b_j$$

$$c_{ij} = a_i b_j = \sum_k a_{ik} b_{kj}$$

Example: Linear Classifier

- For example, to recognize the category (cat, car, or human) of an image, we can process it into a matrix consisting of six pixels, and use the simplest linear classifier:

$$y = wx + b$$

Convert the matrix of this image into a column vector.

w

0	-1	-0.5	0.5	0	0
1	2	3	2	0.5	-0.5
0.5	1	0	-0.5	0	0

x

0.5
0.5
-2
2
4
1

+

b

-1
2
-0.5

=

0.5	Score of the probability of car
3	Score of the probability of cat
-0.75	Score of the probability of human

Common Representation

- Image/ Video
- Text/ Comments
- Times series
- System logs
- Network
- Tabular/ Rating

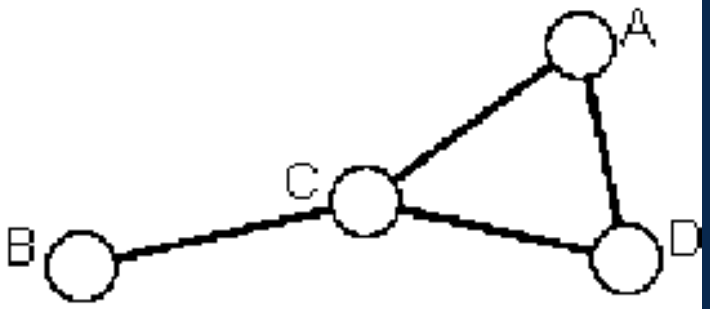
- How can we represent data of different types?
- We represent most data types as a matrix.

Network Data

Nodes

Nodes $\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$

	A	B	C	D
A	0	0	1	1
B	0	0	1	0
C	1	1	0	1
D	1	0	1	0



[Merrill,2008] Merrill et al., Findings from an Organizational Network Analysis to Support Local Public Health Management. Journal of urban health : bulletin of the New York Academy of Medicine.

Text to Matrix

- Document- Word Matrix
- Document 1: “AABBCCAA”
- Document 2: “AABBCCDD”
- Document 3: “BBDDDDDC”

A	B	C	D
4	2	2	0
2	2	2	2
0	2	1	5

Image Data

- In the field of computer vision, an image is stored in the form of matrix, and this matrix is usually converted into a vector in image processing.

$$A_{m*n} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \ddots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

$$A_{m*n} = \begin{bmatrix} a_{11} \\ a_{12} \\ a_{13} \\ \vdots \\ a_{21} \\ \vdots \\ a_{mn} \end{bmatrix}$$

Similarity Computation

- Two vectors x and y

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Dot product

$$x \,.\, y = (x_1 y_1 + x_2 y_2 + \cdots + x_n y_n)$$

- Norm(length) of a vector

$$\|x\| = (x.x)^{1/2} = (x_1 x_1 + x_2 x_2 + \cdots + x_n x_n)^{1/2}$$

- Similarity between two vectors x and y

$$sim(x \,.\, y) = x.y/(\|x\|\|y\|)$$

Determinant

- The determinant is a scalar value that can be computed from the elements of a square matrix. The determinant of a matrix A is denoted $\det(A), \det A, or\ |A|$.
- $\det(A)^{-1} = \frac{1}{\det(A)}$ telling us if the matrix is invertible.
- For 2 by 2 matrix: $|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = (ad - bc)$
- For 3 by 3 matrix:

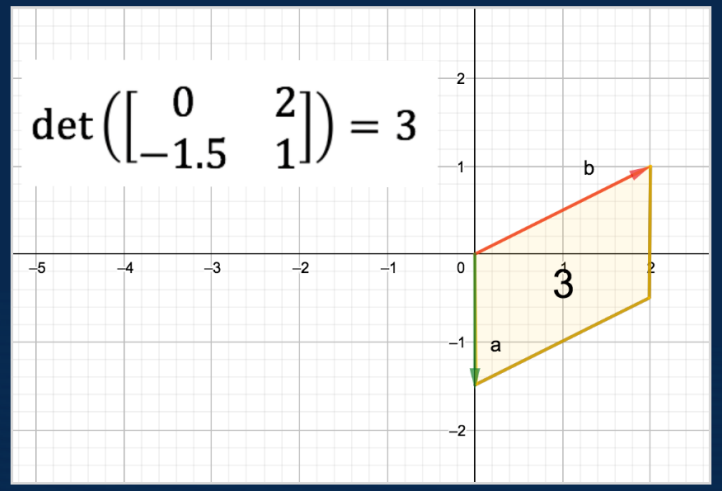
$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

$$= \mathrm{a(ei - fh) - b(di - fg) + c(dh - eg)}$$

$$= aei + bfg + cdh - ceg - bdi - afh$$

Example: Determinant

- Calculation of the second-order determinant: $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$
- The value of the second order determinant $D = \det\left(\begin{bmatrix} 0 & 2 \\ -1.5 & 1.0 \end{bmatrix}\right)$ indicates the signed area of the parallelogram spanned by vector $a = (0, -1.5)^T$ and $b = (2,1)^T$.



Contents

- **Mathematics and AI**
  - **Linear Algebra**
    - Basics Concepts and operations
    - Linear transformation and matrix
    - Matrix Decompositions
  - **Statistics and Probability**
- **Optimization Problems**

Linear Transformation

Definition: Suppose that V_n is an n-dimensional linear space and U_m be an m-dimensional linear space. T, which is a mapping from V_n to U_m, can be regarded as a linear mapping or linear transformation if the following two conditions are satisfied:

(1) For any two elements $\alpha_1, \alpha_2 \in V_n$ (hence $\alpha_1 + \alpha_2 \in V_n$), we have
$T(\alpha_1 + \alpha_2) = T(\alpha_1) + T(\alpha_2)$
(2) For any element $\alpha \in V_n, \lambda \in R$ (hence $\lambda\alpha \in V_n$), we have

$$T(\lambda\alpha) = \lambda T(\alpha)$$

Linear algebra can be regarded as a study of discussing space transformation and vector motion, which are both achieved by linear transformation.
Application of linear transformation: In deep learning, linear transformation is most commonly used to enhance the image or speech dataset through matrix multiplication. For example, existing images are translated, rotated, or scaled to generate new images.

Examples — Matrix and Motion (1)

For an arbitrary point P that moves to the point $P' = A_i \mathrm{P}\ (i = 1,2,3, \dots)$, what transformation do the following matrices represent, respectively?

$$(1) A_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad (2)\ A_2 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \qquad (3) A_3 = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

Solution: For any arbitrary point $\mathrm{P} = \begin{pmatrix} x \\ y \end{pmatrix}$, we have

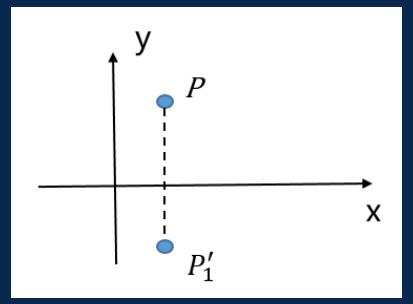
(1) $P_1{}' = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ -y \end{pmatrix}$

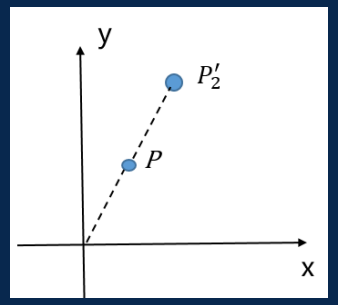
(2) $P_2{}' = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda y \end{pmatrix}$

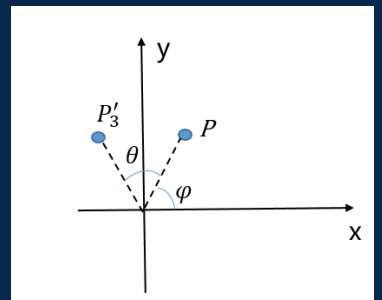
(3) $P_3{}' = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{pmatrix} r\cos(\rho) \\ r\sin(\rho) \end{pmatrix} = \begin{bmatrix} r\cos(\theta + \rho) \\ r\sin(\theta + \rho) \end{bmatrix}$

Examples — Matrix and Motion (2)

- In the transformation defined by A_1, the point P lands on its reflection point P_1' across the x-axis.
- In the transformation defined by A_2, the point P_2' lands on the line that passes the point P and the origin. The scale factor of the distance from P to the origin is the value of λ.
- In the transformation defined by A_3, the point P rotates ϑ degree around the origin and lands on the point P_3'.







In the preceding examples, matrix A that defines the transformation of point P is the transformation matrix. The transformation defined by A_1 is the reflection or mirroring transformation. The transformation defined by A_2 is the scaling transformation, and λ is the scale factor. The transformation defined by A_3 is the rotation transformation.

In a linear space, one matrix represents one linear transformation, such as rotation, scaling, and reflection. Vectors are transformed by matrix multiplication.

Matrix Transposition

- Transposed matrix: The transpose of a matrix is an operator which flips a matrix over its diagonal, that is, it switches the row and column indices of the matrix by producing another matrix denoted as A^T (also written as A').

Example:

$$A=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \Rightarrow A^T=\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \end{bmatrix}.$$

- Nature of a transposed matrix:
  - $(A^T)^T = A$
  - $(\lambda A)^T = \lambda A^T$
  - $(A+B)^T = A^T + B^T$
  - $(AB)^T = B^T A^T$

Diagonal Matrix

Diagonal matrix: a matrix in which the entries outside the main diagonal are all zeroes. It is usually represented by $diag(\lambda_1, \lambda_2, \cdots, \lambda_n)$, and denoted as

$$D = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}$$

Properties of the diagonal matrix:

- The sum, difference, product, and power of the diagonal matrix are those of the entries on the main diagonal.
- The inverse of the diagonal matrix is denoted as $D^{-1} = \begin{bmatrix} {\lambda_1}^{-1} & 0 & \dots & 0 \\ 0 & {\lambda_2}^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & {\lambda_3}^{-1} \end{bmatrix}$.

The multiplication or inversion of diagonal matrices is efficient and require less computing. (The matrix inversion is only applicable to square matrices.) Therefore, diagonal matrices are required in some cases to reduce computational cost.

Identity, Inverse and Orthogonal Matrix

- **Identity matrix** $\mathrm{I_n}$: $n \times n$ square matrix with 1 on the main diagonal are 1, while others are 0.
  - $\mathrm{I_n}$ is an example of diagonal matrix
  - If A is a square matrix, $AI = IA = A$

$$\mathrm{I_n} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- **Inverse Matrix**: The **matrix inverse** of A is denoted as $\mathrm{A^{-1}}$, and it is defined as the matrix such that $\mathrm{A^{-1}A = I_n}$.
- Orthogonal matrix: If $AA^T = A^TA = I_n$ in the square matrix $A = (a_{ij})_{n\times n}$, A is an orthogonal matrix. That is, $A^{-1} = A^T$.

Symmetric Matrix

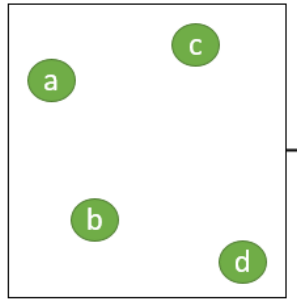
- Symmetric matrix: If $A^T = A$ $(a_{ij} = a_{ji})$ in square matrix $A = (a_{ij})_{n\times n}$, A is a symmetric matrix.

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

For example, the distance matrix and covariance matrix are both symmetric matrices.

$$\Sigma = \begin{bmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{bmatrix}$$

Covariance matrix



	a	b	c	d
a	0	1.2	1.1	2.8
b	1.2	0	2.2	1.3
c	1.1	2.2	0	2.5
d	2.8	1.3	2.5	0

Distance matrix

Contents

- **Mathematics and AI**
  - **Linear Algebra**
    - Basics Concepts and operations
    - Linear transformation and matrix
    - Matrix Decompositions
  - **Statistics and Probability**
  - **Optimization Problems**

Linear Independence

- Every vector can be written as linear combination of some finitely “special” vectors.
- These called basis-vectors.

$$S = \begin{bmatrix} 4 \\ 3 \\ 2 \end{bmatrix} = 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

- Linear independence: a set of vectors is linearly independent if any element of the set can not be expressed as a linear combination of the others.
- The columns are not linearly independent:

$$S = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 5 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Rank of Matrix

- The rank of a matrix is defined as the maximum number of linearly independent column vectors in the matrix or the maximum number of linearly independent row vectors in the matrix.
- A rank 2 matrix:

$$S = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}$$

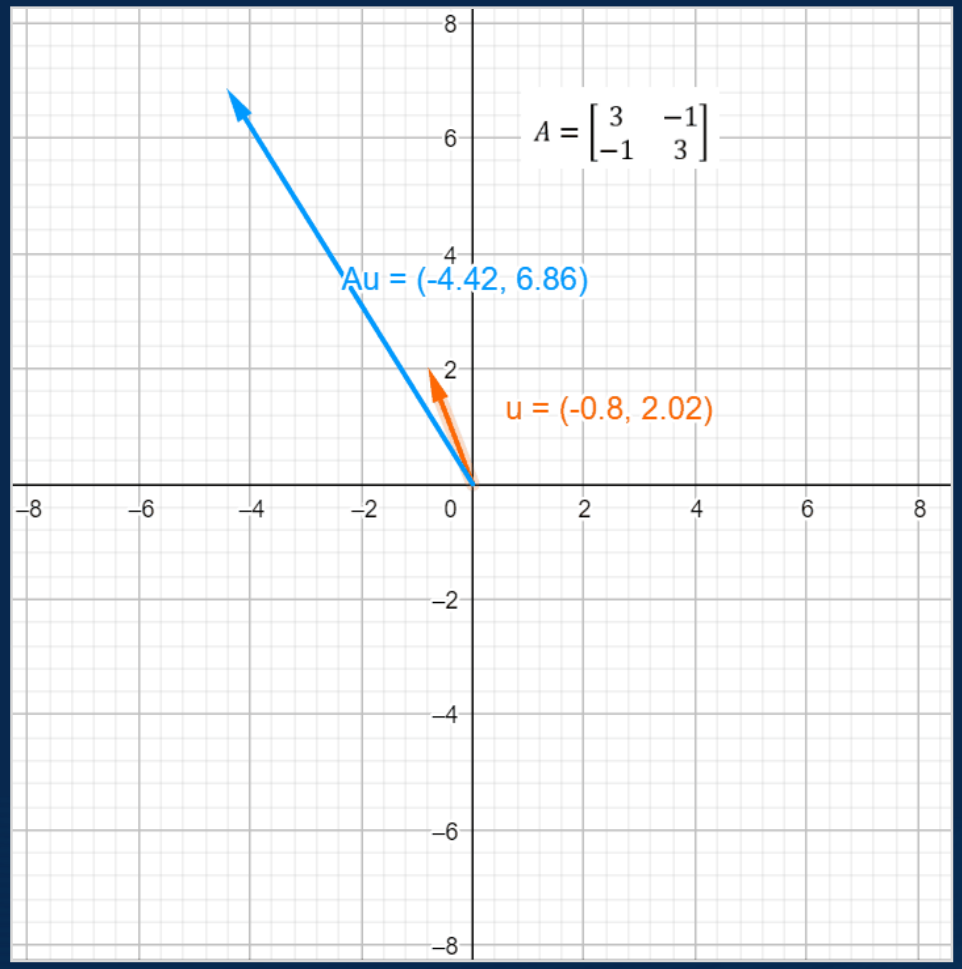
Eigenvalue and Eigenvector

- Definition: For any square matrix A n-by-n over the field K. If a non-zero vector α over the field K^n satisfies

- $A\alpha = \lambda\alpha, \lambda \epsilon K,$

  Where λ is called an eigenvalue of A. α is called an eigenvector of matrix A with eigenvalue λ.

- One matrix represents one linear transformation, such as rotation, scaling, and reflection. Vectors are transformed through matrix multiplication. As shown in the figure, if one or some vectors only scale without the generating the rotation or projection result, these vectors are the eigenvectors of this matrix, and the scaling factor is an eigenvalue.



Eigenvalues and Eigenvectors

- How to seek the eigenvalue and eigenvector of matrix A:

$$A\alpha = \lambda\alpha$$

$$\Leftrightarrow A\alpha - \lambda\alpha = 0$$

$$\Leftrightarrow (A - \lambda I)\alpha = 0$$

$$\alpha \neq 0 \Leftrightarrow |A - \lambda I| = 0$$

$$\Leftrightarrow \begin{bmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn}-\lambda \end{bmatrix} = 0,$$

  where $|A - \lambda I| = 0$ is the characteristic equation of matrix A. The root of this characteristic equation is the eigenvalue λ. It is substituted into the equation $A\alpha = \lambda\alpha$ to obtain the eigenvector α.

- The process of seeking eigenvalues and eigenvectors involves obtaining determinants. Therefore, it is clear that only square matrices have eigenvalues and eigenvectors.

Example:

Example: Find the eigenvalues and eigenvectors of the matrix $A = \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix}$.

Solution: $|A| = \begin{vmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = (4-\lambda)(2-\lambda)$. Then $\lambda_1 = 2$ and $\lambda_2 = 4$.

Taking $\lambda_1 = 2$, the corresponding eigenvector satisfies $\begin{bmatrix} 3-2 & -1 \\ -1 & 3-2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, we find that $x_1 = x_2$.

Therefore, the corresponding eigenvector is $p_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. When $\lambda_1 = 2$, the eigenvector is $kp_1 (k \neq 0)$.

Taking $\lambda_2 = 4$, we find that $x_1 = -x_2$. The eigenvector is $p_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. When $\lambda_2 = 4$, the eigenvector is $kp_2 (k \neq 0)$.

Eigen Decomposition

- If A is symmetric, all its eigenvalues are real and all its eigenvectors are orthonormal, $u_i^T u_j = \delta_{ij}$
- Hence $U^T U = UU^T = I, |U| = 1$.
- Eigen decomposition decomposes a matrix A into a multiplication of a matrix of eigenvectors U and a diagonal matrix of eigenvalues Λ.

$$A = U\Lambda U^T = \sum_{i-1}^{n} \lambda_i u_i u_i^T$$

Principle component

- Given a data matrix $A \in \mathbb{R}^{n \times d}$, the principle components of A are the eigenvectors of UU^T
- Principle components analysis(PCA) for X is to find the eigenvectors and eigenvalues of the matrix UU^T

Singular Value Decomposition

- Singular value decomposition (SVD) is a way to factorize a matrix into singular vectors and singular values. The SVD allows us to write the matrix $A = (a_{ij})_{m\times n}$ as a product of three matrices:

$$A = U\Sigma V^T,$$

where $U = (b_{ij})_{m\times m}$, $\Sigma = (c_{ij})_{m\times n}$, $V^T = (d_{ij})_{n\times n}$. The matrices U and V are both defined to be orthogonal matrices. The columns of U are known as the left-singular vectors. The columns of V are known as the right-singular vectors. The matrix Σ is defined to be a diagonal matrix. Note that Σ is not necessarily a square matrix. The elements along the diagonal of Σ are known as the singular values of matrix A. The singular values are arranged in descending order.

Singular Value Decomposition

For given m x n matrix A

$$A = U\Sigma V^T$$

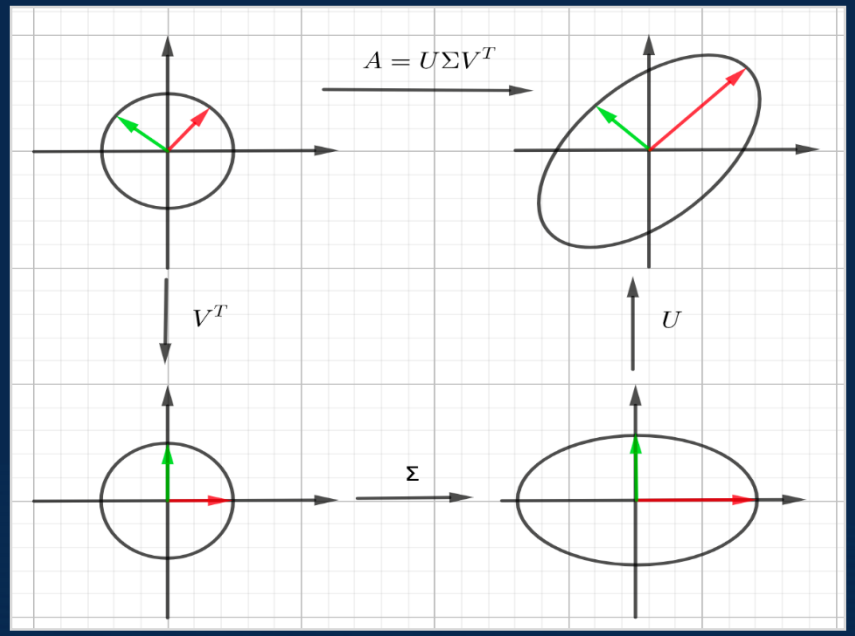
$$A = [\vec{u}_1 \quad \vec{u}_2 \quad \cdots \quad \vec{u}_m] \begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma n \end{pmatrix} \begin{bmatrix} \vec{v}_1^T \\ \vec{v}_2^T \\ \vdots \\ \vec{v}_n^T \end{bmatrix}$$

$$A = u_1\sigma_1\vec{v}_1^T + u_2\sigma_2\vec{v}_2^T + u_1\sigma_1\vec{v}_1^T + \cdots + u_n\sigma_n\vec{v}_n^T$$

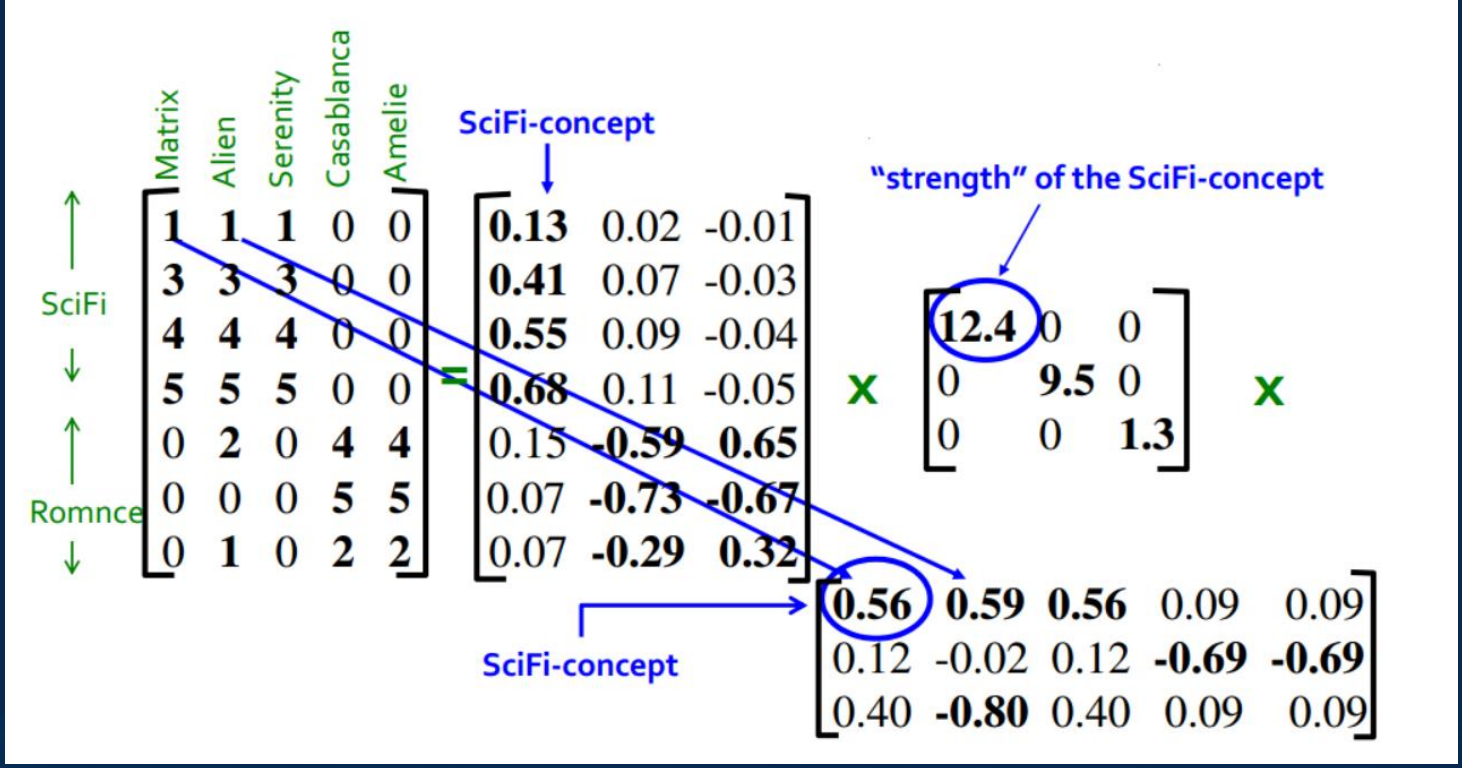
- $U \in m \times n$ maps m points to a n-dimensional *concept space*, which can be seen as a condensed version of the original n-dimensional feature space, as if the n concepts grouped together "similar" features.
- $\Sigma \in n \times n$ diagonal elements can be interpreted as the strength of the concepts in the *concept space*. i.e. how much can the data be summarized by the first concept? and so forth.
- $V^T \in n \times n$ maps concepts back to features. i.e. which features adhere to each concept the most?

Geometric Interpretation of the SVD

- The SVD can find an orthogonal basis in a space and map this orthogonal basis to the image space through matrix multiplication. The singular value is the corresponding stretch factor.
- SVD factorizes the effects of rotation, scaling, and projection that are together described by the matrix.



Example 1—Users to movies



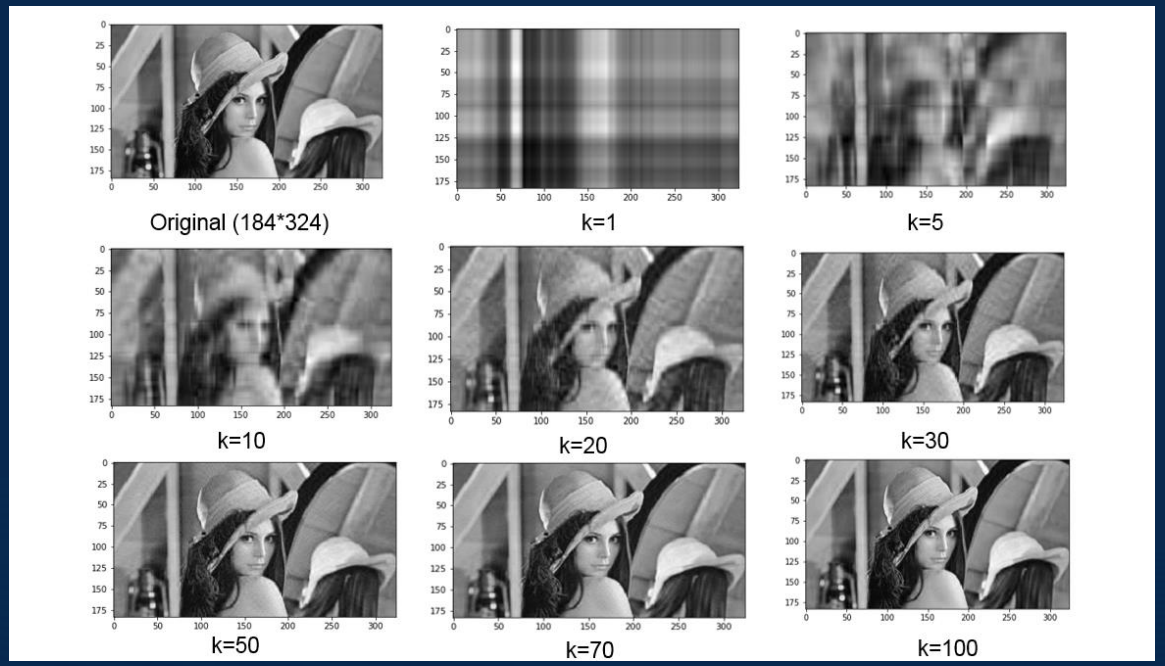
$A = U\Sigma V^T$

'movies', 'users' and 'concepts'

- U: user-to-concept similarity matrix
- V:movie-to-concept similarity matrix
- Σ :its diagonal element 'strength' of each concepts

Example 2—Image Compression

- The input image is a 184 x 324 grayscale image with a data volume of 184 x 324 = 59616. When compressing this image by SVD, if we keep the first 70 components of this image to represent it, its data volume will be 184 x 70 (left-singular vector) + 70 (singular value) + 70 x 324 (right-singular vector) = 35660.



Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

HUAWEI