Mathematics for Artificial Intelligence Transformation and Applications

Andres Mendez-Vazquez

April 9, 2020

<ロト < 回 ト < 巨 ト < 巨 ト ミ の < () 1/140

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation
- 3 Linear Regression
 - The Simplest Functions
 - Splitting the Space
 - Defining the Decision Surface
 - Properties of the Hyperplane $\boldsymbol{w}^T \boldsymbol{x} + w_0$
 - Augmenting the Vector
 - Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
 - The Error Idea
 - The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Singular Value Decomposition

- Introduction
- Image Compression



< ロ > < 同 > < 回 > < 回 >

Outline

Linear Transformation

Introduction

- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





イロト イヨト イヨト

We can go further

We can think on the matrix A as a function!!!



We can go further

We can think on the matrix A as a function!!!

In general

A function f whose domain \mathbb{R}^n and defines a rule that associate $\bm{x}\in\mathbb{R}^n$ to a vector $\bm{y}\in\mathbb{R}^m$

$$\boldsymbol{y} = f\left(\boldsymbol{x}
ight)$$
 equivalently $f: \mathbb{R}^{n} \longrightarrow \mathbb{R}^{m}$

We can go further

We can think on the matrix A as a function!!!

In general

A function f whose domain \mathbb{R}^n and defines a rule that associate $\bm{x}\in\mathbb{R}^n$ to a vector $\bm{y}\in\mathbb{R}^m$

$$\boldsymbol{y} = f\left(\boldsymbol{x}\right)$$
 equivalently $f: \mathbb{R}^{n} \longrightarrow \mathbb{R}^{m}$

イロト イヨト イヨト

4/140

We like the second expression because

1 It is easy to identify the domain \mathbb{R}^n

It is easy to find the range \mathbb{R}^m

We can go further

We can think on the matrix A as a function!!!

In general

A function f whose domain \mathbb{R}^n and defines a rule that associate $\bm{x}\in\mathbb{R}^n$ to a vector $\bm{y}\in\mathbb{R}^m$

$$\boldsymbol{y} = f\left(\boldsymbol{x}\right)$$
 equivalently $f: \mathbb{R}^{n} \longrightarrow \mathbb{R}^{m}$

イロト イヨト イヨト

4/140

We like the second expression because

- **()** It is easy to identify the domain \mathbb{R}^n
- 2 It is easy to find the range \mathbb{R}^m

Examples

$f: \mathbb{R} \to \mathbb{R}^3$ $f(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} t \\ 3t^2 + 1 \\ \sin(t) \end{pmatrix}$

This are called parametric functions

 Depending on the context, it could represent the position or the velocity of a mass point.



Examples

$$f: \mathbb{R} \to \mathbb{R}^3$$

$$f(t) = \begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = \begin{pmatrix} t \\ 3t^2 + 1 \\ sin(t) \end{pmatrix}$$

This are called parametric functions

• Depending on the context, it could represent the position or the velocity of a mass point.



< ロ > < 回 > < 回 > < 回 > < 回 >

Outline

Linear Transformation

Introduction

Functions that can be defined using matrices

- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regressior

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

- Introduction
- Image Compression



イロト イヨト イヨト

A Classic Example

We have

if A is a $m\times n$, we can use A to define a function.

We will call them

 $f_A: \mathbb{R}^n \to \mathbb{R}^m$

In other words

 $f_A(\boldsymbol{x}) = Ax$



ヘロト ヘロト ヘヨト ヘヨト

A Classic Example

We have

if A is a $m\times n$, we can use A to define a function.

We will call them

$$f_A: \mathbb{R}^n \to \mathbb{R}^m$$

In other words

 $f_{A}\left(\boldsymbol{x}\right)=Ax$



イロト イロト イヨト イヨト

A Classic Example

We have

if A is a $m\times n$, we can use A to define a function.

We will call them

$$f_A: \mathbb{R}^n \to \mathbb{R}^m$$

In other words

$$f_A\left(\boldsymbol{x}\right) = Ax$$



イロン イ団 とく ヨン イヨン

Example

Let

$$A_{2\times3} = \left(\begin{array}{rrr} 1 & 2 & 3 \\ 4 & 5 & 6 \end{array}\right)$$

This allows to define



We have

• For each vector $x \in \mathbb{R}^3$ to the vector $Ax \in \mathbb{R}^2$



・ロト ・ 日 ト ・ 日 ト ・ 日 ト ・

Example

Let

$$A_{2\times3} = \left(\begin{array}{rrr} 1 & 2 & 3\\ 4 & 5 & 6 \end{array}\right)$$

This allows to define

$$f_A \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y + z \\ 4x + 5y + 6z \end{pmatrix}$$

We have

ullet For each vector $oldsymbol{x} \in \mathbb{R}^3$ to the vector $Aoldsymbol{x} \in \mathbb{R}^3$



Example

Let

$$A_{2\times3} = \left(\begin{array}{rrr} 1 & 2 & 3\\ 4 & 5 & 6 \end{array}\right)$$

This allows to define

$$f_A \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + 2y + z \\ 4x + 5y + 6z \end{pmatrix}$$

We have

• For each vector $oldsymbol{x} \in \mathbb{R}^3$ to the vector $Aoldsymbol{x} \in \mathbb{R}^2$



イロン イボン イヨン トヨ

Outline

Linear Transformation

Introduction

Functions that can be defined using matrices

Linear Functions

- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regressior

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





イロト イヨト イヨト

We have

Definition

A function $f:\mathbb{R}^n\to\mathbb{R}^m$ is said to be linear if

9
$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

2
$$f(c\boldsymbol{x}) = cf(\boldsymbol{x})$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and for all the scalars c.

A linear function f is also known as a linear transformation.



We have

Definition

A function $f:\mathbb{R}^n\to\mathbb{R}^m$ is said to be linear if

1
$$f(x_1 + x_2) = f(x_1) + f(x_2)$$

2
$$f(c\boldsymbol{x}) = cf(\boldsymbol{x})$$

for all $x_1, x_2 \in \mathbb{R}^n$ and for all the scalars c.

Thus

A linear function f is also known as a linear transformation.



We have the following proposition

Proposition

 $f: \mathbb{R}^n \to \mathbb{R}^m$ is linear \iff for all vectors $x_1, x_2 \in \mathbb{R}^n$ and for all the scalars c_1, c_2 :

$$f(c_1 x_1 + c_2 x_2) = c_1 f(x_1) + c_2 f(x_2)$$

Proof

Any idea?



We have the following proposition

Proposition

 $f: \mathbb{R}^n \to \mathbb{R}^m$ is linear \iff for all vectors $x_1, x_2 \in \mathbb{R}^n$ and for all the scalars c_1, c_2 :

$$f(c_1 x_1 + c_2 x_2) = c_1 f(x_1) + c_2 f(x_2)$$

Proof

Any idea?



Proof

If $A_{m \times n}$ is a matrix, f_A is a linear transformation

How?

First

$f_{A}(x_{1}+x_{2}) = A(x_{1}+x_{2}) = Ax_{1} + Ax_{2} = f_{A}(x_{1}) + f_{A}(x_{2})$

Second

What about $f_A(c m{x}_1)$?



Proof

If $A_{m \times n}$ is a matrix, f_A is a linear transformation

How?

First

$$f_A(x_1 + x_2) = A(x_1 + x_2) = Ax_1 + Ax_2 = f_A(x_1) + f_A(x_2)$$

Second

What about $f_A\left(c oldsymbol{x}_1
ight)?$



Proof

If $A_{m \times n}$ is a matrix, f_A is a linear transformation

How?

First

$$f_A(x_1 + x_2) = A(x_1 + x_2) = Ax_1 + Ax_2 = f_A(x_1) + f_A(x_2)$$

Second

What about $f_A(c\boldsymbol{x}_1)$?



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions

Kernel and Range

- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

- Introduction
- Image Compression



< ロ > < 回 > < 回 > < 回 > < 回 >

We have

Definition (Actually related the null-space)

If $f:\mathbb{R}^n\to\mathbb{R}^m$ is linear, the kernel of f is defined by

$$Ker\left(f\right) = \left\{\boldsymbol{v} \in \mathbb{R}^{n} | f\left(\boldsymbol{v}\right) = 0\right\}$$

Definition

If $f: \mathbb{R}^n \to \mathbb{R}^m$ is linear, the range of f is defined by

 $Range\left(f
ight)=\left\{oldsymbol{y}\in\mathbb{R}^{m}|oldsymbol{y}=f\left(oldsymbol{x}
ight) ext{ for somer }oldsymbol{x}\in\mathbb{R}^{n}
ight\}$



We have

Definition (Actually related the null-space)

If $f:\mathbb{R}^n\to\mathbb{R}^m$ is linear, the kernel of f is defined by

$$Ker\left(f\right) = \left\{\boldsymbol{v} \in \mathbb{R}^{n} | f\left(\boldsymbol{v}\right) = 0\right\}$$

Definition

If $f:\mathbb{R}^n\to\mathbb{R}^m$ is linear, the range of f is defined by

$$Range\left(f
ight) = \left\{oldsymbol{y} \in \mathbb{R}^{m} | oldsymbol{y} = f\left(oldsymbol{x}
ight) \text{ for somer } oldsymbol{x} \in \mathbb{R}^{n}
ight\}$$



We have also the following Spaces

Row Space

We have that the span of the row vectors of A form a subspace.

Column Space

We have that the span of the column vectors of $A_{
m t}$ also, form a subspace.



We have also the following Spaces

Row Space

We have that the span of the row vectors of A form a subspace.

Column Space

We have that the span of the column vectors of A, also, form a subspace.



イロト イヨト イヨト

From This

It can be shown that

 $Ker\left(f
ight)$ is a subspace of \mathbb{R}^{n}

Alsc

 $Range\left(f
ight)$ is a subspace of \mathbb{R}^{n}



From This

It can be shown that

 $Ker\left(f
ight)$ is a subspace of \mathbb{R}^{n}

Also

 $Range\left(f\right)$ is a subspace of \mathbb{R}^{m}



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range

The Matrix of a Linear Transformation

- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

- Introduction
- Image Compression



< ロ > < 回 > < 回 > < 回 > < 回 >

Assume the following



$$oldsymbol{e}_1 = egin{pmatrix} 1 \ 0 \ dots \ 0 \end{pmatrix}_{n imes 1}, oldsymbol{e}_3 = egin{pmatrix} 0 \ 1 \ dots \ 0 \end{pmatrix}_{n imes 1}, ..., oldsymbol{e}_n = egin{pmatrix} 0 \ 0 \ dots \ dots \ 1 \end{pmatrix}_{n imes 1}$$

Then any vector $x\in$ 1

$$x=egin{pmatrix}x_1\x_2\dots\x_n\end{pmatrix}=x_1e_1+x_2e_2+\ldots+x_ne_n$$

Cinvestav

Assume the following



$$oldsymbol{e}_1 = egin{pmatrix} 1 \ 0 \ dots \ 0 \ \end{pmatrix}_{n imes 1}, oldsymbol{e}_3 = egin{pmatrix} 0 \ 1 \ dots \ 0 \ \end{pmatrix}_{n imes 1}, ..., oldsymbol{e}_n = egin{pmatrix} 0 \ 0 \ dots \ dots \ 1 \ \end{pmatrix}_{n imes 1}$$

Then any vector $oldsymbol{x} \in \mathbb{R}^n$

$$oldsymbol{x} = \left(egin{array}{c} x_1 \ x_2 \ dots \ x_n \end{array}
ight) = x_1oldsymbol{e}_1 + x_2oldsymbol{e}_2 + \ldots + x_noldsymbol{e}_n$$

イロト イロト イヨト

Cinu



Applying f

$$f(\mathbf{x}) = x_1 f(\mathbf{e}_1) + x_2 f(\mathbf{e}_2) + \dots + x_n f(\mathbf{e}_n)$$

A linear combination of elements

 $\left\{ f\left(oldsymbol{e}_{1}
ight) ,f\left(oldsymbol{e}_{2}
ight) ,...,f\left(oldsymbol{e}_{n}
ight)
ight\}$

They are column vectors in \mathbb{R}^{r}

 $A = \left(f\left(\boldsymbol{e}_{1}\right)|f\left(\boldsymbol{e}_{2}\right)|...|f\left(\boldsymbol{e}_{n}\right)\right)_{m \times n}$



イロト イロト イヨト イヨト



Applying f

$$f(\mathbf{x}) = x_1 f(\mathbf{e}_1) + x_2 f(\mathbf{e}_2) + \dots + x_n f(\mathbf{e}_n)$$

A linear combination of elements

{
$$f(\boldsymbol{e}_{1}), f(\boldsymbol{e}_{2}), ..., f(\boldsymbol{e}_{n})$$
}

They are column vectors in ${\mathbb R}$

 $A = \left(f\left(e_{1}\right)|f\left(e_{2}\right)|...|f\left(e_{n}\right)\right)_{m \times n}$





Applying f

$$f(\mathbf{x}) = x_1 f(\mathbf{e}_1) + x_2 f(\mathbf{e}_2) + \dots + x_n f(\mathbf{e}_n)$$

A linear combination of elements

{
$$f(\boldsymbol{e}_{1}), f(\boldsymbol{e}_{2}), ..., f(\boldsymbol{e}_{n})$$
}

They are column vectors in \mathbb{R}^m

$$A = (f(\boldsymbol{e}_1) | f(\boldsymbol{e}_2) | ... | f(\boldsymbol{e}_n))_{m \times m}$$



Thus, we have

Finally, we have

$$f(\boldsymbol{x}) = (f(\boldsymbol{e}_1) | f(\boldsymbol{e}_2) | ... | f(\boldsymbol{e}_n)) \boldsymbol{x} = A \boldsymbol{x}$$

Definition

 The matrix A defined above for the function f is called the matrix of f in the standard basis.



Thus, we have

Finally, we have

$$f(\boldsymbol{x}) = (f(\boldsymbol{e}_1) | f(\boldsymbol{e}_2) | ... | f(\boldsymbol{e}_n)) \boldsymbol{x} = A \boldsymbol{x}$$

Definition

• The matrix ${\cal A}$ defined above for the function f is called the matrix of f in the standard basis.



< ロ > < 回 > < 回 > < 回 > < 回 >

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation

Going Back to Homogeneous Equations

The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

- Introduction
- Image Compression



< ロ > < 回 > < 回 > < 回 > < 回 >

Given an $m \times n$ matrix A

The set of all solutions to the homogeneous equation $A m{x}$

• It is a subspace V of \mathbb{R}^n .

 $A \boldsymbol{x} = 0$

Remember how to prove the subspaces.

 $oldsymbol{x}_2 + oldsymbol{x}_2 \in V$ and $oldsymbol{c}oldsymbol{x} \in V$

• Do you remember?



Given an $m \times n$ matrix A

The set of all solutions to the homogeneous equation $A oldsymbol{x}$

• It is a subspace V of \mathbb{R}^n .

 $A\boldsymbol{x} = 0$

Remember how to prove the subspaces...

$$oldsymbol{x}_2 + oldsymbol{x}_2 \in V$$
 and $coldsymbol{x} \in V$

• Do you remember?



イロト イボト イヨト イヨト

Then, we have

Definition

 $\bullet\,$ This important subspace is called the null space of A, and is denoted Null(A)

is also known as

$oldsymbol{x}_{H}=\{oldsymbol{x}|Aoldsymbol{x}=0\}$



< ロ > < 回 > < 回 > < 回 > < 回 >

Then, we have

Definition

 $\bullet\,$ This important subspace is called the null space of A, and is denoted Null(A)

It is also known as

$$\boldsymbol{x}_{H} = \{\boldsymbol{x} | A \boldsymbol{x} = 0\}$$

Knowing that $\operatorname{Range}(f)$ and $\operatorname{Ker}(f)$ are sub-spaces

Which ones they are?

Any Idea?

Range(*f*

The column space of the matrix A.

 $\operatorname{Ker}(f)$

It is the null space of A.



Knowing that $\operatorname{Range}(f)$ and $\operatorname{Ker}(f)$ are sub-spaces

Which ones they are?

Any Idea?

 $\mathsf{Range}(f)$

The column space of the matrix A.

Ker(

It is the null space of A.



Knowing that $\operatorname{Range}(f)$ and $\operatorname{Ker}(f)$ are sub-spaces

Which ones they are?

Any Idea?

 $\mathsf{Range}(f)$

The column space of the matrix A.

 $\mathsf{Ker}(f)$

It is the null space of A.



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation
- 3 Linear Regressior
 - The Simplest Functions
 - Splitting the Space
 - Defining the Decision Surface
 - Properties of the Hyperplane $w^T x + w_0$
 - Augmenting the Vector
 - Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
 - The Error Idea
 - The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

- Introduction
- Image Compression



We have a nice theorem

Dimension Theorem

Let $f: \mathbb{R}^n \to \mathbb{R}^m$ be linear. Then

 $dim\left(domain\left(f\right)\right) = dim\left(Range\left(f\right)\right) + dim\left(Ker\left(f\right)\right)$

Where

The dimension of V , written dim(V), is the number of elements in any basis of V.



We have a nice theorem

Dimension Theorem

Let $f: \mathbb{R}^n \to \mathbb{R}^m$ be linear. Then

$$dim\left(domain\left(f\right)\right) = dim\left(Range\left(f\right)\right) + dim\left(Ker\left(f\right)\right)$$

Where

The dimension of V , written $\dim(V),$ is the number of elements in any basis of V.



Rank and Nullity of a Matrix

Definition

• The rank of the matrix A is the dimension of the row space of A, and is denoted R(A).

Example

• The rank of $I_{n imes n}$ is n.



Rank and Nullity of a Matrix

Definition

• The rank of the matrix A is the dimension of the row space of A, and is denoted R(A).

Example

• The rank of $I_{n \times n}$ is n.



Then

Theorem

The rank of a matrix in Gauss-Jordan form is equal to the number of leading variables.

Proof

• In the G form of a matrix, every non-zero row has a leading 1, which is the only non-zero entry in its column.

Then

 No elementary row operation can zero out a leading 1, so these non-zero rows are linearly independent.



< ロ > < 回 > < 回 > < 回 > < 回 >

Then

Theorem

The rank of a matrix in Gauss-Jordan form is equal to the number of leading variables.

Proof

• In the G form of a matrix, every non-zero row has a leading 1, which is the only non-zero entry in its column.

Then

No elementary row operation can zero out a leading 1, so these non-zero rows are linearly independent.



Then

Theorem

The rank of a matrix in Gauss-Jordan form is equal to the number of leading variables.

Proof

• In the G form of a matrix, every non-zero row has a leading 1, which is the only non-zero entry in its column.

Then

• No elementary row operation can zero out a leading 1, so these non-zero rows are linearly independent.



Thus

We have

• Since all the other rows are zero, the dimension of the row space of the Gauss-Jordan form is equal to the number of leading 1's.

Finally

This is the same as the number of leading variables. Q.E.D.



Thus

We have

• Since all the other rows are zero, the dimension of the row space of the Gauss-Jordan form is equal to the number of leading 1's.

Finally

• This is the same as the number of leading variables. Q.E.D.



イロト イヨト イヨト イヨト

About the Nullity of the Matrix

Definition

 The nullity of the matrix A is the dimension of the null space of A, and is denoted by dim [N(A)].

Example

• The nullity of *I* is 0.



About the Nullity of the Matrix

Definition

 The nullity of the matrix A is the dimension of the null space of A, and is denoted by dim [N(A)].

Example

• The nullity of I is 0.



< ロ > < 回 > < 回 > < 回 > < 回 >

About the Nullity of the Matrix

Definition

• The nullity of the matrix A is the dimension of the null space of A, and is denoted by dim [N(A)].

Example

• The nullity of I is 0.



イロト イヨト イヨト

Number of Free Variables

Theorem

The nullity of a matrix in Gauss-Jordan form is equal to the number of free variables.

Proof

• Suppose A is $m \times n$, and that the Gauss-Jordan form has j leading variables and k free variables:

$$j+k=n$$



イロト イヨト イヨト

Proof

Then, when computing the solution to the homogeneous equation

• We solve for the first *j* (leading) variables in terms of the remaining *k* free variables:

 s_1,s_2,s_3,\ldots,s_k

l hen

Then the general solution to the homogeneous equation are:

 $s_1 \boldsymbol{v}_1 + s_2 \boldsymbol{v}_2 + s_3 \boldsymbol{v}_3 + \dots + s_k \boldsymbol{v}_k$



イロト イヨト イヨト イヨト

Proof

Then, when computing the solution to the homogeneous equation

• We solve for the first *j* (leading) variables in terms of the remaining *k* free variables:

 s_1,s_2,s_3,\ldots,s_k

l hen

Then the general solution to the homogeneous equation are:

 $s_1 \boldsymbol{v}_1 + s_2 \boldsymbol{v}_2 + s_3 \boldsymbol{v}_3 + \dots + s_k \boldsymbol{v}_k$



イロト イヨト イヨト イヨト

Proof

Then, when computing the solution to the homogeneous equation

• We solve for the first *j* (leading) variables in terms of the remaining *k* free variables:

 s_1,s_2,s_3,\ldots,s_k

Then

• Then the general solution to the homogeneous equation are:

 $s_1 \boldsymbol{v}_1 + s_2 \boldsymbol{v}_2 + s_3 \boldsymbol{v}_3 + \dots + s_k \boldsymbol{v}_k$



Where

The vectors are the Canonical Ones

• Here, a trick!!!

Meaning in $oldsymbol{v}_1$, we have 1, after many 0

• It appears at position j + 1, with zeros afterwards, and so on.

Fherefore the vectors are linearly independents

 $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_k$



Where

The vectors are the Canonical Ones

• Here, a trick!!!

Meaning in $oldsymbol{v}_1$, we have 1, after many 0

• It appears at position j + 1, with zeros afterwards, and so on.

Therefore the vectors are linearly independents

 $oldsymbol{v}_1,oldsymbol{v}_2,oldsymbol{v}_3,\cdots,oldsymbol{v}_k$



Where

The vectors are the Canonical Ones

Here, a trick!!!

Meaning in v_1 , we have 1, after many 0

• It appears at position j + 1, with zeros afterwards, and so on.

Therefore the vectors are linearly independents

 $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \cdots, \boldsymbol{v}_k$



Therefore

They are a basis for the null space of A

And there are k of them, the same as the number of free variables.



Now

Definition

The matrix B is said to be row equivalent to A ($B \sim A$) if B can be obtained from A by a finite sequence of elementary row operations.

In matrix terms

 $B \sim A \Leftrightarrow$ There exist elementary matrices such that

 $B = E_k E_{k-1} E_{k-1} \cdots E_1 A$

If we write $C = E_k E_{k-1} E_{k-1} \cdots E_1$

B is row equivalent to A if B = CA with C invertible.



Now

Definition

The matrix B is said to be row equivalent to A ($B \sim A$) if B can be obtained from A by a finite sequence of elementary row operations.

In matrix terms

 $B \sim A \Leftrightarrow \mathsf{There}\ \mathsf{exist}\ \mathsf{elementary}\ \mathsf{matrices}\ \mathsf{such}\ \mathsf{that}$

$$B = E_k E_{k-1} E_{k-1} \cdots E_1 A$$

B is row equivalent to A if B = CA with C invertible



A D > A D > A D > A D >

Now

Definition

The matrix B is said to be row equivalent to A ($B \sim A$) if B can be obtained from A by a finite sequence of elementary row operations.

In matrix terms

 $B \sim A \Leftrightarrow$ There exist elementary matrices such that

$$B = E_k E_{k-1} E_{k-1} \cdots E_1 A$$

If we write $C = E_k E_{k-1} E_{k-1} \cdots E_1$

B is row equivalent to A if B = CA with C invertible.



Then, we have

Theorem

If $B \sim A$, then Null(B) = Null(A).

[heorem]

If $B \sim A$, then the row space of B is identical to that of A.

Summarizing

Row operations change neither the row space nor the null space of A.



Then, we have

Theorem

If
$$B \sim A$$
, then $Null(B) = Null(A)$.

Theorem

If $B \sim A$, then the row space of B is identical to that of A.

Summarizing

Row operations change neither the row space nor the null space of A.



Then, we have

Theorem

If
$$B \sim A$$
, then $Null(B) = Null(A)$.

Theorem

If $B \sim A$, then the row space of B is identical to that of A.

Summarizing

Row operations change neither the row space nor the null space of A.



イロト イヨト イヨト

Corollaries

Corollary 1

• If R is the Gauss-Jordan form of A, then R has the same null space and row space as A.

• If $B \sim A$, then R(B) = R(A), and N(B) = N(A).



イロト イヨト イヨト イヨト

Corollaries

Corollary 1

• If R is the Gauss-Jordan form of A, then R has the same null space and row space as A.

Corollary 2

• If
$$B \sim A$$
, then $R(B) = R(A)$, and $N(B) = N(A)$.



Then

Theorem

• The number of linearly independent rows of the matrix A is equal to the number of linearly independent columns of A.

 In particular, the rank of A is also equal to the number of linearly independent columns, and hence to the dimension of the column space of A

Therefore

 The number of linearly independent columns of A is then just the number of leading entries in the Gauss-Jordan form of A which is, as we know, the same as the rank of A.



Then

Theorem

• The number of linearly independent rows of the matrix A is equal to the number of linearly independent columns of A.

Thus

• In particular, the rank of A is also equal to the number of linearly independent columns, and hence to the dimension of the column space of A

Therefore

The number of linearly independent columns of A is then just the number of leading entries in the Gauss-Jordan form of A which is, as we know, the same as the rank of A.

38 / 140

Then

Theorem

• The number of linearly independent rows of the matrix A is equal to the number of linearly independent columns of A.

Thus

• In particular, the rank of A is also equal to the number of linearly independent columns, and hence to the dimension of the column space of A

Therefore

• The number of linearly independent columns of A is then just the number of leading entries in the Gauss-Jordan form of A which is, as we know, the same as the rank of A.



イロト イヨト イヨト イヨー

Proof of the theorem (Dimension Theorem)

First

• The rank of A is the same as the rank of the Gauss-Jordan form of A which is equal to the number of leading entries in the Gauss-Jordan form.

Additionally

• The dimension of the null space is equal to the number of free variables in the reduced echelon (Gauss-Jordan) form of A.

l hen

We know further that the number of free variables plus the number of leading entries is exactly the number of columns.



< ロ > < 同 > < 回 > < 回 >

Proof of the theorem (Dimension Theorem)

First

• The rank of A is the same as the rank of the Gauss-Jordan form of A which is equal to the number of leading entries in the Gauss-Jordan form.

Additionally

• The dimension of the null space is equal to the number of free variables in the reduced echelon (Gauss-Jordan) form of A.

We know further that the number of free variables plus the number of leading entries is exactly the number of columns.



< ロ > < 同 > < 回 > < 回 >

Proof of the theorem (Dimension Theorem)

First

• The rank of A is the same as the rank of the Gauss-Jordan form of A which is equal to the number of leading entries in the Gauss-Jordan form.

Additionally

• The dimension of the null space is equal to the number of free variables in the reduced echelon (Gauss-Jordan) form of A.

Then

We know further that the number of free variables plus the number of leading entries is exactly the number of columns.





We have

$dim\left(domain\left(f\right)\right)=dim\left(Range\left(f\right)\right)+dim\left(Ker\left(f\right)\right)$



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations Introduction

- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





イロト イヨト イヨト

As we know

Many Times

We want to obtain a maximum or a minimum of a cost function expressed in terms of matrices....

We need then to define matrix derivatives

Thus, this discussion is useful in Machine Learning.



As we know

Many Times

We want to obtain a maximum or a minimum of a cost function expressed in terms of matrices....

We need then to define matrix derivatives

Thus, this discussion is useful in Machine Learning.



Basic Definition

Let $\psi(\boldsymbol{x}) = \boldsymbol{y}$

Where \boldsymbol{y} is an *m*-element vector, and \boldsymbol{x} is an *n*-element vector

Then, we define the derivative with respect to a vector





イロト イロト イヨト イヨト

Basic Definition

Let $\psi(\boldsymbol{x}) = \boldsymbol{y}$

Where \boldsymbol{y} is an *m*-element vector, and \boldsymbol{x} is an *n*-element vector

Then, we define the derivative with respect to a vector

$$rac{\partial oldsymbol{y}}{\partial oldsymbol{x}} = egin{bmatrix} rac{\partial y_1}{\partial x_1} & rac{\partial y_1}{\partial x_2} & \cdots & rac{\partial y_1}{\partial x_n} \ rac{\partial y_2}{\partial x_1} & rac{\partial y_2}{\partial x_2} & \cdots & rac{\partial y_2}{\partial x_n} \ dots & dots & \ddots & dots \ rac{\partial y_m}{\partial x_1} & rac{\partial y_m}{\partial x_2} & \cdots & rac{\partial y_m}{\partial x_n} \end{bmatrix}$$



イロト イロト イヨト イヨト

What is this

The Matrix denotes the $m\times n$ matrix of first order partial derivatives

• Such a matrix is called the Jacobian matrix of the transformation $\psi\left(\pmb{x}
ight).$

Then, we can get our first ideas on derivatives

For Linear Transformations.



What is this

The Matrix denotes the $m\times n$ matrix of first order partial derivatives

• Such a matrix is called the Jacobian matrix of the transformation $\psi\left(\pmb{x}
ight).$

Then, we can get our first ideas on derivatives

• For Linear Transformations.



イロト イヨト イヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

Introduction

Derivative of a Linear Transformation

Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

Derivative of $\boldsymbol{y} = A\boldsymbol{x}$

Theorem

• Let y = Ax where y is a $m \times 1$, x is a $n \times 1$, A is a $m \times n$ and A does not depend on x, then

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = A$$

Each i^{th} element of \boldsymbol{y} is given by

$$y_i = \sum_{k=1}^N a_{ik} x_k$$

We have that

$$\frac{\partial y_i}{\partial x_j} = a_{ij}$$

for all i=1,2,...,m and j=1,2,...,n

Hence

Cinvestav < ロ > < 合 > < き > < き > き う へ (~ 47 / 140

Each i^{th} element of y is given by

$$y_i = \sum_{k=1}^N a_{ik} x_k$$

We have that

$$\frac{\partial y_i}{\partial x_j} = a_{ij}$$

for all i = 1, 2, ..., m and j = 1, 2, ..., n

Hence



Each i^{th} element of \boldsymbol{y} is given by

$$y_i = \sum_{k=1}^N a_{ik} x_k$$

We have that

$$\frac{\partial y_i}{\partial x_j} = a_{ij}$$

for all i=1,2,...,m and j=1,2,...,n

Hence

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = A$$

Cinvestav ・ロト ・ (アト・ミト・ミト・ミー シへへ 47/140

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regressi

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

Derivative of $y^T A x$

Theorem

 $\bullet\,$ Let the scalar α be defined by

$$\alpha = \boldsymbol{y}^T A \boldsymbol{x}$$

where

 $m{y}$ is a $m imes 1,\,m{x}$ is a $n imes 1,\,A$ is a m imes n and A does not depend on $m{x}$ and $m{y},$ then

$$rac{\partial lpha}{\partial oldsymbol{x}} = oldsymbol{y}^T A$$
 and $rac{\partial lpha}{\partial oldsymbol{y}} = oldsymbol{x}^T A^T$



イロト イヨト イヨト イヨト

Derivative of $y^T A x$

Theorem

• Let the scalar α be defined by

$$\alpha = \boldsymbol{y}^T A \boldsymbol{x}$$

where

 ${\bm y}$ is a $m\times 1,\,{\bm x}$ is a $n\times 1,\,A$ is a $m\times n$ and A does not depend on ${\bm x}$ and ${\bm y},$ then

$$rac{\partial lpha}{\partial oldsymbol{x}} = oldsymbol{y}^T A ext{ and } rac{\partial lpha}{\partial oldsymbol{y}} = oldsymbol{x}^T A^T$$



イロト イヨト イヨト

Define

$$\boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

Note

$$\alpha = \boldsymbol{w}^T \boldsymbol{x}$$

By the previous theorem

$$\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

In a similar way, you can prove the other statement.



Define

$$\boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

Note

$$\boldsymbol{\alpha} = \boldsymbol{w}^T \boldsymbol{x}$$

By the previous theorem

$$\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

In a similar way, you can prove the other statement.



Define

$$\boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

Note

$$\alpha = \boldsymbol{w}^T \boldsymbol{x}$$

By the previous theorem

$$\frac{\partial \alpha}{\partial \boldsymbol{x}} = \boldsymbol{w}^T = \boldsymbol{y}^T \boldsymbol{A}$$

In a similar way, you can prove the other statement.



・ロト ・ 日 ト ・ 日 ト ・ 日 ト ・

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

The Simplest Functions

- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

Introduction





イロト 不得 トイヨト イヨト

What is it?

First than anything, we have a parametric model!!!

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

Note: $\boldsymbol{w}^T \boldsymbol{x}$ is also know as dot product

In the case of ${\mathbb R}$

We have:

$$g\left(oldsymbol{x}
ight)=\left(w_{1},w_{2}
ight)\left(egin{array}{c} x_{1}\ x_{2} \end{array}
ight)+w_{0}=w_{1}x_{1}+w_{2}x_{2}+w_{0}$$



< ロ > < 回 > < 回 > < 回 > < 回 >

(1)

What is it?

First than anything, we have a parametric model!!!

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0$$

Note: $\boldsymbol{w}^T \boldsymbol{x}$ is also know as dot product

In the case of \mathbb{R}^2

We have:

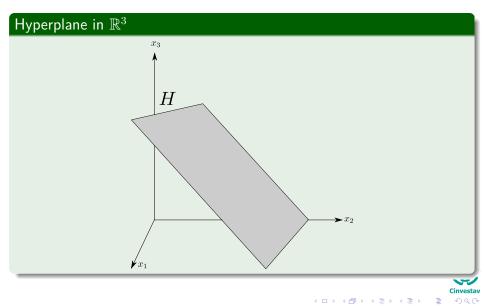
$$g(\mathbf{x}) = (w_1, w_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + w_0 = w_1 x_1 + w_2 x_2 + w_0$$
(2)



イロト イヨト イヨト

(1)

Example



Outline

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

Linear Regression

The Simplest Functions

Splitting the Space

- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
- The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Introduction

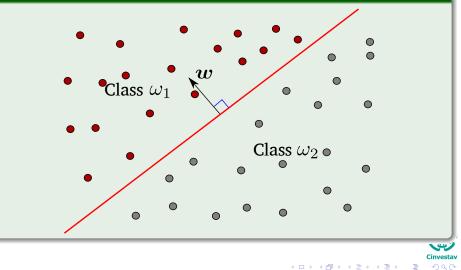




イロト 不得 トイヨト イヨト

Splitting The Space \mathbb{R}^2

Using a simple straight line (Hyperplane)



55/140

Splitting the Space?

For example, assume the following vector $oldsymbol{w}$ and constant w_0

$$w = (-1,2)^T$$
 and $w_0 = 0$

Hyperplane

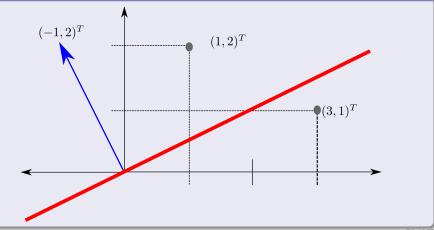


Splitting the Space?

For example, assume the following vector \boldsymbol{w} and constant w_0

$$oldsymbol{w}=(-1,2)^T$$
 and $w_0=0$

Hyperplane



Then, we have

The following results

$$g\left(\left(\begin{array}{c}1\\2\end{array}\right)\right) = (-1,2)\left(\begin{array}{c}1\\2\end{array}\right) = -1 \times 1 + 2 \times 2 = 3$$
$$g\left(\left(\begin{array}{c}3\\1\end{array}\right)\right) = (-1,2)\left(\begin{array}{c}3\\1\end{array}\right) = -1 \times 3 + 2 \times 1 = -1$$

YES!!! We have a positive side and a negative side!!!



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space

Defining the Decision Surface

- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
- The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

Introduction





イロト 不得 トイヨト イヨト

The Decision Surface

The equation g(x) = 0 defines a decision surface

Separating the elements in classes, ω_1 and ω_2 .

When g(x) is linear the decision surface is an hyperplane

Now assume $oldsymbol{x}_1$ and $oldsymbol{x}_2$ are both on the decision surface

 $w^T x_1 + w_0 = 0$ $w^T x_2 + w_0 = 0$

Thus

$$oldsymbol{w}^Toldsymbol{x}_1+w_0=oldsymbol{w}^Toldsymbol{x}_2+w_0$$

Cinvestav ≧ ∽ < 59 / 140

A D > A D > A D > A D >

The Decision Surface

The equation g(x) = 0 defines a decision surface

Separating the elements in classes, ω_1 and ω_2 .

When g(x) is linear the decision surface is an hyperplane

Now assume x_1 and x_2 are both on the decision surface

 $\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0$ $\boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0$

Thus

$$oldsymbol{w}^Toldsymbol{x}_1+w_0=oldsymbol{w}^Toldsymbol{x}_2+w_0$$

Cinvestav ∽ < (~ 59 / 140

The Decision Surface

The equation g(x) = 0 defines a decision surface

Separating the elements in classes, ω_1 and ω_2 .

When g(x) is linear the decision surface is an hyperplane

Now assume \boldsymbol{x}_1 and \boldsymbol{x}_2 are both on the decision surface

 $\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = 0$ $\boldsymbol{w}^T \boldsymbol{x}_2 + w_0 = 0$

Thus

$$w^T x_1 + w_0 = w^T x_2 + w_0$$

Cinvestav つ へ へ 59 / 140

A D > A D > A D > A D >

(3)

Defining a Decision Surface

Then

$$\boldsymbol{w}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2 \right) = 0$$

A B +
 A B +
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

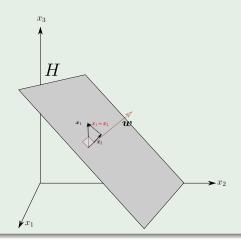
(4)

10

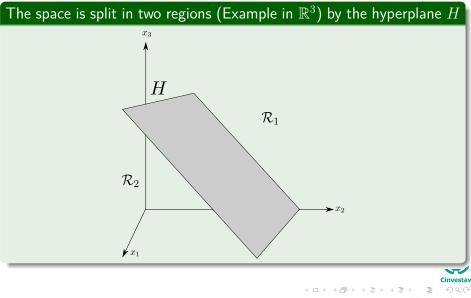
Therefore

$oldsymbol{x}_1 - oldsymbol{x}_2$ lives in the hyperplane i.e. it is perpendicular to $oldsymbol{w}^T$

- Remark: any vector in the hyperplane is a linear combination of elements in a basis
- Therefore any vector in the plane is perpendicular to w^T



Therefore



62/140

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface

• Properties of the Hyperplane $oldsymbol{w}^Toldsymbol{x}+w_0$

- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

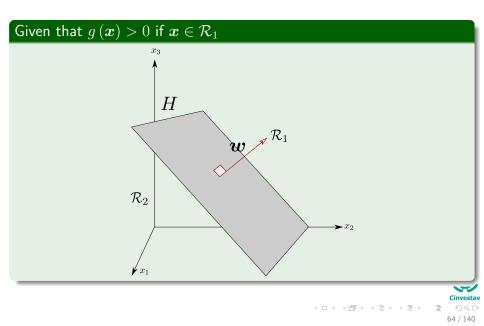
Introduction





イロト 不得 トイヨト イヨト

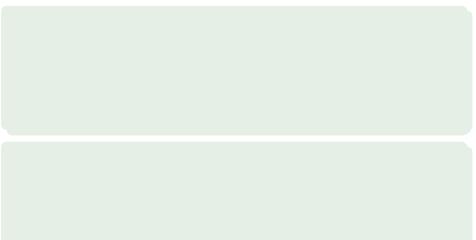
Some Properties of the Hyperplane



We can say the following

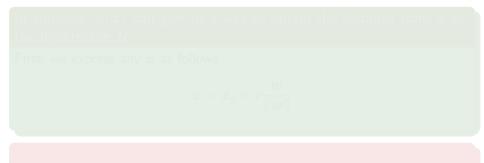
• Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.

Any $oldsymbol{x} \in \mathcal{R}_2$ is on the negative side of I



We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.



We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, g(x) can give us a way to obtain the distance from x to the hyperplane H

First, we express any $m{x}$ as follows

$$\boldsymbol{x} = \boldsymbol{x}_p + r \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

 $ullet \, oldsymbol{x}_p$ is the normal projection of $oldsymbol{x}$ onto H .

- r is the desired distance
 - Positive, if x is in the positive side
 - Negative, if x is in the negative side

We can say the following

- Any $x \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, g(x) can give us a way to obtain the distance from x to the hyperplane H

First, we express any $m{x}$ as follows

$$oldsymbol{x} = oldsymbol{x}_p + r rac{oldsymbol{w}}{\|oldsymbol{w}\|}$$

Where

- x_p is the normal projection of x onto H.
 - Positive, if x is in the positive side
 - \blacktriangleright Negative, if x is in the negative side

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, g(x) can give us a way to obtain the distance from x to the hyperplane H

First, we express any $m{x}$ as follows

$$oldsymbol{x} = oldsymbol{x}_p + r rac{oldsymbol{w}}{\|oldsymbol{w}\|}$$

Where

- x_p is the normal projection of x onto H.
- r is the desired distance

legative, if x is in the positive side

We can say the following

- Any $\boldsymbol{x} \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, g(x) can give us a way to obtain the distance from x to the hyperplane H

First, we express any $m{x}$ as follows

$$oldsymbol{x} = oldsymbol{x}_p + r rac{oldsymbol{w}}{\|oldsymbol{w}\|}$$

Where

- x_p is the normal projection of x onto H.
- r is the desired distance
 - \blacktriangleright Positive, if x is in the positive side

We can say the following

- Any $x \in \mathcal{R}_1$ is on the positive side of H.
- Any $\boldsymbol{x} \in \mathcal{R}_2$ is on the negative side of H.

In addition, g(x) can give us a way to obtain the distance from x to the hyperplane H

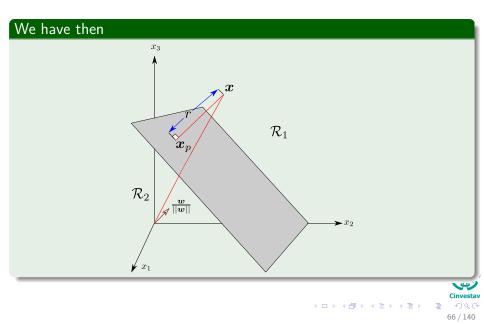
First, we express any $m{x}$ as follows

$$oldsymbol{x} = oldsymbol{x}_p + r rac{oldsymbol{w}}{\|oldsymbol{w}\|}$$

Where

- x_p is the normal projection of x onto H.
- r is the desired distance
 - Positive, if x is in the positive side
 - Negative, if x is in the negative side

We have something like this



Since $g\left(\boldsymbol{x_{p}}\right)=0$

We have that

$$g\left(\boldsymbol{x}\right) = g\left(\boldsymbol{x}_{p} + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$

$$= g\left(\boldsymbol{x}_{p} + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$

$$= g\left(\boldsymbol{x}_{p}\right) + r\frac{\|\boldsymbol{w}\|}{\|\boldsymbol{w}\|}$$

67/140

Since $g\left(\boldsymbol{x_{p}}\right)=0$

We have that

$$g(\boldsymbol{x}) = g\left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$
$$= \boldsymbol{w}^T \left(\boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_0$$



Since $g\left(\boldsymbol{x_{p}}\right)=0$

We have that

$$g(\boldsymbol{x}) = g\left(\boldsymbol{x}_{p} + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$$
$$= \boldsymbol{w}^{T}\left(\boldsymbol{x}_{p} + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + w_{0}$$
$$= \boldsymbol{w}^{T}\boldsymbol{x}_{p} + w_{0} + r\frac{\boldsymbol{w}^{T}\boldsymbol{w}}{\|\boldsymbol{w}\|}$$



Since $g\left(\boldsymbol{x_{p}}\right)=0$

g

We have that

$$\begin{aligned} \mathbf{x}(\mathbf{x}) &= g\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \\ &= \mathbf{w}^T \left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= \mathbf{w}^T \mathbf{x}_p + w_0 + r\frac{\mathbf{w}^T \mathbf{w}}{\|\mathbf{w}\|} \\ &= g\left(\mathbf{x}_p\right) + r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \end{aligned}$$

Then, we have $r = \frac{g(x)}{\|w\|}$ (5)

Since $g(\boldsymbol{x_p}) = 0$

g

We have that

$$\begin{aligned} \mathbf{w}(\mathbf{x}) &= g\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \\ &= \mathbf{w}^T \left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= \mathbf{w}^T \mathbf{x}_p + w_0 + r\frac{\mathbf{w}^T \mathbf{w}}{\|\mathbf{w}\|} \\ &= g\left(\mathbf{x}_p\right) + r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \\ &= r \|\mathbf{w}\| \end{aligned}$$

Then, we have

67/140

Since $g(\boldsymbol{x_p}) = 0$

9

We have that

$$\begin{aligned} \mathbf{v}(\mathbf{x}) &= g\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) \\ &= \mathbf{w}^T \left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_0 \\ &= \mathbf{w}^T \mathbf{x}_p + w_0 + r\frac{\mathbf{w}^T \mathbf{w}}{\|\mathbf{w}\|} \\ &= g\left(\mathbf{x}_p\right) + r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \\ &= r \|\mathbf{w}\| \end{aligned}$$

Then, we have

$$r = rac{g\left(oldsymbol{x}
ight)}{\|oldsymbol{w}\|}$$

67/140

(5)

The distance from the origin to ${\cal H}$

$$r = \frac{g\left(\mathbf{0}\right)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^{T}\left(\mathbf{0}\right) + w_{0}}{\|\boldsymbol{w}\|} = \frac{w_{0}}{\|\boldsymbol{w}\|}$$
(6)





2

ヘロト ヘロト ヘヨト ヘヨト

The distance from the origin to ${\cal H}$

$$r = \frac{g\left(\mathbf{0}\right)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^{T}\left(\mathbf{0}\right) + w_{0}}{\|\boldsymbol{w}\|} = \frac{w_{0}}{\|\boldsymbol{w}\|}$$
(6)

Remarks

• If $w_0 > 0$, the origin is on the positive side of H.

• If $w_0 = 0$, the hyperplane has the homogeneous form $w^T x$ and hyperplane passes through the origin.



< ロ > < 同 > < 回 > < 回 >

The distance from the origin to ${\cal H}$

$$r = \frac{g(\mathbf{0})}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T(\mathbf{0}) + w_0}{\|\mathbf{w}\|} = \frac{w_0}{\|\mathbf{w}\|}$$
(6)

Remarks

- If $w_0 > 0$, the origin is on the positive side of H.
- If $w_0 < 0$, the origin is on the negative side of H.



< ロ > < 同 > < 回 > < 回 >

The distance from the origin to ${\cal H}$

$$r = \frac{g\left(\mathbf{0}\right)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^{T}\left(\mathbf{0}\right) + w_{0}}{\|\boldsymbol{w}\|} = \frac{w_{0}}{\|\boldsymbol{w}\|}$$
(6)

Remarks

- If $w_0 > 0$, the origin is on the positive side of H.
- If $w_0 < 0$, the origin is on the negative side of H.
- If $w_0 = 0$, the hyperplane has the homogeneous form $w^T x$ and hyperplane passes through the origin.



< ロ > < 同 > < 回 > < 回 >

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$

Augmenting the Vector

- Least Squared Error Procedure
 The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

Introduction





イロト 不得 トイヨト イヨト

We would like w_0 as part of the dot product by making $x_0 = 1$

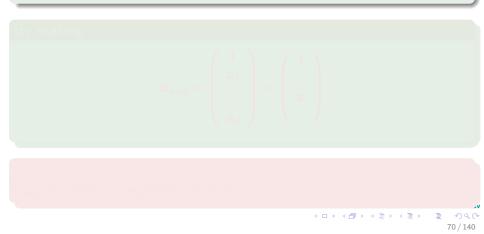
$$g\left(\boldsymbol{x}\right) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = \min\left(\sum_{i=1}^{d} w_i x_i\right) = \min\left(\sum_{i=1}^{d} w_i x_i\right)$$

イロト イヨト イヨト イヨト

70/140

We would like w_0 as part of the dot product by making $x_0 = 1$

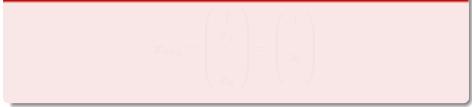
$$g(\mathbf{x}) = w_0 \times 1 + \sum_{i=1}^{d} w_i x_i = w_0 \times x_0 + \sum_{i=1}^{d} w_i x_i = \sum_{i=1}^{d$$



We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\mathbf{x}) = w_0 \times 1 + \sum_{i=1}^d w_i x_i = w_0 \times x_0 + \sum_{i=1}^d w_i x_i = \sum_{i=0}^d w_i x_i$$
(7)

By making



A D > A D > A D > A D >

70/140

Where

 x_{aug} is called an augmented feature vector

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\mathbf{x}) = w_0 \times 1 + \sum_{i=1}^d w_i x_i = w_0 \times x_0 + \sum_{i=1}^d w_i x_i = \sum_{i=0}^d w_i x_i$$
 (7)

By making

$$oldsymbol{x}_{aug} = egin{pmatrix} 1 \ x_1 \ dots \ x_d \end{pmatrix} = egin{pmatrix} 1 \ x \ x \end{pmatrix}$$

Where

 x_{aug} is called an augmented feature vector.

We would like w_0 as part of the dot product by making $x_0 = 1$

$$g(\mathbf{x}) = w_0 \times 1 + \sum_{i=1}^d w_i x_i = w_0 \times x_0 + \sum_{i=1}^d w_i x_i = \sum_{i=0}^d w_i x_i$$
 (7)

By making

$$oldsymbol{x}_{aug} = egin{pmatrix} 1 \ x_1 \ dots \ x_d \end{pmatrix} = egin{pmatrix} 1 \ x \ x \end{pmatrix}$$

イロト イヨト イヨト

70/140

Where

 x_{aug} is called an augmented feature vector.

In a similar way

We have the augmented weight vector

<ロ> (日本) (日本) (日本)

In a similar way

We have the augmented weight vector

$$oldsymbol{w}_{aug} = \left(egin{array}{c} w_0 \ w_1 \ dots \ w_d \end{array}
ight) = \left(egin{array}{c} w_0 \ w \ w \end{array}
ight)$$

Remarks

• The addition of a constant component to \boldsymbol{x} preserves all the distance relationship between samples.



イロト イヨト イヨト

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \left(egin{array}{c} w_0 \ w_1 \ dots \ w_d \end{array}
ight) = \left(egin{array}{c} w_0 \ w \ w \end{array}
ight)$$

Remarks

- The addition of a constant component to \boldsymbol{x} preserves all the distance relationship between samples.
- The resulting x_{aug} vectors, all lie in a *d*-dimensional subspace which is the *x*-space itself.

イロト イヨト イヨト

71/140

More Remarks

In addition

The hyperplane decision surface \widehat{H} defined by

$$\boldsymbol{w}_{aug}^T \boldsymbol{x}_{aug} = 0$$

passes through the origin in x_{aug} -space.

Even Though

The corresponding hyperplane H can be in any position of the x-space.



More Remarks

In addition

The hyperplane decision surface \widehat{H} defined by

$$\boldsymbol{w}_{aug}^T \boldsymbol{x}_{aug} = 0$$

passes through the origin in x_{aug} -space.

Even Though

The corresponding hyperplane H can be in any position of the x-space.



イロト イボト イヨト イヨト

More Remarks

In addition

The distance from ${\pmb y}$ to \widehat{H} is:

$$\frac{\boldsymbol{w}_{aug}^{T}\boldsymbol{x}_{aug}\Big|}{\left\|\boldsymbol{w}_{aug}\right\|} = \frac{\left|g\left(\boldsymbol{x}_{aug}\right)\right|}{\left\|\boldsymbol{w}_{aug}\right\|}$$

・ロト ・日下・ ・ ヨト・

Now

Is $\|w\| \le \|w_{aug}\|$ • Ideas? $\sqrt{\sum_{i=1}^d w_i^2} \le \sqrt{\sum_{i=1}^d w_i^2 + w_0^2}$

This mapping is quite useful

Because we only need to find a weight vector $oldsymbol{w}_{aug}$ instead of finding the weight vector $oldsymbol{w}$ and the threshold $w_0.$



Now

Is $\|w\| \le \|w_{aug}\|$ • Ideas? $\sqrt{\sum_{i=1}^d w_i^2} \le \sqrt{\sum_{i=1}^d w_i^2 + w_0^2}$

This mapping is quite useful

Because we only need to find a weight vector w_{aug} instead of finding the weight vector w and the threshold w_0 .



イロト イヨト イヨト イヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector

Least Squared Error Procedure

- The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

Introduction





イロト イヨト イヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector

Least Squared Error Procedure The Geometry of a Two-Category Linearly-Separable Case

- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





Suppose, we have

n samples $x_1, x_2, ..., x_n$ some labeled ω_1 and some labeled ω_2 .



Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2 .

We want a vector weight ${\boldsymbol w}$ such that

•
$$oldsymbol{w}^Toldsymbol{x}_i > 0$$
, if $oldsymbol{x}_i \in \omega_1$.

I he name of this weight vector

It is called a separating vector or solution vector.



イロト イヨト イヨト イヨト

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2 .

We want a vector weight ${\boldsymbol w}$ such that

•
$$oldsymbol{w}^Toldsymbol{x}_i > 0$$
, if $oldsymbol{x}_i \in \omega_1$

•
$$oldsymbol{w}^Toldsymbol{x}_i < 0$$
, if $oldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.



イロト イヨト イヨト イヨト

Suppose, we have

n samples $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$ some labeled ω_1 and some labeled ω_2 .

We want a vector weight ${oldsymbol w}$ such that

•
$$oldsymbol{w}^Toldsymbol{x}_i > 0$$
, if $oldsymbol{x}_i \in \omega_1$

•
$$oldsymbol{w}^Toldsymbol{x}_i < 0$$
, if $oldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.



Imagine that your problem has two classes ω_1 and ω_2 in \mathbb{R}^2

- They are linearly separable!!!
- You require to label them.



Imagine that your problem has two classes ω_1 and ω_2 in \mathbb{R}^2

- They are linearly separable!!!
- ② You require to label them.

We have a problem!!!

Which is the problem?



Imagine that your problem has two classes ω_1 and ω_2 in \mathbb{R}^2

- They are linearly separable!!!
- 2 You require to label them.

We have a problem!!!

Which is the problem?

Thus, what distance each point has to the hyperplane



< ロ > < 同 > < 回 > < 回 >

Imagine that your problem has two classes ω_1 and ω_2 in \mathbb{R}^2

- They are linearly separable!!!
- 2 You require to label them.

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?



イロト イヨト イヨト イヨト

Label the Classes • $\omega_1 \Longrightarrow +1$ • $\omega_2 \Longrightarrow -1$



Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

1 if
$$\boldsymbol{x} \in \omega_1$$
 then $y_{ideal} = g_{ideal} (\boldsymbol{x}) = +1$.

(investav cinvestav イロト イクト イミト イミト 王 今 Q (~ 79/140

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

• if
$$x \in \omega_1$$
 then $y_{ideal} = g_{ideal}(x) = +1$.

2) if
$$x \in \omega_2$$
 then $y_{ideal} = g_{ideal} (x) = -1$.

Label the Classes

- $\omega_1 \Longrightarrow +1$
- $\omega_2 \Longrightarrow -1$

We produce the following labels

() if
$$oldsymbol{x} \in \omega_1$$
 then $y_{ideal} = g_{ideal}\left(oldsymbol{x}
ight) = +1.$

2) if
$$x \in \omega_2$$
 then $y_{ideal} = g_{ideal} (x) = -1$.

Remark: We have a problem with this labels!!!



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 The Geometry of a Two-Category Linearly-Separable Case

The Error Idea

The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Singular Value Decomposition

Introduction





イロト 不得 トイヨト イヨト

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise} \left(\boldsymbol{x} \right) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e$$

Where the ϵ

It has a $e \sim N\left(\mu, \sigma^2
ight)$

Thus, we can do the following

$$y_{noise} = g_{noise}\left(x
ight) = g_{ideal}\left(x
ight) + e$$



イロト イヨト イヨト イヨト

(8)

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e$$

Where the e

It has a $e \sim N\left(\mu, \sigma^2\right)$

Thus, we can do the following

$$y_{noise} = g_{noise}\left(oldsymbol{x}
ight) = g_{ideal}\left(oldsymbol{x}
ight) + e$$

イロト イロト イヨト イヨト



(8)

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise} \left(\boldsymbol{x} \right) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e$$

Where the e

It has a $e \sim N\left(\mu, \sigma^2\right)$

Thus, we can do the following

$$y_{noise} = g_{noise}\left(\boldsymbol{x}\right) = g_{ideal}\left(\boldsymbol{x}\right) + e$$
 (9)



(8)

Thus, we have

What to do?

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right)$$

Graphically

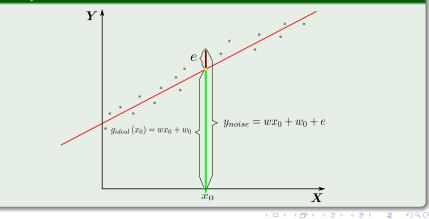


Thus, we have

What to do?

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{10}$$

Graphically



82/140

Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal}\left(oldsymbol{x}
ight)$$

We will see

How the geometry will solve the problem with using these labels.



Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal} \left(\boldsymbol{x} \right) \tag{12}$$

We will see

How the geometry will solve the problem with using these labels.



Then, we have

A TRICK... Quite a good one!!! Instead of using y_{noise}

$$e = y_{noise} - g_{ideal}\left(\boldsymbol{x}\right) \tag{11}$$

We use y_{ideal}

$$e = y_{ideal} - g_{ideal} \left(\boldsymbol{x} \right) \tag{12}$$

We will see

How the geometry will solve the problem with using these labels.



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

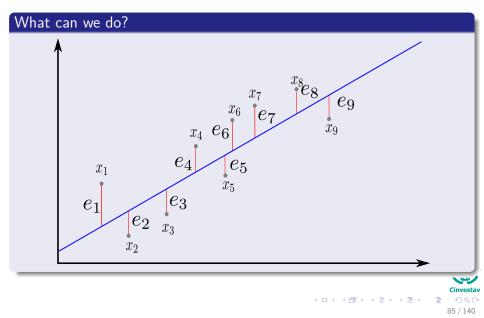
Introduction





イロト 不得 トイヨト イヨト

Here, we have multiple errors



Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}_i))^2$$
(13)

Remark: This is know as the Least Squared Error cost function



< ロ > < 回 > < 回 > < 回 > < 回 >

Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}_i))^2$$
(13)

Remark: This is know as the Least Squared Error cost function

Generalizing

• The dimensionality of each sample (data point) is *d*.



Sum Over All the Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}_i))^2$$
(13)

Remark: This is know as the Least Squared Error cost function

Generalizing

- The dimensionality of each sample (data point) is d.
- You can extend each vector sample to be ${m x}^T=({f 1},{m x}').$



イロト イロト イヨト イヨト

We can use a trick

The following function

We can rewrite the error equation as

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}_i))^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T \boldsymbol{w})^2$$
(14)



イロト イロト イヨト イヨト

We can use a trick

The following function

We can rewrite the error equation as

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}_i))^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{x}_i^T \boldsymbol{w})^2$$
(14)



э

イロト イヨト イヨト イヨト

Furthermore

Then stacking all the possible estimations into the product Data Matrix and weight vector

	$\begin{pmatrix} 1 \end{pmatrix}$	$(\boldsymbol{x}_1)_1$	 $(oldsymbol{x}_1)_j$	•••	$(\boldsymbol{x}_1)_d$)	(w_1)
	1 :		÷		÷	$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$
Xw =	1	$(oldsymbol{x}_i)_1$	$(oldsymbol{x}_i)_j$		$(oldsymbol{x}_i)_d$	w_3
	1 :		÷		:	
	$\left(1 \right)$	$(oldsymbol{x}_N)_1$	 $(oldsymbol{x}_N)_j$		$\left(oldsymbol{x}_{N} ight) _{d}$	$\left(w_{d+1} \right)$



< ロ > < 回 > < 回 > < 回 > < 回 >

Note about other representations

We could have $oldsymbol{x}^T = (x_1, x_2,, x_d, 1)$ thus											
X =	$(oldsymbol{x}_i)_1$		$egin{array}{cccc} (oldsymbol{x}_1)_j & \cdots & \ dots & \ \dots &$	$\stackrel{arepsilon}{(oldsymbol{x}_i)_d}$: 1 :	(15)					



Then, we have the following trick with $oldsymbol{X}$

With the Data Matrix

$$\boldsymbol{X}w = \begin{pmatrix} \boldsymbol{x}_{1}^{T}\boldsymbol{w} \\ \boldsymbol{x}_{2}^{T}\boldsymbol{w} \\ \boldsymbol{x}_{3}^{T}\boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_{N}^{T}\boldsymbol{w} \end{pmatrix}$$
(16)



Therefore

We have that

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_4 \end{pmatrix} - \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{w} \\ \boldsymbol{x}_2^T \boldsymbol{w} \\ \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix}$$

Cinvestav

・ロン ・回と ・ヨン ・ヨン

Then, we have the following equality

$$\begin{pmatrix} y_1 - x_1^T w & y_2 - x_2^T w & y_3 - x_3^T w & \cdots & y_4 - x_N^T w \\ \vdots & \vdots & \vdots \\ y_4 - x_N^T w & \end{pmatrix} = \sum_{i=1}^N \begin{pmatrix} y_i - x_i^T w \\ y_3 - x_1^T w \\ \vdots \\ y_4 - x_N^T w \end{pmatrix} = \sum_{i=1}^N \begin{pmatrix} y_i - x_i^T w \\ y_i - x_i^T w \end{pmatrix}^2$$

Therefore

We have that

$$egin{pmatrix} y_1 \ y_2 \ y_3 \ dots \ y_4 \end{pmatrix} = egin{pmatrix} x_1^T oldsymbol{w} \ x_2^T oldsymbol{w} \ x_3^T oldsymbol{w} \ dots \ x_N^T oldsymbol{w} \end{pmatrix} = egin{pmatrix} y_1 - oldsymbol{x}_1^T oldsymbol{w} \ y_2 - oldsymbol{x}_2^T oldsymbol{w} \ y_3 - oldsymbol{x}_3^T oldsymbol{w} \ dots \ y_3 - oldsymbol{x}_3^T oldsymbol{w} \ dots \ do$$

Then, we have the following equality

91/140

Then, we have

The following equality

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w} \right)^2 = \left(\boldsymbol{y} - \boldsymbol{X} \boldsymbol{w} \right)^T \left(\boldsymbol{y} - \boldsymbol{X} \boldsymbol{w} \right) = \| \boldsymbol{y} - \boldsymbol{X} \boldsymbol{w} \|_2^2 \qquad (17)$$



We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}$$
 (18)



We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}$$
 (18)

Now

ullet Derive with respect to w



We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y} + \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w}$$
 (18)

Now

- ullet Derive with respect to w
- Assume that $\boldsymbol{X}^T \boldsymbol{X}$ is invertible



< ロ > < 回 > < 回 > < 回 > < 回 >

Therefore

We have the following equivalences

$$\frac{d\boldsymbol{w}^{T}A\boldsymbol{w}}{d\boldsymbol{w}} = \boldsymbol{w}^{T}\left(A + A^{T}\right), \ \frac{d\boldsymbol{w}^{T}A}{d\boldsymbol{w}} = A^{T}$$
(19)

Now given that the transpose of a number is the number itself

$$oldsymbol{y}^Toldsymbol{X}oldsymbol{w} = egin{bmatrix} oldsymbol{y}^Toldsymbol{X}oldsymbol{w}\end{bmatrix}^T = oldsymbol{w}^Toldsymbol{X}^Toldsymbol{y}$$

イロン イロン イヨン イヨン

Therefore

We have the following equivalences

$$\frac{d\boldsymbol{w}^{T}A\boldsymbol{w}}{d\boldsymbol{w}} = \boldsymbol{w}^{T}\left(A + A^{T}\right), \ \frac{d\boldsymbol{w}^{T}A}{d\boldsymbol{w}} = A^{T}$$
(19)

Now given that the transpose of a number is the number itself

$$\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} = \left[\boldsymbol{y}^T \boldsymbol{X} \boldsymbol{w} \right]^T = \boldsymbol{w}^T \boldsymbol{X}^T \boldsymbol{y}$$



We have then

$$\frac{d\left(\boldsymbol{y}^{T}\boldsymbol{y}-2\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{y}+\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{w}\right)}{d\boldsymbol{w}}=-2\boldsymbol{y}^{T}\boldsymbol{X}+\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}+\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right)$$



We have then

$$\frac{d\left(\boldsymbol{y}^{T}\boldsymbol{y} - 2\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{y} + \boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{w}\right)}{d\boldsymbol{w}} = -2\boldsymbol{y}^{T}\boldsymbol{X} + \boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X} + \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right)$$
$$= -2\boldsymbol{y}^{T}\boldsymbol{X} + 2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)$$

Making this equal to the zero row vector

We have then

$$\frac{d\left(\boldsymbol{y}^{T}\boldsymbol{y} - 2\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{y} + \boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{w}\right)}{d\boldsymbol{w}} = -2\boldsymbol{y}^{T}\boldsymbol{X} + \boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X} + \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right)$$
$$= -2\boldsymbol{y}^{T}\boldsymbol{X} + 2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)$$

Making this equal to the zero row vector

$$-2\boldsymbol{y}^{T}\boldsymbol{X}+2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)=0$$

We apply the transpose

$$egin{bmatrix} \left[-2oldsymbol{y}^Toldsymbol{X}+2oldsymbol{w}^Toldsymbol{X}^Toldsymbol{X}
ight]^T=[0]^T\ -2oldsymbol{X}^Toldsymbol{y}+2ig(oldsymbol{X}^Toldsymbol{X}ig)oldsymbol{w}=0 \end{tabular}$$
 (column vector)

We have then

$$\frac{d\left(\boldsymbol{y}^{T}\boldsymbol{y} - 2\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{y} + \boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{w}\right)}{d\boldsymbol{w}} = -2\boldsymbol{y}^{T}\boldsymbol{X} + \boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X} + \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right)$$
$$= -2\boldsymbol{y}^{T}\boldsymbol{X} + 2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)$$

Making this equal to the zero row vector

$$-2\boldsymbol{y}^{T}\boldsymbol{X}+2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)=0$$

We apply the transpose

$$\left[-2\boldsymbol{y}^{T}\boldsymbol{X}+2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right]^{T}=\left[0\right]^{T}$$

 $\boldsymbol{y}+2\left(\boldsymbol{X}^{T}\boldsymbol{X}
ight)\boldsymbol{w}=0$ (column vector)

We have then

$$\frac{d\left(\boldsymbol{y}^{T}\boldsymbol{y} - 2\boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{y} + \boldsymbol{w}^{T}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{w}\right)}{d\boldsymbol{w}} = -2\boldsymbol{y}^{T}\boldsymbol{X} + \boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X} + \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\right)$$
$$= -2\boldsymbol{y}^{T}\boldsymbol{X} + 2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)$$

Making this equal to the zero row vector

$$-2\boldsymbol{y}^{T}\boldsymbol{X}+2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)=0$$

We apply the transpose

$$\begin{bmatrix} -2\boldsymbol{y}^{T}\boldsymbol{X} + 2\boldsymbol{w}^{T}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right) \end{bmatrix}^{T} = \begin{bmatrix} 0 \end{bmatrix}^{T}$$
$$-2\boldsymbol{X}^{T}\boldsymbol{y} + 2\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)\boldsymbol{w} = 0 \text{ (column vector)}$$

95/146

Solving for \boldsymbol{w}

We have then

$$\boldsymbol{w} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$
(20)

Note: $X^T X$ is always positive semi-definite. If it is also invertible, it is positive definite.

Thus, How we get the discriminant function?

Any Ideas?



Solving for \boldsymbol{w}

We have then

$$\boldsymbol{w} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$
(20)

Note: $X^T X$ is always positive semi-definite. If it is also invertible, it is positive definite.

Thus, How we get the discriminant function?

Any Ideas?



The Final Discriminant Function

Very Simple!!!

$$g(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{w} = \boldsymbol{x}^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$
(21)

メロト メロト メヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

Karhunen-Loeve Transform

- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decompositior

Introduction





イロト イヨト イヨト イヨト

Also Known as Karhunen-Loeve Transform

Setup

• Consider a data set of observations $\{ {m x}_n \}$ with n=1,2,...,N and ${m x}_{m n} \in R^d.$

Goal

Project data onto space with dimensionality m < d (We assume m is given)



Also Known as Karhunen-Loeve Transform

Setup

• Consider a data set of observations $\{ {m x}_n \}$ with n=1,2,...,N and ${m x}_{m n} \in R^d.$

Goal

Project data onto space with dimensionality m < d (We assume m is given)



イロト イヨト イヨト イヨト

Dimensional Variance

Remember the Variance Sample in $\ensuremath{\mathbb{R}}$

$$VAR(X) = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N - 1}$$
(22)





< ロ > < 回 > < 回 > < 回 > < 回 >

Dimensional Variance

Remember the Variance Sample in $\ensuremath{\mathbb{R}}$

$$VAR(X) = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N - 1}$$
(22)

You can do the same in the case of two variables X and Y

$$COV(x,y) = \frac{\sum_{i=1}^{N} (x_i - \overline{x}) (y_i - \overline{y})}{N - 1}$$
(23)

< ロ > < 回 > < 回 > < 回 > < 回 >



Now, Define

Given the data

$$oldsymbol{x}_1, oldsymbol{x}_2, ..., oldsymbol{x}_N$$

where
$$x_i$$
 is a column vector

Construct the sample mean

$$\overline{oldsymbol{x}} = rac{1}{N}\sum_{i=1}^N oldsymbol{x}_i$$

Center data

$$oldsymbol{x}_1-\overline{oldsymbol{x}},oldsymbol{x}_2-\overline{oldsymbol{x}},...,oldsymbol{x}_N-\overline{oldsymbol{x}}$$

Cinvestav Ξ ∽ へ ペ 101 / 140

イロト イヨト イヨト イヨト

(24)

Now, Define

Given the data

$$oldsymbol{x}_1, oldsymbol{x}_2, ..., oldsymbol{x}_N$$

where
$$x_i$$
 is a column vector

Construct the sample mean

$$\overline{oldsymbol{x}} = rac{1}{N}\sum_{i=1}^N oldsymbol{x}_i$$

Center data

$oldsymbol{x}_1-\overline{oldsymbol{x}},oldsymbol{x}_2-\overline{oldsymbol{x}},...,oldsymbol{x}_N-\overline{oldsymbol{x}}$



・ロト ・ 日 ト ・ 日 ト ・ 日 ト ・

(24)

(25)

Now, Define

Given the data

$$oldsymbol{x}_1, oldsymbol{x}_2, ..., oldsymbol{x}_N$$

where
$$x_i$$
 is a column vector

Construct the sample mean

$$\overline{oldsymbol{x}} = rac{1}{N}\sum_{i=1}^N oldsymbol{x}_i$$

Center data

$$oldsymbol{x}_1-\overline{oldsymbol{x}},oldsymbol{x}_2-\overline{oldsymbol{x}},...,oldsymbol{x}_N-\overline{oldsymbol{x}}$$



2

イロト イロト イヨト イヨト

(24)

(25)

Build the Sample Mean

The Covariance Matrix

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$
(27)

Properties

- The ijth value of S is equivalent to σ_{ij}^2
-) The iith value of S is equivalent to $\sigma^2_{ii}.$



Build the Sample Mean

The Covariance Matrix

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$
(27)

Properties

- The *ij*th value of S is equivalent to σ_{ij}^2 .
- 2 The *ii*th value of S is equivalent to σ_{ii}^2 .



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

Karhunen-Loeve Transform

Projecting the Data

- Lagrange Multipliers
- The Process
- Example

5 Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

Using \boldsymbol{S} to Project Data

For this we use a $oldsymbol{u}_1$

ullet with $oldsymbol{u}_1^Toldsymbol{u}_1=1,$ an orthonormal vector

Question

• What is the Sample Variance of the Projected Data?



Using \boldsymbol{S} to Project Data

For this we use a u_1

• with $oldsymbol{u}_1^Toldsymbol{u}_1=1,$ an orthonormal vector

Question

What is the Sample Variance of the Projected Data?



Using \boldsymbol{S} to Project Data

For this we use a $oldsymbol{u}_1$

• with $oldsymbol{u}_1^Toldsymbol{u}_1=1,$ an orthonormal vector

Question

• What is the Sample Variance of the Projected Data?



イロト イヨト イヨト イヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

Karhunen-Loeve Transform
 Projecting the Data

Lagrange Multipliers

- The Process
- Example

5 Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

Thus we have

Variance of the projected data

$$\frac{1}{N-1}\sum_{i=1}^{N} \left[\boldsymbol{u}_{1}\boldsymbol{x}_{i} - \boldsymbol{u}_{1}\overline{\boldsymbol{x}}\right] = \boldsymbol{u}_{1}^{T}S\boldsymbol{u}_{1}$$
(28)

Use Lagrange Multipliers to Maximize

$$\boldsymbol{u}_1^T \boldsymbol{S} \boldsymbol{u}_1 + \lambda_1 \left(1 - \boldsymbol{u}_1^T \boldsymbol{u}_1 \right) \tag{29}$$

イロト イヨト イヨト イヨト



Thus we have

Variance of the projected data

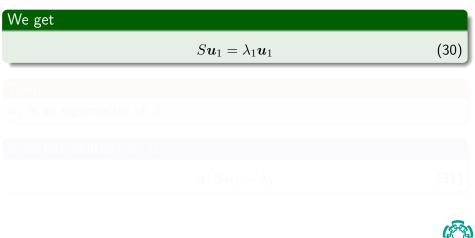
$$\frac{1}{N-1}\sum_{i=1}^{N} \left[\boldsymbol{u}_{1}\boldsymbol{x}_{i} - \boldsymbol{u}_{1}\overline{\boldsymbol{x}}\right] = \boldsymbol{u}_{1}^{T}S\boldsymbol{u}_{1}$$
(28)

Use Lagrange Multipliers to Maximize

$$\boldsymbol{u}_{1}^{T} \boldsymbol{S} \boldsymbol{u}_{1} + \lambda_{1} \left(1 - \boldsymbol{u}_{1}^{T} \boldsymbol{u}_{1} \right)$$
(29)



Derive by \boldsymbol{u}_1





ヘロト ヘロト ヘヨト ヘヨト

Derive by \boldsymbol{u}_1

$S \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$

Then

We get

 \boldsymbol{u}_1 is an eigenvector of S.

If we left-multiply by $oldsymbol{u}_1$

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1$$

イロト イヨト イヨト イヨト



(30)

Derive by \boldsymbol{u}_1

We get

$$S\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1$$

(30)

Then

 \boldsymbol{u}_1 is an eigenvector of S.

If we left-multiply by $oldsymbol{u}_1$

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \tag{31}$$



What about the second eigenvector \boldsymbol{u}_2

We have the following optimization problem

$$\max \ \boldsymbol{u}_2^T S \boldsymbol{u}_2$$

s.t. $\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1$
 $\boldsymbol{u}_2^T \boldsymbol{u}_1 = 0$

Lagrangian

 $L\left(oldsymbol{u}_{2},\lambda_{1},\lambda_{2}
ight)=oldsymbol{u}_{2}^{T}Soldsymbol{u}_{2}-\lambda_{1}\left(oldsymbol{u}_{2}^{T}oldsymbol{u}_{2}-1
ight)-\lambda_{2}\left(oldsymbol{u}_{2}^{T}oldsymbol{u}_{1}-0
ight)$



What about the second eigenvector \boldsymbol{u}_2

We have the following optimization problem

$$\begin{array}{l} \max \ \boldsymbol{u}_2^T S \boldsymbol{u}_2 \\ \text{s.t.} \ \boldsymbol{u}_2^T \boldsymbol{u}_2 = 1 \\ \boldsymbol{u}_2^T \boldsymbol{u}_1 = 0 \end{array}$$

Lagrangian

$$L\left(oldsymbol{u}_{2},\lambda_{1},\lambda_{2}
ight)=oldsymbol{u}_{2}^{T}Soldsymbol{u}_{2}-\lambda_{1}\left(oldsymbol{u}_{2}^{T}oldsymbol{u}_{2}-1
ight)-\lambda_{2}\left(oldsymbol{u}_{2}^{T}oldsymbol{u}_{1}-0
ight)$$



Explanation

First the constrained minimization

• We want to to maximize $oldsymbol{u}_2^TSoldsymbol{u}_2$

Given that the second eigenvector is orthonormal

ullet We have then $oldsymbol{u}_2^Toldsymbol{u}_2=1$

Under orthonormal vectors

• The covariance goes to zero $cov(u_1, u_2) = u_2^T S u_1 = u_2 \lambda_1 u_1 = \lambda_1 u_1^T u_2 = 0$



イロン イロン イヨン イヨン

Explanation

First the constrained minimization

• We want to to maximize $oldsymbol{u}_2^T S oldsymbol{u}_2$

Given that the second eigenvector is orthonormal

• We have then $oldsymbol{u}_2^Toldsymbol{u}_2=1$

Under orthonormal vectors

• The covariance goes to zero $cov(\boldsymbol{u}_1, \boldsymbol{u}_2) = \boldsymbol{u}_2^T S \boldsymbol{u}_1 = \boldsymbol{u}_2 \lambda_1 \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$



イロン イロン イヨン イヨン

Explanation

First the constrained minimization

• We want to to maximize $oldsymbol{u}_2^T S oldsymbol{u}_2$

Given that the second eigenvector is orthonormal

• We have then $oldsymbol{u}_2^Toldsymbol{u}_2=1$

Under orthonormal vectors

• The covariance goes to zero $cov(\boldsymbol{u}_1, \boldsymbol{u}_2) = \boldsymbol{u}_2^T S \boldsymbol{u}_1 = \boldsymbol{u}_2 \lambda_1 \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$



イロト イヨト イヨト

Meaning

The PCA's are perpendicular

$$L\left(\boldsymbol{u}_{2},\lambda_{1},\lambda_{2}\right) = \boldsymbol{u}_{2}^{T}S\boldsymbol{u}_{2} - \lambda_{1}\left(\boldsymbol{u}_{2}^{T}\boldsymbol{u}_{2} - 1\right) - \lambda_{2}\left(\boldsymbol{u}_{2}^{T}\boldsymbol{u}_{1} - 0\right)$$

The the derivative with respect to $oldsymbol{u}_{2}$

$$rac{\partial L\left(oldsymbol{u}_2,\lambda_1,\lambda_2
ight)}{\partialoldsymbol{u}_2}=Soldsymbol{u}_2-\lambda_1oldsymbol{u}_2-\lambda_2oldsymbol{u}_1=0$$

Then, we left multiply $oldsymbol{u}_1$

$$oldsymbol{u}_1^T S oldsymbol{u}_2 - \lambda_1 oldsymbol{u}_1^T oldsymbol{u}_2 - \lambda_2 oldsymbol{u}_1^T oldsymbol{u}_1 = 0$$



Meaning

The PCA's are perpendicular

$$L\left(\boldsymbol{u}_{2},\lambda_{1},\lambda_{2}\right) = \boldsymbol{u}_{2}^{T}S\boldsymbol{u}_{2} - \lambda_{1}\left(\boldsymbol{u}_{2}^{T}\boldsymbol{u}_{2} - 1\right) - \lambda_{2}\left(\boldsymbol{u}_{2}^{T}\boldsymbol{u}_{1} - 0\right)$$

The the derivative with respect to $oldsymbol{u}_2$

$$\frac{\partial L\left(\boldsymbol{u}_{2},\lambda_{1},\lambda_{2}\right)}{\partial \boldsymbol{u}_{2}}=S\boldsymbol{u}_{2}-\lambda_{1}\boldsymbol{u}_{2}-\lambda_{2}\boldsymbol{u}_{1}=0$$

Then, we left multiply a

$$oldsymbol{u}_1^T oldsymbol{S}oldsymbol{u}_2 - \lambda_1 oldsymbol{u}_1^T oldsymbol{u}_2 - \lambda_2 oldsymbol{u}_1^T oldsymbol{u}_1 = 0$$



< ロ > < 回 > < 回 > < 回 > < 回 >

Meaning

The PCA's are perpendicular

$$L(\boldsymbol{u}_2, \lambda_1, \lambda_2) = \boldsymbol{u}_2^T S \boldsymbol{u}_2 - \lambda_1 \left(\boldsymbol{u}_2^T \boldsymbol{u}_2 - 1 \right) - \lambda_2 \left(\boldsymbol{u}_2^T \boldsymbol{u}_1 - 0 \right)$$

The the derivative with respect to $oldsymbol{u}_2$

$$\frac{\partial L\left(\boldsymbol{u}_{2},\lambda_{1},\lambda_{2}\right)}{\partial \boldsymbol{u}_{2}} = S\boldsymbol{u}_{2} - \lambda_{1}\boldsymbol{u}_{2} - \lambda_{2}\boldsymbol{u}_{1} = 0$$

Then, we left multiply $oldsymbol{u}_1$

$$\boldsymbol{u}_1^T S \boldsymbol{u}_2 - \lambda_1 \boldsymbol{u}_1^T \boldsymbol{u}_2 - \lambda_2 \boldsymbol{u}_1^T \boldsymbol{u}_1 = 0$$



Then, we have that

Something Notable

$$0 - 0 - \lambda_2 = 0$$

We have

$$S\boldsymbol{u}_2 - \lambda_2 \boldsymbol{u}_2 = 0$$

Implying

• \boldsymbol{u}_2 is the eigenvector of S with second largest eigenvalue λ_2



Then, we have that

Something Notable

$$0 - 0 - \lambda_2 = 0$$

We have $Soldsymbol{u}_2 - \lambda_2oldsymbol{u}_2 = 0$

Implying

ullet $oldsymbol{u}_2$ is the eigenvector of S with second largest eigenvalue λ_2



Then, we have that

Something Notable

$$0 - 0 - \lambda_2 = 0$$

We have $Soldsymbol{u}_2 - \lambda_2oldsymbol{u}_2 = 0$

Implying

• u_2 is the eigenvector of S with second largest eigenvalue λ_2 .



Thus

Variance will be the maximum when

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \tag{32}$$

is set to the largest eigenvalue. Also know as the First Principal Component

By Induction

It is possible for *M*-dimensional space to define *M* eigenvectors $u_1, u_2, ..., u_M$ of the data covariance S corresponding to $\lambda_1, \lambda_2, ..., \lambda_M$ that maximize the variance of the projected data.

Computational Cost

- Full eigenvector decomposition $O\left(d^3
 ight)$
- $igodoldsymbol{igo$
- Use the Expectation Maximization Algorithm

Thus

Variance will be the maximum when

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \tag{32}$$

is set to the largest eigenvalue. Also know as the First Principal Component

By Induction

It is possible for M-dimensional space to define M eigenvectors $u_1, u_2, ..., u_M$ of the data covariance S corresponding to $\lambda_1, \lambda_2, ..., \lambda_M$ that maximize the variance of the projected data.

Computational Cost

-) Full eigenvector decomposition $O\left(d^3
 ight)$
- ullet Power Method $O\left(Md^2
 ight)$ "Golub and Van Loan, 1996)"
- Use the Expectation Maximization Algorithm

Thus

Variance will be the maximum when

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \tag{32}$$

is set to the largest eigenvalue. Also know as the First Principal Component

By Induction

It is possible for M-dimensional space to define M eigenvectors $u_1, u_2, ..., u_M$ of the data covariance S corresponding to $\lambda_1, \lambda_2, ..., \lambda_M$ that maximize the variance of the projected data.

Computational Cost

- **1** Full eigenvector decomposition $O\left(d^3\right)$
- 2 Power Method $O\left(Md^2\right)$ "Golub and Van Loan, 1996)"
- **③** Use the Expectation Maximization Algorithm

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers

The Process

Example

Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

We have the following steps

Determine covariance matrix

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$
(33)

Generate the decomposition

$$S = U\Sigma U^T$$

With

• Eigenvalues in Σ and eigenvectors in the columns of U.



We have the following steps

Determine covariance matrix

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$
(33)

Generate the decomposition

$$S = U\Sigma U^T$$

With

 \circ Eigenvalues in Σ and eigenvectors in the columns of U.



< ロ > < 回 > < 回 > < 回 > < 回 >

We have the following steps

Determine covariance matrix

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}) (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$$
(33)

Generate the decomposition

$$S = U\Sigma U^T$$

With

• Eigenvalues in Σ and eigenvectors in the columns of U.



Project samples x_i into subspaces dim=k

$$z_i = U_K^T \boldsymbol{x}_i$$

• With U_k is a matrix with k columns



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Singular Value Decomposition

Introduction

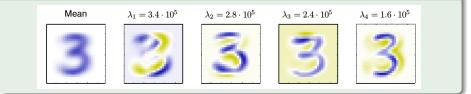




< ロ > < 回 > < 回 > < 回 > < 回 >

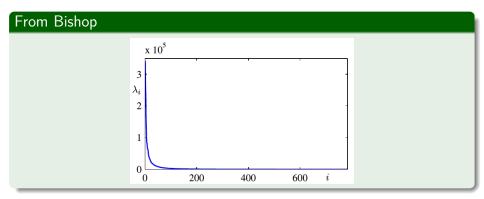
Example

From Bishop





Example





Example

From Bishop





・ロト ・四ト ・ヨト ・ヨト

Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regression

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Singular Value Decomposition Introduction Image Compression



< ロ > < 回 > < 回 > < 回 > < 回 >

What happened with no-square matrices

We can still diagonalize it

Thus, we can obtain certain properties.

We want to avoid the problems with

 $S^{-1}AS$

The eigenvectors in S have three big problems

- They are usually not orthogonal...
- There are not always enough eigenvectors.
- $Ax = \lambda x$ requires A to be square.



What happened with no-square matrices

We can still diagonalize it

Thus, we can obtain certain properties.

We want to avoid the problems with

 $S^{-1}AS$

The eigenvectors in S have three big problems

- They are usually not orthogonal.
- There are not always enough eigenvectors.
- $A m{x} = \lambda m{x}$ requires A to be square.



< ロ > < 回 > < 回 > < 回 > < 回 >

What happened with no-square matrices

We can still diagonalize it

Thus, we can obtain certain properties.

We want to avoid the problems with

 $S^{-1}AS$

The eigenvectors in S have three big problems

- They are usually not orthogonal.
- Interest of the second seco
- $Ax = \lambda x$ requires A to be square.



< ロ > < 同 > < 回 > < 回 >

Therefore, we can look at the following problem

We have a series of vectors

 $\{ {m x}_1, {m x}_2, ..., {m x}_d \}$

Then imagine a set of projection vectors and differences

 $\{oldsymbol{eta}_1,oldsymbol{eta}_2,...,oldsymbol{eta}_d\}$ and $\{oldsymbol{lpha}_1,oldsymbol{lpha}_2,...,oldsymbol{lpha}_d\}$

We want to know a little bit of the relations between them

 After all, we are looking at the possibility of using them for our problem



Therefore, we can look at the following problem

We have a series of vectors

$$\{m{x}_1, m{x}_2, ..., m{x}_d\}$$

Then imagine a set of projection vectors and differences

$$\{oldsymbol{eta}_1,oldsymbol{eta}_2,...,oldsymbol{eta}_d\}$$
 and $\{oldsymbol{lpha}_1,oldsymbol{lpha}_2,...,oldsymbol{lpha}_d\}$

We want to know a little bit of the relations between them

 After all, we are looking at the possibility of using them for our problem



Therefore, we can look at the following problem

We have a series of vectors

 $\{ {m x}_1, {m x}_2, ..., {m x}_d \}$

Then imagine a set of projection vectors and differences

$$\{oldsymbol{eta}_1,oldsymbol{eta}_2,...,oldsymbol{eta}_d\}$$
 and $\{oldsymbol{lpha}_1,oldsymbol{lpha}_2,...,oldsymbol{lpha}_d\}$

We want to know a little bit of the relations between them

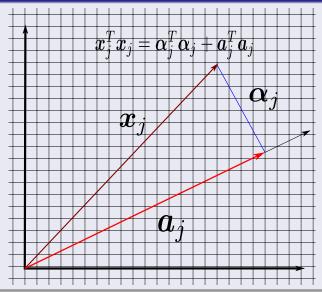
 After all, we are looking at the possibility of using them for our problem



イロト イヨト イヨト

Using the Hypotenuse

A little bit of Geometry, we get



Therefore

We have two possible quantities for each j

$$oldsymbol{lpha}_j^T oldsymbol{lpha}_j = oldsymbol{x}_j^T oldsymbol{x}_j - oldsymbol{a}_j^T oldsymbol{a}_j \ oldsymbol{a}_j^T oldsymbol{a}_j = oldsymbol{x}_j^T oldsymbol{x}_j - oldsymbol{lpha}_j^T oldsymbol{lpha}_j$$

Then, we can minimize and maximize given that $x_{i}^{\prime}x_{i}$ is a constant

$$\min \sum_{j=1}^n oldsymbol{lpha}_j^T oldsymbol{lpha}_j \ \max \sum_{j=1}^n oldsymbol{a}_j^T oldsymbol{a}_j$$



イロン イロン イヨン イヨン

Therefore

We have two possible quantities for each j

$$egin{aligned} & oldsymbol{lpha}_j^Toldsymbol{lpha}_j = oldsymbol{x}_j^Toldsymbol{x}_j - oldsymbol{a}_j^Toldsymbol{a}_j \ & oldsymbol{a}_j^Toldsymbol{a}_j = oldsymbol{x}_j^Toldsymbol{x}_j - oldsymbol{lpha}_j^Toldsymbol{lpha}_j \ & oldsymbol{a}_j \ & oldsymbol{a}_j \ & oldsymbol{a}_j = oldsymbol{x}_j^Toldsymbol{x}_j - oldsymbol{lpha}_j^Toldsymbol{lpha}_j \ & oldsymbol{a}_j \ & oldsym$$

Then, we can minimize and maximize given that $x_i^T x_j$ is a constant

$$\min \sum_{j=1}^n oldsymbol{lpha}_j^T oldsymbol{lpha}_j \ \max \sum_{j=1}^n oldsymbol{a}_j^T oldsymbol{a}_j$$

Cinvestav ≣ ∽ ۹ ↔ 124 / 140

イロト イロト イヨト イヨト

Actually this is know as the dual problem (Weak Duality)

An example of this

 $\min \ \boldsymbol{w}^T \boldsymbol{x} \\ s.t \mathsf{A} \boldsymbol{x} \leq \boldsymbol{b} \\ \boldsymbol{x} \geq 0$

Then, using what is know as slack variabless

 $A\boldsymbol{x} + A'\boldsymbol{x} = b$

Each row lives in the column space, but the y, lives in the column space

 $(Am{x}+A'm{x})_i
ightarrow y_i$ and $m{x}' \ge 0$.

Cinvestav ∽ < (~ 125 / 140

Actually this is know as the dual problem (Weak Duality)

An example of this

$$\begin{array}{l} \min \ \boldsymbol{w}^T \boldsymbol{x} \\ s.t \mathsf{A} \boldsymbol{x} \leq \boldsymbol{b} \\ \boldsymbol{x} \geq 0 \end{array}$$

Then, using what is know as slack variables

 $A\boldsymbol{x} + A'\boldsymbol{x} = b$

Each row lives in the column space, but the y_i lives in the column space

 $\left(Aoldsymbol{x}+A'oldsymbol{x}
ight)_{oldsymbol{i}}
ightarrow y_{oldsymbol{i}}$ and $oldsymbol{x}'\geq 0$.

Cinvestav ∽ < (~ 125 / 140

ヘロト 人間 ト 人 田 ト 人 田 ト

Actually this is know as the dual problem (Weak Duality)

An example of this

$$\begin{array}{l} \min \ \boldsymbol{w}^T \boldsymbol{x} \\ s.t \mathsf{A} \boldsymbol{x} \leq \boldsymbol{b} \\ \boldsymbol{x} \geq 0 \end{array}$$

Then, using what is know as slack variables

$$A\boldsymbol{x} + A'\boldsymbol{x} = b$$

Each row lives in the column space, but the y_i lives in the column space

$$(A\boldsymbol{x} + A'\boldsymbol{x})_i o y_i \text{ and } \boldsymbol{x}' \ge 0$$

Cinvestav ∽ < ↔ 125 / 140

イロン イロン イヨン イヨン

Then, we have that

Example		
	$\left[\begin{array}{rrr} 0 & 0 \\ 0 & 1 \\ 1 & 0 \end{array}\right]$	

Element in the column space of dimensionality have three dimensions

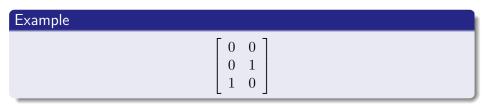
• But in the row space their dimension is 2

Properties



・ロト ・四ト ・ヨト ・ヨト

Then, we have that



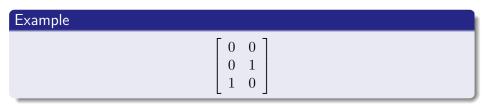
Element in the column space of dimensionality have three dimensions

• But in the row space their dimension is 2



< ロ > < 回 > < 回 > < 回 > < 回 >

Then, we have that



Element in the column space of dimensionality have three dimensions

• But in the row space their dimension is 2

Properties



イロト イヨト イヨト イヨト

We have then

Stack such vectors that in the *d*-dimensional space

$\bullet\,$ In a matrix A of $n\times d$

$$\mathbf{A} = \left[egin{array}{c} oldsymbol{a}_1^T \ oldsymbol{a}_2^T \ dots \ oldsymbol{a}_n^T \end{array}
ight]$$

The matrix works as a Projection Matrix

 We are looking for a unit vector v such that length of the projection is maximized.



We have then

Stack such vectors that in the *d*-dimensional space

$\bullet~$ In a matrix A~ of $n\times d$

$$A = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_n^T \end{bmatrix}$$

The matrix works as a Projection Matrix

• We are looking for a unit vector v such that length of the projection is maximized.



Why? Do you remember the Projection to a single vector p?

Definition of the projection under unitary vector

$$oldsymbol{p} = rac{oldsymbol{v}^Toldsymbol{a}_i}{oldsymbol{v}^Toldsymbol{v}}oldsymbol{v} = \left[oldsymbol{v}^Toldsymbol{a}_i
ight]oldsymbol{v}$$

Therefore the length of the projected vector is

 $\left\| \left[oldsymbol{v}^T oldsymbol{a}_i
ight] oldsymbol{v}
ight\| = \left| oldsymbol{v}^T oldsymbol{a}_i
ight|$



Why? Do you remember the Projection to a single vector p?

Definition of the projection under unitary vector

$$oldsymbol{p} = rac{oldsymbol{v}^Toldsymbol{a}_i}{oldsymbol{v}^Toldsymbol{v}}oldsymbol{v} = \left[oldsymbol{v}^Toldsymbol{a}_i
ight]oldsymbol{v}$$

Therefore the length of the projected vector is

$$\left\| \begin{bmatrix} oldsymbol{v}^T oldsymbol{a}_i \end{bmatrix} oldsymbol{v}
ight\| = \left| oldsymbol{v}^T oldsymbol{a}_i
ight|$$



Then

Thus with a little bit of notation

$$A oldsymbol{v} = \left[egin{array}{c} oldsymbol{a}_1^T \ oldsymbol{a}_2^T \ dots \ oldsymbol{a}_d^T \end{array}
ight] oldsymbol{v} = \left[egin{array}{c} oldsymbol{a}_1^T oldsymbol{v} \ oldsymbol{a}_2^T oldsymbol{v} \ dots \ oldsymbol{a}_2^T oldsymbol{v} \ dots \ oldsymbol{a}_d^T oldsymbol{v} \end{array}
ight]$$

Therefore

$$\left\|Av
ight\|=\sqrt{\sum\limits_{i=1}^{d}\left(a_{i}^{T}v
ight)^{2}}$$



◆□▶ ◆圖▶ ◆臣▶ ◆臣▶

Then

Thus with a little bit of notation

$$A\boldsymbol{v} = \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{a}_d^T \end{bmatrix} \boldsymbol{v} = \begin{bmatrix} \boldsymbol{a}_1^T \boldsymbol{v} \\ \boldsymbol{a}_2^T \boldsymbol{v} \\ \vdots \\ \boldsymbol{a}_d^T \boldsymbol{v} \end{bmatrix}$$

Therefore

$$\left\|Aoldsymbol{v}
ight\|=\sqrt{\sum_{i=1}^{d}\left(a_{i}^{T}oldsymbol{v}
ight)^{2}}$$



イロト イヨト イヨト イヨト



It is possible to ask to maximize the longitude of such vector (Singular Vector)

$$oldsymbol{v}_1 = rg\max_{\|oldsymbol{v}\|=1} \|Aoldsymbol{v}\|$$

Then, we can define the following singular values

 $\sigma_1\left(A\right) = \|A\boldsymbol{v}_1\|$





It is possible to ask to maximize the longitude of such vector (Singular Vector)

$$oldsymbol{v}_1 = rg\max_{\|oldsymbol{v}\|=1} \|Aoldsymbol{v}\|$$

Then, we can define the following singular value

 $\sigma_1\left(A\right) = \|A\boldsymbol{v}_1\|$



This is known as

Definition

- The **best-fit line problem** describes the problem of finding the best line for a set of data points, where the quality of the line is measured by the sum of squared (perpendicular) distances of the points to the line.
 - Remember, we are looking at the dual problem....

Generalization

 This can be transferred to higher dimensions: One can find the best-fit d-dimensional subspace, so the subspace which minimizes the sum of the squared distances of the points to the subspace



< ロ > < 同 > < 回 > < 回 >

This is known as

Definition

- The **best-fit line problem** describes the problem of finding the best line for a set of data points, where the quality of the line is measured by the sum of squared (perpendicular) distances of the points to the line.
 - Remember, we are looking at the dual problem....

Generalization

• This can be transferred to higher dimensions: One can find the best-fit *d*-dimensional subspace, so the subspace which minimizes the sum of the squared distances of the points to the subspace



< ロ > < 同 > < 回 > < 回)

Then, in a Greedy Fashion

The second singular vector $oldsymbol{v}_2$

$$oldsymbol{v}_2 = rg\max_{oldsymbol{v}ot v_1, \|oldsymbol{v}\|=1} \|Aoldsymbol{v}\|$$

hem you go through this process

Stop when we have found all the following vectors:

 $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r$

As singular vectors and



Then, in a Greedy Fashion

The second singular vector \boldsymbol{v}_2

$$oldsymbol{v}_2 = rg\max_{oldsymbol{v}ot oldsymbol{v}_1, \|oldsymbol{v}\| = 1} \|Aoldsymbol{v}\|$$

Them you go through this process

• Stop when we have found all the following vectors:

 $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r$

As singular vectors and





Then, in a Greedy Fashion

The second singular vector \boldsymbol{v}_2

$$oldsymbol{v}_2 = rg\max_{oldsymbol{v}ot oldsymbol{v}_1, \|oldsymbol{v}\| = 1} \|Aoldsymbol{v}\|$$

Them you go through this process

• Stop when we have found all the following vectors:

 $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r$

As singular vectors and

イロト イボト イヨト イヨト

Proving that the strategy is good

Theorem

• Let A be an $n \times d$ matrix where $v_1, v_2, ..., v_r$ are the singular vectors defined above. For $1 \le k \le r$, let V_k be the subspace spanned by $v_1, v_2, ..., v_k$. Then for each k, V_k is the best-fit k-dimensional subspace for A.



Proof

For k = 1

• What about k = 2? Let W be a best-fit 2- dimensional subspace for A.

For any basis $oldsymbol{w}_1,oldsymbol{w}_2$ of W

|Aw₁|² + |Aw₂|² is the sum of the squared lengths of the projections of the rows of A to W.

Now, choose a basis w_1, w_2 so that w_2 is perpendicular to $oldsymbol{v}$

• This can be a unit vector perpendicular to v₁ projection in W.



イロト イヨト イヨト イヨト

Proof

For k = 1

• What about k = 2? Let W be a best-fit 2- dimensional subspace for A.

For any basis $\boldsymbol{w}_1, \boldsymbol{w}_2$ of W

• $|Aw_1|^2 + |Aw_2|^2$ is the sum of the squared lengths of the projections of the rows of A to W.

Now, choose a basis w_1, w_2 so that w_2 is perpendicular to v_2

This can be a unit vector perpendicular to v₁ projection in W.



Proof

For k = 1

• What about k = 2? Let W be a best-fit 2- dimensional subspace for A.

For any basis $\boldsymbol{w}_1, \boldsymbol{w}_2$ of W

• $|Aw_1|^2 + |Aw_2|^2$ is the sum of the squared lengths of the projections of the rows of A to W.

Now, choose a basis $oldsymbol{w}_1, oldsymbol{w}_2$ so that $oldsymbol{w}_2$ is perpendicular to $oldsymbol{v}_1$

• This can be a unit vector perpendicular to v_1 projection in W.



Do you remember $\boldsymbol{v}_1 = \arg \max_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$?

Therefore

$$|A oldsymbol{w}_1|^2 \leq |A oldsymbol{v}_1|^2$$
 and $|A oldsymbol{w}_2|^2 \leq |A oldsymbol{v}_2|^2$

Then

$$|A\boldsymbol{w}_1|^2 + |A\boldsymbol{w}_2|^2 \le |A\boldsymbol{v}_1|^2 + |A\boldsymbol{v}_2|^2$$

In a similar way for k

V_k is at least as good as W and hence is optimal.



Do you remember $\boldsymbol{v}_1 = \arg \max_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$?

Therefore

$$|A oldsymbol{w}_1|^2 \leq |A oldsymbol{v}_1|^2$$
 and $|A oldsymbol{w}_2|^2 \leq |A oldsymbol{v}_2|^2$

Then

$$|A\boldsymbol{w}_{1}|^{2} + |A\boldsymbol{w}_{2}|^{2} \le |A\boldsymbol{v}_{1}|^{2} + |A\boldsymbol{v}_{2}|^{2}$$

In a similar way for *k*

V_k is at least as good as W and hence is optimal.



Do you remember $\boldsymbol{v}_1 = \arg \max_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$?

Therefore

$$|A oldsymbol{w}_1|^2 \leq |A oldsymbol{v}_1|^2$$
 and $|A oldsymbol{w}_2|^2 \leq |A oldsymbol{v}_2|^2$

Then

$$|A\boldsymbol{w}_{1}|^{2} + |A\boldsymbol{w}_{2}|^{2} \le |A\boldsymbol{v}_{1}|^{2} + |A\boldsymbol{v}_{2}|^{2}$$

In a similar way for k > 2

• V_k is at least as good as W and hence is optimal.



Remarks

Every Matrix has a singular value decomposition

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

Where **Where**

• The columns of U are an orthonormal basis for the column space.



Remarks

Every Matrix has a singular value decomposition

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

Where

- The columns of U are an orthonormal basis for the column space.
- $\bullet\,$ The columns of V are an orthonormal basis for the row space.



Remarks

Every Matrix has a singular value decomposition

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

Where

- $\bullet\,$ The columns of U are an orthonormal basis for the column space.
- The columns of V are an orthonormal basis for the row space.
- The Σ is diagonal and the entries on its diagonal $\sigma_i = \Sigma_{ii}$ are positive real numbers, called the singular values of A.



Properties of the Singular Value Decomposition

First

The eigenvalues of the symmetric matrix $A^T A$ are equal to the square of the singular values of A

$$A^TA = V\Sigma U^T U^T \Sigma V^T = V\Sigma^2 V^T$$

Second

The rank of a matrix is equal to the number of non-zero singular values.



Properties of the Singular Value Decomposition

First

The eigenvalues of the symmetric matrix $A^T A$ are equal to the square of the singular values of A

$$A^T A = V \Sigma U^T U^T \Sigma V^T = V \Sigma^2 V^T$$

Second

The rank of a matrix is equal to the number of non-zero singular values.



Outline

Linear Transformation

- Introduction
- Functions that can be defined using matrices
- Linear Functions
- Kernel and Range
- The Matrix of a Linear Transformation
- Going Back to Homogeneous Equations
- The Rank-Nullity Theorem

Derivative of Transformations

- Introduction
- Derivative of a Linear Transformation
- Derivative of a Quadratic Transformation

3 Linear Regressior

- The Simplest Functions
- Splitting the Space
- Defining the Decision Surface
- Properties of the Hyperplane $w^T x + w_0$
- Augmenting the Vector
- Least Squared Error Procedure
 - The Geometry of a Two-Category Linearly-Separable Case
- The Error Idea
- The Final Error Equation

Principal Component Analysis

- Karhunen-Loeve Transform
- Projecting the Data
- Lagrange Multipliers
- The Process
- Example

Singular Value Decomposition

Introduction





< ロ > < 回 > < 回 > < 回 > < 回 >

Singular Value Decomposition as Sums

The singular value decomposition can be viewed as a sum of rank 1 matrices

$$A = A_1 + A_2 + \dots + A_R \tag{34}$$







Singular Value Decomposition as Sums

The singular value decomposition can be viewed as a sum of rank 1 matrices

$$A = A_1 + A_2 + \dots + A_R \tag{34}$$

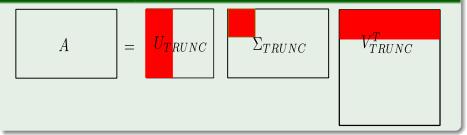
イロト イヨト イヨト

Why?

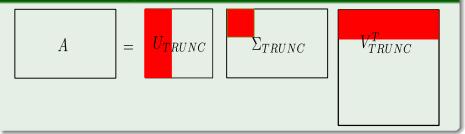
$$\boldsymbol{u}_{1}\boldsymbol{A} = \boldsymbol{U} \begin{pmatrix} \sigma_{1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_{2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \sigma_{R} \end{pmatrix} \boldsymbol{V}^{T} = \begin{pmatrix} \boldsymbol{u}_{1} & \boldsymbol{u}_{2} & \cdots & \boldsymbol{u}_{R} \end{pmatrix} \begin{pmatrix} \sigma_{1}\boldsymbol{v}_{1}^{T} \\ \sigma_{2}\boldsymbol{v}_{2}^{T} \\ \vdots \\ \sigma_{R}\boldsymbol{v}_{R}^{T} \end{pmatrix}$$
$$= \sigma_{1}\boldsymbol{u}_{1}\boldsymbol{v}_{1}^{T} + \sigma_{2}\boldsymbol{u}_{2}\boldsymbol{v}_{2}^{T} + \cdots + \sigma_{R}\boldsymbol{u}_{R}\boldsymbol{v}_{R}^{T}$$



Truncating the singular value decomposition allows us to represent the matrix with less parameters



Truncating the singular value decomposition allows us to represent the matrix with less parameters



For a 512×512

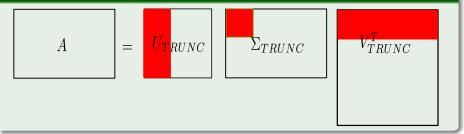
- Full Representation $512 \times 512 = 262, 144$
- Rank 10 approximation $512 \times 10 + 10 + 10 \times 512 = 10,250$

• Rank 40 approximation $512 \times 40 + 40 + 40 \times 512 = 41,000$

• Rank 80 approximation $512 \times 80 + 80 + 80 \times 512 = 82,000$

< (1) > < = >

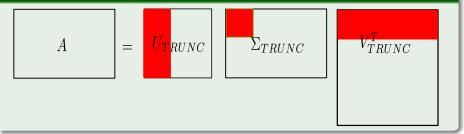
Truncating the singular value decomposition allows us to represent the matrix with less parameters



For a 512×512

- Full Representation $512 \times 512 = 262, 144$
- Rank 10 approximation $512 \times 10 + 10 + 10 \times 512 = 10,250$
- Rank 40 approximation $512 \times 40 + 40 + 40 \times 512 = 41,000$

Truncating the singular value decomposition allows us to represent the matrix with less parameters



For a 512×512

- Full Representation $512 \times 512 = 262, 144$
- Rank 10 approximation $512 \times 10 + 10 + 10 \times 512 = 10,250$
- Rank 40 approximation $512 \times 40 + 40 + 40 \times 512 = 41,000$
- Rank 80 approximation $512 \times 80 + 80 + 80 \times 512 = 82,000$