Introduction to Machine Learning

Introduction to Linear Classifiers

Andres Mendez-Vazquez

January 16, 2023

Outline

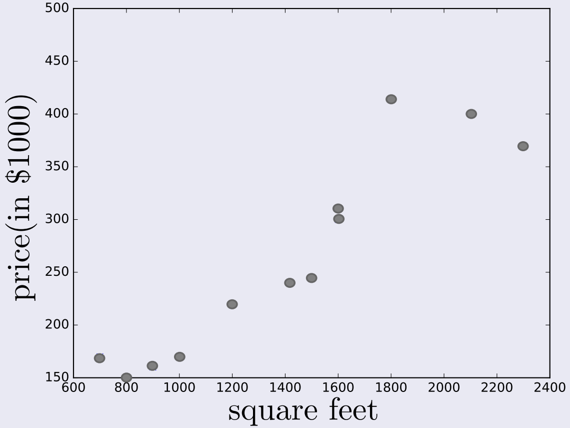
Cinvestav

Outline

Many Times, we have things as regression

We have this kind of data sets (House Prices per Square Feet)

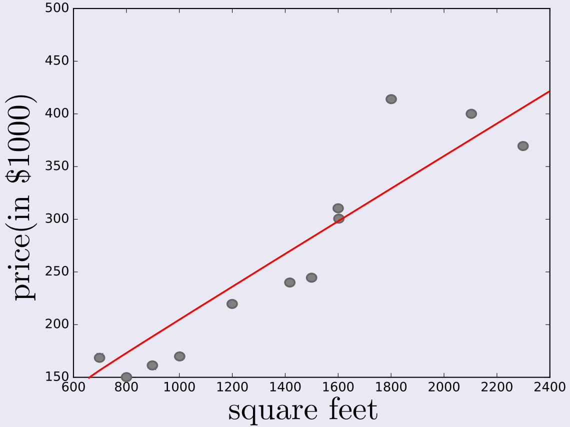
$$\begin{pmatrix} \text{Squared Feet} \\ \text{Price} \end{pmatrix} \rightarrow \begin{pmatrix} 2104 \\ 400 \end{pmatrix} \begin{pmatrix} 1800 \\ 460 \end{pmatrix} \begin{pmatrix} 1600 \\ 300 \end{pmatrix} \begin{pmatrix} 2300 \\ 370 \end{pmatrix} \cdots$$



Thus

We can adjust a line/hyperplane to be able to forecast prices

$$\begin{pmatrix} \text{Squared Feet} \\ \text{Price} \end{pmatrix} \rightarrow \begin{pmatrix} 2104 \\ 400 \end{pmatrix} \begin{pmatrix} 1800 \\ 460 \end{pmatrix} \begin{pmatrix} 1600 \\ 300 \end{pmatrix} \begin{pmatrix} 2300 \\ 370 \end{pmatrix} \cdots$$



Thus, Our Objective

To find such hyperplane

- To do forecasting on the prices of a house given its surface!!!

Here, where "Learning" Machine Learning style comes around

- Basically, the process defined in Machine Learning!!!

Thus, Our Objective

To find such hyperplane

- To do forecasting on the prices of a house given its surface!!!

Here, where "Learning" Machine Learning style comes around

- Basically, the process defined in Machine Learning!!!

Outline

Cinvestav

Regression[1, 2]

Intuition

- The regression model is a procedure that allows to estimate certain relationship that relates two or more variables with an output.

We have two types

- Linear Regression
- Non-Linear Regression

Regression[1, 2]

Intuition

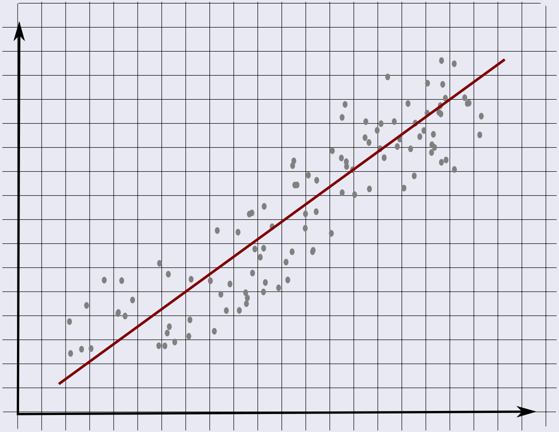
- The regression model is a procedure that allows to estimate certain relationship that relates two or more variables with an output.

We have two types

- Linear Regression
- Non-Linear Regression

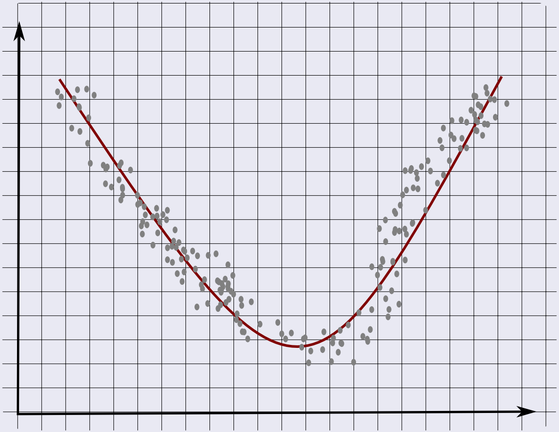
Linear Regression

We have something like



Non-Linear Regression

We have something like



As an Approximation

It is clear that in these univariate cases, we have

$$\{(x_i, y_i)\}_{i=1}^{N} \text{ with } x_i, y_i \in \mathbb{R}$$

Data to try to approximate by

$$\min_f \otimes_{i=1}^{N} g\{f(x_i) \oplus y_i\}$$

Where

- $\otimes, \oplus$ are binary operators
- g, f are functions

As an Approximation

It is clear that in these univariate cases, we have

$$\{(x_i, y_i)\}_{i=1}^{N} \text{ with } x_i, y_i \in \mathbb{R}$$

Data to try to approximate by

$$\min_{f} \otimes_{i=1}^{N} g\left\{f\left(x_i\right) \oplus y_i\right\}$$

Where

- $\otimes, \oplus$ are binary operators
- g, f are functions

As an Approximation

It is clear that in these univariate cases, we have

$$\{(x_i, y_i)\}_{i=1}^{N} \text{ with } x_i, y_i \in \mathbb{R}$$

Data to try to approximate by

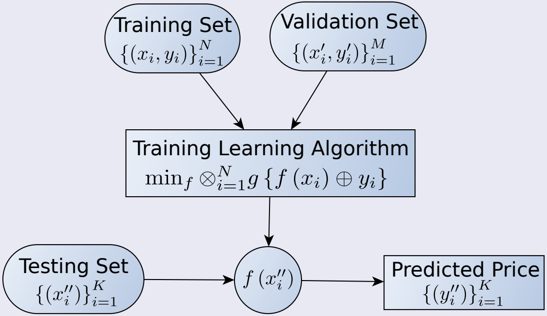
$$\min_f \otimes_{i=1}^{N} g \{f(x_i) \oplus y_i\}$$

Where

- $\otimes, \oplus$ are binary operators
- g, f are functions

Then, in Supervised Training[2]

We have the following process (x_i, y_i)



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 $\mathbb{R}^2$

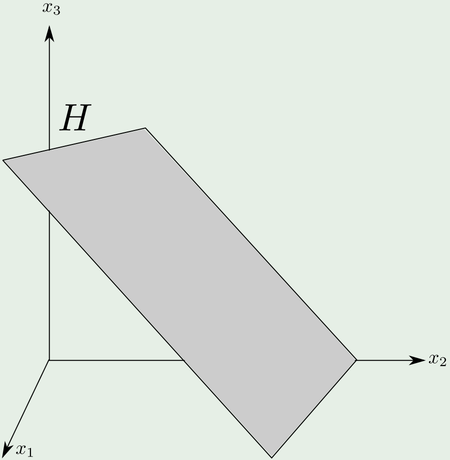
We have:

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

Cinvestav

Example

Hyperplane in $\mathbb{R}^3$

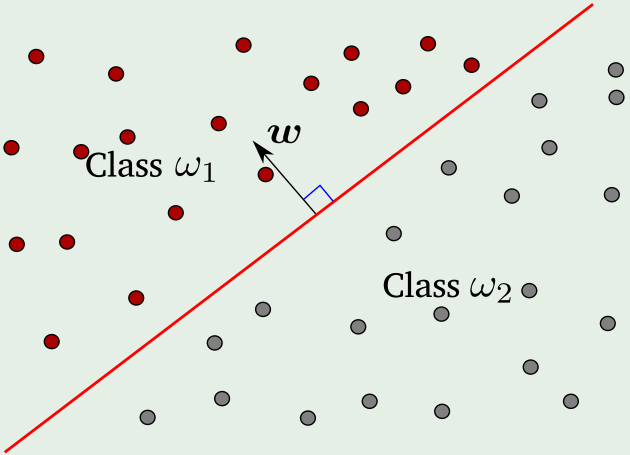


Outline

Cinvestav

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

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

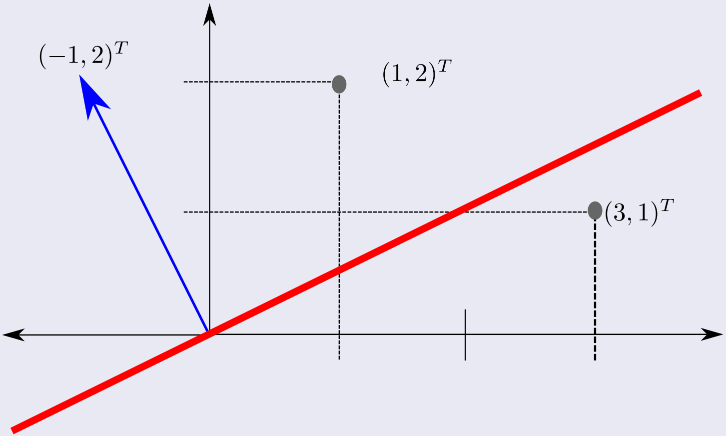
Hyperplane

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[3]

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[3]

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

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

The Decision Surface[3]

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

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

Defining a Decision Surface

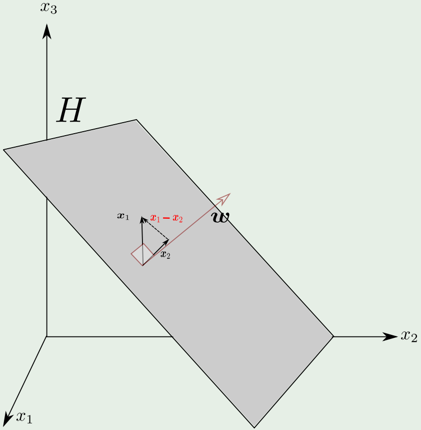
Then

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

Therefore

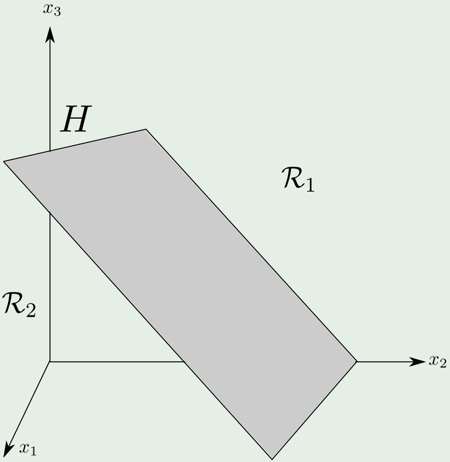
$\boldsymbol{x}_1 - \boldsymbol{x}_2$ lives in the hyperplane i.e. it is perpendicular to $\boldsymbol{w}^T$

- Remark: any vector in the hyperplane is a linear combination of elements in the plane.
- **Therefore any vector in the plane is perpendicular to $\boldsymbol{w}^T$**



Therefore

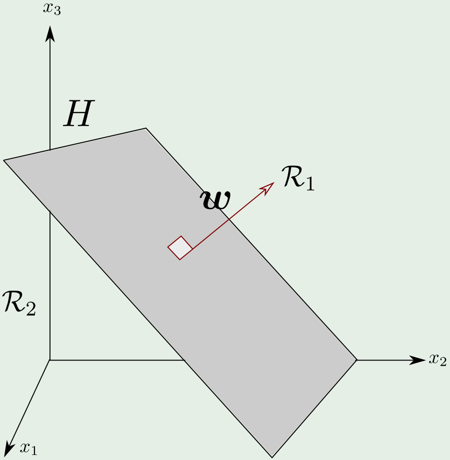
The space is split in two regions (Example in $\mathbb{R}^3$) by the hyperplane H



Outline

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.

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

- 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 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

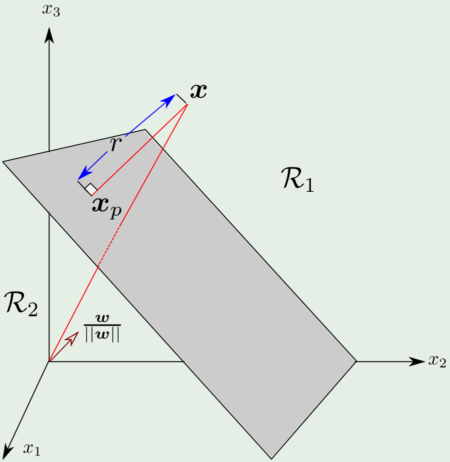
$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

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

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

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

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

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

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

Now

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

We have that

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

Then, we have

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

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

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

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

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.

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

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

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

Cinvestav

Outline

Cinvestav

We want to solve the independence of w_0

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

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

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 $x_0 = 1$

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

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

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

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

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

Remarks

- The addition of a constant component to 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.

Cinvestav

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.

Cinvestav

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.

Cinvestav

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:

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

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

Remember

Our original function

$$\min_{f} \otimes_{i=1}^{N} g\left\{f\left(x_i\right) \oplus y_i\right\}$$

Outline

Cinvestav

Initial Supposition - The Binary Problem

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 - The Binary Problem

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 - The Binary Problem

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 - The Binary Problem

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.

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?

Cinvestav

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?

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

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

We produce the following labels

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

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!!!

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

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

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

Where the e

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

Thus, we can do the following

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

Now, What?

Assume true function f is given by

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

Where the e

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

Thus, we can do the following

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

Thus, we have

What to do?

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

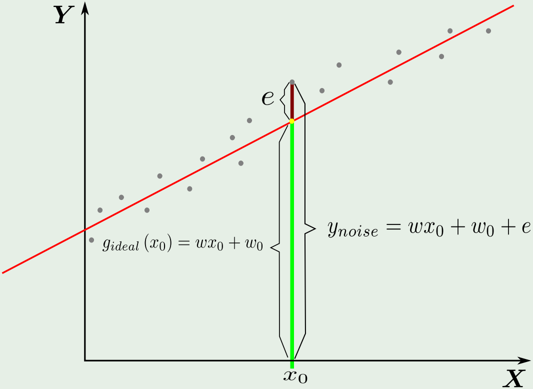
Graphically

Thus, we have

What to do?

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

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}

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

We use y_{ideal}

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

We will see

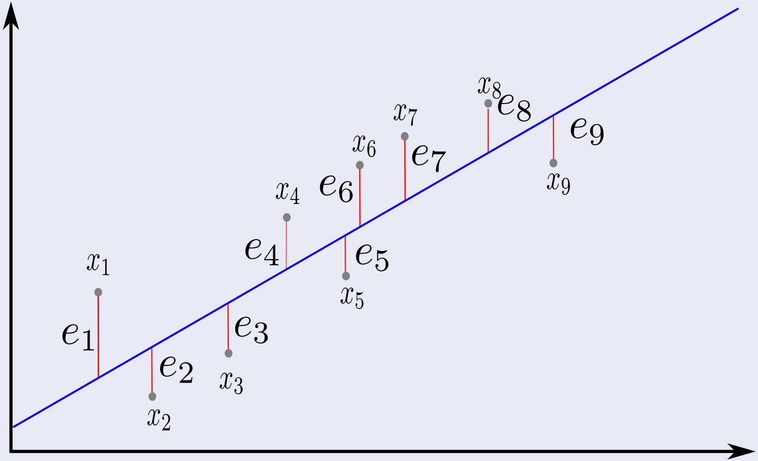
How the geometry will solve the problem with using these labels.

Outline

Cinvestav

Here, we have multiple errors

What can we do?



Sum Over All the Errors

We can do the following

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

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

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

Outline

Cinvestav

Assume that $x \in \mathbb{R}$

Then we have that the function looks like

$$f(x) = b_0 + b_1 x$$

Therefore the lose function looks like

$$L\left(b_1, b_2, \{x_i, y_i\}_{i=1}^{N}\right) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - b_0 - b_1 x]^2$$

Assume that $x \in \mathbb{R}$

Then we have that the function looks like

$$f(x) = b_0 + b_1 x$$

Therefore the lose function looks like

$$L\left(b_1, b_2, \{x_i, y_i\}_{i=1}^{N}\right) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \left[y_i - b_0 - b_1 x\right]^2$$

Cinvestav

Then, you use simple derivatives

Then, you have derivatives with respect to b_0

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_0} = -2 \sum_{i=1}^{N} \left[y_i - b_0 - b_1 x \right] = 0$$

Derivatives with respect to b_1

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_1} = -2 \sum_{i=1}^{N} \left[y_i - b_0 - b_1 x \right] x = 0$$

Then, you use simple derivatives

Then, you have derivatives with respect to b_0

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_0} = -2 \sum_{i=1}^{N} \left[y_i - b_0 - b_1 x \right] = 0$$

Derivatives with respect to b_1

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial b_1} = -2 \sum_{i=1}^{N} \left[y_i - b_0 - b_1 x \right] x = 0$$

Cinvestav

Previous equations are known as normal equations

Solving them

$$b_o = \overline{y} - b_1 \overline{x}$$

$$b_1 = \frac{\sum_{i=1}^{N} \left[x_i - \overline{x}\right]\left[y_i - \overline{y}\right]}{\sum_{i=1}^{N} \left[x_i - \overline{x}\right]^2}$$

Cinvestav

Outline

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

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

Cinvestav

We can use a trick

The following function

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

We can rewrite the error equation as

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

Cinvestav

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

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

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

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

Cinvestav

Then, we have

The following equality

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

We can expand our quadratic formula!!!

Thus

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

Now

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

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 $\boldsymbol{X}^T\boldsymbol{X}$ is invertible

Outline

Some Basic Definitions

Transpose of a Matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}^T = \begin{pmatrix} a_1 & a_2 & a_2 \end{pmatrix}$$

Some Basic Definitions

Transpose of a Matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} \\ a_{12} & a_{22} & a_{32} \\ a_{13} & a_{23} & a_{33} \end{pmatrix}$$

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}^T = \begin{pmatrix} a_1 & a_2 & a_2 \end{pmatrix}$$

Cinvestav

Additionally

We have [4]

Given A and B matrices:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

Given vectors x, y and a matrix A such that you can multiply them:

- $x^T A y = \left[x^T A y\right]^T = y^T A^T x$ given that the transpose of a number is the number itself.

Additionally

We have [4]

Given A and B matrices:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

Given vectors x, y and a matrix A such that you can multiply them:

- $x^T A y = \left[x^T A y\right]^T = y^T A^T x$ given that the transpose of a number is the number itself.

Additionally

We have [4]

Given A and B matrices:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

Given vectors x, y and a matrix A such that you can multiply them:

- $x^T A y = \left[x^T A y \right]^T = y^T A^T x$ given that the transpose of a number is the number itself.

Additionally

We have [4]

Given A and B matrices:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

Given vectors x, y and a matrix A such that you can multiply them:

- $x^T A y = \left[x^T A y\right]^T = y^T A^T x$ given that the transpose of a number is the number itself.

Additionally

We have [4]

Given A and B matrices:

- $(A+B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

Given vectors x, y and a matrix A such that you can multiply them:

- $x^T A y = \left[x^T A y\right]^T = y^T A^T x$ given that the transpose of a number is the number itself.

Some Basic Definitions for

Derivative on Matrices

$$\frac{dA\boldsymbol{x}}{d\boldsymbol{x}} = \frac{d\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}{d\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}$$

Cinvestav

Therefore

We have

$$\frac{d\begin{pmatrix} a_{11}x_1+ & a_{12}x_2+ & a_{13}x_3 \\ a_{21}x_1+ & a_{22}x_2+ & a_{23}x_3 \\ a_{31}x_1+ & a_{32}x_2+ & a_{33}x_3 \end{pmatrix}}{d\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}} = \dots$$

$$\begin{pmatrix} \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_1} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_2} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_3} \\ \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_1} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_2} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_3} \\ \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_1} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_2} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_3} \end{pmatrix} = \dots$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Therefore

We have

$$\frac{d\begin{pmatrix} a_{11}x_1 + & a_{12}x_2 + & a_{13}x_3 \\ a_{21}x_1 + & a_{22}x_2 + & a_{23}x_3 \\ a_{31}x_1 + & a_{32}x_2 + & a_{33}x_3 \end{pmatrix}}{d\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}} = \ldots$$

$$\begin{pmatrix} \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_1} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_2} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_3} \\ \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_1} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_2} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_3} \\ \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_1} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_2} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_3} \end{pmatrix} = \ldots$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Therefore

We have

$$\frac{d\begin{pmatrix} a_{11}x_1+ & a_{12}x_2+ & a_{13}x_3 \\ a_{21}x_1+ & a_{22}x_2+ & a_{23}x_3 \\ a_{31}x_1+ & a_{32}x_2+ & a_{33}x_3 \end{pmatrix}}{d\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}} = \dots$$

$$\begin{pmatrix} \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_1} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_2} & \frac{d(a_{11}x_1+a_{12}x_2+a_{13}x_3)}{dx_3} \\ \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_1} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_2} & \frac{d(a_{21}x_1+a_{22}x_2+a_{23}x_3)}{dx_3} \\ \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_1} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_2} & \frac{d(a_{31}x_1+a_{32}x_2+a_{33}x_3)}{dx_3} \end{pmatrix} = \dots$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Therefore

We have the following equivalences

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

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

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

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

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

Cinvestav

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

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

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

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

We have then

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

Making this equal to the zero row vector

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

We apply the transpose

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

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

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

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

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

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

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

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

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

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

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

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

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?

The Final Discriminant Function

Very Simple!!!

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

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $\boldsymbol{w} \in image\left(\boldsymbol{X}\right)$, then there is some $\boldsymbol{v} \in \mathbb{R}^{n}$ such that $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v}$.
- Hence, $\boldsymbol{X}^{+}\boldsymbol{w} = \boldsymbol{X}^{+}\boldsymbol{X}\boldsymbol{v} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{v} = \boldsymbol{v}$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\,(X) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $\boldsymbol{w} \in image\,(\boldsymbol{X})$, then there is some $\boldsymbol{v} \in \mathbb{R}^{n}$ such that $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v}$.
- Hence, $\boldsymbol{X}^{+}\boldsymbol{w} = \boldsymbol{X}^{+}\boldsymbol{X}\boldsymbol{v} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{v} = \boldsymbol{v}$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $w \in image\left(X\right)$, then there is some $v \in \mathbb{R}^{n}$ such that $w = Xv$.
- Hence, $\boldsymbol{X}^{+}w = \boldsymbol{X}^{+}\boldsymbol{X}v = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}v = v$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\,(X) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $w \in image\,(X)$, then there is some $v \in \mathbb{R}^{n}$ such that $w = Xv$.
- Hence, $\boldsymbol{X}^{+}w = \boldsymbol{X}^{+}\boldsymbol{X}v = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}v = v$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\,(X) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $w \in image\,(X)$, then there is some $v \in \mathbb{R}^{n}$ such that $w = Xv$.
- Hence, $\boldsymbol{X}^{+}w = \boldsymbol{X}^{+}\boldsymbol{X}v = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}v = v$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\,(X) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $\boldsymbol{w} \in image\,(\boldsymbol{X})$, then there is some $\boldsymbol{v} \in \mathbb{R}^{n}$ such that $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v}$.
- Hence, $\boldsymbol{X}^{+}\boldsymbol{w} = \boldsymbol{X}^{+}\boldsymbol{X}\boldsymbol{v} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{v} = \boldsymbol{v}$

Pseudo-inverse of a Matrix

Definition

Suppose that $X \in \mathbb{R}^{m \times n}$ and $rank\left(X\right) = m$. We call the matrix

$$\boldsymbol{X}^{+} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$$

the pseudo inverse of X.

Reason

$\boldsymbol{X}^{+}$ inverts $\boldsymbol{X}$ on its image

What?

- First a definition
  - If $\boldsymbol{w} \in image\left(\boldsymbol{X}\right)$, then there is some $\boldsymbol{v} \in \mathbb{R}^{n}$ such that $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v}$.
- Hence, $\boldsymbol{X}^{+}\boldsymbol{w} = \boldsymbol{X}^{+}\boldsymbol{X}\boldsymbol{v} = \left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{v} = \boldsymbol{v}$

Outline

Cinvestav

We have that

The Data Matrix

$$\boldsymbol{X} \in \mathbb{R}^{N \times (d+1)}$$

$$\boldsymbol{x}_i \in \mathbb{R}^d$$

We have that

The Data Matrix

$$\boldsymbol{X} \in \mathbb{R}^{N\times(d+1)}$$

$$\boldsymbol{x}_i \in \mathbb{R}^d$$

We have that

The Data Matrix

$$\boldsymbol{X} \in \mathbb{R}^{N\times(d+1)}$$

$$\boldsymbol{x_i} \in \mathbb{R}^{d}$$

The projected elements by the matrix

Definition $Image\left(\boldsymbol{X}\right)$

- The column space of a matrix $\boldsymbol{X}$ is the span (set of all possible linear combinations) of its column vectors.

$$Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_1^{col}, ..., \boldsymbol{X}_{d+1}^{col}\right\}$$

  - In other words, the image of a matrix $\boldsymbol{X}$ is all the vectors $\boldsymbol{X}\boldsymbol{v} \in \mathbb{R}^N$ with $\boldsymbol{v} \in \mathbb{R}^{d+1}$

Cinvestav

The Data Samples

The Data Samples

$$\boldsymbol{x_i} \in \mathbb{R}^d$$

Additionally, we have that

The Weight Vector $\boldsymbol{w}$

$$\boldsymbol{w} \in \mathbb{R}^{d+1}$$

What about the column space of X and the ideal input vector $\boldsymbol{y}$

$$X_i^{col}, \boldsymbol{y} \in \mathbb{R}^N$$

Cinvestav

Additionally, we have that

The Weight Vector $\boldsymbol{w}$

$$\boldsymbol{w} \in \mathbb{R}^{d+1}$$

What about the column space of $\boldsymbol{X}$ and the ideal input vector $\boldsymbol{y}$

$$\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{col}}, \boldsymbol{y} \in \mathbb{R}^{N}$$

We can now see where $\boldsymbol{y}$ is being projected

Basically $\boldsymbol{y}$, the list of real inputs is being projected into

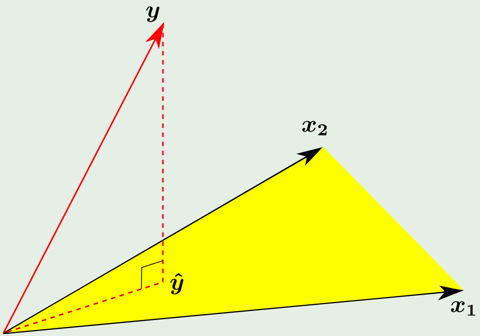
$$span\left\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\right\} \tag{22}$$

- by function $\widehat{\boldsymbol{y}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Cinvestav

Geometrically

Given a $\boldsymbol{y}$, you obtain a projected $\widehat{\boldsymbol{y}}$ through the projection function $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$

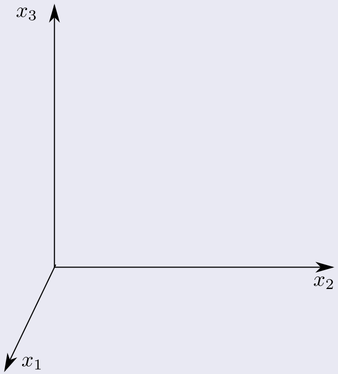


Outline

Cinvestav

Why? Assume that you are in $\mathbb{R}^3$

Something like



Simple but complex

A simple question[5]

- What are the projections of $b = (2, 3, 4)$ onto the z axis and the xy plane?
- Can we use matrices to talk about these projections?

First

We must have a projection matrix P with the following property:

$$P^2 = P$$

Why?

Ideas?

Simple but complex

A simple question[5]

- What are the projections of $b = (2, 3, 4)$ onto the z axis and the xy plane?
- Can we use matrices to talk about these projections?

First

We must have a projection matrix P with the following property:

$$P^2 = P$$

Why?

Ideas?

Simple but complex

A simple question[5]

- What are the projections of $b = (2, 3, 4)$ onto the z axis and the xy plane?
- Can we use matrices to talk about these projections?

First

We must have a projection matrix P with the following property:

$$P^2 = P$$

Why?

Ideas?

Cinvestav

Then, the Projection $P\boldsymbol{b}$

First

When $\boldsymbol{b}$ is projected onto a line, its projection $\boldsymbol{p}$ is the part of $\boldsymbol{b}$ along that line.

Second

When $\boldsymbol{b}$ is projected onto a plane, its projection $\boldsymbol{p}$ is the part of the plane.

Then, the Projection $P\boldsymbol{b}$

First

When $\boldsymbol{b}$ is projected onto a line, its projection $\boldsymbol{p}$ is the part of $\boldsymbol{b}$ along that line.

Second

When $\boldsymbol{b}$ is projected onto a plane, its projection $\boldsymbol{p}$ is the part of the plane.

In our case

The Projection Matrices for the coordinate systems

$$P_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, P_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, P_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Example

We have the following vector $\boldsymbol{b} = (2,3,4)^T$

Onto the $\boldsymbol{z}$ axis:

$$P_1 \boldsymbol{b} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}$$

What about the plane xy

Any idea?

Example

We have the following vector $\boldsymbol{b} = (2, 3, 4)^T$

Onto the $\boldsymbol{z}$ axis:

$$P_1 \boldsymbol{b} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix}$$

What about the plane xy

Any idea?

Cinvestav

We have something more complex

Something Notable

$$P_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Then

$$P_4 b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}$$

We have something more complex

Something Notable

$$P_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Then

$$P_4 \boldsymbol{b} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 0 \end{pmatrix}$$

Cinvestav

Assume the following

We have that

$\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n$ in $\mathbb{R}^m$.

Assume they are linearly independent

They span a subspace, we want projections into the subspace

We want to project b into such subspace

How do we do it?

Assume the following

We have that

$\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n$ in $\mathbb{R}^m$.

Assume they are linearly independent

They span a subspace, we want projections into the subspace

We want to project b into such subspace

How do we do it?

Assume the following

We have that

$\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n$ in $\mathbb{R}^m$.

Assume they are linearly independent

They span a subspace, we want projections into the subspace

We want to project $\boldsymbol{b}$ into such subspace

How do we do it?

This is the important part

Problem

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Something Notable

With $n = 1$ (only one vector a_1) this projection onto a line.

This line is the column space of A

Basically the columns are spanned by a single column.

This is the important part

Problem

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Something Notable

With $n = 1$ (only one vector a_1) this projection onto a line.

This line is the column space of A

Basically the columns are spanned by a single column.

This is the important part

Problem

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Something Notable

With $n = 1$ (only one vector a_1) this projection onto a line.

This line is the column space of A

Basically the columns are spanned by a single column.

Cinvestav

In General

The matrix has n columns $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n$

The combinations in $\mathbb{R}^m$ are vectors $A\boldsymbol{x}$ in the column space

We are looking for the particular combination

The nearest to the original $\boldsymbol{b}$

$$\boldsymbol{p} = A\widehat{\boldsymbol{x}}$$

In General

The matrix has n columns $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n$

The combinations in $\mathbb{R}^m$ are vectors $A\boldsymbol{x}$ in the column space

We are looking for the particular combination

The nearest to the original $\boldsymbol{b}$

$$\boldsymbol{p} = A\widehat{\boldsymbol{x}}$$

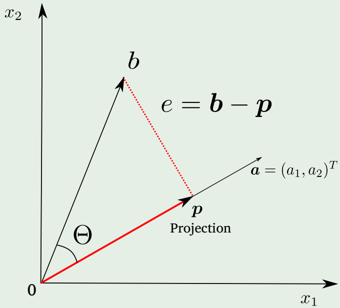
First

We look at the simplest case

The projection into a line...

With a little of Geometry

We have the following



Therefore

Using the fact that the projection is equal to

$$\boldsymbol{p} = x\boldsymbol{a}$$

Then, the error is equal to

$$\boldsymbol{e} = \boldsymbol{b} - x\boldsymbol{a}$$

We have that $\boldsymbol{a} \cdot \boldsymbol{e} = 0$

$$\boldsymbol{a} \cdot \boldsymbol{e} = \boldsymbol{a} \cdot (\boldsymbol{b} - x\boldsymbol{a}) = \boldsymbol{a} \cdot \boldsymbol{b} - x\boldsymbol{a} \cdot \boldsymbol{a} = 0$$

Therefore

Using the fact that the projection is equal to

$$\boldsymbol{p} = x\boldsymbol{a}$$

Then, the error is equal to

$$\boldsymbol{e} = \boldsymbol{b} - x\boldsymbol{a}$$

We have that $\boldsymbol{a} \cdot \boldsymbol{e} = 0$

$$\boldsymbol{a} \cdot \boldsymbol{e} = \boldsymbol{a} \cdot (\boldsymbol{b} - x\boldsymbol{a}) = \boldsymbol{a} \cdot \boldsymbol{b} - x\boldsymbol{a} \cdot \boldsymbol{a} = 0$$

Therefore

Using the fact that the projection is equal to

$$\boldsymbol{p} = x\boldsymbol{a}$$

Then, the error is equal to

$$\boldsymbol{e} = \boldsymbol{b} - x\boldsymbol{a}$$

We have that $\boldsymbol{a} \cdot \boldsymbol{e} = 0$

$$\boldsymbol{a} \cdot \boldsymbol{e} = \boldsymbol{a} \cdot (\boldsymbol{b} - x\boldsymbol{a}) = \boldsymbol{a} \cdot \boldsymbol{b} - x\boldsymbol{a} \boldsymbol{\cdot} \boldsymbol{a} \boldsymbol{=} \boldsymbol{0}$$

Cinvestav

Therefore

We have that

$$x = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\boldsymbol{a} \cdot \boldsymbol{a}} = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\boldsymbol{a}^T \boldsymbol{a}}$$

Or something quite simple

$$p = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\boldsymbol{a}^T \boldsymbol{a}} \boldsymbol{a}$$

Cinvestav

Therefore

We have that

$$x = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\boldsymbol{a} \cdot \boldsymbol{a}} = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\boldsymbol{a}^T \boldsymbol{a}}$$

Or something quite simple

$$\boldsymbol{p} = \frac{\boldsymbol{a}^T \boldsymbol{b}}{\boldsymbol{a}^T \boldsymbol{a}} \boldsymbol{a}$$

Cinvestav

By the Law of Cosines

Something Notable

$$\left\| \boldsymbol{a} - \boldsymbol{b} \right\|^2 = \left\| \boldsymbol{a} \right\|^2 + \left\| \boldsymbol{b} \right\|^2 - 2 \left\| \boldsymbol{a} \right\| \left\| \boldsymbol{b} \right\| \cos \Theta$$

We have

The following product

$$\boldsymbol{a}\cdot\boldsymbol{a}-2\boldsymbol{a}\cdot\boldsymbol{b}+\boldsymbol{b}\cdot\boldsymbol{b}=\left\|\boldsymbol{a}\right\|^{2}+\left\|\boldsymbol{b}\right\|^{2}-2\left\|\boldsymbol{a}\right\|\left\|\boldsymbol{b}\right\|\cos\Theta$$

Then

$$\boldsymbol{a}\cdot\boldsymbol{b}=\left\|\boldsymbol{a}\right\|\left\|\boldsymbol{b}\right\|\cos\Theta$$

We have

The following product

$$\boldsymbol{a} \cdot \boldsymbol{a} - 2\boldsymbol{a} \cdot \boldsymbol{b} + \boldsymbol{b} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2 - 2\,\|\boldsymbol{a}\|\,\|\boldsymbol{b}\|\cos\Theta$$

Then

$$\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|\cos\Theta$$

With Length

Using the Norm

$$\|\boldsymbol{p}\| = \left|\frac{\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}}\right| \|\boldsymbol{a}\| = \left|\frac{\|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \cos\Theta}{\|\boldsymbol{a}\|^2}\right| \|\boldsymbol{a}\| = \|\boldsymbol{b}\| \, |\cos\Theta|$$

Example

Project

$$\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ onto } \boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

Find

$$\boldsymbol{p} = x\boldsymbol{a}$$

Example

Project

$$\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ onto } \boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

Find

$$\boldsymbol{p} = x\boldsymbol{a}$$

Cinvestav

What about the Projection Matrix in general

We have

$$\boldsymbol{p} = \boldsymbol{a}x = \frac{\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}} = P\boldsymbol{b}$$

Then

$$P = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}$$

What about the Projection Matrix in general

We have

$$\boldsymbol{p} = \boldsymbol{a}x = \frac{\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{b}}{\boldsymbol{a}^T\boldsymbol{a}} = P\boldsymbol{b}$$

Then

$$P = \frac{\boldsymbol{a}\boldsymbol{a}^T}{\boldsymbol{a}^T\boldsymbol{a}}$$

Example

Find the projection matrix for

$$\boldsymbol{b} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ onto } \boldsymbol{a} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$$

What about the general case?

We have that

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Now you need a vector

Find the vector $\boldsymbol{x}$,find the projection $\boldsymbol{p} = A\boldsymbol{x}$,find the matrix P.

Again, the error is perpendicular to the space

$$\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$$

What about the general case?

We have that

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Now you need a vector

Find the vector $\boldsymbol{x}$,find the projection $\boldsymbol{p} = A\boldsymbol{x}$,find the matrix P.

Again, the error is perpendicular to the space

$$\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$$

What about the general case?

We have that

Find the combination $\boldsymbol{p} = x_1\boldsymbol{a}_1 + x_2\boldsymbol{a}_2 + \cdots + x_n\boldsymbol{a}_n$ closest to vector $\boldsymbol{b}$.

Now you need a vector

Find the vector $\boldsymbol{x}$,find the projection $\boldsymbol{p} = A\boldsymbol{x}$,find the matrix P.

Again, the error is perpendicular to the space

$$\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$$

Cinvestav

Therefore

The error $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$

$$\boldsymbol{a}_1^T \left(\boldsymbol{b} - A\boldsymbol{x}\right) = 0$$
$$\vdots$$
$$\boldsymbol{a}_n^T \left(\boldsymbol{b} - A\boldsymbol{x}\right) = 0$$

Or

$$\begin{bmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_n^T \end{bmatrix} \left[\boldsymbol{b} - A\boldsymbol{x}\right] = 0$$

Therefore

The error $\boldsymbol{e} = \boldsymbol{b} - A\boldsymbol{x}$

$$\begin{aligned} \boldsymbol{a}_1^T \left(\boldsymbol{b} - A\boldsymbol{x}\right) &= 0 \\ &\vdots \\ \boldsymbol{a}_n^T \left(\boldsymbol{b} - A\boldsymbol{x}\right) &= 0 \end{aligned}$$

Or

$$\begin{bmatrix} \boldsymbol{a}_1^T \\ \vdots \\ \boldsymbol{a}_n^T \end{bmatrix} \left[\boldsymbol{b} - A\boldsymbol{x}\right] = 0$$

Therefore

The Matrix with those rows is A^T

$$A^T\left(\boldsymbol{b}-A\boldsymbol{x}\right)=0$$

Therefore

$$A^T\boldsymbol{b}-A^TA\boldsymbol{x}=0$$

Or the most know form

$$\boldsymbol{x}=\left(A^TA\right)^{-1}A^T\boldsymbol{b}$$

Therefore

The Matrix with those rows is A^T

$$A^T\left(\boldsymbol{b}-A\boldsymbol{x}\right)=0$$

Therefore

$$A^T\boldsymbol{b}-A^TA\boldsymbol{x}=0$$

Or the most know form

$$\boldsymbol{x}=\left(A^TA\right)^{-1}A^T\boldsymbol{b}$$

Therefore

The Matrix with those rows is A^T

$$A^T (\boldsymbol{b} - A\boldsymbol{x}) = 0$$

Therefore

$$A^T\boldsymbol{b} - A^T A\boldsymbol{x} = 0$$

Or the most know form

$$\boldsymbol{x} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}$$

Therefore

The Projection is

$$\boldsymbol{p} = A\boldsymbol{x} = A\left(A^T A\right)^{-1} A^T \boldsymbol{b}$$

Therefore

$$P = A\left(A^T A\right)^{-1} A^T$$

Therefore

The Projection is

$$\boldsymbol{p} = A\boldsymbol{x} = A\left(A^{T}A\right)^{-1}A^{T}\boldsymbol{b}$$

Therefore

$$P = A\left(A^{T}A\right)^{-1}A^{T}$$

Cinvestav

The key step was $A^T\left[\boldsymbol{b}-A\boldsymbol{x}\right]=0$

Linear algebra gives this "normal equation"

1. Our subspace is the column space of A.
2. The error vector $\boldsymbol{b}-A\boldsymbol{x}$ is perpendicular to that column space.
3. Therefore $\boldsymbol{b}-A\boldsymbol{x}$ is in the nullspace of A^T

When A has independent columns, $A^T A$ is invertible

Theorem

$A^T A$ is invertible if and only if Ahas linearly independent columns.

Proof

Consider the following

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

Here, Ax is in the null space of A^T

- Remember the column space and null space of A^T are orthogonal complements.

And Ax an element in the column space of A

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

Cinvestav

Proof

Consider the following

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

Here, Ax is in the null space of A^T

- Remember the column space and null space of A^T are orthogonal complements.

And Ax an element in the column space of A

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

Proof

Consider the following

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

Here, Ax is in the null space of A^T

- Remember the column space and null space of A^T are orthogonal complements.

And Ax an element in the column space of A

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

Proof

If A has linearly independent columns

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

Then, the null space

$$Null\left(A^T A\right) = \{0\}$$

i.e $A^T A$ is full rank

- Then, $A^T A$ is invertible...

Proof

If A has linearly independent columns

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

Then, the null space

$$Null\left(A^T A\right) = \{0\}$$

i.e $A^T A$ is full rank

- Then, $A^T A$ is invertible...

Proof

If A has linearly independent columns

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

Then, the null space

$$Null\left(A^T A\right) = \{0\}$$

i.e $A^T A$ is full rank

- Then, $A^T A$ is invertible...

Cinvestav

Finally

Theorem

- When A has independent columns, $A^T A$ is square, symmetric and invertible.

Outline

Cinvestav

Geometric Interpretation

We have

The image of the mapping:

$$h : \boldsymbol{w} \longmapsto \boldsymbol{X}\boldsymbol{w}$$

$$h : \mathbb{R}^{d+1} \longmapsto \mathbb{R}^{N}$$

is a linear subspace of $\mathbb{R}^{N}$.

What about $\boldsymbol{w}$?

$\boldsymbol{w}$ moves through all points in $\mathbb{R}^{d+1}$ when being generated

- Thus, the function value $h(\boldsymbol{w}) = \boldsymbol{X}\boldsymbol{w}$ can move through all points in the image space:

$$image(\boldsymbol{X}) = span\left\{\boldsymbol{X}_1^{col}, \boldsymbol{X}_2^{col}, ..., \boldsymbol{X}_{d+1}^{col}\right\}$$

Additionally, each $\boldsymbol{w}$ defines one point in $span\left\{\boldsymbol{X}_1^{col}, \boldsymbol{X}_2^{col}, ..., \boldsymbol{X}_{d+1}^{col}\right\} \subseteq \mathbb{R}^N$

$$h(\boldsymbol{w}) = \boldsymbol{X}\boldsymbol{w} = \sum_{i=1}^{d+1} w_i \boldsymbol{X}_i^{col}.$$

Cinvestav

What about $\boldsymbol{w}$?

$\boldsymbol{w}$ moves through all points in $\mathbb{R}^{d+1}$ when being generated

- Thus, the function value $h(\boldsymbol{w}) = \boldsymbol{X}\boldsymbol{w}$ can move through all points in the image space:

$$image(\boldsymbol{X}) = span\left\{\boldsymbol{X}_1^{col}, \boldsymbol{X}_2^{col}, ..., \boldsymbol{X}_{d+1}^{col}\right\}$$

Additionally, each $\boldsymbol{w}$ defines one point in $span\left\{\boldsymbol{X}_1^{col}, \boldsymbol{X}_2^{col}, ..., \boldsymbol{X}_{d+1}^{col}\right\} \subseteq \mathbb{R}^N$

$$h(\boldsymbol{w}) = \boldsymbol{X}\boldsymbol{w} = \sum_{i=1}^{d+1} w_i \boldsymbol{X}_i^{col}.$$

Cinvestav

What about the optimality of $\boldsymbol{w}$?

We have a composition of functions that are convex

$$
\begin{aligned}
f(\boldsymbol{w}) &= \boldsymbol{w}^T \boldsymbol{x} \\
g(t) &= (y - t) \\
h(e) &= \sum_{i=1}^{n} e^2
\end{aligned}
$$

- Making the Least Squared Error a Convex function with a single minimum!!!

The derivative method produces a $\widehat{w}$

- Such that $\widehat{w}$ minimizes the distance $d(y, image(X))$.

What about the optimality of $\boldsymbol{w}$?

We have a composition of functions that are convex

$$
\begin{aligned}
f(\boldsymbol{w}) &= \boldsymbol{w}^T \boldsymbol{x} \\
g(t) &= (y - t) \\
h(e) &= \sum_{i=1}^{n} e^2
\end{aligned}
$$

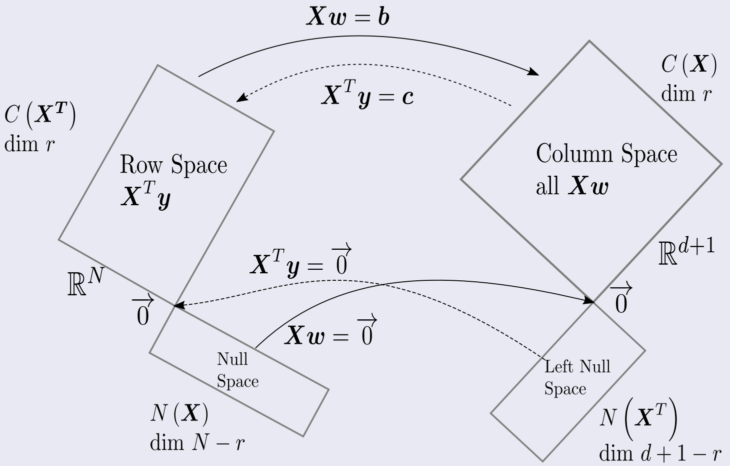
- Making the Least Squared Error a Convex function with a single minimum!!!

The derivative method produces a $\widehat{\boldsymbol{w}}$

- Such that $\widehat{\boldsymbol{w}}$ minimizes the distance $d(\boldsymbol{y}, image(\boldsymbol{X}))$.

This comes from the following representation

Given a matrix $\boldsymbol{X}$ ("Linear Algebra and Its Applications" by Hilbert Strang)



Outline

This Resolve Our Problem

With the Labels being chosen at the beginning

Question? Did you noticed the following?

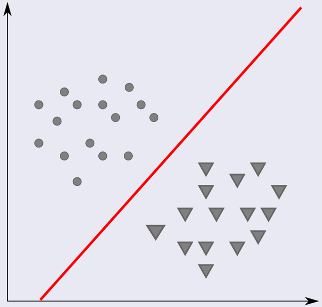
We assume a similar number of elements in both classes

This Resolve Our Problem

With the Labels being chosen at the beginning

Question? Did you noticed the following?

We assume a similar number of elements in both classes



Outline

Cinvestav

Multi-Class Solution

What to do?

1. We might reduce the problem to $c-1$ two-class problems.
2. We might use $\frac{c(c-1)}{2}$ linear discriminants, one for every pair of classes.

However

Multi-Class Solution

What to do?

1. We might reduce the problem to $c-1$ two-class problems.
2. We might use $\frac{c(c-1)}{2}$ linear discriminants, one for every pair of classes.

However

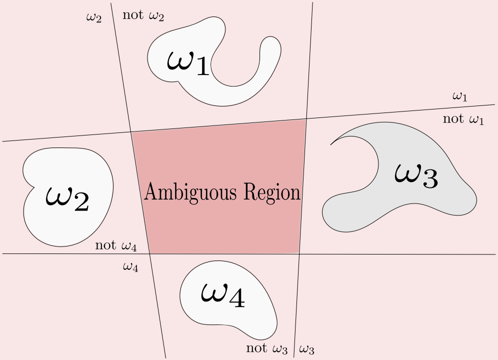
Cinvestav

Multi-Class Solution

What to do?

1. We might reduce the problem to $c-1$ two-class problems.
2. We might use $\frac{c(c-1)}{2}$ linear discriminants, one for every pair of classes.

However



What to Do?

Define c linear discriminant functions

$$g_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_{i0} \text{ for } i = 1, ..., c \tag{23}$$

This is known as a linear machine

Rule: if $g_k(\boldsymbol{x}) > g_j(\boldsymbol{x})$ for all $j \neq k \Longrightarrow \boldsymbol{x} \in \omega_k$

What to Do?

Define c linear discriminant functions

$$g_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_{i0} \text{ for } i = 1, ..., c \tag{23}$$

This is known as a linear machine

Rule: if $g_k(\boldsymbol{x}) > g_j(\boldsymbol{x})$ for all $j \neq k \implies \boldsymbol{x} \in \omega_k$

Nice Properties (It can be proved!!!)

1. Decision Regions are Singly Connected.
2. Decision Regions are Convex.

What to Do?

Define c linear discriminant functions

$$g_i(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_{i0} \text{ for } i = 1, ..., c \tag{23}$$

This is known as a **linear machine**

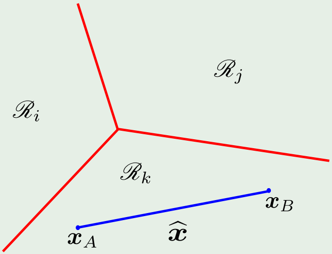
Rule: if $g_k(\boldsymbol{x}) > g_j(\boldsymbol{x})$ for all $j \neq k \Longrightarrow \boldsymbol{x} \in \omega_k$

Nice Properties (It can be proved!!!)

1. Decision Regions are Singly Connected.
2. Decision Regions are Convex.

Proof of Properties

Proof



Actually quite simple

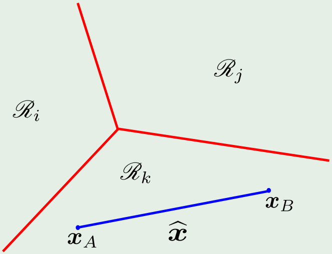
Given

$$y = \lambda x_A + (1 - \lambda) x_B$$

with $\lambda \in (0, 1)$.

Proof of Properties

Proof



Actually quite simple

Given

$$\boldsymbol{y} = \lambda \boldsymbol{x}_A + (1 - \lambda)\, \boldsymbol{x}_B$$

with $\lambda \in (0, 1)$.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B\right) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda) w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda) g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda) g_j(\boldsymbol{x}_A) \\
&> g_j\left(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B\right) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- $\boldsymbol{y}$ belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of $\boldsymbol{y}$.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\,\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda)\,w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)\,g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)\,g_j(\boldsymbol{x}_A) \\
&> g_j\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\,\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda)\, w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)\, g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)\, g_j(\boldsymbol{x}_A) \\
&> g_j\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\,\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda)\,w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)\,g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)\,g_j(\boldsymbol{x}_A) \\
&> g_j\left(\lambda \boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda) w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda) g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda) g_j(\boldsymbol{x}_A) \\
&> g_j(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) + w_0 \\
&= \lambda \boldsymbol{w}^T \boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\boldsymbol{w}^T \boldsymbol{x}_B + (1-\lambda) w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda) g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda) g_j(\boldsymbol{x}_A) \\
&> g_j(\lambda \boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda\boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) + w_0 \\
&= \lambda\boldsymbol{w}^T\boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\,\boldsymbol{w}^T\boldsymbol{x}_B + (1-\lambda)\,w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)\,g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)\,g_j(\boldsymbol{x}_A) \\
&> g_j\left(\lambda\boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- y belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of y.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T\left(\lambda\boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right) + w_0\\
&= \lambda\boldsymbol{w}^T\boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\,\boldsymbol{w}^T\boldsymbol{x}_B + (1-\lambda)\,w_0\\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)\,g_k(\boldsymbol{x}_A)\\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)\,g_j(\boldsymbol{x}_A)\\
&> g_j\left(\lambda\boldsymbol{x}_A + (1-\lambda)\,\boldsymbol{x}_B\right)\\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- $\boldsymbol{y}$ belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of $\boldsymbol{y}$.

Proof of Properties

We know that

$$
\begin{aligned}
g_k(\boldsymbol{y}) &= \boldsymbol{w}^T(\lambda\boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) + w_0 \\
&= \lambda\boldsymbol{w}^T\boldsymbol{x}_A + \lambda w_0 + (1-\lambda)\boldsymbol{w}^T\boldsymbol{x}_B + (1-\lambda)w_0 \\
&= \lambda g_k(\boldsymbol{x}_A) + (1-\lambda)g_k(\boldsymbol{x}_A) \\
&> \lambda g_j(\boldsymbol{x}_A) + (1-\lambda)g_j(\boldsymbol{x}_A) \\
&> g_j(\lambda\boldsymbol{x}_A + (1-\lambda)\boldsymbol{x}_B) \\
&> g_j(\boldsymbol{y})
\end{aligned}
$$

For all $j \neq k$

Or...

- $\boldsymbol{y}$ belongs to an area k defined by the rule!!!
- This area is **Convex** and **Singly Connected** because the definition of $\boldsymbol{y}$.

However!!!

No so nice properties!!!

- **It limits the power of classification for multi-objective function.**

How do we train this Linear Machine?

We know that each ω_k class is described by

$$g_k(\boldsymbol{x}) = \boldsymbol{w}_{\boldsymbol{k}}^{\boldsymbol{T}}\boldsymbol{x} + w_0 \text{ where } k = 1, ..., c$$

We then design a single machine

$$g(x) = W^T x \quad (24)$$

How do we train this Linear Machine?

We know that each ω_k class is described by

$$g_k(\boldsymbol{x}) = \boldsymbol{w_k^T x} + w_0 \text{ where } k = 1, ..., c$$

We then design a single machine

$$g(\boldsymbol{x}) = \boldsymbol{W^T x} \tag{24}$$

Where

We have the following

$$
\boldsymbol{W}^T = \begin{pmatrix} 1 & w_{11} & w_{12} & \cdots & w_{1d} \\ 1 & w_{21} & w_{22} & \cdots & w_{2d} \\ 1 & w_{31} & w_{32} & \cdots & w_{3d} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w_{c1} & w_{c2} & \cdots & w_{cd} \end{pmatrix} \tag{25}
$$

What about the labels?

OK, we know how to do with 2 classes, What about many classes?

Where

We have the following

$$\boldsymbol{W}^T = \begin{pmatrix} 1 & w_{11} & w_{12} & \cdots & w_{1d} \\ 1 & w_{21} & w_{22} & \cdots & w_{2d} \\ 1 & w_{31} & w_{32} & \cdots & w_{3d} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & w_{c1} & w_{c2} & \cdots & w_{cd} \end{pmatrix} \tag{25}$$

What about the labels?

OK, we know how to do with 2 classes, What about many classes?

How do we train this Linear Machine?

Use a vector $\boldsymbol{t}_i$ with dimensionality c to identify each element at each class

We have then the following dataset

$$\{\boldsymbol{x}_i, \boldsymbol{t}_i\} \text{ for } i = 1, 2, ..., N$$

We build the following Matrix of Vectors

$$T = \begin{pmatrix} \boldsymbol{t}_1^T \\ \boldsymbol{t}_2^T \\ \vdots \\ \boldsymbol{t}_{N-1}^T \\ \boldsymbol{t}_N^T \end{pmatrix} \quad (26)$$

How do we train this Linear Machine?

Use a vector $\boldsymbol{t}_i$ with dimensionality c to identify each element at each class

We have then the following dataset

$$\{\boldsymbol{x}_i, \boldsymbol{t}_i\} \text{ for } i = 1, 2, ..., N$$

We build the following Matrix of Vectors

$$\boldsymbol{T} = \begin{pmatrix} \boldsymbol{t}_1^T \\ \boldsymbol{t}_2^T \\ \vdots \\ \boldsymbol{t}_{N-1}^T \\ \boldsymbol{t}_N^T \end{pmatrix} \tag{26}$$

Examples for the $\boldsymbol{t}_i$

Vectors like (One Shot Representation)

$$x_i \neq 0, i \text{ Class} \rightarrow \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Another possible vector

$$x_i \neq -1, i \text{ Class} \rightarrow \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

Examples for the $\boldsymbol{t}_i$

Vectors like (One Shot Representation)

$$x_i \neq 0, i \text{ Class} \rightarrow \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Another possible vector

$$x_i \neq -1, i \text{ Class} \rightarrow \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

Cinvestav

Thus, we create the following Matrix

A Matrix containing all the required information

$$\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T} \tag{27}$$

Where we have the following vector

$$\left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] \tag{28}$$

Remark: It is the vector result of multiplication of row i of $\boldsymbol{X}$ against $\boldsymbol{W}$ on $\boldsymbol{X}\boldsymbol{W}$.

That is compared to the vector $\boldsymbol{t}_i^T$ on $\boldsymbol{T}$ by using the subtraction of vectors

$$\boldsymbol{e}_i = \left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] - \boldsymbol{t}_i^T \tag{29}$$

Thus, we create the following Matrix

A Matrix containing all the required information

$$\boldsymbol{X}\boldsymbol{W} - \boldsymbol{T} \tag{27}$$

Where we have the following vector

$$\left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] \tag{28}$$

Remark: It is the vector result of multiplication of row i of $\boldsymbol{X}$ against $\boldsymbol{W}$ on $\boldsymbol{X}\boldsymbol{W}$.

That is compared to the vector $\boldsymbol{t}_i^T$ on $\boldsymbol{T}$ by using the subtraction of vectors

$$\boldsymbol{e}_i = \left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] - \boldsymbol{t}_i^T \tag{29}$$

Thus, we create the following Matrix

A Matrix containing all the required information

$$\boldsymbol{X}\boldsymbol{W} - \boldsymbol{T} \tag{27}$$

Where we have the following vector

$$\left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] \tag{28}$$

Remark: It is the vector result of multiplication of row i of $\boldsymbol{X}$ against $\boldsymbol{W}$ on $\boldsymbol{X}\boldsymbol{W}$.

That is compared to the vector $\boldsymbol{t}_i^T$ on $\boldsymbol{T}$ by using the subtraction of vectors

$$e_i = \left[\boldsymbol{x}_i^T\boldsymbol{w}_1, \boldsymbol{x}_i^T\boldsymbol{w}_2, \boldsymbol{x}_i^T\boldsymbol{w}_3, ..., \boldsymbol{x}_i^T\boldsymbol{w}_c\right] - \boldsymbol{t}_i^T \tag{29}$$

Thus, we create the following Matrix

A Matrix containing all the required information

$$\boldsymbol{XW} - \boldsymbol{T} \tag{27}$$

Where we have the following vector

$$\left[\boldsymbol{x}_i^T \boldsymbol{w}_1, \boldsymbol{x}_i^T \boldsymbol{w}_2, \boldsymbol{x}_i^T \boldsymbol{w}_3, ..., \boldsymbol{x}_i^T \boldsymbol{w}_c\right] \tag{28}$$

Remark: It is the vector result of multiplication of row i of $\boldsymbol{X}$ against $\boldsymbol{W}$ on $\boldsymbol{XW}$.

That is compared to the vector $\boldsymbol{t}_i^T$ on $\boldsymbol{T}$ by using the subtraction of vectors

$$e_i = \left[\boldsymbol{x}_i^T \boldsymbol{w}_1, \boldsymbol{x}_i^T \boldsymbol{w}_2, \boldsymbol{x}_i^T \boldsymbol{w}_3, ..., \boldsymbol{x}_i^T \boldsymbol{w}_c\right] - \boldsymbol{t}_i^T \tag{29}$$

What do we want?

We want the quadratic error

$$\frac{1}{2}e_i^2$$

This specific quadratic errors are at the diagonal of the matrix

$$(\boldsymbol{XW}-\boldsymbol{T})^T(\boldsymbol{XW}-\boldsymbol{T})$$

We can use the trace function to generate the desired total error of

$$J(\cdot)=\frac{1}{2}\sum_{i=1}^{N}e_i^2 \tag{30}$$

What do we want?

We want the quadratic error

$$\frac{1}{2}e_i^2$$

This specific quadratic errors are at the diagonal of the matrix

$$\left(\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T}\right)^T\left(\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T}\right)$$

We can use the trace function to generate the desired total error of

$$J\left(\cdot\right)=\frac{1}{2}\sum_{i=1}^{N}e_i^2 \tag{30}$$

What do we want?

We want the quadratic error

$$\frac{1}{2}e_i^2$$

This specific quadratic errors are at the diagonal of the matrix

$$\left(\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T}\right)^T\left(\boldsymbol{X}\boldsymbol{W}-\boldsymbol{T}\right)$$

We can use the trace function to generate the desired total error of

$$J\left(\cdot\right)=\frac{1}{2}\sum_{i=1}^{N}e_i^2 \tag{30}$$

Cinvestav

Then

The trace allows to express the total error

$$J\left(\boldsymbol{W}\right)=\frac{1}{2}Trace\left\{\left(\boldsymbol{XW}-\boldsymbol{T}\right)^{T}\left(\boldsymbol{XW}-\boldsymbol{T}\right)\right\} \tag{31}$$

Thus, we have by the same derivative method

$$\boldsymbol{W}=\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}\boldsymbol{T}=\boldsymbol{X}^{+}\boldsymbol{T} \tag{32}$$

Cinvestav

Then

The trace allows to express the total error

$$J(\boldsymbol{W}) = \frac{1}{2} Trace\left\{(\boldsymbol{XW} - \boldsymbol{T})^T (\boldsymbol{XW} - \boldsymbol{T})\right\} \tag{31}$$

Thus, we have by the same derivative method

$$\boldsymbol{W} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T\boldsymbol{T} = \boldsymbol{X}^{+}\boldsymbol{T} \tag{32}$$

How do we obtain the discriminant?

Thus, we obtain the discriminant

$$g(\boldsymbol{x}) = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x} = \boldsymbol{T}^{\boldsymbol{T}}\left(\boldsymbol{X}^{+}\right)^{\boldsymbol{T}}\boldsymbol{x} \tag{33}$$

Cinvestav

Outline

Cinvestav

Outline

Let me show you the covariance matrix

We have in matrix notation

$$S = \frac{1}{N-1}\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)^T\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)$$

Thus, $X^T X$

It looks a lot like a covariance matrix

Actually, the dependency observed in matrix $X^T X$ between its columns!!!

- It is the same dependency observed between the features in the data X after the featured have been centered by $\overline{x}$.

Let me show you the covariance matrix

We have in matrix notation

$$S = \frac{1}{N-1}\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)^T\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)$$

Thus, $\boldsymbol{X}^T\boldsymbol{X}$

It looks a lot like a covariance matrix

Actually, the dependency observed in matrix $X^T X$ between its columns!!!

- It is the same dependency observed between the features in the data X after the featured have been centered by $\overline{x}$.

Let me show you the covariance matrix

We have in matrix notation

$$S = \frac{1}{N-1}\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)^T\left(X - \mathbf{1}\overline{\boldsymbol{x}}^T\right)$$

Thus, $\boldsymbol{X}^T\boldsymbol{X}$

It looks a lot like a covariance matrix

Actually, the dependency observed in matrix $\boldsymbol{X}^T\boldsymbol{X}$ between its columns!!!

- It is the same dependency observed between the features in the data $\boldsymbol{X}$ after the featured have been centered by $\overline{\boldsymbol{x}}$.

Cinvestav

Thus

We can apply a similar analysis...

- To obtain some of the possible cases that make $X^T X$ singular

A Classical One

- If there is a interdependence between features
  - Meaning some feature is an exact linear combination of the other features.
  - The $X^T X$ matrix of the features will be singular.

Cinvestav

Thus

We can apply a similar analysis...

- To obtain some of the possible cases that make $X^T X$ singular

A Classical One

- If there is a interdependence between features
  - Meaning some feature is an exact linear combination of the other features.
  - The $X^T X$ matrix of the features will be singular.

When does this happen?

First

Number of features is equal or greater than the number of samples.

Second

Two or more features sum up to a constant

- For example, $x_2 - 5x_{10} = 0$

Third

Two features are identical or differ merely in mean or variance.

When does this happen?

First

Number of features is equal or greater than the number of samples.

Second

Two or more features sum up to a constant

- For example, $x_2 - 5x_{10} = 0$

Third

Two features are identical or differ merely in mean or variance.

When does this happen?

First

Number of features is equal or greater than the number of samples.

Second

Two or more features sum up to a constant

- For example, $x_2 - 5x_{10} = 0$

Third

Two features are identical or differ merely in mean or variance.

Nevertheless

The least squares coefficients $\widehat{\boldsymbol{w}}$ are not uniquely defined.

- The fitted values $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{w}}$ are still the projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

Additionally

Duplicate observations in a data set

- It will lead the matrix toward singularity.

Cautionary Tale

- When doing some sort of imputation (Adding missing features), it is always beneficial (from both statistical and mathematical view) to add some noise to the imputed data.

This can happen in the preprocessing phase

- Be careful.

Additionally

Duplicate observations in a data set

- It will lead the matrix toward singularity.

Cautionary Tale

- When doing some sort of imputation (Adding missing features), it is always beneficial (from both statistical and mathematical view) to add some noise to the imputed data.

This can happen in the preprocessing phase

- Be careful.

Additionally

Duplicate observations in a data set

- It will lead the matrix toward singularity.

Cautionary Tale

- When doing some sort of imputation (Adding missing features), it is always beneficial (from both statistical and mathematical view) to add some noise to the imputed data.

This can happen in the preprocessing phase

- Be careful.

Also

It can happen also that

- $\boldsymbol{X}^T\boldsymbol{X}$ could be almost not invertible, making Least Squares numerically unstable.

Statistical consequence

- High variance of predictions.

Also

It can happen also that

- $\boldsymbol{X}^T\boldsymbol{X}$ could be almost not invertible, making Least Squares numerically unstable.

Statistical consequence

- High variance of predictions.

When can this happen?

The non-full-rank case occurs

- Most often when one or more qualitative (Categorical Variables/Dummy Variables) inputs are coded in a redundant fashion.

How do we solve this?

- Re-encode or dropping redundant columns in X.

Most regression software packages

- They detect these redundancies and automatically implement some strategies for removing them.

When can this happen?

The non-full-rank case occurs

- Most often when one or more qualitative (Categorical Variables/Dummy Variables) inputs are coded in a redundant fashion.

How do we solve this?

- Re-encode or dropping redundant columns in $\boldsymbol{X}$.

Most regression software packages

- They detect these redundancies and automatically implement some strategies for removing them.

Cinvestav

When can this happen?

The non-full-rank case occurs

- Most often when one or more qualitative (Categorical Variables/Dummy Variables) inputs are coded in a redundant fashion.

How do we solve this?

- Re-encode or dropping redundant columns in $\boldsymbol{X}$.

Most regression software packages

- They detect these redundancies and automatically implement some strategies for removing them.

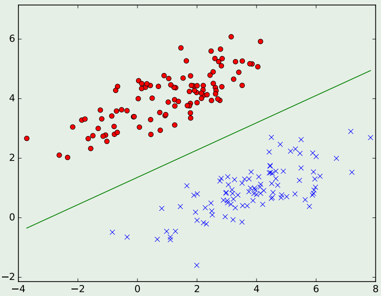
Outline

Cinvestav

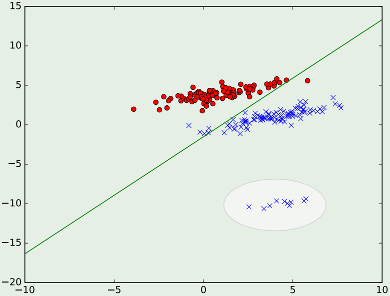
Issues with Least Squares

Problem with Outliers

No Outliers



Outliers



Outline

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Cinvestav

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Cinvestav

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Cinvestav

Problems with a High Number of Dimensions

In Many Modern Problems

- Many dimensions/features/predictors (possibly thousands).

Only a few of these may be important

- It needs some form of feature selection.
- Possible some type of regularization.

Why?

- Least Square Error Regression treats all dimensions equally.
- Relevant dimensions might be averaged with irrelevant ones.

Outline

Cinvestav

Outline

Cinvestav

We will start using some statistics

We want to obtain sampling properties for $\widehat{\boldsymbol{w}}$

For this remember:

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

For this assume,

- The observations y_i are uncorrelated and have constant variance σ^2.
- The x_i are fixed = not random.

We will start using some statistics

We want to obtain sampling properties for $\widehat{\boldsymbol{w}}$

For this remember:

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

For this assume,

- The observations y_i are uncorrelated and have constant variance σ^2.
- The $\boldsymbol{x}_i$ are fixed = not random.

Then, we have the variance-covariance matrix

We have

$$Var\left(\widehat{\boldsymbol{w}}\right) = Var\left[\left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T y}\right]$$

We have the following equivalence

$$Var\left(Ay\right) = AVar\left(y\right)A^T$$

Then, we have the variance-covariance matrix

We have

$$Var\left(\widehat{\boldsymbol{w}}\right) = Var\left[\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}\right]$$

We have the following equivalence

$$Var\left(A\boldsymbol{y}\right) = AVar\left(\boldsymbol{y}\right)A^{T}$$

Cinvestav

Therefore

Something Notable

$$
\begin{aligned}
Var\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Var\left(\boldsymbol{y}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2 I\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

Given that

$$
Var\left(y\right) = \begin{bmatrix} Var\left(y_1\right) & Cov\left(y_1, y_2\right) & \cdots & Cov\left(y_1, y_N\right) \\ Cov\left(y_2, y_1\right) & \cdots Var\left(y_2\right) & \cdots & Cov\left(y_2, y_N\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(y_N, y_1\right) & Cov\left(y_N, y_2\right) & \cdots & Var\left(y_N\right) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \sigma^2 \end{bmatrix}
$$

Therefore

Something Notable

$$
\begin{aligned}
Var\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Var\left(\boldsymbol{y}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2 I\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

Given that

$$
Var\left(\boldsymbol{y}\right) = \begin{bmatrix} Var\left(y_1\right) & Cov\left(y_1, y_2\right) & \cdots & Cov\left(y_1, y_N\right) \\ Cov\left(y_2, y_1\right) & \cdots Var\left(y_2\right) & \cdots & Cov\left(y_2, y_N\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(y_N, y_1\right) & Cov\left(y_N, y_2\right) & \cdots & Var\left(y_N\right) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \sigma^2 \end{bmatrix}
$$

Therefore

Something Notable

$$
\begin{aligned}
Var\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Var\left(\boldsymbol{y}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2 I\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

Given that

$$
Var\left(\boldsymbol{y}\right) = \begin{bmatrix} Var\left(y_1\right) & Cov\left(y_1, y_2\right) & \cdots & Cov\left(y_1, y_N\right) \\ Cov\left(y_2, y_1\right) & \cdots Var\left(y_2\right) & \cdots & Cov\left(y_2, y_N\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(y_N, y_1\right) & Cov\left(y_N, y_2\right) & \cdots & Var\left(y_N\right) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \sigma^2 \end{bmatrix}
$$

Therefore

Something Notable

$$
\begin{aligned}
Var\left[\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T Var\left(\boldsymbol{y}\right)\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\\
&= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\sigma^2 I\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\\
&= \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
\end{aligned}
$$

Given that

$$
Var\left(\boldsymbol{y}\right) = \begin{bmatrix} Var\left(y_1\right) & Cov\left(y_1, y_2\right) & \cdots & Cov\left(y_1, y_N\right)\\ Cov\left(y_2, y_1\right) & \cdots Var\left(y_2\right) & \cdots & Cov\left(y_2, y_N\right)\\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(y_N, y_1\right) & Cov\left(y_N, y_2\right) & \cdots & Var\left(y_N\right)\end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & & \sigma^2\end{bmatrix}
$$

Cinvestav

Thus

Typically, we can use the following unbiased estimator

$$\widehat{\sigma}^2 = \frac{1}{N-d-1}\sum_{i=1}^{N}(y_i - \widehat{y}_i)$$

- Which is an unbiased estimator $E\left[\widehat{\sigma}^2\right] = \sigma^2$.

If we have the following relation

$$Y = E\left(Y|X_1, X_2, ..., X_d\right) + \epsilon$$

Where

- $\epsilon \sim N\left(0, \sigma^2\right)$

Thus

Typically, we can use the following unbiased estimator

$$\widehat{\sigma}^2 = \frac{1}{N-d-1}\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)$$

- Which is an unbiased estimator $E\left[\widehat{\sigma}^2\right] = \sigma^2$.

If we have the following relation

$$Y = E\left(Y|X_1, X_2, ..., X_d\right) + \epsilon$$

Where

- $\epsilon \sim N\left(0, \sigma^2\right)$

Thus

Typically, we can use the following unbiased estimator

$$\widehat{\sigma}^2 = \frac{1}{N-d-1}\sum_{i=1}^{N}(y_i - \widehat{y}_i)$$

- Which is an unbiased estimator $E\left[\widehat{\sigma}^2\right] = \sigma^2$.

If we have the following relation

$$Y = E\left(Y|X_1, X_2, ..., X_d\right) + \epsilon$$

Where

- $\epsilon \sim N\left(0, \sigma^2\right)$

Cinvestav

Then

We have

$$\widehat{\boldsymbol{w}} \sim N\left(\boldsymbol{w}, \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

Thus, we can be a little bit smart

$$H_0 : w_j = 0$$
$$H_1 : w_j \neq 0$$

To test for Hypothesis $w_j = 0$, we get the following $z-$score

$$z_j = \frac{\widehat{w}_j - w_j}{\widehat{\sigma}\sqrt{v_j}} = \frac{\widehat{w}_j}{\widehat{\sigma}\sqrt{v_j}} \text{ with } v_j \text{ the } j^{th} \text{ diagonal element at } \left(X^TX\right)^{-1}$$

Then

We have

$$\widehat{\boldsymbol{w}} \sim N\left(\boldsymbol{w}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$$

Thus, we can be a little bit smart

$$\begin{aligned} H_0 : & w_j = 0 \\ H_1 : & w_j \neq 0 \end{aligned}$$

To test for Hypothesis $w_j = 0$, we get the following $z-$score

$$z_j = \frac{\widehat{w}_j - w_j}{\widehat{\sigma}\sqrt{v_j}} = \frac{\widehat{w}_j}{\widehat{\sigma}\sqrt{v_j}} \text{ with } v_j \text{ the } j^{th} \text{ diagonal element at } \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$$

Then

We have

$$\widehat{\boldsymbol{w}} \sim N\left(\boldsymbol{w}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

Thus, we can be a little bit smart

$$\begin{aligned} H_0 : & w_j = 0 \\ H_1 : & w_j \neq 0 \end{aligned}$$

To test for Hypothesis $w_j = 0$, we get the following $z-$score

$$z_j = \frac{\widehat{w}_j - w_j}{\widehat{\sigma}\sqrt{v_j}} = \frac{\widehat{w}_j}{\widehat{\sigma}\sqrt{v_j}}$$ with v_j the j^{th} diagonal element at $\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$

Therefore

$z_j \sim t_{N-d-1}$ a t-student distribution

- Therefore, a large(absolute) value of z_j will lead to rejection of the Null Hypothesis

Therefore

You can use the simple rule:

- Accept H_0 remove the feature
- Reject H_0 keep the feature

However

There are still more techniques for feature selection quite more advanced...

Cinvestav

Therefore

$z_j \sim t_{N-d-1}$ a t-student distribution

- Therefore, a large(absolute) value of z_j will lead to rejection of the Null Hypothesis

Therefore

You can use the simple rule:

- Accept H_0 remove the feature
- Reject H_0 keep the feature

However

There are still more techniques for feature selection quite more advanced...

Therefore

$z_j \sim t_{N-d-1}$ a t-student distribution

- Therefore, a large(absolute) value of z_j will lead to rejection of the Null Hypothesis

Therefore

You can use the simple rule:

- Accept H_0 remove the feature
- Reject H_0 keep the feature

However

There are still more techniques for feature selection quite more advanced...

Therefore

$z_j \sim t_{N-d-1}$ a t-student distribution

- Therefore, a large(absolute) value of z_j will lead to rejection of the Null Hypothesis

Therefore

You can use the simple rule:

- Accept H_0 remove the feature
- Reject H_0 keep the feature

However

There are still more techniques for feature selection quite more advanced...

Therefore

$z_j \sim t_{N-d-1}$ a t-student distribution

- Therefore, a large(absolute) value of z_j will lead to rejection of the Null Hypothesis

Therefore

You can use the simple rule:

- Accept H_0 remove the feature
- Reject H_0 keep the feature

However

There are still more techniques for feature selection quite more advanced...

Outline

What to Do About Numerical Stability?

Definition

- A matrix which is not invertible is also called a **singular** matrix.
- A matrix which is invertible (not singular) is called **regular**.

What is the Meaning?

Imagine the following in $\mathbb{R}^3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Given that the columns are vectors

They span a subspace for those column vectors in $\mathbb{R}^3$

$$span\left\{ \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}, \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right\}$$

What to Do About Numerical Stability?

Definition

- A matrix which is not invertible is also called a **singular** matrix.
- A matrix which is invertible (not singular) is called **regular**.

What is the Meaning?

Imagine the following in $\mathbb{R}^3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Given that the columns are vectors

They span a subspace for those column vectors in $\mathbb{R}^3$

$$span\left\{ \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}, \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right\}$$

What to Do About Numerical Stability?

Definition

- A matrix which is not invertible is also called a **singular** matrix.
- A matrix which is invertible (not singular) is called **regular**.

What is the Meaning?

Imagine the following in $\mathbb{R}^3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Given that the columns are vectors

They span a subspace for those column vectors in $\mathbb{R}^3$

$$span\left\{ \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}, \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right\}$$

What to Do About Numerical Stability?

Definition

- A matrix which is not invertible is also called a **singular** matrix.
- A matrix which is invertible (not singular) is called **regular**.

What is the Meaning?

Imagine the following in $\mathbb{R}^3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Given that the columns are vectors

They span a subspace for those column vectors in $\mathbb{R}^3$

$$span \left\{ \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}, \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix}, \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right\}$$

Relation with the Rank

If a matrix is singular

Its Rank is less than 3, i.e :

1. The subspace is squashed into a plane.
2. The subspace is squashed into a line.
3. The subspace in the WORST CASE into a point.

Relation with the Rank

If a matrix is singular

Its Rank is less than 3, i.e :

1. The subspace is squashed into a plane.
2. The subspace is squashed into a line.
3. The subspace in the WORST CASE into a point.

Relation with the Rank

If a matrix is singular

Its Rank is less than 3, i.e :

1. The subspace is squashed into a plane.
2. The subspace is squashed into a line.
3. The subspace in the WORST CASE into a point.

Relation with the Rank

If a matrix is singular

Its Rank is less than 3, i.e :

1. The subspace is squashed into a plane.
2. The subspace is squashed into a line.
3. The subspace in the WORST CASE into a point.

Remember

That, we have

$$\boldsymbol{v} = \lambda_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \lambda_2 \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} + \lambda_3 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix}$$

Thus, if for example, the matrix projects into a plane

$$\boldsymbol{v} = \lambda_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \lambda_2 \left[\alpha_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \alpha_2 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right] + \lambda_3 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix}$$

$$= c_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + c_2 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \quad \text{with } c_1 = \lambda_1 + \alpha_1 \lambda_2,\ c_2 = \alpha_2 \lambda_2 + \lambda_3$$

Remember

That, we have

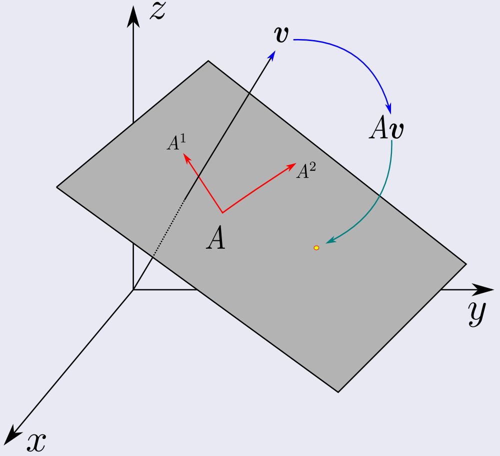
$$\boldsymbol{v} = \lambda_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \lambda_2 \begin{pmatrix} a_{12} \\ a_{22} \\ a_{32} \end{pmatrix} + \lambda_3 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix}$$

Thus, if for example, the matrix projects into a plane

$$\begin{aligned} \boldsymbol{v} =& \lambda_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \lambda_2 \left[\alpha_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + \alpha_2 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \right] + \lambda_3 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \\ =& c_1 \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} + c_2 \begin{pmatrix} a_{13} \\ a_{23} \\ a_{33} \end{pmatrix} \text{ with } c_1 = \lambda_1 + \alpha_1 \lambda_2, c_2 = \alpha_2 \lambda_2 + \lambda_3 \end{aligned}$$

For Example

We have a squashing into a plane



Computational Intuition

First Intuition

A singular matrix maps an entire linear subspace into a single point.

Second Intuitions

If a matrix maps points far away from each other to points very close to each other, it almost behaves like a singular matrix.

Computational Intuition

First Intuition

A singular matrix maps an entire linear subspace into a single point.

Second Intuitions

If a matrix maps points far away from each other to points very close to each other, it almost behaves like a singular matrix.

Thus

Mapping is related to the eigenvalues!!!

- **Large positive eigenvalues $\Rightarrow$ the mapping is large!!!**
- **Small positive eigenvalues $\Rightarrow$ the mapping is small!!!**

Thus

Mapping is related to the eigenvalues!!!

- **Large positive eigenvalues $\Rightarrow$ the mapping is large!!!**
- **Small positive eigenvalues $\Rightarrow$ the mapping is small!!!**

There is a statement to support this

All this comes from the following statement

A positive semi-definite matrix A is singular $\iff$ smallest eigenvalue is 0

Consequence for Statistics

If a statistical prediction involves the inverse of an almost-singular matrix, the predictions become unreliable (high variance).

There is a statement to support this

All this comes from the following statement

A positive semi-definite matrix A is singular $\iff$ smallest eigenvalue is 0

Consequence for Statistics

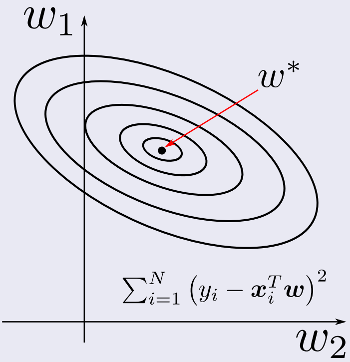
If a statistical prediction involves the inverse of an almost-singular matrix, the predictions become unreliable (high variance).

Outline

What can be done?

What could be the problem?

- Imagine that you finish with an over-fitting at the optimal $\boldsymbol{w}^*$



Overfitting?

Basically (Intuition)

- $\boldsymbol{x}_i^T \boldsymbol{w}^* \approx y_i$

Then

- You are quite good with the training data
- But Really bad with the validation and testing data

We need to pull the optimal in some way!!!

IDEAS?

Overfitting?

Basically (Intuition)

- $\boldsymbol{x}_i^T \boldsymbol{w}^* \approx y_i$

Then

- You are quite good with the training data
- But Really bad with the validation and testing data

We need to pull the optimal in some way!!!

IDEAS?

Cinvestav

Overfitting?

Basically (Intuition)

- $\boldsymbol{x}_i^T \boldsymbol{w}^* \approx y_i$

Then

- You are quite good with the training data
- But Really bad with the validation and testing data

We need to pull the optimal in some way!!!

IDEAS?

How do we integrate this solution to the Least Squared Error Solution?

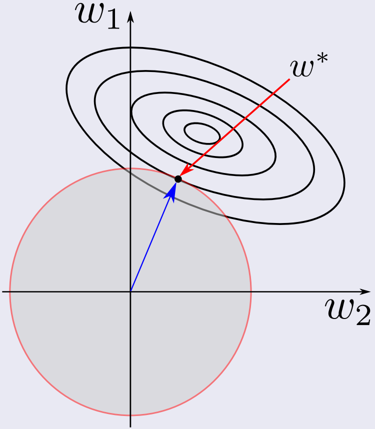
We modify it by adding en extra parameter and tweak the λ

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}\right)^2 + \lambda\sum_{i=1}^{d+1} w_i^2 \tag{34}$$

How do we integrate this solution to the Least Squared Error Solution?

Geometrically Equivalent to pulling away the optimal, it is known as Ridge Regression

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}\right)^2 + \lambda\sum_{i=1}^{d+1} w_i^2$$



Outline

Cinvestav

Something quite interesting

The w_i in the vector $\boldsymbol{w}^*$ are related to the eigenvalues in $\boldsymbol{X}^T\boldsymbol{X}$

- Thus, we can tweak the eigenvalues to obtain a similar effect than in the Ridge Regression

$$\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}\right)^2 + \lambda\sum_{i=1}^{d+1} w_i^2 \tag{35}$$

It is equivalent to avoid eigenvalues to become zero!!!

Thus, we can do the following given that $X^T X$ is positive definite

Assume that $\xi_1, \xi_2, ..., \xi_{d+1}$ are eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{d+1}$

We have

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\xi_i = \lambda_i \xi_i \text{ for all } i = 1, ..., d+1 \tag{36}$$

Given that $\boldsymbol{X}^T\boldsymbol{X}$ is singular, some λ_i is equal to 0.

Very Simple, add a convenient λ

$$\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)\xi_i = (\lambda_i + \lambda)\,\xi_i \tag{37}$$

i.e. $\lambda_i + \lambda$ is an eigenvalue for $\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)$.

It is equivalent to avoid eigenvalues to become zero!!!

Thus, we can do the following given that $X^T X$ is positive definite

Assume that $\xi_1, \xi_2, ..., \xi_{d+1}$ are eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{d+1}$

We have

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\xi_i = \lambda_i \xi_i \text{ for all } i = 1, ..., d+1 \tag{36}$$

Given that $\boldsymbol{X}^T\boldsymbol{X}$ is singular, some λ_i is equal to 0.

Very Simple, add a convenient λ

$$\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)\xi_i = (\lambda_i + \lambda)\,\xi_i \tag{37}$$

i.e. $\lambda_i + \lambda$ is an eigenvalue for $\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)$.

It is equivalent to avoid eigenvalues to become zero!!!

Thus, we can do the following given that $X^T X$ is positive definite

Assume that $\xi_1, \xi_2, ..., \xi_{d+1}$ are eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with eigenvalues $\lambda_1, \lambda_2, ..., \lambda_{d+1}$

We have

$$\left(\boldsymbol{X}^T\boldsymbol{X}\right)\xi_i = \lambda_i\xi_i \text{ for all } i = 1, ..., d+1 \tag{36}$$

Given that $\boldsymbol{X}^T\boldsymbol{X}$ is singular, some λ_i is equal to 0.

Very Simple, add a convenient λ

$$\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)\xi_i = (\lambda_i + \lambda)\,\xi_i \tag{37}$$

i.e. $\lambda_i + \lambda$ is an eigenvalue for $\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)$.

What does this mean?

Something Notable

You can control the singularity by detecting the smallest eigenvalue.

Thus

We add an appropriate tunning value λ.

What does this mean?

Something Notable

You can control the singularity by detecting the smallest eigenvalue.

Thus

We add an appropriate tunning value λ.

Ridge Regression

Ridge Regression

It tries to make least squares more robust if $\boldsymbol{X}^T\boldsymbol{X}$ is almost singular.

Process

1. Find the eigenvalues of X^TX
2. If all of them are bigger enough than zero we are fine!!!
3. Find the smallest one, then tune if necessary.
4. Build $\widehat{\boldsymbol{w}}^{Ridge} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Ridge Regression

Ridge Regression

It tries to make least squares more robust if $\boldsymbol{X}^T\boldsymbol{X}$ is almost singular.

Process

1. Find the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$
2. If all of them are bigger enough than zero we are fine!!!
3. Find the smallest one, then tune if necessary.
4. Build $\widehat{w}^{Ridge} = \left(X^TX + \lambda I\right)^{-1} X^T y$.

Ridge Regression

Ridge Regression

It tries to make least squares more robust if $\boldsymbol{X}^T\boldsymbol{X}$ is almost singular.

Process

1. Find the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$
2. If all of them are bigger enough than zero we are fine!!!
3. Find the smallest one, then tune if necessary.
4. Build $\widehat{w}^{Ridge} = \left(X^TX + \lambda I\right)^{-1} X^T y$.

Ridge Regression

Ridge Regression

It tries to make least squares more robust if $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is almost singular.

Process

1. Find the eigenvalues of $\boldsymbol{X}^{T}\boldsymbol{X}$
2. If all of them are bigger enough than zero we are fine!!!
3. Find the smallest one, then tune if necessary.
4. Build $\widehat{w}^{Ridge} = \left(X^{T}X + \lambda I\right)^{-1} X^{T}y.$

Ridge Regression

Ridge Regression

It tries to make least squares more robust if $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is almost singular.

Process

1. Find the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$
2. If all of them are bigger enough than zero we are fine!!!
3. Find the smallest one, then tune if necessary.
4. Build $\widehat{\boldsymbol{w}}^{Ridge} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Outline

Cinvestav

Exercises

Duda and Hart

Chapter 5

- 1, 3, 4, 7, 13, 17

Bishop

Chapter 4

- 4.1, 4.4, 4.7,

Hastie-Tibishirani

Chapter 3 - Problems

- Ex 3.5
- Ex 3.6

Exercises

Duda and Hart

Chapter 5

- 1, 3, 4, 7, 13, 17

Bishop

Chapter 4

- 4.1, 4.4, 4.7,

Hastie-Tibishirani

Chapter 3 - Problems

- Ex 3.5
- Ex 3.6

Exercises

Duda and Hart

Chapter 5

- 1, 3, 4, 7, 13, 17

Bishop

Chapter 4

- 4.1, 4.4, 4.7,

Hastie-Tibishirani

Chapter 3 - Problems

- Ex 3.5
- Ex 3.6

Exercises

Theodoridis

Chapter 3 - Problems

F. E. Harrell Jr, *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

J. E. Gentle, "Matrix algebra," *Springer texts in statistics, Springer, New York, NY, doi*, vol. 10, pp. 978–0, 2007.

G. Strang, *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2016.