Basic Linear Algebra for Machine Learning

Maryam Ramezani
maryam.ramezani@sharif.edu

Machine Learning Mathematics

01

Linear Algebra

Provides the foundation for manipulating data in high-dimensional spaces, essential for vector operations in machine learning.

02

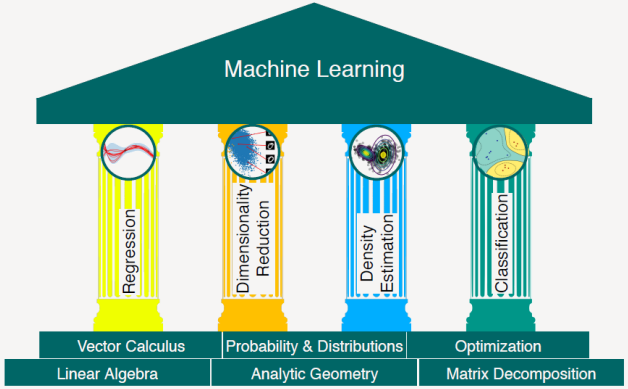
Statistics & Probability

Offers tools to model uncertainty and make inferences about data, forming the backbone of many machine learning algorithms.

03

Optimization

Focuses on finding the best parameters for a model by minimizing or maximizing an objective function.



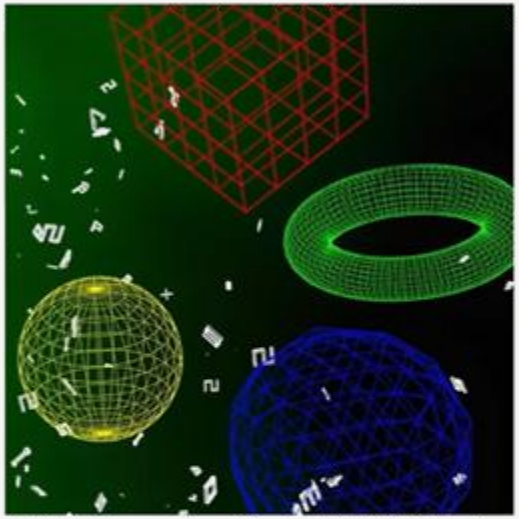
Introduction

Lets think about a question!

Famous Topics of Linear Algebra in ML

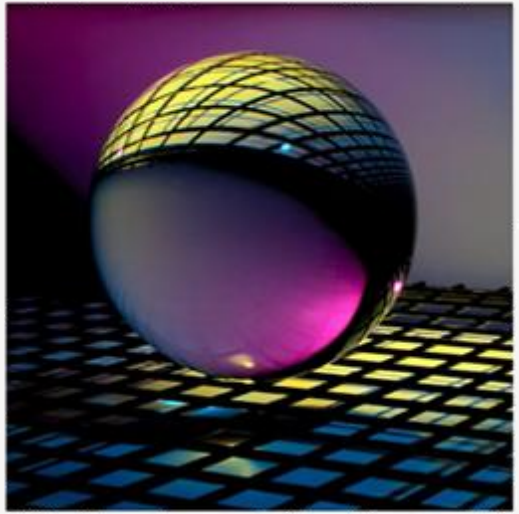
Data Representations

Use basic ideas of linear algebra to represent data in a way that computers can understand: vectors.



Vector Embeddings

Learn ways to choose these representations wisely via matrix factorizations.



Dimensionality Reduction

Deal with large-dimensional data using linear maps and their eigenvectors and eigenvalues.

01

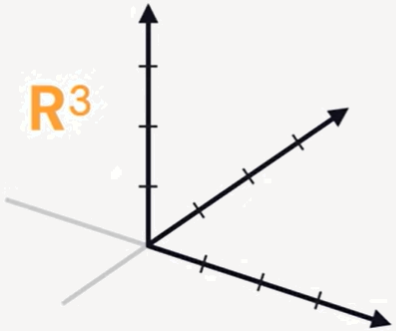
Vector

Basic Concept

Data Representations (Linear Algebra)

- How can we represent data (images, text, user preferences, etc.) in a way that computers can understand?
- Organize information into a vector!
  - A vector is a 1-dimensional array of numbers.
  - It has both a magnitude (length) and a direction
- The totality of a vectors with n entries is an n-dimensional vector space

$$V = \nearrow = \begin{bmatrix} -3 \\ 0.7 \\ 2 \end{bmatrix}$$



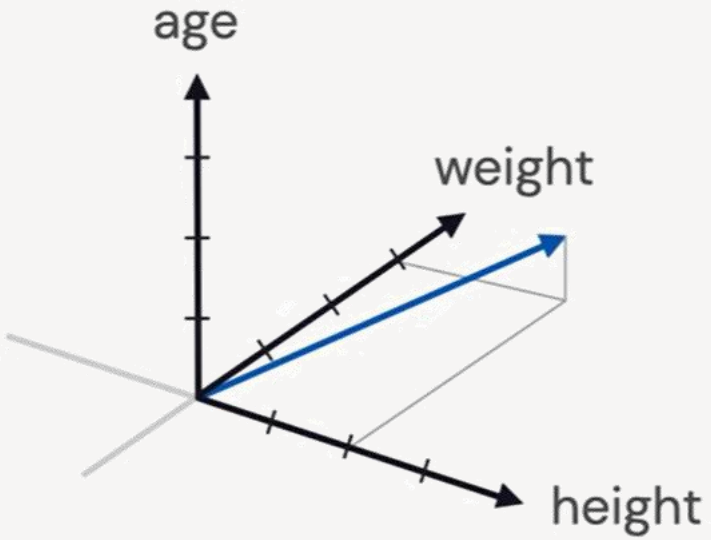
"3-dimensional space" consists of all vectors with 3 entries:

$$\begin{bmatrix} * \\ * \\ * \end{bmatrix}$$

Data Representations (Machine Learning)

- A **feature vector** is a vector whose entries represent the “features” of an object.
- The vector space containing them is called **feature space**.

$$P = \begin{bmatrix} 64 \\ 131 \\ 24 \end{bmatrix} \begin{matrix} \text{height} \\ \text{weight} \\ \text{age} \end{matrix}$$



02

Data Representation Applications

Image

- In black and white images, black and white pixels correspond to 0s and 1s.

- In grayscale pixels are numbers between 0 and 255.

Words and Documents

- Given a collection of documents (e.g. Wikipedia articles), assign to every word a vector whose i^{th} entry is the number of times the word appears in the i^{th} document.
- These vectors can be assemble into a large matrix, useful for latent semantic analytics.

$$\text{dog} = \begin{bmatrix} 0 \\ 7 \\ 0 \\ 0 \\ 51 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \text{Wiki \#1} \\ \text{Wiki \#2} \\ \text{Wiki \#3} \\ \text{Wiki \#4} \\ \text{Wiki \#5} \\ \\ \text{Wiki \#54,000,000} \end{matrix}$$

Latent Semantic Analysis

- In the sub-field of machine learning for working with text data called natural language processing (NLP), it is common to represent documents as large matrices of word occurrences.

	Quick	Brown	Fox	Jumps	Over	Lazy	Dog
The quick brown fox jumps over the lazy dog	1	1	1	1	1	1	1
If the fox is quick he can jump over the dog.	1	0	1	0	1	0	1
Foxes are quick. Dogs are lazy.	0	1	1	0	0	1	1
Can a fox jump over a dog?	0	0	1	1	1	0	1

- Matrix factorization methods, such as the singular-value decomposition can be applied to this sparse matrix. Documents processed in this way are much easier to compare, query, and use as the basis for a supervised machine learning model.

Yes/No or Ratings

- Given users and items (e.g. movies), vectors can indicate if a user has interacted with the item (yes=1, no=0).
- User's rating a number between 0 and 5.

$$\text{user1} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \begin{matrix} \text{No} \\ \text{Yes} \\ \text{No} \\ \text{No} \\ \\ \text{Yes} \\ \text{No} \end{matrix} \qquad \text{user2} = \begin{bmatrix} 0 \\ 5 \\ 0 \\ 3 \\ \vdots \\ 0 \\ 2 \end{bmatrix} \begin{matrix} ? \\ \text{Love} \\ ? \\ \text{Like} \\ \\ ? \\ \text{Dislike} \end{matrix}$$

Categorical (Non-numerical) Data

- Sometimes you work with categorical data in machine learning.
- It is common to encode categorical variables to make them easier to work with and learn by some techniques. A popular encoding for categorical variables is the one hot encoding.
- A one hot encoding is:

Example

$$\text{red}=\begin{bmatrix}1\\0\\0\end{bmatrix} \qquad \text{green}=\begin{bmatrix}0\\1\\0\end{bmatrix} \qquad \text{blue}=\begin{bmatrix}0\\0\\1\end{bmatrix}$$

Categorical (Non-numerical) Data

- One-Hot Encodings (standard basis vector)
  - Assign to each word a vector with one 1 and 0s elsewhere.
  - Suppose our language only has four words:

$$\text{apple} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \quad \text{cat} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \quad \text{house} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix} \quad \text{tiger} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

Drawbacks

- Very sparse vectors.
- Are never **similar**!

How to measure the similarity?

- **Dot Product**
  - The product of numbers is another number.
  - The dot product of vectors is not another vector! It is a number!!

$2 \times 5 = 10$

Numbers

vs

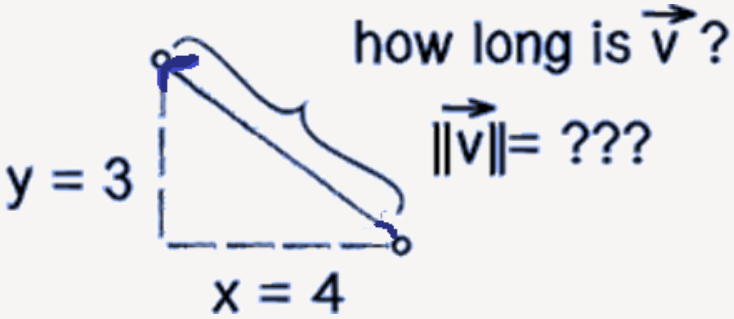
$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 5 \end{bmatrix} = 7$

Vectors | A number

$$\begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix} \cdot \begin{bmatrix} 7 \\ 2 \\ -1 \end{bmatrix} = (1)(7) + (0)(2) + (3)(-1) = 4$$

Length of vector

- Dot product between a vector and itself: magnitude-squared, the **length** squared, or the squared-norm, of the vector.



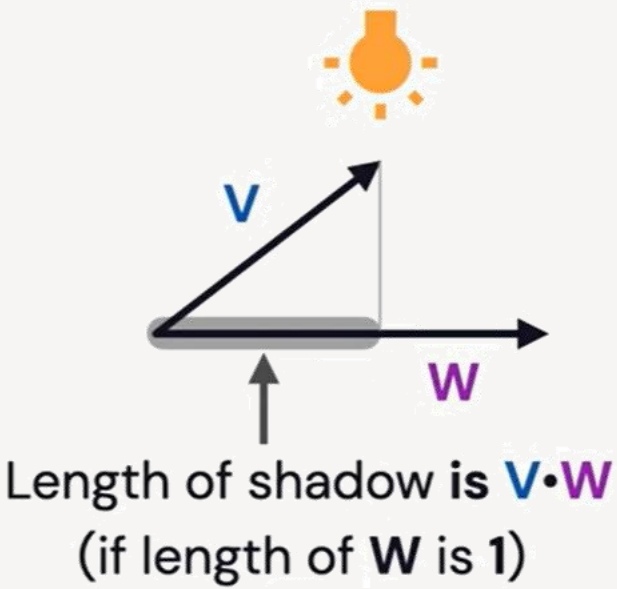
$$v.v = \begin{bmatrix} 4 \\ 3 \end{bmatrix} . \begin{bmatrix} 4 \\ 3 \end{bmatrix} = 16+9=25$$

$$Length(v)=5$$

$$a^T a = \|a\|^2 = \sum_{i=1}^{n} a_i a_i = \sum_{i=1}^{n} a_i^2$$

Dot Product (Geometric Interpretation and Intuition)

- Represents the length of the “shadow” of one vector along another.
- This indicates how similar the two vectors are.



One-Hot Encodings Drawbacks

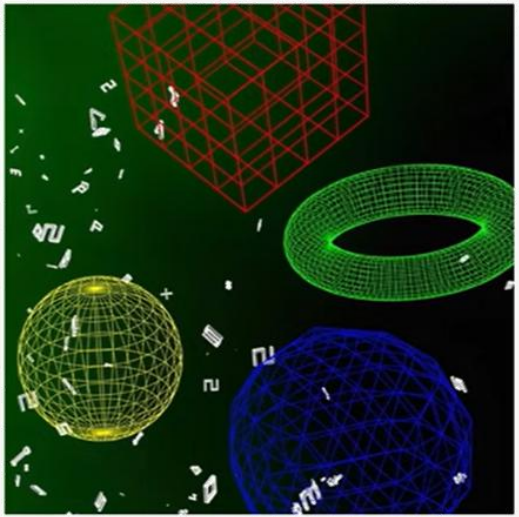
$$\text{apple} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \qquad \text{cat} = \begin{bmatrix}0\\1\\0\\0\end{bmatrix} \qquad \text{house} = \begin{bmatrix}0\\0\\1\\0\end{bmatrix} \qquad \text{tiger} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix}$$

$$\text{apple} \cdot \text{cat} = \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\\0\end{bmatrix} = \mathbf{0} = \begin{bmatrix}0\\0\\0\\1\end{bmatrix} \cdot \begin{bmatrix}0\\1\\0\\0\end{bmatrix} = \text{tiger} \cdot \text{cat}$$

Famous Topics of Linear Algebra in ML

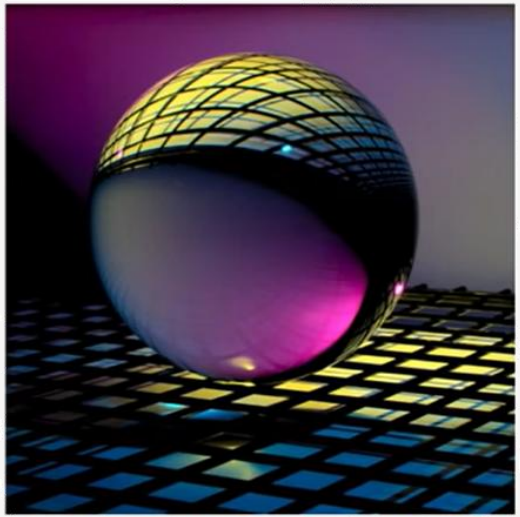
Data Representations

Use basic ideas of linear algebra to represent data in a way that computers can understand: vectors.



Vector Embeddings

Learn ways to choose these representations wisely via matrix factorizations.



Dimensionality Reduction

Deal with large-dimensional data using linear maps and their eigenvectors and eigenvalues.

Vector Embeddings (Linear Algebra)

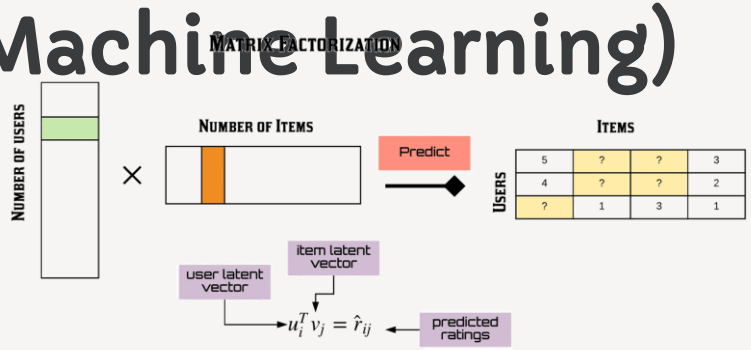
- An embedding of a vector is another vector in a smaller dimensional space.

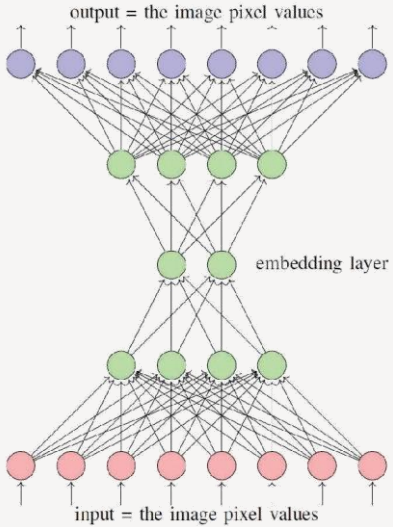
Replace  with 

Vector Embeddings (Machine Learning)

Matrix Factorization

Neural Networks





What is Matrix?

- A **matrix** is a 2-dimensional array of numbers.
- **Matrix is a linear transformation**
  - It represents a particular process of turning one vector into another: stretching, rotating, scaling or something more complex.

$$\begin{bmatrix} 3 & 0 & 7 \\ 1 & -5 & 9 \\ 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

input output

What is Matrix?

- Image Rotation

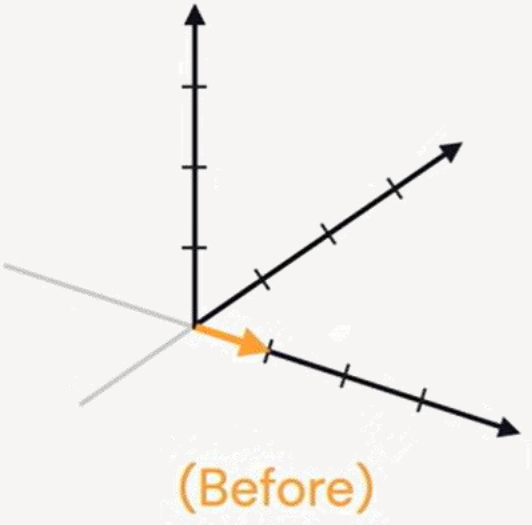
(a) (b)

- Image Scaling

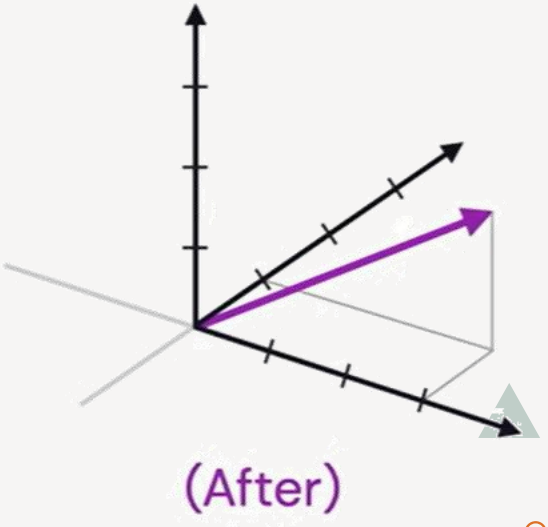
(a) (b)

What is Matrix?

- A **matrix** represents a **transformation** of an entire vector space to another (possibly of different dimensions)



$$\begin{bmatrix} 3 & 0 & 7 \\ 1 & -5 & 9 \\ 2 & 0 & 0 \end{bmatrix}$$



Matrix Factorization

- We can multiply **numbers** and get **number**.
- We can multiply **vectors** by dot product and get **number**.
- We can multiply **matrices** and get a **matrix.**

$$\begin{bmatrix} 3 & 0 & 7 \\ 1 & -5 & 9 \\ 2 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 2 \\ 1 & 0 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -7 & 6 \\ -14 & 2 \\ 0 & 4 \end{bmatrix}$$

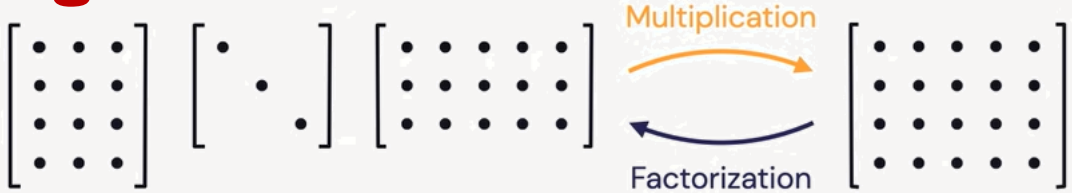
3×3 $\quad$ 3×2 $\quad$ 3×2

- Factorization?

Sizes must match

$$12 \times 3 \times 15 \quad \underset{\text{Factorization}}{\overset{\text{Multiplication}}{\rightleftarrows}} \quad 540$$

In general factorization is HARD!



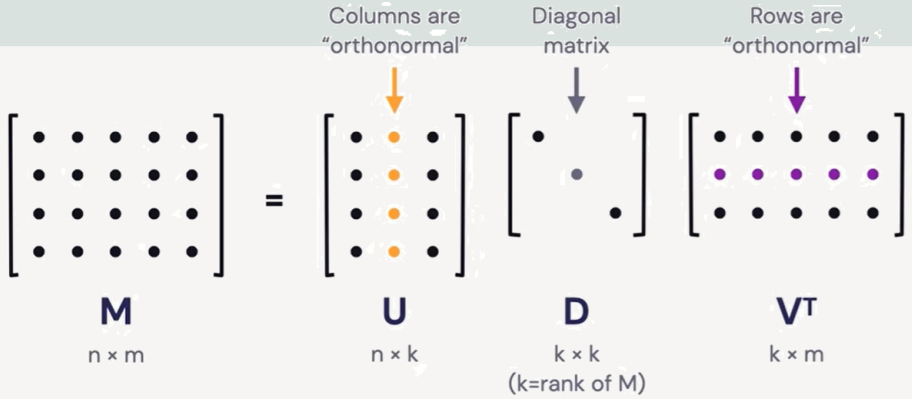
Matrix Factorization (Linear Algebra)

- Fundamental Theorem in Linear Algebra:
  - **Every matrices can be factored!**

Theorem

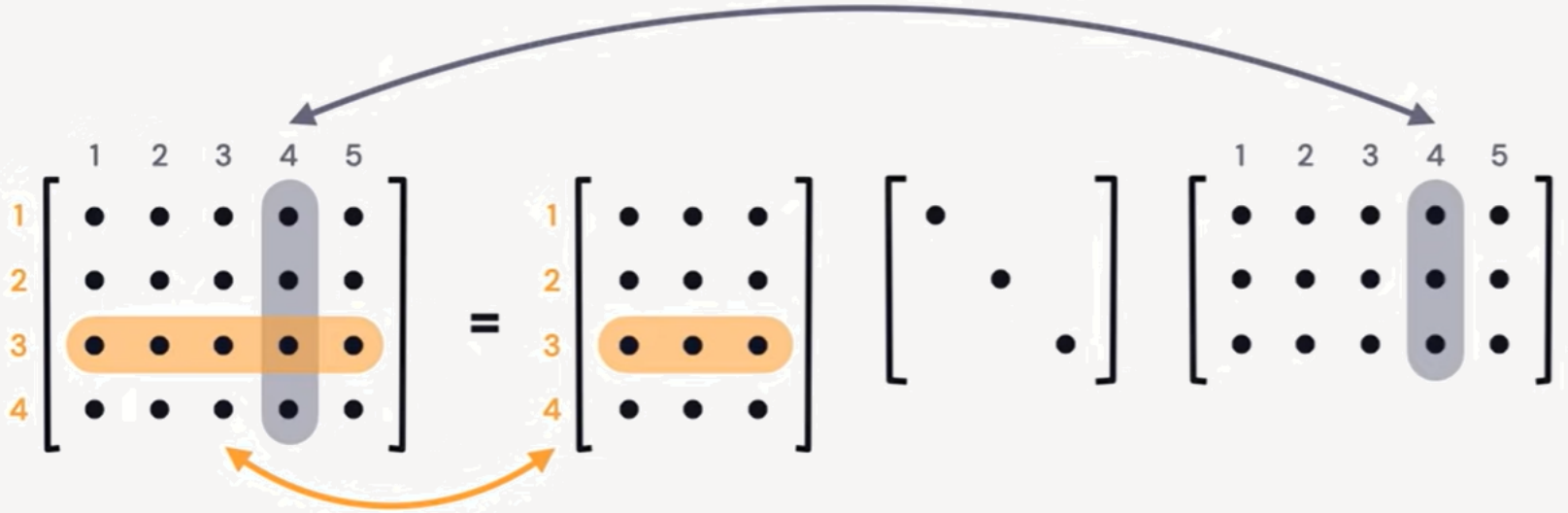
Singular Value Decomposition (SVD)

Every $n \times m$ matrix can be written as a product of three smaller matrices as below:



Matrix Factorization (Linear Algebra)

- It has wide use in linear algebra and can be used directly in applications such as feature selection, visualization, noise reduction, and more.
- The columns/rows of the factors are candidates for embeddings.



03

Vector Embedding Applications

Recommender Systems

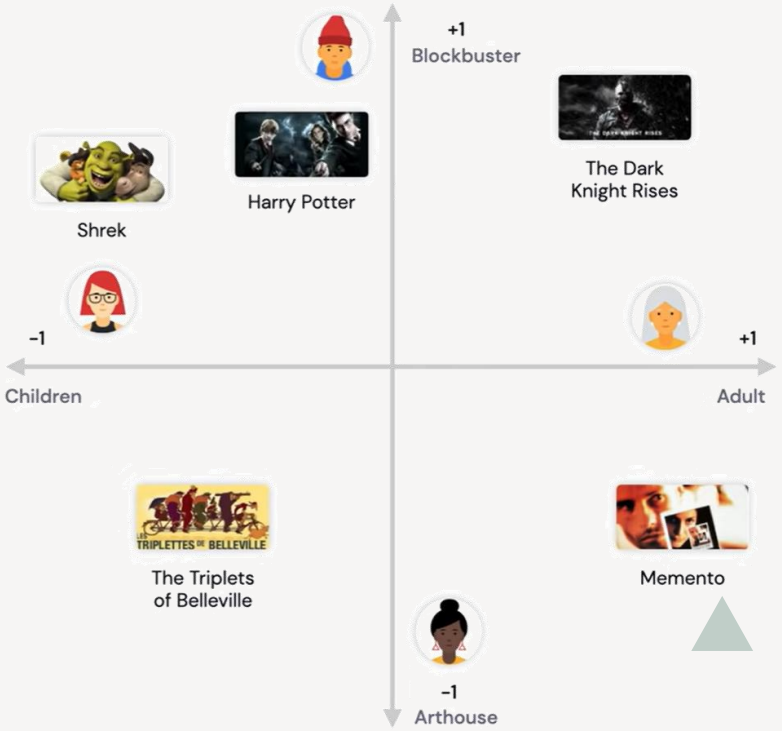
- User – Movie Matrix
  - Checkmarks = watched movie
  - Empty cells = not watched movie

$$\text{User 3} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \text{Shrek} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Recommender Systems

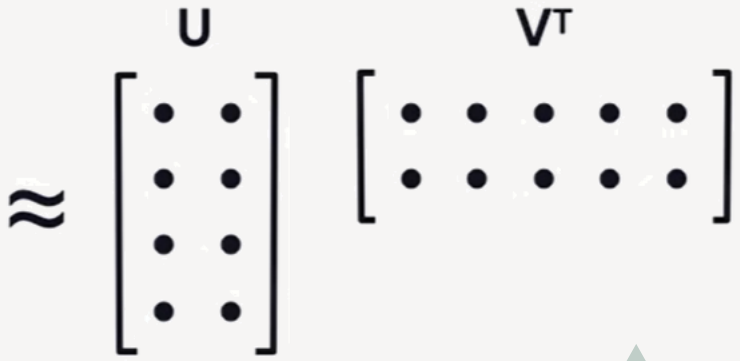
- We dont know the features!
  - Example: 2-dimensional “latent” feature space!

- We want to find the new, smaller dimensional vector representations that capture these features.



Recommender Systems

4 × 5



4 × 2 2 × 5

Recommender Systems

Harry Potter	The Triplets of Belleville	Shrek	The Dark Knight Rises	Memento
✓		✓	✓	
	✓			✓
✓	✓	✓		
			✓	✓

$\approx$

1	.1
-1	0
.2	-1
.1	1

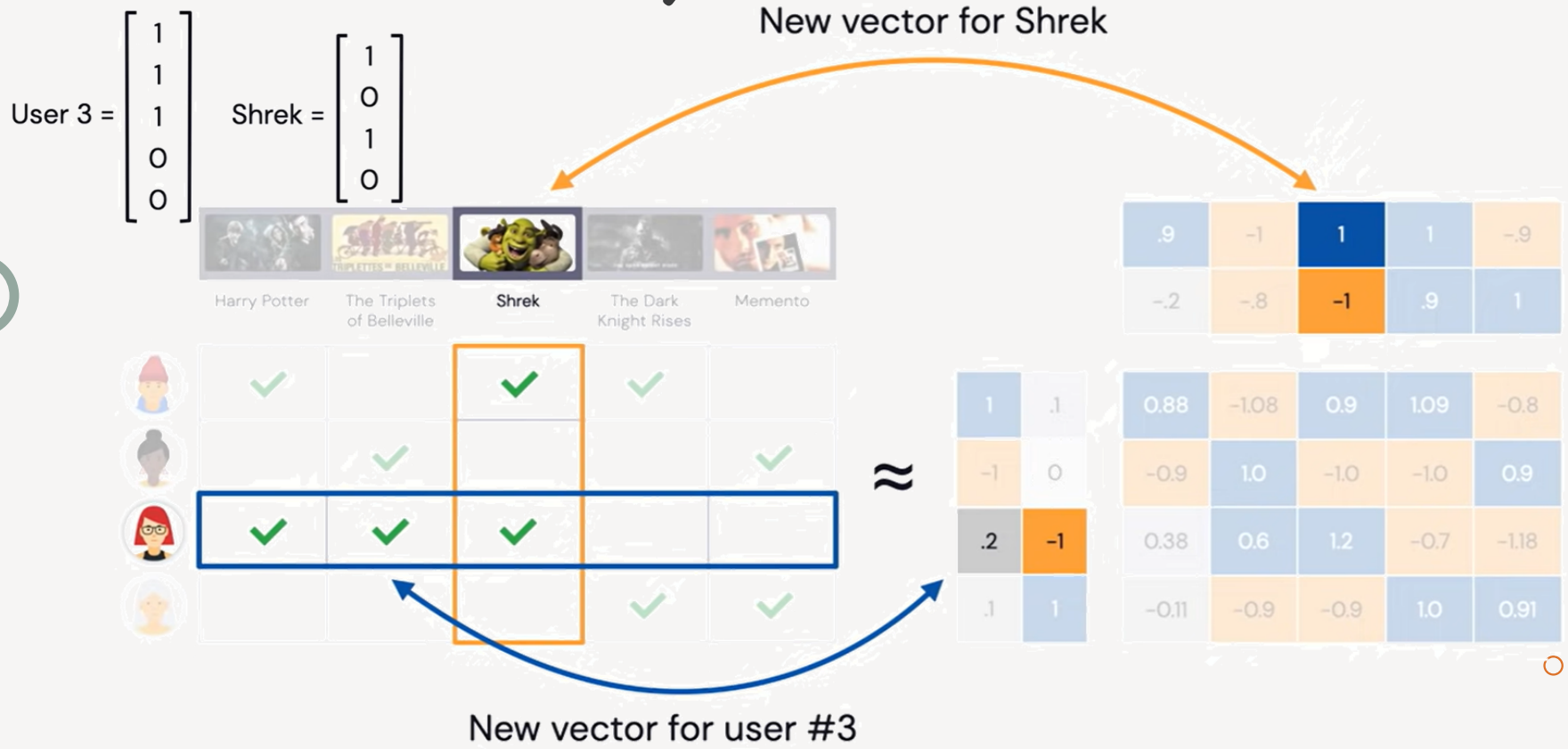
U is a **user** × feature matrix

V^T is a feature × **movie** matrix

.9	-1	1	1	-.9
-.2	-.8	-1	.9	1

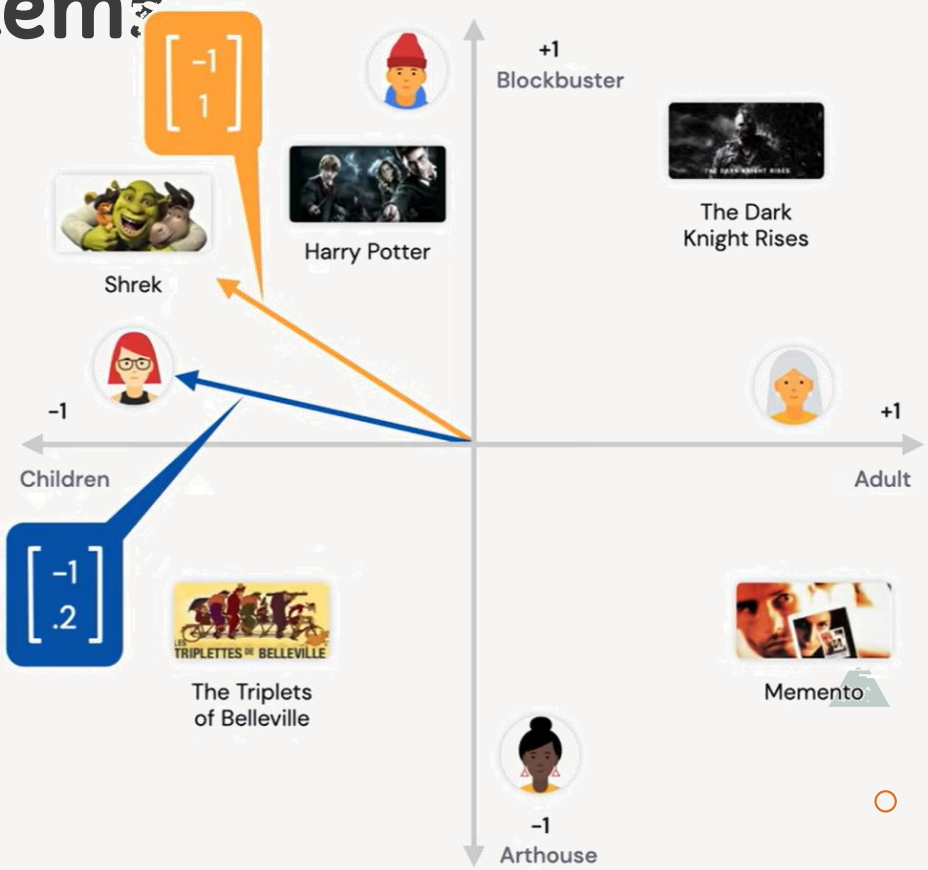
0.88	-1.08	0.9	1.09	-0.8
-0.9	1.0	-1.0	-1.0	0.9
0.38	0.6	1.2	-0.7	-1.18
-0.11	-0.9	-0.9	1.0	0.91

Recommender Systems



Recommender Systems

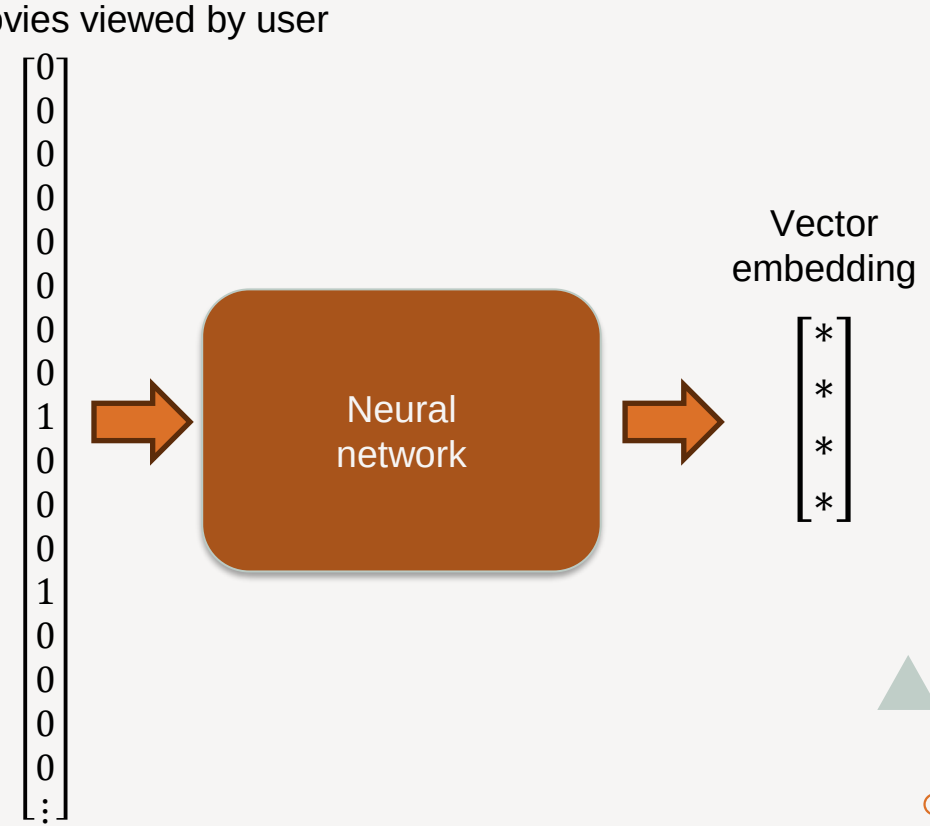
- These two vectors are close!
- The shadow of orange vector onto blue vector is pretty large!



Neural Network

Feed data vector into a Neural network. The output is vector embedding.

Under the hood:
Matrix multiplication *plus* more.

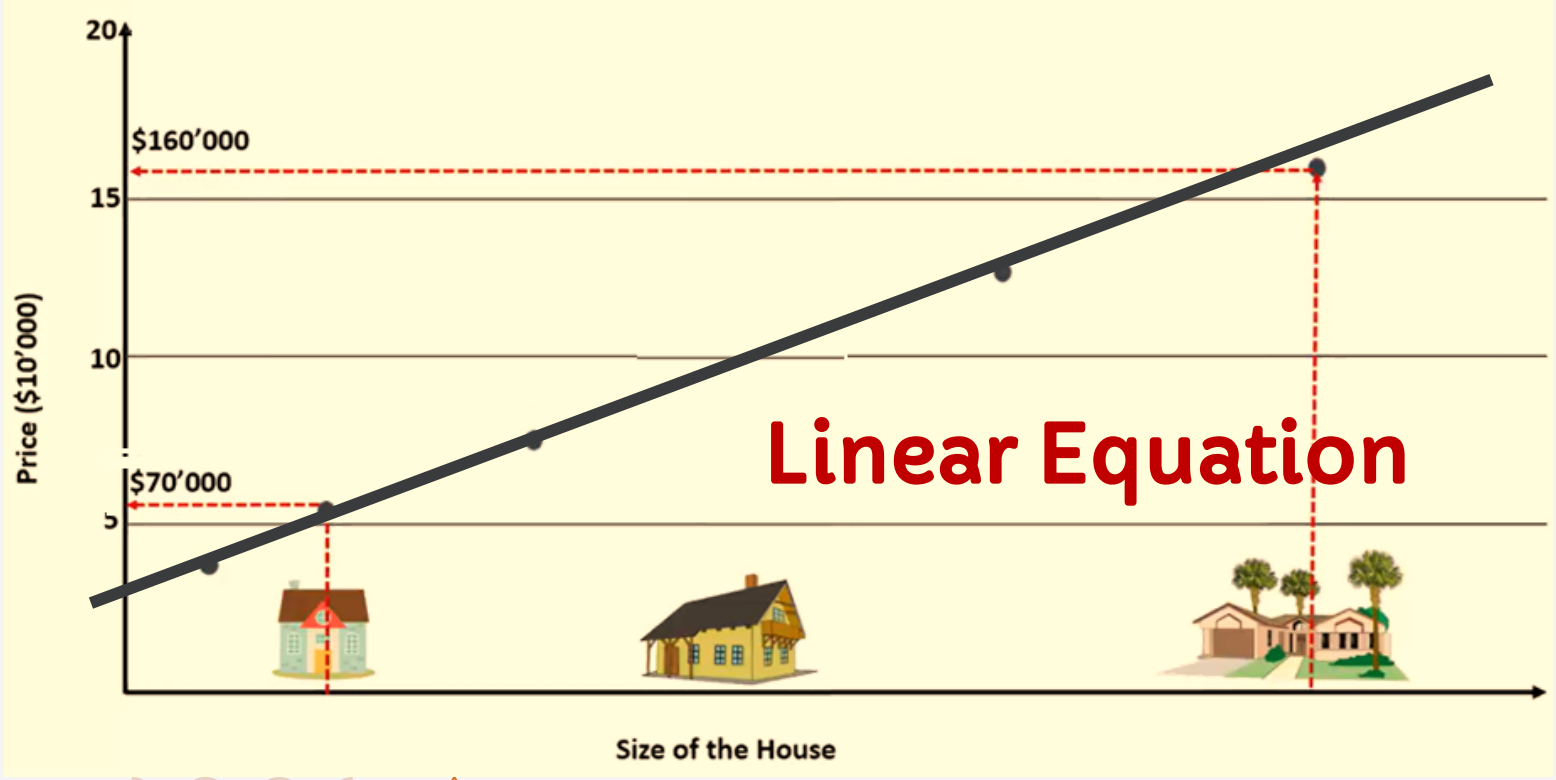


04 Equation with Matrix and Vector Format

House Pricing Example

Price Problem

Price Problem



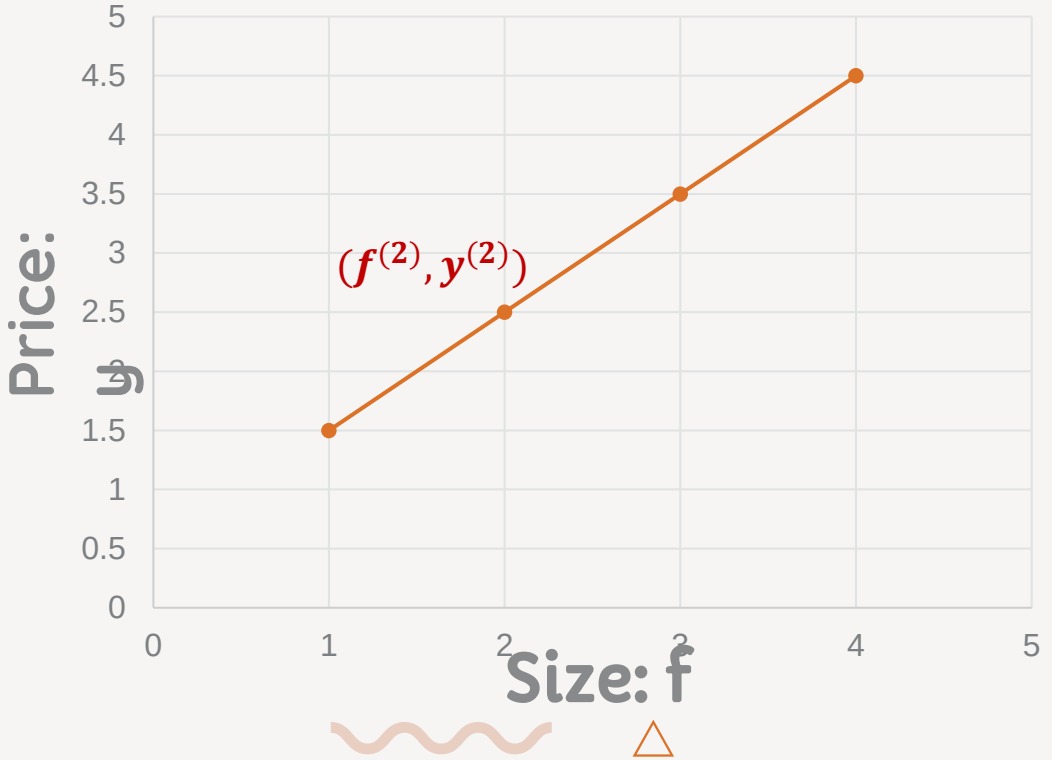
Linear Equation without offset

$$y = af$$

How to convert it to matrix-vecto multiplication?

$$Ax = b$$

Linear Equation with offset



$$y = af + c$$

How to convert it to matrix-vecto multiplication?

$$Ax = b$$

House Features

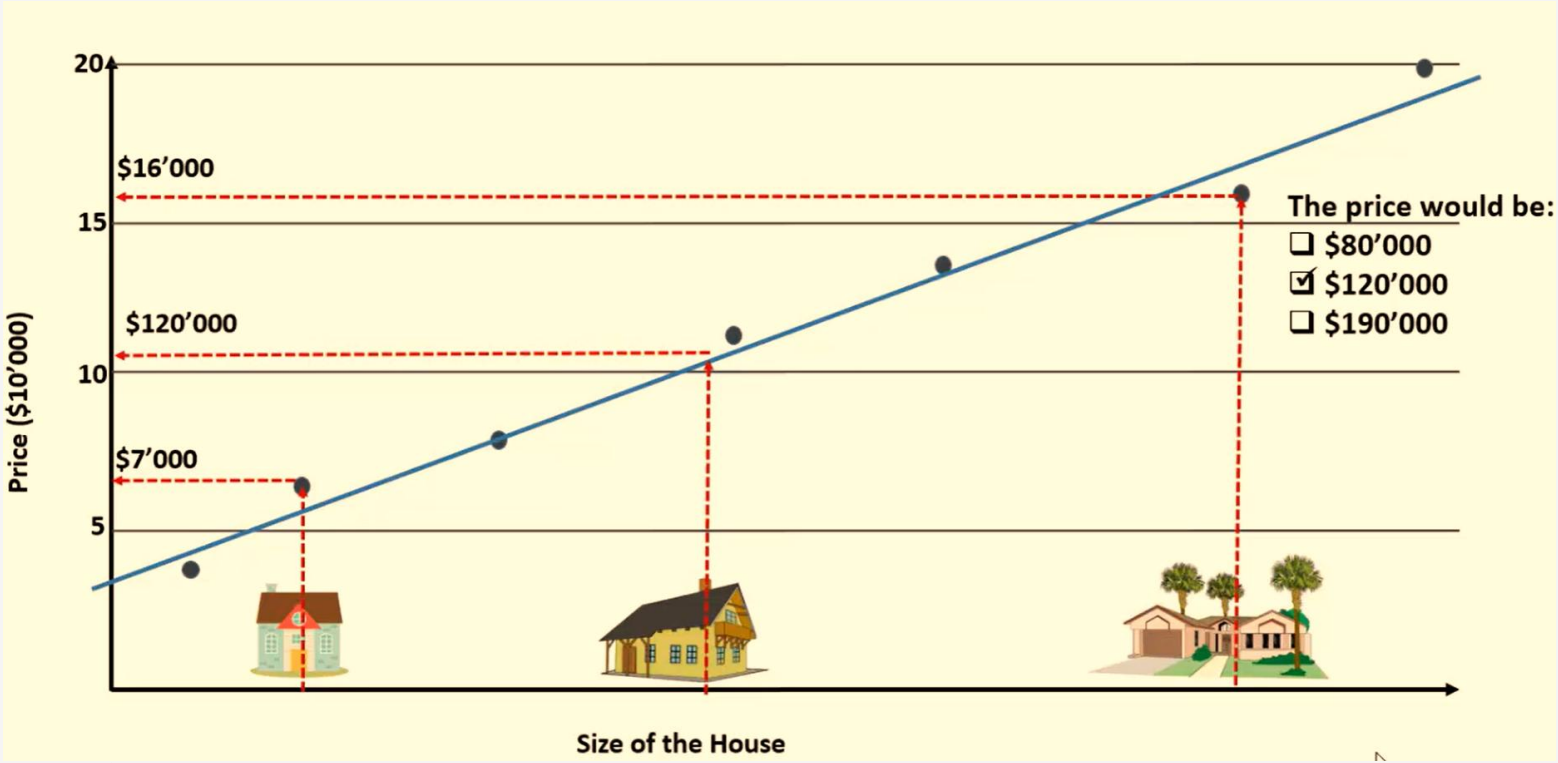
- #Room
- Size
- #Bedroom
- Age
- Address features: Street, Alley, ...
- Size of part1, part2, part3, part4
- Floors
- #Bathrooms
- ...

N number of training data with M featues:

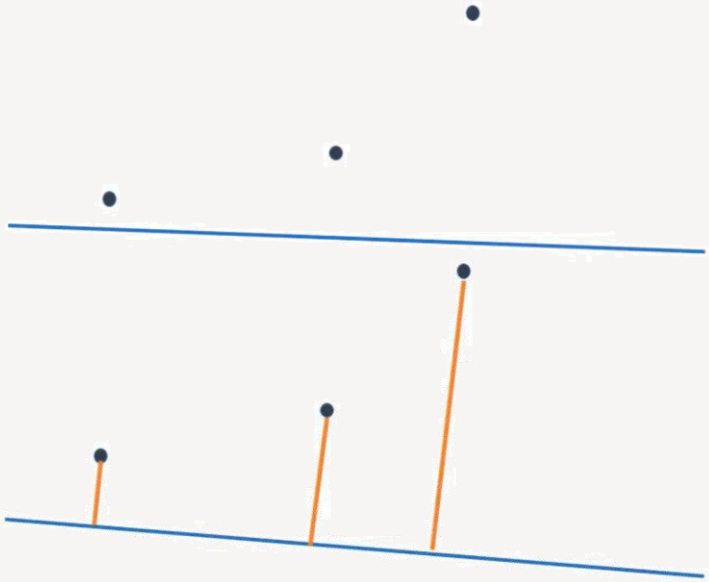
$$A_{N\times M}x_{M\times 1} = b_{N\times 1}$$

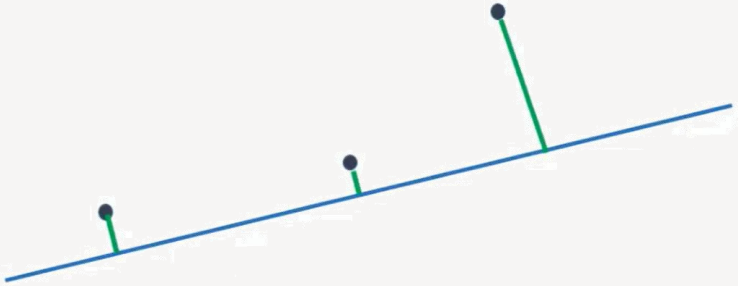
Linear Equation

Least Squares Error Correction



Least Squares Error Correction



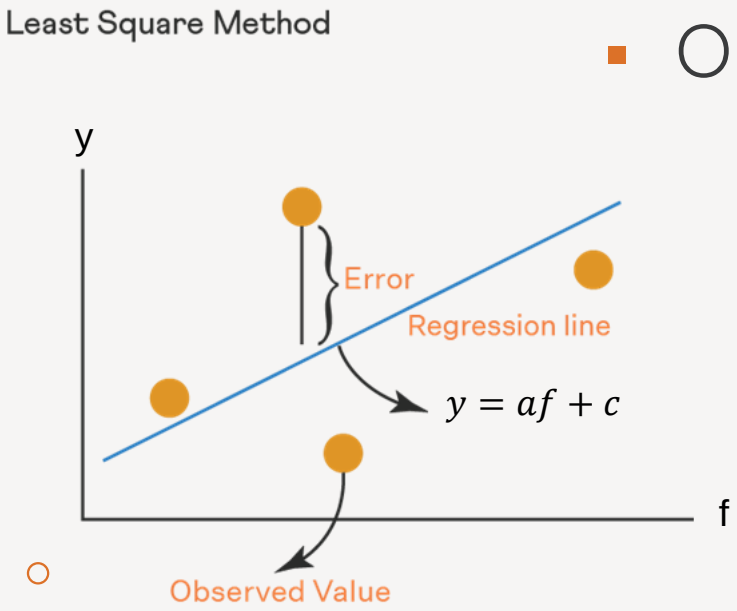


Error 1:

Error 2:

Error 3:

Least Squares



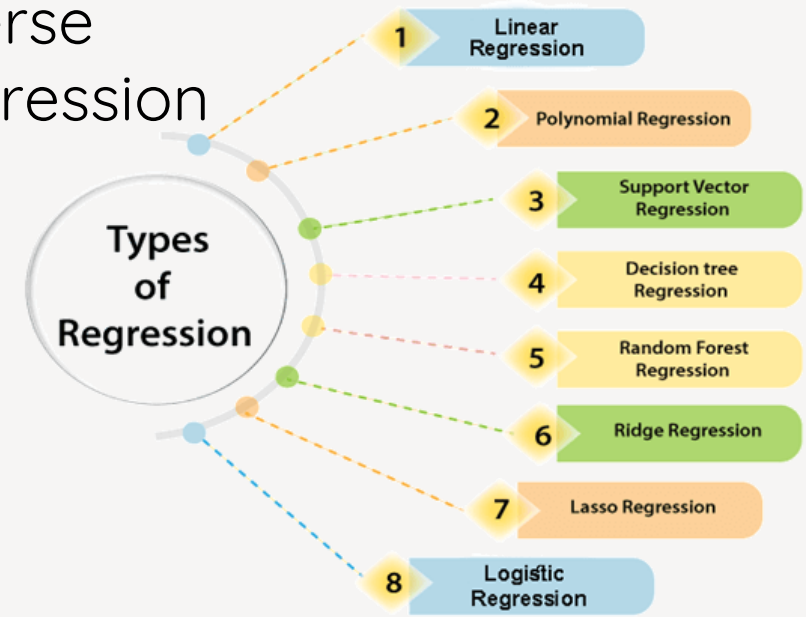
- Objective for “m” features:

$$\hat{y}=a_1 f_1 + a_2 f_2 + \cdots + a_m f_m + c$$

$$\min ||y - \hat{y}||$$

Linear Algebra and Machine Learning Application

- $Ax = b \quad \rightarrow \quad x = A^{-1}b$ Inverse of Matrix/Pseudo Inverse
- Regression



05

Vector Operation

Vector Operations

- Vector-Vector Addition
- Vector-Vector Subtraction
- Scalar-Vector Product
- Vector-Vector Products:
  - $\boldsymbol{x.y}$ is called the inner product or dot product or scalar product of the vectors: $x^T y$ or $y^T x$
    - $< a, b >$ $< a|b >$ (a, b) $a.b$

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

  - Transpose of dot product:
    - $(a.b)^T = (a^T b)^T = (b^T a) = (b.a) = b^T a$
  - Length of vector

Dot Product Properties

- Commutativity
  - The order of the two vector arguments in the inner product does not matter.

$$a^T b = b^T a$$

- Distributivity with vector addition
  - The inner product can be distributed across vector addition.

$$(a + b)^T c = a^T c + b^T c$$
$$a^T (b + c) = a^T b + a^T c$$

Dot Product Properties

- Bilinear (linear in both a and b)

$$a^T(\lambda b + \beta c) = \lambda a^T b + \beta a^T c$$

- Positive Definite:

$$(a.a) = a^T a \geq 0$$

  - 0 only if a itself is a zero vector ($a = \mathbf{0}$)

Dot Product Properties

- Associative
  - Note: the associative law is that parentheses can be moved around, e.g., (x+y)+z = x+(y+z) and x(yz) = (xy)z

1) Associative property of the vector dot product with a scalar (scalar-vector multiplication embedded inside the dot product)

scalar

$$\gamma(\mathbf{u}^{\mathrm{T}}\mathbf{v}) = (\gamma\mathbf{u}^{\mathrm{T}})\mathbf{v} = \mathbf{u}^{\mathrm{T}}(\gamma\mathbf{v}) = (\mathbf{u}^{\mathrm{T}}\mathbf{v})\gamma$$

$$= (\gamma\boldsymbol{u})^{T}\boldsymbol{v} = \gamma\boldsymbol{u}^{\boldsymbol{T}}\boldsymbol{v}$$

Dot Product Properties

- Associative

2) Does vector dot product obey the associative property?

$$\mathbf{u}^{\mathrm{T}}(\mathbf{v}^{\mathrm{T}}\mathbf{w}) \;?\; (\mathbf{u}^{\mathrm{T}}\mathbf{v})^{\mathrm{T}}\mathbf{w}$$

vector–scalar product
row vector

scalar–vector product
column vector

Vector Operations

- Vector-Vector Products:
  - Given two vectors $x \in R^m, y \in R^n$:
    - $x \otimes y = xy^T \in R^{m \times n}$ is called the outer product of the vectors: $(xy^T)_{ij} = x_i y_j$

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}$$

Example

- Represent $A \in R^{m \times n}$ with outer product of two vectors:

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix}$$

Outer Product Properties

- Properties:
  - $(u \otimes v)^T = (v \otimes u)$
  - $(v + w) \otimes u = v \otimes u + w \otimes u$
  - $u \otimes (v + w) = u \otimes v + u \otimes w$
  - $c(v \otimes u) = (cv) \otimes u = v \otimes (cu)$
  - $(u.v) = trace(u \otimes v) \;\; (u, v \in R^n)$
  - $(u \otimes v)w = (v.w)u$

Vector Operations

- Vector-Vector Products:
  - Hadamard
  - Element-wise product

$$c = a \odot b = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ \cdot \\ \cdot \\ a_n b_n \end{bmatrix}$$

- Hadamard product is used in image compression techniques such as JPEG. It is also known as Schur product

- Hadamard Product is used in LSTM (Long Short-Term Memory) cells of Recurrent Neural Networks (RNNs).

Hadamard Product Properties

- Properties:
  - $a \odot b = b \odot a$
  - $a \odot (b \odot c) = (a \odot b) \odot c$
  - $a \odot (b + c) = a \odot b + a \odot c$
  - $(\theta a) \odot b = a \odot (\theta b) = \theta (a \odot b)$
  - $a \odot \mathbf{0} = \mathbf{0} \odot a = \mathbf{0}$

06

Matrix Multiplication

Basic Notation

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

Definition

The linear combinations of m vectors $a_1, \ldots a_m$, each with size n is:

$$\beta_1 a_1 + \cdots + \beta_m a_m$$

where $\beta_1, \ldots, \beta_m$ are scalars and called the coefficients of the linear combination

Matrix-Vector Multiplication

- If we write A by rows, then we can express Ax as,

$$A \in \mathbb{R}^{m \times n} \qquad y = Ax = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}. \qquad a_i^T x = \sum_{j=1}^{n} a_{ij} x_j$$

- If we write A by columns, then we have:

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [a_1]x_1 + [a_2]x_2 + \cdots + [a_n]x_n \, .$$

  - y is a **linear combination** of the columns A.

We will learn in next lectures

columns of A are linearly independent if Ax = 0 implies x = 0

Matrix-Vector Multiplication

It is also possible to multiply on the left by a row vector.

- If we write A by columns, then we can express $x^T A$ as,

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = [x^T a_1 \quad x^T a_2 \quad \cdots \quad x^T a_n]$$

- Expressing A in terms of rows we have:

$$y^T = x^T A = [x_1 \quad x_2 \quad \cdots \quad x_m] \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

$$= x_1[- \quad a_1^T \quad -] + x_2[- \quad a_2^T \quad -] + \ldots + x_m[- \quad a_m^T \quad -]$$

  - y^T is a linear combination of the rows of A.

Matrix-Vector Multiplication

- $A(u+v) = Au + Av$
- $(A+B)u = Au + Bu$
- $(\alpha A)u = \alpha(Au) = A(\alpha u) = \alpha Au$
- $0u = 0$
- $A0 = 0$
- $Iu = u$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix}$$

Example: Write in matrix-vector multiplication

- Column j: $a_j =$
- Row i: $a_i^T =$
- Vector sum of rows of $A =$
- Vector sum of columns of $A =$

$$A = \begin{bmatrix} -1 & 2 & 1 \\ 2 & 0 & -2 \end{bmatrix}$$

Matrix-Matrix Multiplication

Definition

Let A be an $m \times n$ matrix over the field F and let B be an $n \times p$ matrix over F. The product AB is the $m \times p$ matrix C whose i, j entry is:

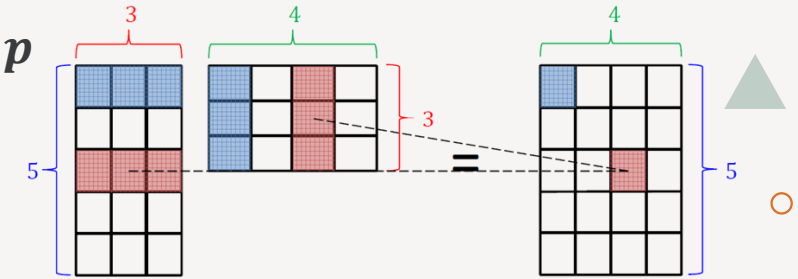
$$C_{ij} = \sum_{r=1}^{n} A_{ir} B_{rj}$$

- $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p} \rightarrow \mathbb{R}^{m \times p}$
  - a_i rows of A, b_j cols of B

$$\boldsymbol{C = AB} \qquad \boldsymbol{for\ 1 \leq i \leq m, 1 \leq j \leq p}$$

$$\boldsymbol{dot\ product(a_i . b_j)}$$

$$\boldsymbol{C_{ij} = a_i^T b_j}$$



Matrix-Matrix Multiplication (different views)

1. As a set of vector-vector products

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix}$$

2. As a sum of outer products

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_n^T & - \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T$$

Matrix-Matrix Multiplication (different views)

3. As a set of matrix-vector products.

$$C = AB = A\begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}$$

Here the ith column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection.

4. As a set of vector-matrix products.

$$C = AB = \begin{bmatrix} - & a_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_m^T & - \end{bmatrix} B = \begin{bmatrix} - & a_1^T B & - \\ - & a_2^T B & - \\ & \vdots & \\ - & a_m^T B & - \end{bmatrix}$$

Matrix-Matrix Multiplication

- Properties:
  - Associative

$$(AB)C = A(BC)$$

  - Distributive

$$A(B + C) = AB + AC$$

  - NOT commutative

$$AB \neq BA$$

- Dimensions may not even be conformable

07

Orthogonal Matrix

Orthogonal Matrix

Definition

- A matrix is orthogonal if its columns are:
  - orthogonal
  - has norm 1

Orthogonal Matrix

Theorem

A matrix is orthogonal if and only if it preserves length and angle.

Proof

Square Orthogonal Matrix

Note

- Columns of A are orthonormal $\leftrightarrow A^T A = I$
- Square matrix with orthonormal columns is an orthogonal matrix
  - Columns and rows are orthonormal vectors
  - $A^T A = A A^T = I$
  - Is necessarily invertible with inverse $A^T = A^{-1}$

Square Orthogonal Matrix

Example

- Identity matrix $I^T I = I$

- Rotation matrix

$$R^T R = \begin{bmatrix} cos\,\theta & sin\,\theta \\ -sin\,\theta & cos\,\theta \end{bmatrix} \begin{bmatrix} cos\,\theta & -sin\,\theta \\ sin\,\theta & cos\,\theta \end{bmatrix} =$$

$$\begin{bmatrix} cos^2\theta + sin^2\theta & -cos\theta sin\theta + sin\theta cos\theta \\ -sin\theta cos\theta + cos\theta sin\theta & sin^2\theta + cos^2\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

Square Orthogonal Matrix

Example

❑Reflection matrix

$$\begin{bmatrix}\cos(2\theta) & \sin(2\theta)\\ \sin(2\theta) & -\cos(2\theta)\end{bmatrix}^T \begin{bmatrix}\cos(2\theta) & \sin(2\theta)\\ \sin(2\theta) & -\cos(2\theta)\end{bmatrix} =$$

$$\begin{bmatrix}\cos^2(2\theta)+\sin^2(2\theta) & cos\,(2\theta)\sin(2\theta)-\sin(2\theta)\cos(2\theta)\\ \sin(2\theta)\cos(2\theta)-\cos(2\theta)\sin(2\theta) & \sin^2(2\theta)+\cos^2(2\theta)\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = I$$

Theorem

Every 2×2 orthogonal matrices can be expressed as Rotation or Reflection, or a composition of them.

proof?

Tall Orthogonal Matrix

- $A \in \mathbb{R}^{m \times n}, m > n$
- The inner products of the columns give the identity, so $\boldsymbol{A^T A = I_n}$

$$(A^T A)_{ij} = \langle a_i, a_j \rangle = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

- AA^T is a **projection matrix** onto the column space of A. If $m>n$ then AA^T is not full rank. Therefore, $\boldsymbol{AA^T \neq I_m}$

$$\text{rank}(AA^T) \leq \min(\text{rank}(A), \text{rank}(A^T)) = \text{rank}(A) \leq n$$

But remember: $AA^T \in \mathbb{R}^{m \times m}$. So **the maximum possible rank of AA^T is n, which is strictly less than m.**

Wide Orthogonal Matrix

- $A \in \mathbb{R}^{m \times n}, m < n$
- The inner products of the rows give the identity, so $\boldsymbol{AA^T = I_m}$

$$(AA^T)_{ij} = \langle a_i, a_j \rangle = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

- $A^T A$ is a **projection matrix** onto the row space of A. If $m<n$ then $A^T A$ is not full rank. Therefore, $\boldsymbol{A^T A \neq I_n}$

$$\text{rank}(A^T A) \leq \min(\text{rank}(A), \text{rank}(A^T)) = \text{rank}(A) \leq m$$

But remember: $A^T A \in \mathbb{R}^{n \times n}$.

So the maximum possible rank of $A^T A$ **is** m**, which is strictly less than** n**.**

Properties of Orthogonal Matrix

Note

If $A \in \mathbb{R}^{m \times n}$ has orthonormal columns, then the linear function $f(x)=Ax$:

- Preserves inner product:

$$(Ax)^T(Ay) = x^T y$$

- Preserves norm:

$$\|Ax\| = \|x\|$$

- Preserves distances:

$$\|Ax - Ay\| = \|x - y\|$$

- Preserves angels:

$$\angle(Ax, Ay) = \arccos\left(\frac{(Ax)T(Ay)}{\|Ax\|\|Ay\|}\right) = \arccos\left(\frac{xTy}{\|x\|\|y\|}\right) = \angle(x, y)$$

This is a mapping with preserving properties of input

08

Linear Transformation

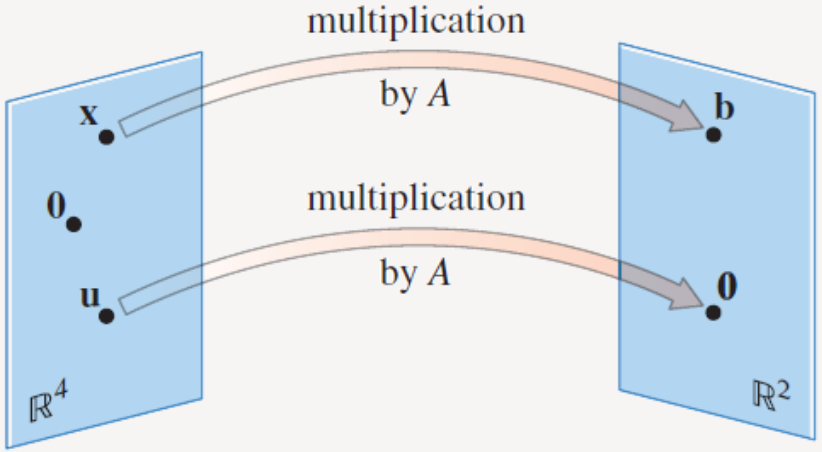
Linear Transformation

- Matrix is a linear transformation: map one vector to another vector

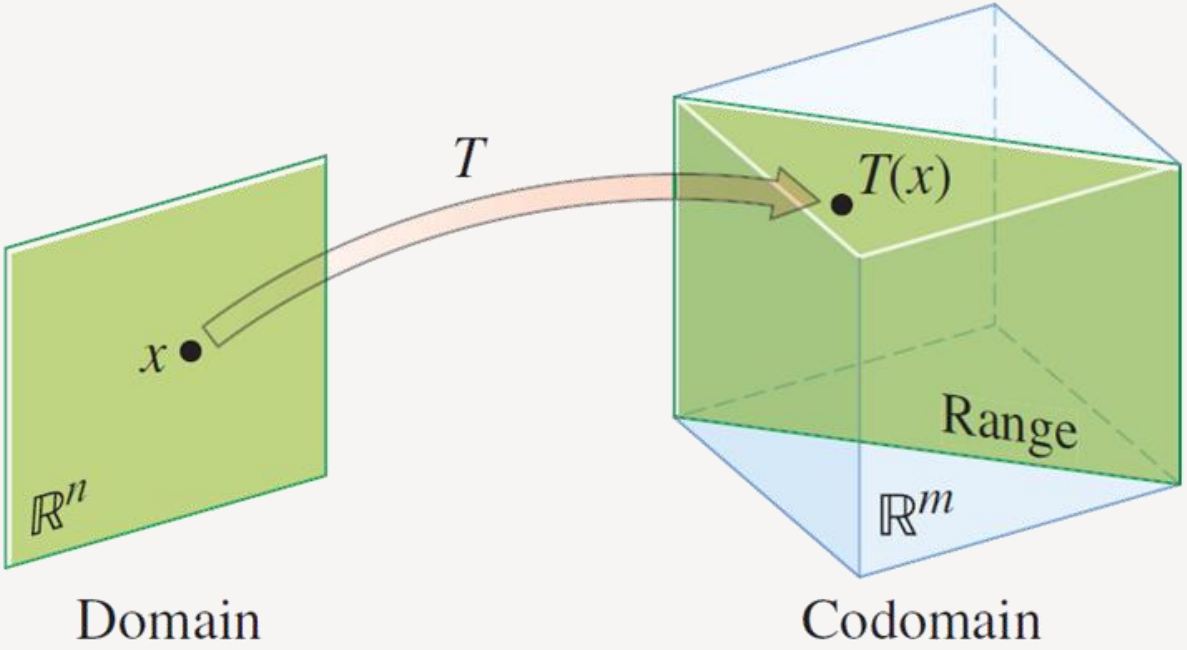
$$A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^{n}, y \in \mathbb{R}^{m}: \qquad y_{m\times 1} = A_{m\times n} x_{n\times 1}$$

$$A : \mathbb{R}^{n} \rightarrow \mathbb{R}^{m}$$

- Input–output



Linear Transformation



Domain, codomain, and range of $T : \mathbb{R}^n \to \mathbb{R}^m$

Linear Mapping

Definition

Let V and W be vector spaces over the field $\mathbb{F}$. A linear transformation (or a linear map) from V into W is a function $\mathbf{T}: V \to W$ that satisfies following properties for all x, y in V and all scalars a in $\mathbb{F}$:

$$T(x+y) = T(x) + T(y)$$
$$T(\alpha x) = \alpha T(x)$$

Notes

- $T(0) = 0$
- Transformation preserves linear combinations

$$T(\alpha_1 x_1 + \cdots + \alpha_n x_n) = \alpha_1 \big(\mathrm{T}(\mathrm{x}_1)\big) + \cdots + \alpha_n \big(\mathrm{T}(\mathrm{x}_n)\big)$$

Linear Mapping

Notes

- The set of linear maps from V to W is denoted by $\mathcal{L}(\mathrm{V}, \mathrm{W})$.
- The set of linear maps from V to V is denoted by $\mathcal{L}(\mathrm{V})$.
  In other words, $\mathcal{L}(\mathrm{V}) = \mathcal{L}(\mathrm{V}, V)$

Linear Mapping

Example

Which are linear mapping?

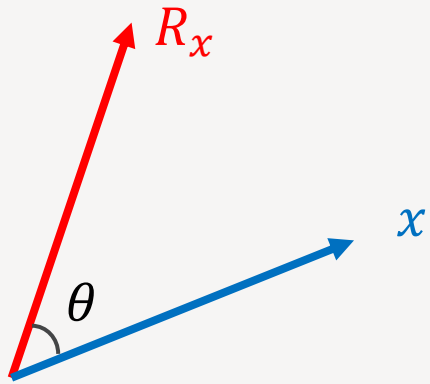
- **zero** map $0 : V \to W$
- **identity** map $I : V \to V$
- Let $T : \mathcal{P}(\mathbb{F}) \to \mathcal{P}(\mathbb{F})$ be the **differentiation** map defined as $T_{\mathcal{P}(z)} = \mathcal{P}(\acute{z})$
- Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ be the map given by $T(x, y) = (x - 2y, 3x + y)$
- $T(x) = e^x$
- $T : \mathbb{F} \to \mathbb{F}$ given by $T(x) = x - 1$

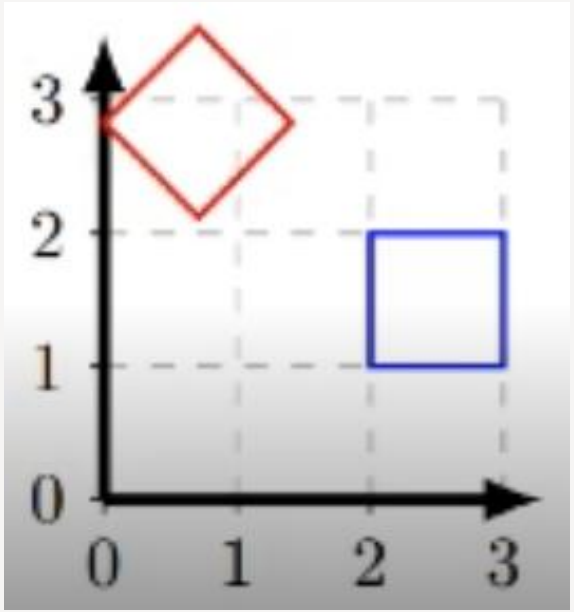
09

Rotation-Projection-Reflection

Rotation with θ degree

- $R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ Why?



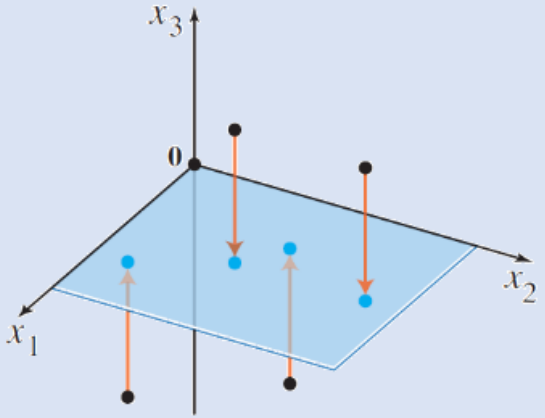


Projection

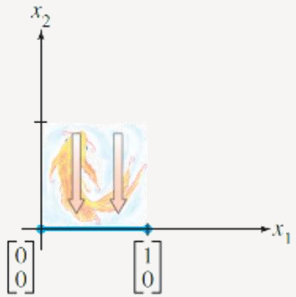
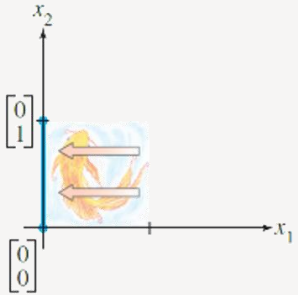
Example

If $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, then the transformation $\mathbf{x} \mapsto A\mathbf{x}$ projects points in $\mathbb{R}^3$ onto the x_1x_2–plane because

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \mapsto \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$$

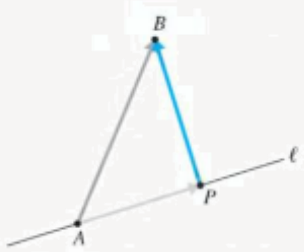


Projection

Transformation	Image of the Unit Square	Standard Matrix
Projection onto the x_1–axis		$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$
Projection onto the x_2–axis		$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$

Projection

- Finding the distance from a point B to line l = Finding the length of line segment BP
- AP: projection of AB onto the line l



Definition

If **u** and **v** are vectors in $\mathbb{R}^n$ and $\mathbf{u} \neq \mathbf{0}$, then the **projection of v onto u** is the vector $proj_{\mathbf{u}}(\mathbf{v})$ defined by

$$proj_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}}\right)\mathbf{u}$$

The projection of **v** onto **u**

Projection on θ Line

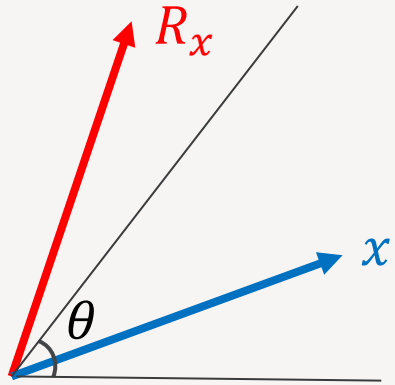
$$P = \begin{bmatrix} \cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{bmatrix}$$

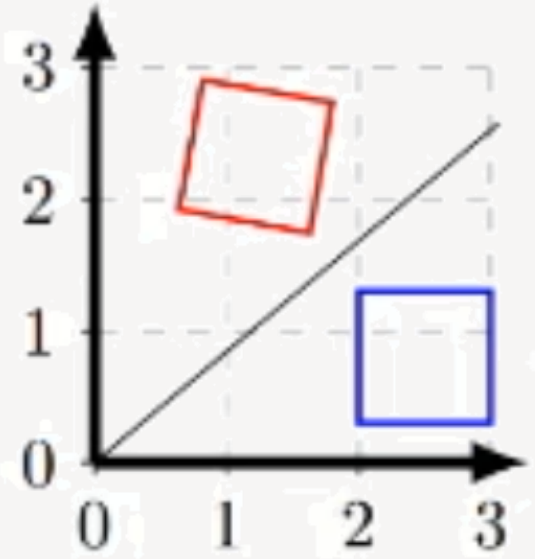
Why?

$$P^2 = P$$

Reflection in θ Line

- $R = \begin{bmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{bmatrix}$





Why?

$$R^2 = I$$

Conclusion

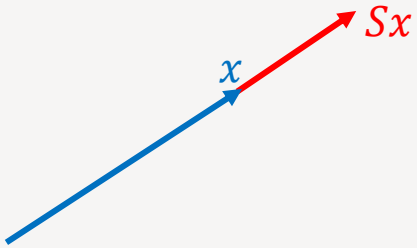
Lemma

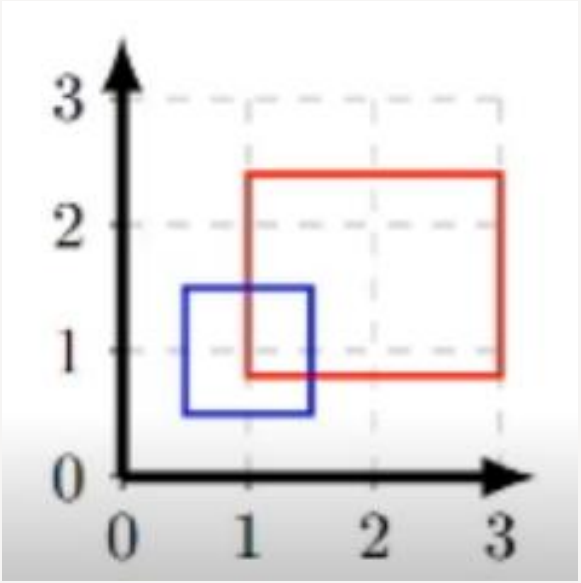
All orthogonal matrices can be expressed as Rotation or Reflection or their combination

Applications

Uniform Scaling

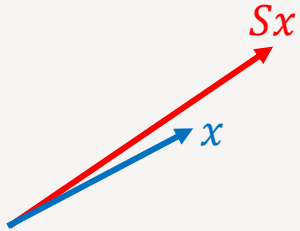
- $S = sI = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix}$

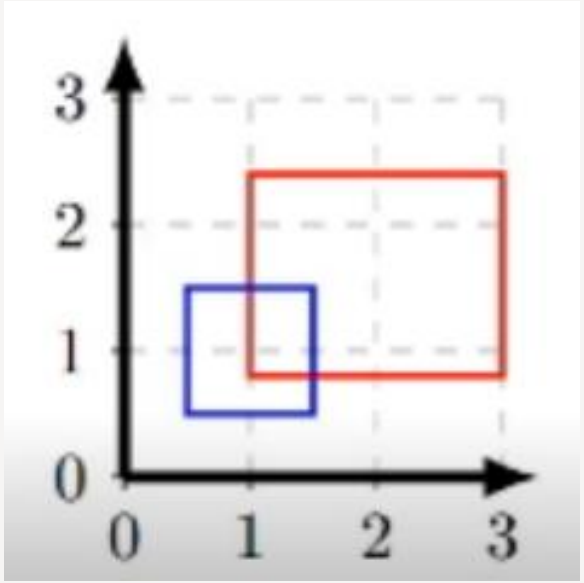




Non-uniform Scaling

- $S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix}$





Shearing

Example

Let $A = \begin{bmatrix} 1 & 3 \\ 0 & 1 \end{bmatrix}$. The transformation $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$

A typical shear matrix is of the form

$$S = \begin{pmatrix} 1 & 0 & 0 & \lambda & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

sheep

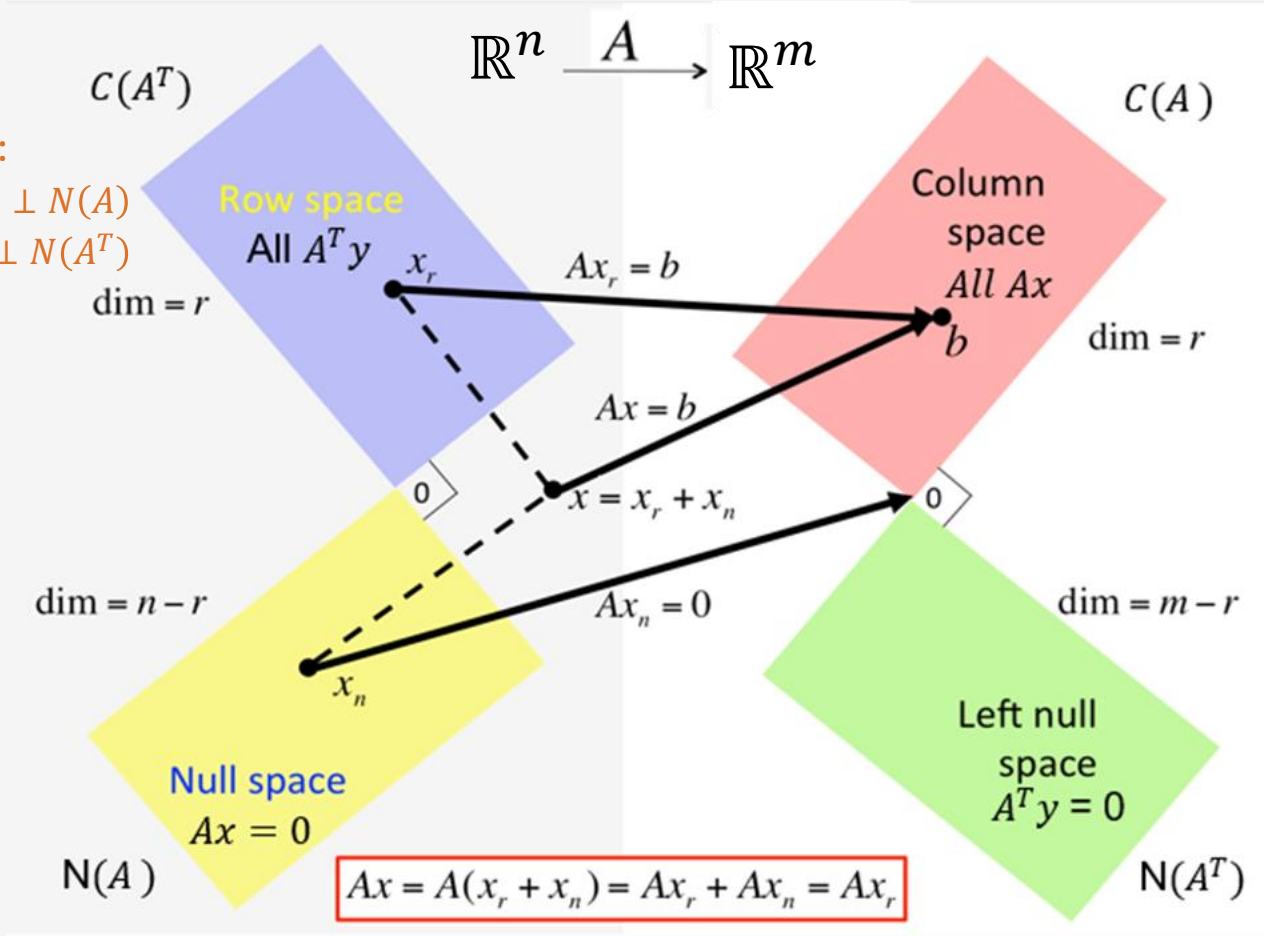
sheared sheep

10

Four Fundamental Subspaces

(of Matrix Space)

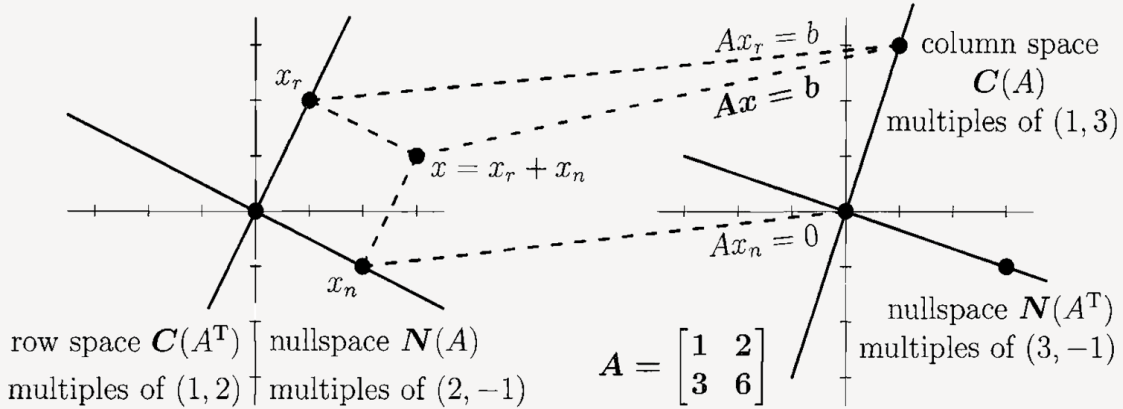
- **Proof that:**
  - $C(A^T) \perp N(A)$
  - $C(A) \perp N(A^T)$



For $A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$ $m = n = 2$, and $r = 1$

- The column space has all multiples of $\begin{bmatrix} 1 \\ 3 \end{bmatrix}$.
- The row space contains all multiples of $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$.
- The null space contains all multiples of $\begin{bmatrix} -2 \\ 1 \end{bmatrix}$.
- The left null space contains all multiples of y = $\begin{bmatrix} -3 \\ 1 \end{bmatrix}$.

The rows of A with coefficients -3 and 1 add to zero, $A^T y = 0$



Here the row space $C(A^T)$ is the multiples of (1,2), the null space $N(A)$ is the multiples of (2, -1), column space $C(A)$ is the multiples of (1,3) and left null space $N(A^T)$ is multiples of (3,-1). Also we can see that Row space is orthogonal to Null space and Column space is orthogonal to Left null space.

11

Determinant

Determinant of a matrix

The determinant of a 2×2 matrix $A = [a_{ij}]$ is the number:
Why???

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}$$

❑ The absolute value of the determinant of a matrix measures how much it expands space when acting as a linear transformation. That is, it is the area (or volume, or hypervolume, depending on the dimension) of the output of the unit square, cube, or hypercube after it is acted upon by the matrix.

Definition

Note

Let $T: \mathbb{R}^2 \to \mathbb{R}^2$ be the linear transformation determined by a 2×2 matrix A. If S is a parallelogram in $\mathbb{R}^2$, then

$$\{area\ of\ T(S)\} = |\det A|.\{area\ of\ S\}$$

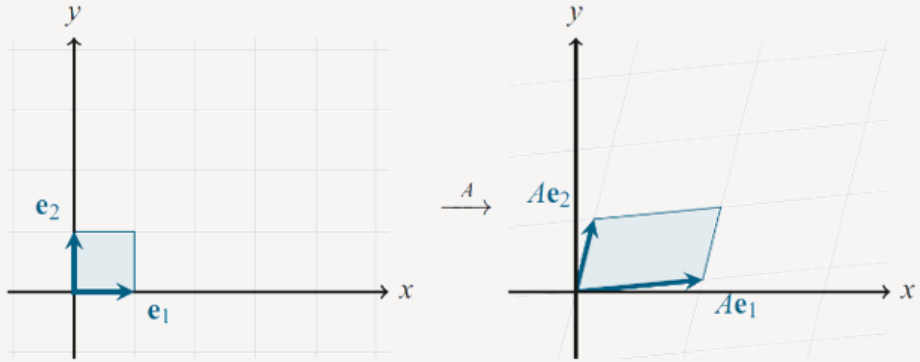
If T is determined by a 3×3 matrix A, and if S is a parallelepiped in $\mathbb{R}^3$, then

$$\{volume\ of\ T(S)\} = |\det A|.\{volume\ of\ S\}$$

Geometric interpretation

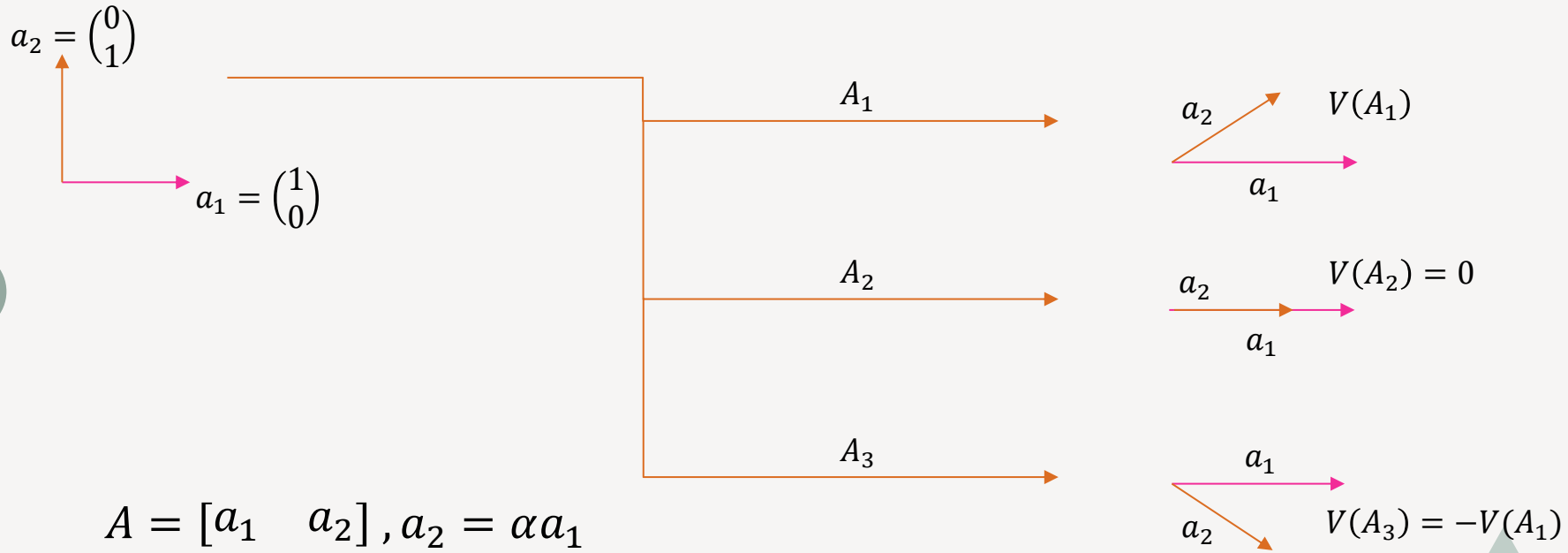
❑ The volume is a n-alternating multilinear map on all n-parallelepipeds such that the volume of standard unit parallelepiped is one.

$$\frac{volume\ of\ output\ region}{volume\ of\ input\ region}$$



A 2×2 matrix A stretches the unit square (with sides e_1 and e_2) into a parallelogram with sides Ae_1 and Ae_2 (the columns of A). The determinant of A is the area of this parallelogram.

Geometric interpretation



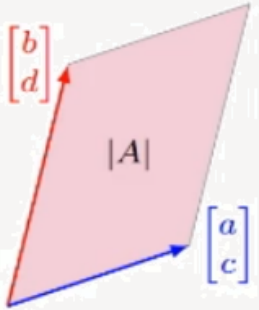
$$A = [a_1 \quad a_2], a_2 = \alpha a_1$$

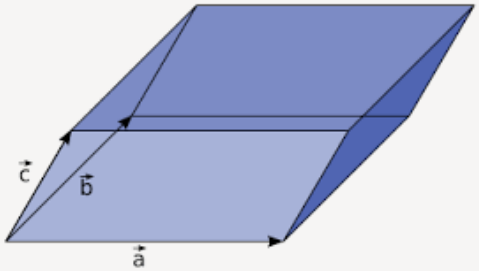
$$V(a_1, a_2) = -V(a_2, a_1)$$

Determinants as Area or Volume

- If A is a 2×2 matrix, the area of the parallelogram determined by the columns of A is det(A)
- If A is a 3×3 matrix, the volume of the parallelepiped determined by the columns of A is det(A)
- Examples:

Volume of $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ It is a rotation with $\boldsymbol{\theta}$ degree





Volume

Definition

Every $\boldsymbol{n}$-dimensional parallelepiped with $\{\boldsymbol{a_1}, \dots, \boldsymbol{a_n}\}$ as legs is associated with a real number, called its volume which has the following properties:

If we stretch a parallelepiped by multiplying one of its legs by a scalar λ, its volume gets multiplied by λ.

If we add a vector ω to i-th legs of a n-dimensional parallelepiped with $\{a_1, \dots, a_i, a_{i+1}, \dots, a_n\}$, then its volume is the sum of the volume from $\{a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_n\}$ and the volume of $\{a_1, \dots, a_{i-1}, \omega, a_{i+1}, \dots, a_n\}$.

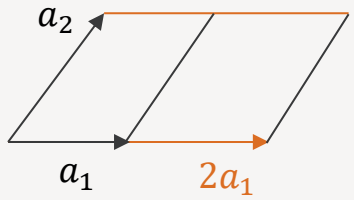
The volume changes sign when two legs are exchanged.

The volume of the parallelepiped with $\{e_1, \dots, e_n\}$ is one.

$$\phi : \underbrace{V \times \cdots \times V}_{n} \to \mathbb{R}$$

Volume Example

- 2D Case
  - $V(a_1, a_2)$
  - $V(2\ a_1, a_2) = 2V(a_1, a_2)$
  - $V(-a_1, a_2) = -\ V(a_1, a_2)$
  - $V(\beta a_2, a_1) = -\ \beta\ V(a_1, a_2)$



12

Norm Space

Vector Norms

- p-norm:

$$\|x\|_p = (|x_1|^p + |x_2|^p + \ldots + |x_n|^p)^{\frac{1}{p}}$$

$$subject\ to\ p \geq 1$$

- What is the shape of $\|x\|_p = 1$?
- Properties?

Norm

Definition (Norm)

- A function $f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is a norm if
  1. $f(x) \geq 0, f(x) = 0 \Leftrightarrow x = 0$ (positivity)
  2. $f(\alpha x) = |\alpha| f(x), \forall \alpha \in \mathbb{R}$ (homogeneity)
  3. $f(x + y) \leq f(x) + f(y)$ (triangle inequality)

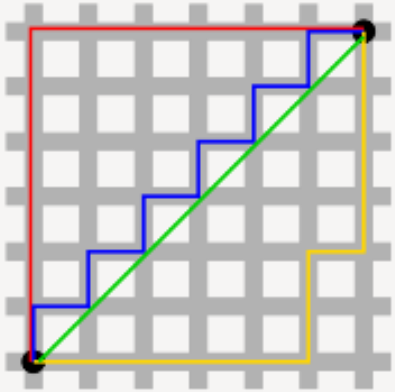
Vector Norms

- 1-norm(l_1):

$$\|x\|_1 = (|x_1| + |x_2| + \ldots + |x_n|)$$

- What is the shape of $\|x\|_1 = 1$?
- The distance between two vectors under the l_1 norm is also referred to as the Manhattan Distance.
- Properties?

l_1 distance between $(0, 1)$ and $(1, 0)$?

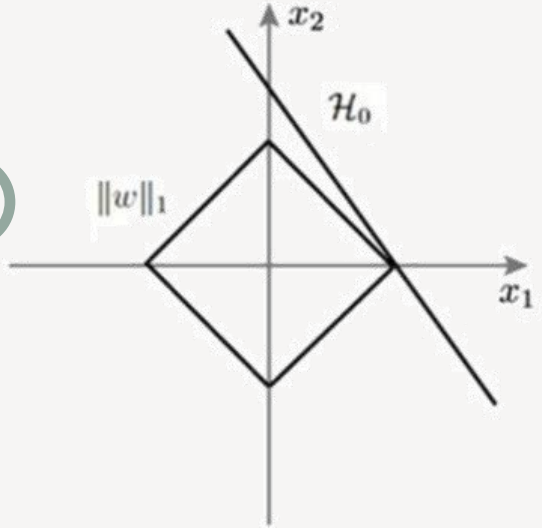


L1 and L2 norm comparisons

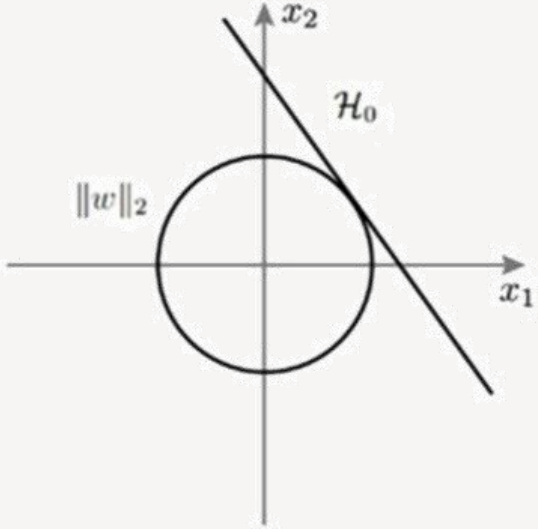
- Robustness is defined as resistance to outliers in a dataset. The more able a model is to ignore extreme values in the data, the more robust it is.
- Stability is defined as resistance to horizontal adjustments. This is the perpendicular opposite of robustness.
- Computational difficulty
- Sparsity

Why is l_1 supposed to lead to sparsity than l_2?

$$\min_x \|x\|_{1\ or\ 2}\ ,\quad subject\ to\ Ax = b$$



l_1 regularization



l_2 regularization

Vector Norms

- ∞-norm(l_∞)(max norm):

$$l_\infty = \max(|x_1|, |x_2|, \dots, |x_n|)$$

- What is the shape of $|x|_\infty = 1$?
- Properties?

Vector Norms

- $\frac{1}{2}$-norm($l_{\frac{1}{2}}$)

- What is the shape of $|x|_{\frac{1}{2}} = 1$?

- Properties?

Vector Norms

- 0-norm(l_0):

$$\|x\|_0 = \lim_{\alpha \to 0^+} \|x\|_\alpha = \left(\sum_{k=1}^{n} |x|^\alpha \right)^{\frac{1}{\alpha}} = \sum_{k=1}^{n} 1_{(0,\infty)}(|x|)$$

- 0-norm, defined as **the number of non-zero elements in a vector**, is an ideal quantity for feature selection. However, minimization of 0-norm is generally regarded as a combinatorially difficult optimization

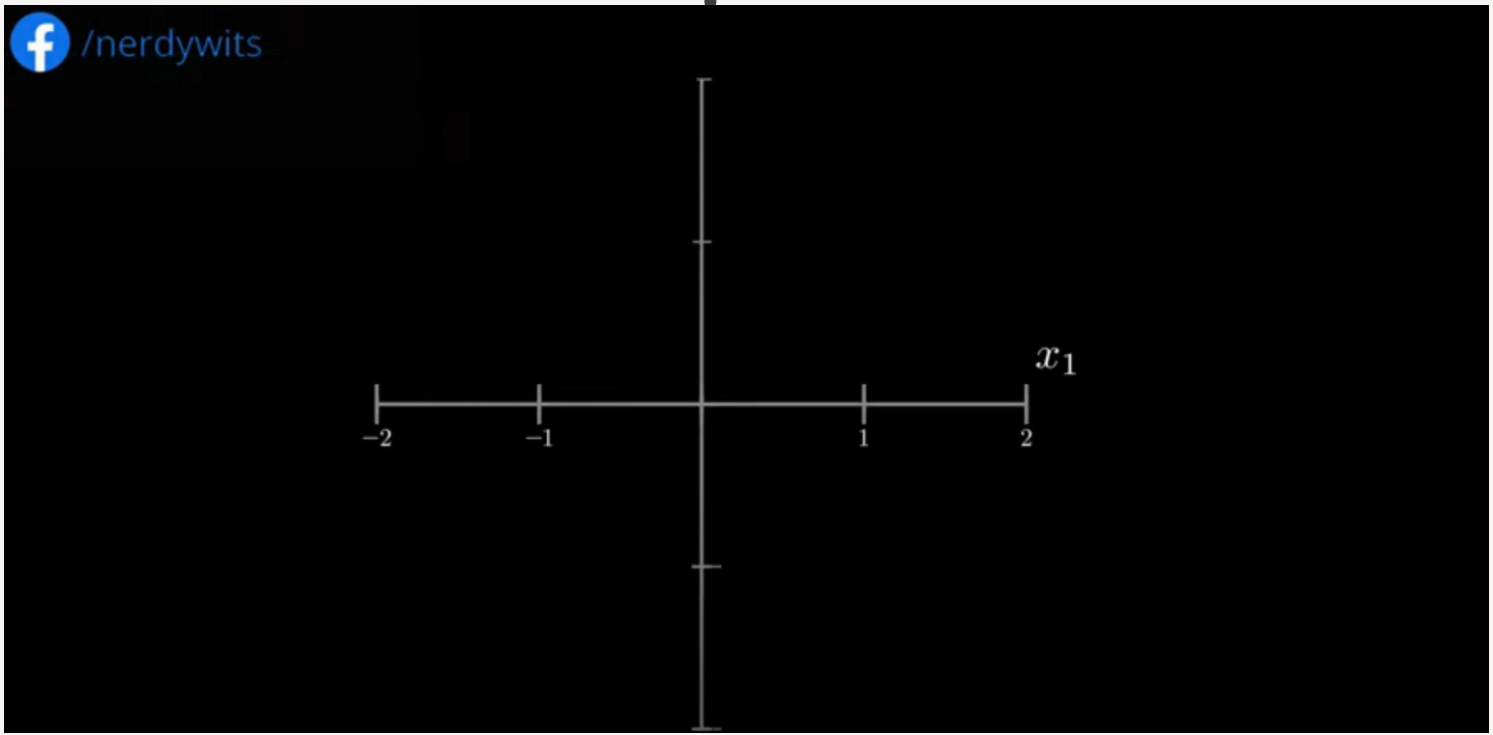
- $\|x\|_0 = \sum_{x_i \neq 0} 1$

Vector Norms

- Is 0-norm a valid norm?

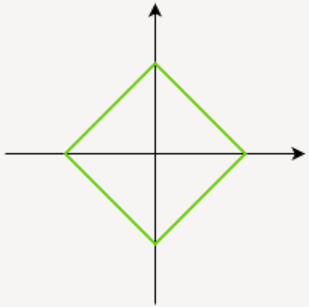
- What is the shape of $\|x\|_0 = 1$?

Vector Norms Shapes

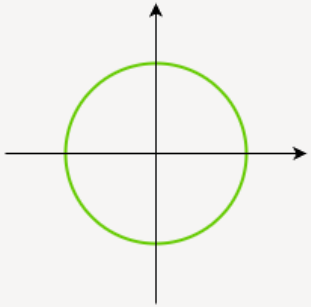


Norms and Convexity

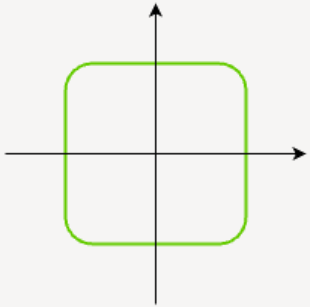
- For $\boldsymbol{p \geq 1}$, $\boldsymbol{l_p}$ norm is convex



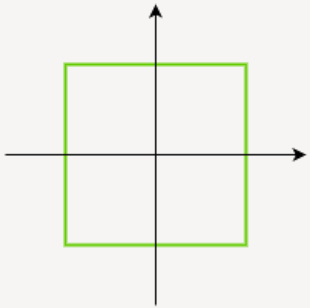
$\|x\|_1 = 1$



$\|x\|_2 = 1$



$\|x\|_p = 1$



$\|x\|_\infty = 1$