Mathematics for Artificial Intelligence

Transformation and Applications

Andres Mendez-Vazquez

April 9, 2020

Outline

Cinvestav

Outline

Cinvestav

Going further than solving $Ax = y$

We can go further

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

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

Going further than solving $A\boldsymbol{x} = \boldsymbol{y}$

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 $\boldsymbol{x} \in \mathbb{R}^n$ to a vector $\boldsymbol{y} \in \mathbb{R}^m$

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

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

Cinvestav

Going further than solving $A\boldsymbol{x} = \boldsymbol{y}$

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 $\boldsymbol{x} \in \mathbb{R}^n$ to a vector $\boldsymbol{y} \in \mathbb{R}^m$

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

We like the second expression because

1. It is easy to identify the domain $\mathbb{R}^n$
2. It is easy to find the range $\mathbb{R}^m$

Going further than solving $A\boldsymbol{x} = \boldsymbol{y}$

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 $\boldsymbol{x} \in \mathbb{R}^n$ to a vector $\boldsymbol{y} \in \mathbb{R}^m$

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

We like the second expression because

1. It is easy to identify the domain $\mathbb{R}^n$
2. It is easy to find the range $\mathbb{R}^m$

Examples

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

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

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.

Cinvestav

Outline

Cinvestav

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

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

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

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

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

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

Cinvestav

Example

Let

$$A_{2\times 3} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

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

Example

Let

$$A_{2\times 3} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

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

Example

Let

$$A_{2\times 3} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

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 $\boldsymbol{x} \in \mathbb{R}^3$ to the vector $A\boldsymbol{x} \in \mathbb{R}^2$

Cinvestav

Outline

Cinvestav

We have

Definition

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

1. $f\left(\boldsymbol{x}_1 + \boldsymbol{x}_2\right) = f\left(\boldsymbol{x}_1\right) + f\left(\boldsymbol{x}_2\right)$
2. $f\left(c\boldsymbol{x}\right) = cf\left(\boldsymbol{x}\right)$

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

Thus

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

We have

Definition

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

1. $f(\boldsymbol{x}_1 + \boldsymbol{x}_2) = f(\boldsymbol{x}_1) + f(\boldsymbol{x}_2)$
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.

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 $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and for all the scalars c_1, c_2:

$$f\left(c_1 \boldsymbol{x}_1 + c_2 \boldsymbol{x}_2\right) = c_1 f\left(\boldsymbol{x}_1\right) + c_2 f\left(\boldsymbol{x}_2\right)$$

Proof

Any idea?

We have the following proposition

Proposition

$f : \mathbb{R}^n \to \mathbb{R}^m$ is linear $\iff$ for all vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$ and for all the scalars c_1, c_2:

$$f\left(c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2\right) = c_1 f\left(\boldsymbol{x}_1\right) + c_2 f\left(\boldsymbol{x}_2\right)$$

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 $f_A(cx_1)$?

Proof

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

How?

First

$$f_A\left(\boldsymbol{x}_1+\boldsymbol{x}_2\right)=A\left(\boldsymbol{x}_1+\boldsymbol{x}_2\right)=A\boldsymbol{x}_1+A\boldsymbol{x}_2=f_A\left(\boldsymbol{x}_1\right)+f_A\left(\boldsymbol{x}_2\right)$$

Second

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

Proof

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

How?

First

$$f_A(\boldsymbol{x}_1+\boldsymbol{x}_2)=A(\boldsymbol{x}_1+\boldsymbol{x}_2)=A\boldsymbol{x}_1+A\boldsymbol{x}_2=f_A(\boldsymbol{x}_1)+f_A(\boldsymbol{x}_2)$$

Second

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

Outline

Cinvestav

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

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

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

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

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.

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.

From This

It can be shown that

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

Also

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

From This

It can be shown that

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

Also

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

Outline

Assume the following

Let

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1}, \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1}, ..., \boldsymbol{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}$$

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

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 e_1 + x_2 e_2 + \dots + x_n e_n$$

Cinvestav

Assume the following

Let

$$\boldsymbol{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1}, \boldsymbol{e}_3 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}_{n\times 1}, ..., \boldsymbol{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}_{n\times 1}$$

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

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1\boldsymbol{e}_1 + x_2\boldsymbol{e}_2 + ... + x_n\boldsymbol{e}_n$$

Then

Applying f

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

A linear combination of elements

$$\{f(\boldsymbol{e}_1), f(\boldsymbol{e}_2), ..., f(\boldsymbol{e}_n)\}$$

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

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

Then

Applying f

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

A linear combination of elements

$$\{f(\boldsymbol{e}_1), f(\boldsymbol{e}_2), \dots, f(\boldsymbol{e}_n)\}$$

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

$$A = \left(f(\boldsymbol{e}_1) | f(\boldsymbol{e}_2) | \dots | f(\boldsymbol{e}_n)\right)_{m \times n}$$

Then

Applying f

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

A linear combination of elements

$$\{f(\boldsymbol{e}_1), f(\boldsymbol{e}_2), ..., f(\boldsymbol{e}_n)\}$$

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

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

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.

Cinvestav

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.

Outline

Cinvestav

Given an $m \times n$ matrix A

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

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

$$A\boldsymbol{x} = 0$$

Remember how to prove the subspaces...

$$\boldsymbol{x}_2 + \boldsymbol{x}_2 \in V \text{ and } c\boldsymbol{x} \in V$$

- Do you remember?

Given an $m \times n$ matrix A

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

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

$$A\boldsymbol{x} = 0$$

Remember how to prove the subspaces...

$$\boldsymbol{x}_2 + \boldsymbol{x}_2 \in V \text{ and } c\boldsymbol{x} \in V$$

- Do you remember?

Cinvestav

Then, we have

Definition

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

It is also known as

$$x_H = \{x | Ax = 0\}$$

Then, we have

Definition

- 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 Range(f) and Ker(f) are sub-spaces

Which ones they are?

Any Idea?

Range(f)

The column space of the matrix A.

Ker(f)

It is the null space of A.

Cinvestav

Knowing that Range(f) and Ker(f) are sub-spaces

Which ones they are?

Any Idea?

Range(f)

The column space of the matrix A.

Ker(f)

It is the null space of A.

Knowing that Range(f) and Ker(f) are sub-spaces

Which ones they are?

Any Idea?

Range(f)

The column space of the matrix A.

Ker(f)

It is the null space of A.

Outline

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.

Cinvestav

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.

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.

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.

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 $dim\left[N(A)\right]$.

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 $dim\left[N(A)\right]$.

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.

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:

$$j + k = n$$

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:

$$s_1 v_1 + s_2 v_2 + s_3 v_3 + \cdots + s_k v_k$$

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:

$$s_1 v_1 + s_2 v_2 + s_3 v_3 + \cdots + s_k v_k$$

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:

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

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

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

Where

The vectors are the Canonical Ones

- Here, a trick!!!

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

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

Therefore the vectors are linearly independents

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

Where

The vectors are the Canonical Ones

- Here, a trick!!!

Meaning in $\boldsymbol{v}_1$, 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

$B \sim A \Leftrightarrow$ 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

$B \sim A \Leftrightarrow$ 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.

Cinvestav

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.

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.

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 $N(B) = N(A)$.

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.

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.

Cinvestav

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.

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.

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.

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.

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.

Finally

We have

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

Cinvestav

Outline

Cinvestav

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

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

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

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Cinvestav

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 $\psi(\boldsymbol{x})$.

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 $\psi(\boldsymbol{x})$.

Then, we can get our first ideas on derivatives

- For Linear Transformations.

Outline

Cinvestav

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

Theorem

- Let $\boldsymbol{y} = A\boldsymbol{x}$ where $\boldsymbol{y}$ is a $m \times 1$, $\boldsymbol{x}$ is a $n \times 1$, A is a $m \times n$ and A does not depend on $\boldsymbol{x}$, then

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

Cinvestav

Proof

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

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

Proof

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

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

Cinvestav

Proof

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

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

Outline

Cinvestav

Derivative of $\boldsymbol{y}^T A\boldsymbol{x}$

Theorem

- Let the scalar α be defined by

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

where

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

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

Derivative of $\boldsymbol{y}^T A \boldsymbol{x}$

Theorem

- Let the scalar α be defined by

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

where

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

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

Proof

Define

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

Note

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

By the previous theorem

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

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

Proof

Define

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

Note

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

By the previous theorem

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

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

Proof

Define

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

Note

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

By the previous theorem

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

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

Outline

Cinvestav

What is it?

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

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 \tag{1}$$

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

In the case of $\mathbb{R}^2$

We have:

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

What is it?

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

Here, we have an hyperplane as a model:

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 \tag{1}$$

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

In the case of $\mathbb{R}^2$

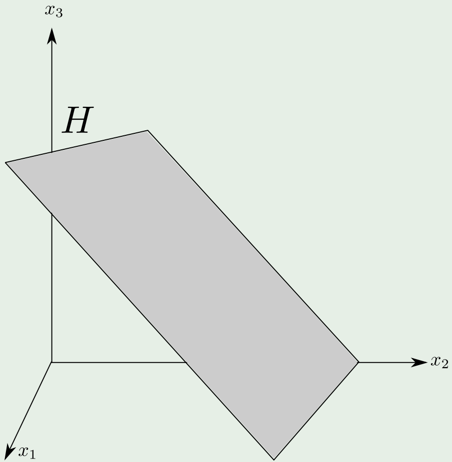
We have:

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

Cinvestav

Example

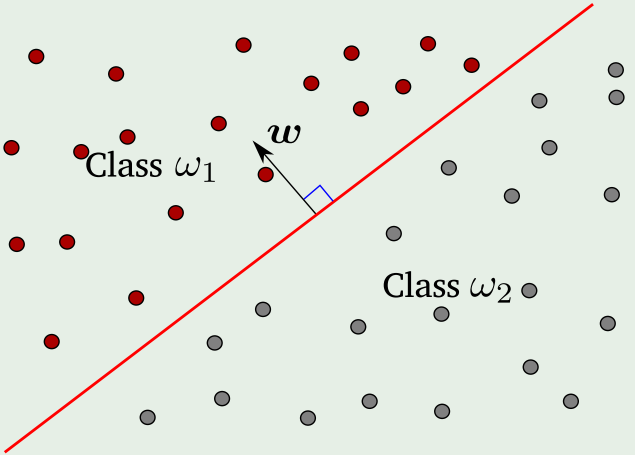
Hyperplane in $\mathbb{R}^3$



Outline

Splitting The Space $\mathbb{R}^2$

Using a simple straight line (Hyperplane)



Splitting the Space?

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

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

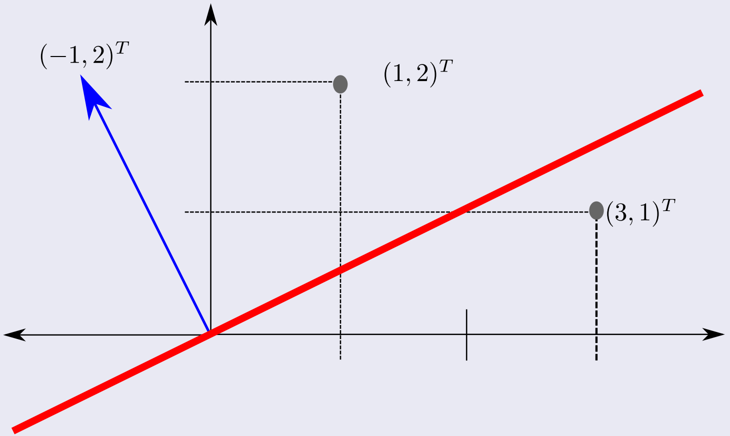
Hyperplane

Splitting the Space?

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

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

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

Cinvestav

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

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

Thus

$$w^T x_1 + w_0 = w^T x_2 + w_0 \quad (3)$$

The Decision Surface

The equation $g\left(x\right)=0$ defines a decision surface

Separating the elements in classes, ω_1 and ω_2.

When $g\left(x\right)$ 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

$$\boldsymbol{w}^T\boldsymbol{x}_1+w_0=\boldsymbol{w}^T\boldsymbol{x}_2+w_0 \quad (3)$$

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

$$\boldsymbol{w}^T \boldsymbol{x}_1 + w_0 = \boldsymbol{w}^T \boldsymbol{x}_2 + w_0 \tag{3}$$

Defining a Decision Surface

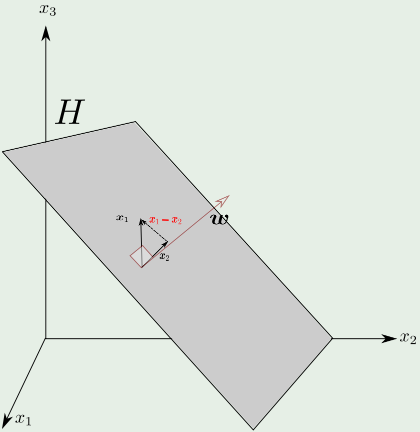
Then

$$\boldsymbol{w}^T \left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right) = 0 \tag{4}$$

Therefore

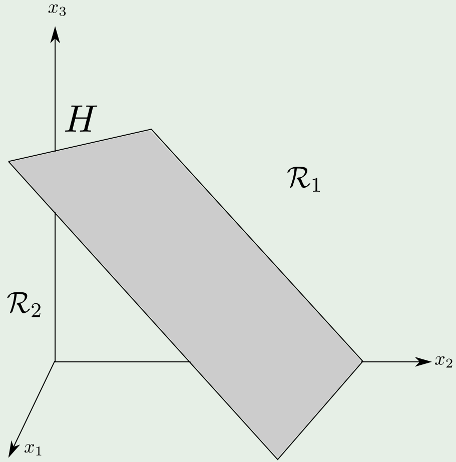
$\boldsymbol{x}_1 - \boldsymbol{x}_2$ lives in the hyperplane i.e. it is perpendicular to $\boldsymbol{w}^T$

- 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

The space is split in two regions (Example in $\mathbb{R}^3$) by the hyperplane H

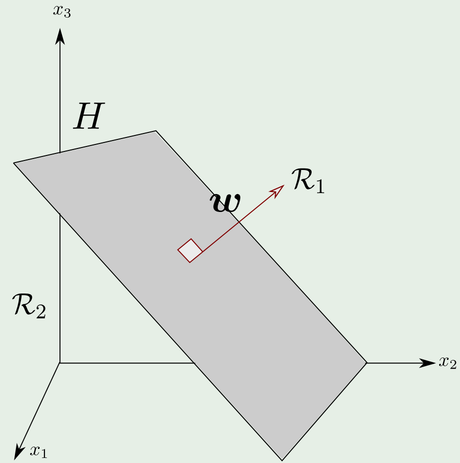


Outline

Cinvestav

Some Properties of the Hyperplane

Given that $g(\boldsymbol{x}) > 0$ if $\boldsymbol{x} \in \mathcal{R}_1$



It is more

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.

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

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

- r is the desired distance
  - Positive, if $\boldsymbol{x}$ is in the positive side
  - Negative, if $\boldsymbol{x}$ is in the negative side

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

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

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

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

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

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
    - Positive, if x is in the positive side
    - Negative, if x is in the negative side

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

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

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if x is in the positive side
  - Negative, if x is in the negative side

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

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

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if $\boldsymbol{x}$ is in the positive side
  - Negative, if $\boldsymbol{x}$ is in the negative side

It is more

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, $g(\boldsymbol{x})$ can give us a way to obtain the distance from $\boldsymbol{x}$ to the hyperplane H

First, we express any $\boldsymbol{x}$ as follows

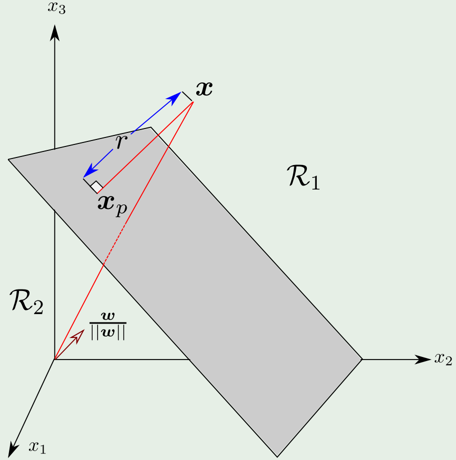
$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

- $\boldsymbol{x}_p$ is the normal projection of $\boldsymbol{x}$ onto H.
- r is the desired distance
  - Positive, if $\boldsymbol{x}$ is in the positive side
  - Negative, if $\boldsymbol{x}$ is in the negative side

We have something like this

We have then



Now

Since $g(\boldsymbol{x}_p) = 0$

We have that

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

Now

Since $g\left(\boldsymbol{x}_p\right)=0$

We have that

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

Then, we have

$$
r = \frac{g\left(\boldsymbol{x}\right)}{\|\boldsymbol{w}\|} \tag{5}
$$

Now

Since $g\left(\boldsymbol{x}_p\right) = 0$

We have that

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

Then, we have

$$
r = \frac{g\left(\boldsymbol{x}\right)}{\|\boldsymbol{w}\|} \tag{5}
$$

Now

Since $g(\boldsymbol{x}_p) = 0$

We have that

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

Then, we have

$$r = \frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \quad (5)$$

Now

Since $g(\boldsymbol{x}_p) = 0$

We have that

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

Then, we have

$$
r = \frac{g(\boldsymbol{x})}{\|\boldsymbol{w}\|} \tag{5}
$$

Now

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

We have that

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

Then, we have

$$
r = \frac{g\left(\boldsymbol{x}\right)}{\|\boldsymbol{w}\|} \tag{5}
$$

In particular

The distance from the origin to H

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

Cinvestav

In particular

The distance from the origin to H

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

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.

Cinvestav

In particular

The distance from the origin to H

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

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.

In particular

The distance from the origin to H

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

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 $\boldsymbol{w}^T\boldsymbol{x}$ and hyperplane passes through the origin.

Cinvestav

Outline

We want to solve the independence of w_0

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

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

We want to solve the independence of w_0

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

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

By making

$$x_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

x_{aug} is called an augmented feature vector.

We want to solve the independence of w_0

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

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

By making

$$x_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ x \end{pmatrix}$$

Where

x_{aug} is called an augmented feature vector.

We want to solve the independence of w_0

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

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

By making

$$\boldsymbol{x}_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \\ \boldsymbol{x} \\ \end{pmatrix}$$

Where

$\boldsymbol{x}_{aug}$ is called an augmented feature vector.

We want to solve the independence of w_0

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

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

By making

$$\boldsymbol{x}_{aug} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

Where

$\boldsymbol{x}_{aug}$ is called an augmented feature vector.

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

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.

In a similar way

We have the augmented weight vector

$$\boldsymbol{w}_{aug} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} = \begin{pmatrix} w_0 \\ \\ \boldsymbol{w} \\ \\ \end{pmatrix}$$

Remarks

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

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 $\boldsymbol{x}_{aug}$-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 $\boldsymbol{x}_{aug}$-space.

Even Though

The corresponding hyperplane H can be in any position of the $\boldsymbol{x}$-space.

More Remarks

In addition

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

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

Now

Is $\|\boldsymbol{w}\| \leq \|\boldsymbol{w}_{aug}\|$

- Ideas?

$$\sqrt{\sum_{i=1}^{d} w_i^2} \leq \sqrt{\sum_{i=1}^{d} w_i^2 + w_0^2}$$

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.

Now

Is $\|\boldsymbol{w}\| \leq \|\boldsymbol{w}_{aug}\|$

- Ideas?

$$\sqrt{\sum_{i=1}^{d} w_i^2} \leq \sqrt{\sum_{i=1}^{d} w_i^2 + w_0^2}$$

This mapping is quite useful

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

Outline

Cinvestav

Outline

Cinvestav

Initial Supposition

Suppose, we have

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

Initial Supposition

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

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Initial Supposition

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

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Initial Supposition

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

- $\boldsymbol{w}^T \boldsymbol{x}_i > 0$, if $\boldsymbol{x}_i \in \omega_1$.
- $\boldsymbol{w}^T \boldsymbol{x}_i < 0$, if $\boldsymbol{x}_i \in \omega_2$.

The name of this weight vector

It is called a separating vector or solution vector.

Now, assume the following

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

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

Now, assume the following

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

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

We have a problem!!!

Which is the problem?

Now, assume the following

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

1. 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?

Now, assume the following

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

1. 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?

A Simple Solution For Our Quandary

Label the Classes

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

A Simple Solution For Our Quandary

Label the Classes

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

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = -1$.

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

A Simple Solution For Our Quandary

Label the Classes

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

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = -1$.

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

A Simple Solution For Our Quandary

Label the Classes

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

We produce the following labels

1. if $\boldsymbol{x} \in \omega_1$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = +1$.
2. if $\boldsymbol{x} \in \omega_2$ then $y_{ideal} = g_{ideal}(\boldsymbol{x}) = -1$.

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

Outline

Cinvestav

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 + e \quad (8)$$

Where the e

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

Thus, we can do the following

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

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e \quad (8)$$

Where the e

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

Thus, we can do the following

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

Now, What?

Assume true function f is given by

$$y_{noise} = g_{noise}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 + e \quad (8)$$

Where the e

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

Thus, we can do the following

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

Thus, we have

What to do?

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

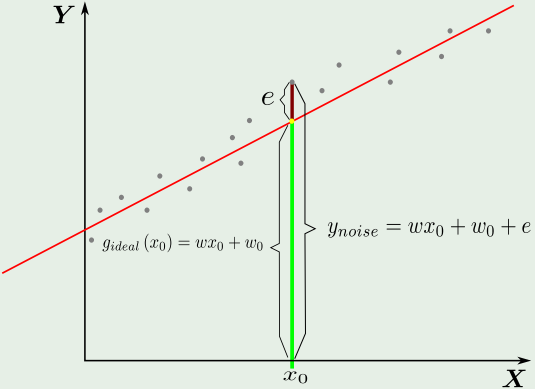
Graphically

Thus, we have

What to do?

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

Graphically



Then, we have

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

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

We use y_{ideal}

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

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}

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

We use y_{ideal}

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

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}

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

We use y_{ideal}

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

We will see

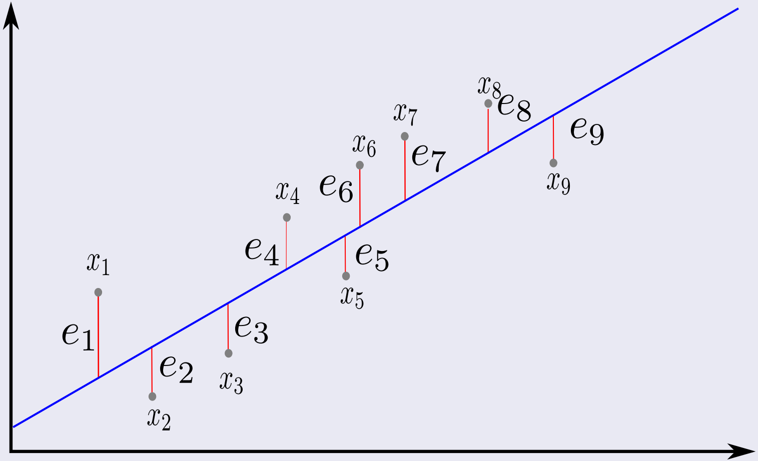
How the geometry will solve the problem with using these labels.

Outline

Cinvestav

Here, we have multiple errors

What can we do?



Sum Over All the Errors

We can do the following

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

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

Cinvestav

Sum Over All the Errors

We can do the following

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

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 $x^T = (1, x')$.

Sum Over All the Errors

We can do the following

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

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 $\boldsymbol{x}^T = (\mathbf{1}, \boldsymbol{x'})$.

Cinvestav

We can use a trick

The following function

$$g_{ideal}\left(\boldsymbol{x}\right) = \begin{pmatrix} 1 & x_1 & x_2 & \dots & x_d \end{pmatrix} \begin{pmatrix} w_0 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix} = \boldsymbol{x}^T \boldsymbol{w}$$

We can rewrite the error equation as

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

We can use a trick

The following function

$$g_{ideal}\left(\boldsymbol{x}\right)=\begin{pmatrix} 1 & x_1 & x_2 & \dots & x_d \end{pmatrix}\begin{pmatrix} w_0 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{pmatrix}=\boldsymbol{x}^T\boldsymbol{w}$$

We can rewrite the error equation as

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

Furthermore

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

$$\boldsymbol{X}\boldsymbol{w} = \begin{pmatrix} 1 & (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{d+1} \end{pmatrix}$$

Note about other representations

We could have $\boldsymbol{x}^T = (x_1, x_2, ..., x_d, 1)$ thus

$$\boldsymbol{X} = \begin{pmatrix} (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d & 1 \\ & & \vdots & & \vdots & \vdots \\ (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d & 1 \\ & & \vdots & & \vdots & \vdots \\ (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d & 1 \end{pmatrix} \quad (15)$$

Cinvestav

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

With the Data Matrix

$$\boldsymbol{X}w = \begin{pmatrix} \boldsymbol{x}_1^T \boldsymbol{w} \\ \boldsymbol{x}_2^T \boldsymbol{w} \\ \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} \quad (16)$$

Therefore

We have that

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

Then, we have the following equality

$$\begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} & y_2 - \boldsymbol{x}_2^T \boldsymbol{w} & y_3 - \boldsymbol{x}_3^T \boldsymbol{w} & \cdots & y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w} \right)^2$$

Therefore

We have that

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

Then, we have the following equality

$$\begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} & y_2 - \boldsymbol{x}_2^T \boldsymbol{w} & y_3 - \boldsymbol{x}_3^T \boldsymbol{w} & \cdots & y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} \begin{pmatrix} y_1 - \boldsymbol{x}_1^T \boldsymbol{w} \\ y_2 - \boldsymbol{x}_2^T \boldsymbol{w} \\ y_3 - \boldsymbol{x}_3^T \boldsymbol{w} \\ \vdots \\ y_4 - \boldsymbol{x}_N^T \boldsymbol{w} \end{pmatrix} = \sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w} \right)^2$$

Then, we have

The following equality

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

We can expand our quadratic formula!!!

Thus

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

We can expand our quadratic formula!!!

Thus

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

Now

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

We can expand our quadratic formula!!!

Thus

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

Now

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

Cinvestav

Therefore

We have the following equivalences

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

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

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

Therefore

We have the following equivalences

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

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

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

Then, when we derive by $\boldsymbol{w}$

We have then

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

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

Then, when we derive by $\boldsymbol{w}$

We have then

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

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

Making this equal to the zero row vector

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

$$\left[-2y^T X + 2w^T\left(X^T X\right)\right]^T$$

$$-2X^T y + 2\left(X^T X\right) w$$

Then, when we derive by $\boldsymbol{w}$

We have then

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

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

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

Then, when we derive by $\boldsymbol{w}$

We have then

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

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

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

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

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

Then, when we derive by $\boldsymbol{w}$

We have then

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

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

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

Solving for $\boldsymbol{w}$

We have then

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

Note: $\boldsymbol{X}^T\boldsymbol{X}$ 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

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

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

Thus, How we get the discriminant function?

Any Ideas?

Cinvestav

The Final Discriminant Function

Very Simple!!!

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

Cinvestav

Outline

Cinvestav

Also Known as Karhunen-Loeve Transform

Setup

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

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 $\{\boldsymbol{x}_n\}$ with $n = 1, 2, ..., N$ and $\boldsymbol{x_n} \in R^d$.

Goal

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

Dimensional Variance

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

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

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

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

Dimensional Variance

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

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

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

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

Cinvestav

Now, Define

Given the data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \tag{24}$$

where $\boldsymbol{x}_i$ is a column vector

Construct the sample mean

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \tag{25}$$

Center data

$$\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \boldsymbol{x}_2 - \overline{\boldsymbol{x}}, ..., \boldsymbol{x}_N - \overline{\boldsymbol{x}} \tag{26}$$

Now, Define

Given the data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \tag{24}$$

where $\boldsymbol{x}_i$ is a column vector

Construct the sample mean

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \tag{25}$$

Center data

$$\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \boldsymbol{x}_2 - \overline{\boldsymbol{x}}, ..., \boldsymbol{x}_N - \overline{\boldsymbol{x}} \tag{26}$$

Now, Define

Given the data

$$\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N \tag{24}$$

where $\boldsymbol{x}_i$ is a column vector

Construct the sample mean

$$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \tag{25}$$

Center data

$$\boldsymbol{x}_1 - \overline{\boldsymbol{x}}, \boldsymbol{x}_2 - \overline{\boldsymbol{x}}, ..., \boldsymbol{x}_N - \overline{\boldsymbol{x}} \tag{26}$$

Cinvestav

Build the Sample Mean

The Covariance Matrix

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

Properties

1. The ijth value of S is equivalent to σ_{ij}^2.
2. The iith value of S is equivalent to σ_{ii}^2.

Build the Sample Mean

The Covariance Matrix

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

Properties

1. The ijth value of S is equivalent to σ_{ij}^2.
2. The iith value of S is equivalent to σ_{ii}^2.

Outline

Cinvestav

Using S to Project Data

For this we use a u_1

- with $u_1^T u_1 = 1$, an orthonormal vector

Question

- What is the Sample Variance of the Projected Data?

Using S to Project Data

For this we use a $\boldsymbol{u}_1$

- with $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$, an orthonormal vector

Question

- What is the Sample Variance of the Projected Data?

Using S to Project Data

For this we use a $\boldsymbol{u}_1$

- with $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$, an orthonormal vector

Question

- What is the Sample Variance of the Projected Data?

Outline

Thus we have

Variance of the projected data

$$\frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{u}_1\boldsymbol{x}_i-\boldsymbol{u}_1\overline{\boldsymbol{x}}\right]=\boldsymbol{u}_1^T S\boldsymbol{u}_1 \tag{28}$$

Use Lagrange Multipliers to Maximize

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

Thus we have

Variance of the projected data

$$\frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{u}_1\boldsymbol{x}_i - \boldsymbol{u}_1\overline{\boldsymbol{x}}\right] = \boldsymbol{u}_1^T S \boldsymbol{u}_1 \quad (28)$$

Use Lagrange Multipliers to Maximize

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

Derive by $\boldsymbol{u}_1$

We get

$$S\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \tag{30}$$

Then

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

If we left-multiply by $\boldsymbol{u}_1$

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

Derive by $\boldsymbol{u}_1$

We get

$$S\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \tag{30}$$

Then

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

If we left-multiply by $\boldsymbol{u}_1$

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

Derive by $\boldsymbol{u}_1$

We get

$$S\boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1 \tag{30}$$

Then

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

If we left-multiply by $\boldsymbol{u}_1$

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

Cinvestav

What about the second eigenvector $\boldsymbol{u}_2$

We have the following optimization problem

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

Lagrangian

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

What about the second eigenvector $\boldsymbol{u}_2$

We have the following optimization problem

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

Lagrangian

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

Cinvestav

Explanation

First the constrained minimization

- We want to to maximize $\boldsymbol{u}_2^T S \boldsymbol{u}_2$

Given that the second eigenvector is orthonormal

- We have then $\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1$

Under orthonormal vectors

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

Explanation

First the constrained minimization

- We want to to maximize $\boldsymbol{u}_2^T S \boldsymbol{u}_2$

Given that the second eigenvector is orthonormal

- We have then $\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1$

Under orthonormal vectors

- The covariance goes to zero

$$cov\left(\boldsymbol{u}_1, \boldsymbol{u}_2\right) = \boldsymbol{u}_2^T S \boldsymbol{u}_1 = \boldsymbol{u}_2 \lambda_1 \boldsymbol{u}_1 = \lambda_1 \boldsymbol{u}_1^T \boldsymbol{u}_2 = 0$$

Explanation

First the constrained minimization

- We want to to maximize $\boldsymbol{u}_2^T S \boldsymbol{u}_2$

Given that the second eigenvector is orthonormal

- We have then $\boldsymbol{u}_2^T \boldsymbol{u}_2 = 1$

Under orthonormal vectors

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

Cinvestav

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 $\boldsymbol{u}_2$

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

Then, we left multiply $\boldsymbol{u}_1$

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

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 $\boldsymbol{u}_2$

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

Then, we left multiply $\boldsymbol{u}_1$

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

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 $\boldsymbol{u}_2$

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

Then, we left multiply $\boldsymbol{u}_1$

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

Cinvestav

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

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

Implying

- 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

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

Implying

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

Thus

Variance will be the maximum when

$$u_1^T S u_1 = \lambda_1 \quad (32)$$

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

Thus

Variance will be the maximum when

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \quad (32)$$

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 $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M$ 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

Thus

Variance will be the maximum when

$$\boldsymbol{u}_1^T S \boldsymbol{u}_1 = \lambda_1 \quad (32)$$

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 $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_M$ 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

We have the following steps

Determine covariance matrix

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

Generate the decomposition

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

With

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

Cinvestav

We have the following steps

Determine covariance matrix

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

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

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

Generate the decomposition

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

With

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

Then

Project samples $\boldsymbol{x}_i$ into subspaces dim$=k$

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

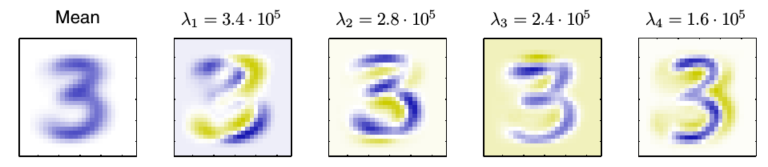
- With U_k is a matrix with k columns

Outline

Cinvestav

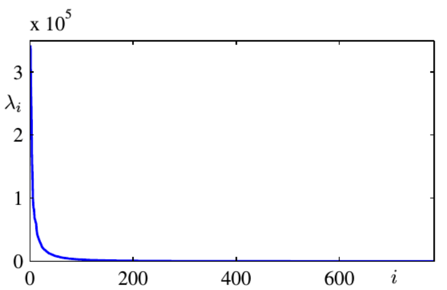
Example

From Bishop



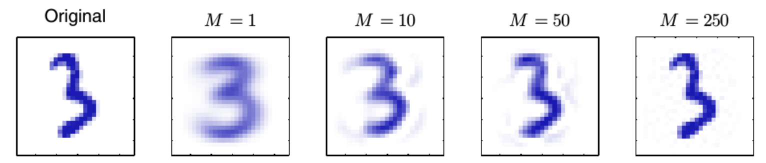
Example

From Bishop



Example

From Bishop



Outline

Cinvestav

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

1. They are usually not orthogonal.
2. There are not always enough eigenvectors.
3. $Ax = \lambda x$ 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

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

Cinvestav

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

1. They are usually not orthogonal.
2. There are not always enough eigenvectors.
3. $A\boldsymbol{x} = \lambda\boldsymbol{x}$ requires A to be square.

Cinvestav

Therefore, we can look at the following problem

We have a series of vectors

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

Then imagine a set of projection vectors and differences

$$\{\beta_1, \beta_2, ..., \beta_d\} \text{ and } \{\alpha_1, \alpha_2, ..., \alpha_d\}$$

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

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

Then imagine a set of projection vectors and differences

$$\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_d\} \text{ and } \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d\}$$

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

Cinvestav

Therefore, we can look at the following problem

We have a series of vectors

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

Then imagine a set of projection vectors and differences

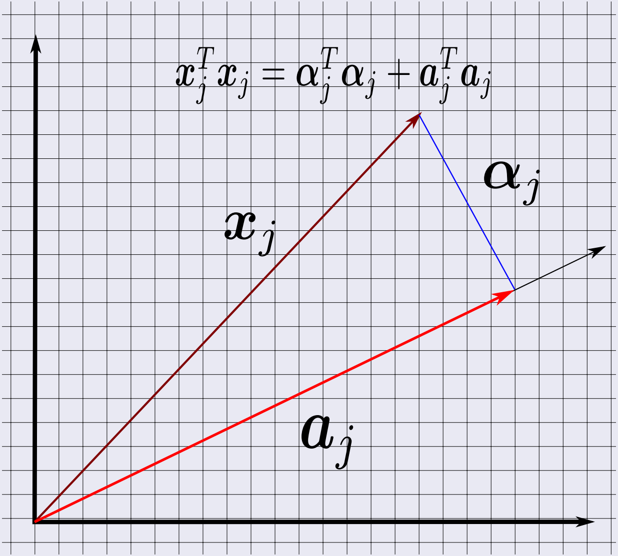
$$\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_d\} \text{ and } \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, ..., \boldsymbol{\alpha}_d\}$$

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

Using the Hypotenuse

A little bit of Geometry, we get



Therefore

We have two possible quantities for each j

$$
\begin{aligned}
\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j &= \boldsymbol{x}_j^T \boldsymbol{x}_j - \boldsymbol{a}_j^T \boldsymbol{a}_j \\
\boldsymbol{a}_j^T \boldsymbol{a}_j &= \boldsymbol{x}_j^T \boldsymbol{x}_j - \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j
\end{aligned}
$$

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

$$
\min \sum_{j=1}^{n} \alpha_j^T \alpha_j
$$

$$
\max \sum_{j=1}^{n} a_j^T a_j
$$

Therefore

We have two possible quantities for each j

$$\boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j = \boldsymbol{x}_j^T \boldsymbol{x}_j - \boldsymbol{a}_j^T \boldsymbol{a}_j$$
$$\boldsymbol{a}_j^T \boldsymbol{a}_j = \boldsymbol{x}_j^T \boldsymbol{x}_j - \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j$$

Then, we can minimize and maximize given that $\boldsymbol{x}_j^T \boldsymbol{x}_j$ is a constant

$$\min \sum_{j=1}^{n} \boldsymbol{\alpha}_j^T \boldsymbol{\alpha}_j$$
$$\max \sum_{j=1}^{n} \boldsymbol{a}_j^T \boldsymbol{a}_j$$

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

An example of this

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

Then, using what is know as slack variables

$$Ax + A'x = b$$

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

$$\left(Ax + A'x\right)_i \to y_i \text{ and } x' \geq 0$$

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

An example of this

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

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

$$\left(A\boldsymbol{x} + A'\boldsymbol{x}\right)_i \to y_i \text{ and } \boldsymbol{x}' \geq 0$$

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

An example of this

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

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

$$\left(A\boldsymbol{x} + A'\boldsymbol{x}\right)_i \rightarrow y_i \text{ and } \boldsymbol{x}' \geq 0$$

Then, we have that

Example

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

Element in the column space of dimensionality have three dimensions

- But in the row space their dimension is 2

Properties

Then, we have that

Example

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

Element in the column space of dimensionality have three dimensions

- But in the row space their dimension is 2

Properties

Then, we have that

Example

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

Element in the column space of dimensionality have three dimensions

- But in the row space their dimension is 2

Properties

We have then

Stack such vectors that in the d-dimensional space

- 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.

We have then

Stack such vectors that in the d-dimensional space

- 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 $\boldsymbol{v}$ such that length of the projection is maximized.

Why? Do you remember the Projection to a single vector $\boldsymbol{p}$?

Definition of the projection under unitary vector

$$\boldsymbol{p} = \frac{\boldsymbol{v}^T \boldsymbol{a}_i}{\boldsymbol{v}^T \boldsymbol{v}} \boldsymbol{v} = \left[\boldsymbol{v}^T \boldsymbol{a}_i \right] \boldsymbol{v}$$

Therefore the length of the projected vector is

$$\left\| \left[\boldsymbol{v}^T \boldsymbol{a}_i \right] \boldsymbol{v} \right\| = \left| \boldsymbol{v}^T \boldsymbol{a}_i \right|$$

Why? Do you remember the Projection to a single vector $\boldsymbol{p}$?

Definition of the projection under unitary vector

$$\boldsymbol{p} = \frac{\boldsymbol{v}^T \boldsymbol{a}_i}{\boldsymbol{v}^T \boldsymbol{v}} \boldsymbol{v} = \left[\boldsymbol{v}^T \boldsymbol{a}_i \right] \boldsymbol{v}$$

Therefore the length of the projected vector is

$$\left\| \left[\boldsymbol{v}^T \boldsymbol{a}_i \right] \boldsymbol{v} \right\| = \left| \boldsymbol{v}^T \boldsymbol{a}_i \right|$$

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

$$
\|A\boldsymbol{v}\| = \sqrt{\sum_{i=1}^{d} \left(\boldsymbol{a}_i^T \boldsymbol{v}\right)^2}
$$

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

$$\|A\boldsymbol{v}\| = \sqrt{\sum_{i=1}^{d} \left(\boldsymbol{a}_i^T \boldsymbol{v}\right)^2}$$

Then

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

$$\boldsymbol{v}_1 = \arg \max_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$$

Then, we can define the following singular value

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

Then

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

$$\boldsymbol{v}_1 = \arg\max_{\|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$$

Then, we can define the following singular value

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

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

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

Then, in a Greedy Fashion

The second singular vector $\boldsymbol{v}_2$

$$\boldsymbol{v}_2 = \arg \max_{\boldsymbol{v} \perp \boldsymbol{v}_1, \|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$$

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

$$\arg \max_{\substack{\boldsymbol{v} \perp \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r \\ \|\boldsymbol{v}\| = 1}} \|A\boldsymbol{v}\|$$

Then, in a Greedy Fashion

The second singular vector $\boldsymbol{v}_2$

$$\boldsymbol{v}_2 = \arg \max_{\boldsymbol{v} \perp \boldsymbol{v}_1, \|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$$

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

$$\arg \max_{\substack{\boldsymbol{v} \perp \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r \\ \|\boldsymbol{v}\| = 1}} \|A\boldsymbol{v}\|$$

Then, in a Greedy Fashion

The second singular vector $\boldsymbol{v}_2$

$$\boldsymbol{v}_2 = \arg \max_{\boldsymbol{v} \perp \boldsymbol{v}_1, \|\boldsymbol{v}\|=1} \|A\boldsymbol{v}\|$$

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

$$\arg \max_{\substack{\boldsymbol{v} \perp \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r \\ \|\boldsymbol{v}\| = 1}} \|A\boldsymbol{v}\|$$

Proving that the strategy is good

Theorem

- Let A be an $n \times d$ matrix where $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_r$ are the singular vectors defined above. For $1 \leq k \leq r$, let V_k be the subspace spanned by $\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_k$. Then for each k, V_k is the best-fit k-dimensional subspace for A.

Cinvestav

Proof

For $k = 1$

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

For any basis w_1, 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_1

- This can be a unit vector perpendicular to v_1 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

- $|A\boldsymbol{w}_1|^2 + |A\boldsymbol{w}_2|^2$ is the sum of the squared lengths of the projections of the rows of A to W.

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

- This can be a unit vector perpendicular to $\boldsymbol{v}_1$ 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

- $|A\boldsymbol{w}_1|^2 + |A\boldsymbol{w}_2|^2$ is the sum of the squared lengths of the projections of the rows of A to W.

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

- This can be a unit vector perpendicular to $\boldsymbol{v}_1$ projection in W.

Cinvestav

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

Therefore

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

Then

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

In a similar way for $k > 2$

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

Cinvestav

Do you remember $v_1 = \arg\max_{\|v\|=1} \|Av\|$?

Therefore

$$|Aw_1|^2 \leq |Av_1|^2 \text{ and } |Aw_2|^2 \leq |Av_2|^2$$

Then

$$|Aw_1|^2 + |Aw_2|^2 \leq |Av_1|^2 + |Av_2|^2$$

In a similar way for $k > 2$

- 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

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

Then

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

In a similar way for $k > 2$

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

Cinvestav

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.

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.
- 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.

Cinvestav

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.
- 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

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

Second

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

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

Singular Value Decomposition as Sums

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

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

Why?

$$u_1 A = U \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_R \end{pmatrix} V^T = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_R \end{pmatrix} \begin{pmatrix} \sigma_1 \boldsymbol{v}_1^T \\ \sigma_2 \boldsymbol{v}_2^T \\ \vdots \\ \sigma_R \boldsymbol{v}_R^T \end{pmatrix}$$

$$= \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^T + \cdots + \sigma_R \boldsymbol{u}_R \boldsymbol{v}_R^T$$

Singular Value Decomposition as Sums

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

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

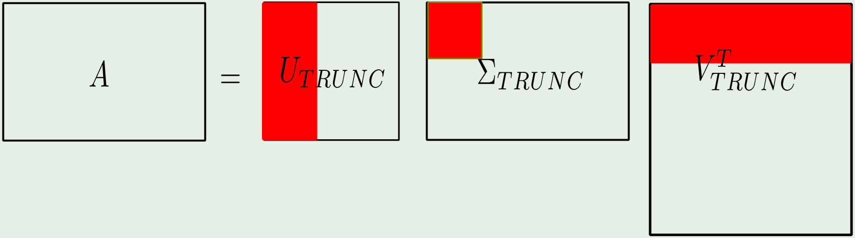
Why?

$$\begin{aligned} \boldsymbol{u}_1 A = U \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_R \end{pmatrix} V^T = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_R \end{pmatrix} \begin{pmatrix} \sigma_1 \boldsymbol{v}_1^T \\ \sigma_2 \boldsymbol{v}_2^T \\ \vdots \\ \sigma_R \boldsymbol{v}_R^T \end{pmatrix} \\ = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^T + \cdots + \sigma_R \boldsymbol{u}_R \boldsymbol{v}_R^T \end{aligned}$$

Cinvestav

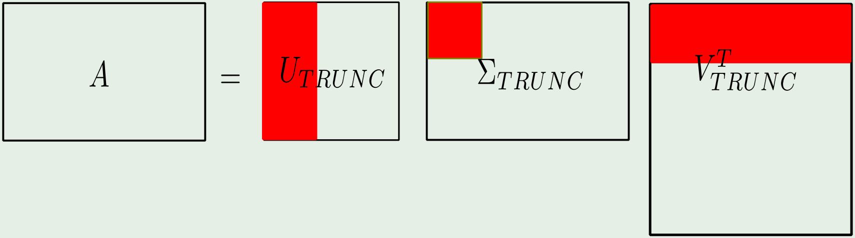
Truncating

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



Truncating

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

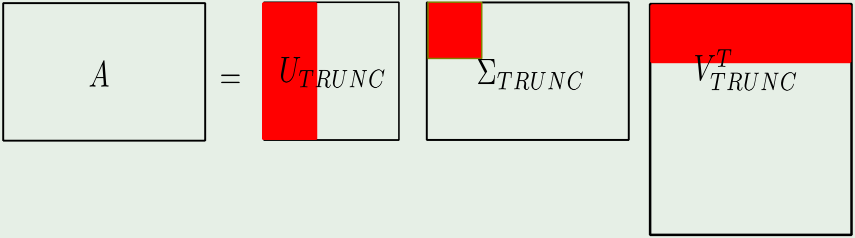


For a 512×512

- 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 $512{\times}80 + 80 + 80 \times 512 = 82,000$

Truncating

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

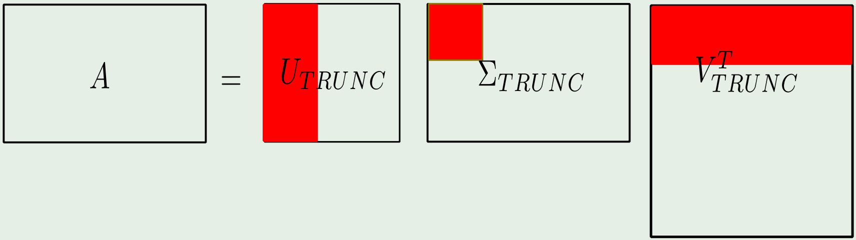


For a 512×512

- 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 $512 \times 80 + 80 + 80 \times 512 = 82,000$

Truncating

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 $512{\times}80 + 80 + 80 \times 512 = 82,000$