The Data Science Cycle

Unsupervised Learning - Clustering

Andres Mendez-Vazquez

September 21, 2016

Outline

DataLab
Data Science Community

Outline

What is it?

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

Here, we have an hyperplane as a model:

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

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

We have the following function:

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

What is it?

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

Here, we have an hyperplane as a model:

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

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

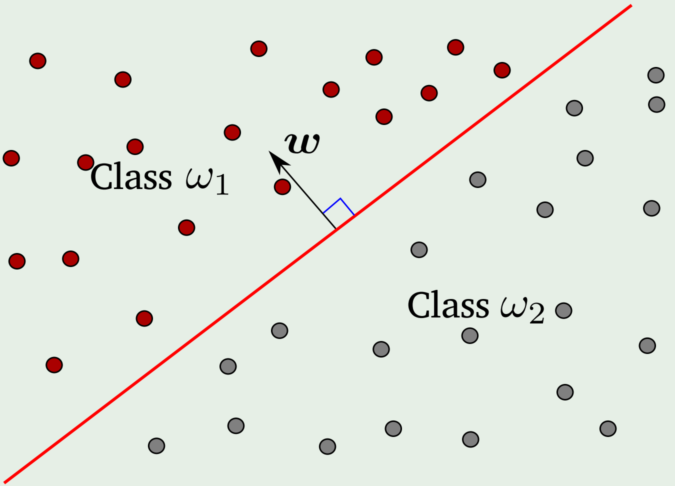
We have the following function:

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

Outline

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

Using a simple straight line



Outline

Defining a Decision Surface

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

Separating the elements in classes, ω_1 and ω_2.

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

Given 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

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

Defining a Decision Surface

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

Separating the elements in classes, ω_1 and ω_2.

When $g\left(x\right)$ is linear the decision surface is an hyperplane

Given $\boldsymbol{x}_1$ 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

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

Given $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are both on the decision surface:

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

Thus

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

Defining a Decision Surface

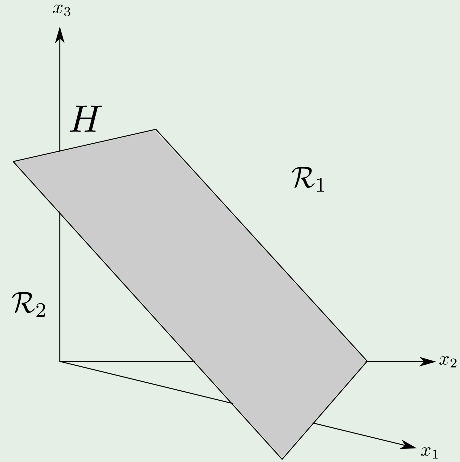
Thus

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

Remark: Any vector in the hyperplane is perpendicular to $\boldsymbol{w}^T$ i.e. $\boldsymbol{w}^T$ is normal to the hyperplane.

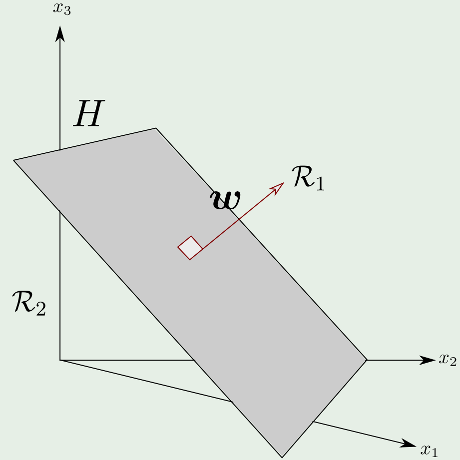
Therefore

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



Some Properties of the Hyperplane

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



It is more

We can say the following

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

It is more

We can say the following

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

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

First, we express any x as follows

$$x = x_p + r\frac{w}{\|w\|}$$

- 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

- 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 $\boldsymbol{x}$ 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

It is more

We can say the following

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

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

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

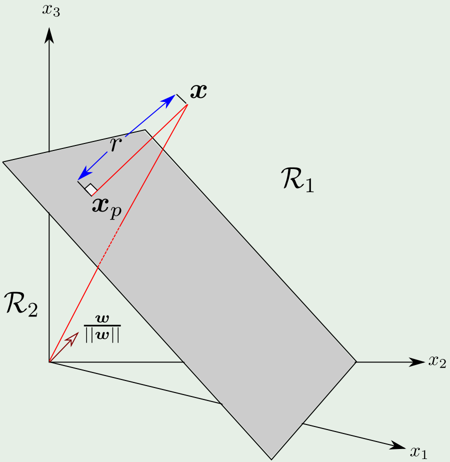
$$\boldsymbol{x} = \boldsymbol{x}_p + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$$

Where

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

We have something like this

We have then



Now

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

We have that

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

Now

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

We have that

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

Then, we have

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

Now

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

We have that

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

Then, we have

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

Now

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

We have that

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

Then, we have

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

Now

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

We have that

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

Then, we have

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

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

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

In particular

The distance from the origin to H

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

In particular

The distance from the origin to H

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

Remarks

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

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.

DataLab
Data Science Community

In addition...

If we do the following

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

By making

$$x_0 = 1 \text{ and } \boldsymbol{y} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

Where

$\boldsymbol{y}$ is called an augmented feature vector.

In addition...

If we do the following

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

By making

$$x_0 = 1 \text{ and } \boldsymbol{y} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \\ \boldsymbol{x} \\ \\ \end{pmatrix}$$

Where

y is called an augmented feature vector.

In addition...

If we do the following

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

By making

$$x_0 = 1 \text{ and } \boldsymbol{y} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 1 \\ \\ \boldsymbol{x} \end{pmatrix}$$

Where

$\boldsymbol{y}$ is called an augmented feature vector.

In a similar way

We have the augmented weight vector

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

In a similar way

We have the augmented weight vector

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

Remarks

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

In a similar way

We have the augmented weight vector

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

Remarks

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

Outline

DataLab
Data Science Community

Initial Setup

Important

We get away from our initial normalization of the samples!!!

Now, we are going to use the method know as

Minimum Squared Error

Initial Setup

Important

We get away from our initial normalization of the samples!!!

Now, we are going to use the method know as

Minimum Squared Error

Now, assume the following

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

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

Now, assume the following

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

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

We have a problem!!!

Which is the problem?

Now, assume the following

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

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

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?

Now, assume the following

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

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

We have a problem!!!

Which is the problem?

We do not know the hyperplane!!!

Thus, what distance each point has to the hyperplane?

DataLab
Data Science Community

A Simple Solution For Our Quandary

Label the Classes

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

A Simple Solution For Our Quandary

Label the Classes

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

We produce the following labels

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

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

DataLab
Data Science Community

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!!!

DataLab

Data Science Community

A Simple Solution For Our Quandary

Label the Classes

- $\omega_1 \implies +1$
- $\omega_2 \implies -1$

We produce the following labels

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

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

Outline

DataLab
Data Science Community

Now, What?

Assume true function f is given by

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

Where the ϵ

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

Thus, we can do the following

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

Now, What?

Assume true function f is given by

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

Where the ϵ

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

Thus, we can do the following

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

Now, What?

Assume true function f is given by

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

Where the ϵ

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

Thus, we can do the following

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

Thus, we have

What to do?

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

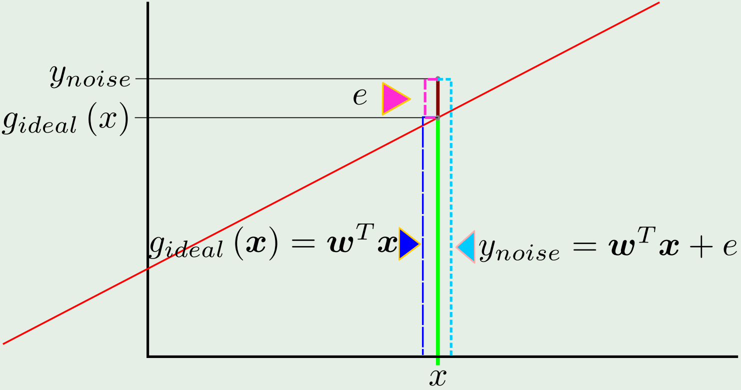
Graphically

Thus, we have

What to do?

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

Graphically



Outline

Sum Over All Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (y_i - g_{ideal}(\boldsymbol{x}))^2 \tag{11}$$

Remark: Know as least squares (Fitting the vertical offset!!!)

Sum Over All Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x})\right)^2 \tag{11}$$

Remark: Know as least squares (Fitting the vertical offset!!!)

Generalize

If

- The dimensionality of each sample (data point) is d,
- You can extend each vector sample to be $\boldsymbol{x}^T = (1, \boldsymbol{x}')$,
- We have:

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w}\right)^2 = (\boldsymbol{y} - X\boldsymbol{w})^T (\boldsymbol{y} - X\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 \tag{12}$$

Sum Over All Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x})\right)^2 \tag{11}$$

Remark: Know as least squares (Fitting the vertical offset!!!)

Generalize

If

- The dimensionality of each sample (data point) is d,
- You can extend each vector sample to be $\boldsymbol{x}^T = (\boldsymbol{1}, \boldsymbol{x'})$,
- We have:

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}\right)^2 = (\boldsymbol{y} - X\boldsymbol{w})^T (\boldsymbol{y} - X\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|_2^2 \tag{12}$$

Sum Over All Errors

We can do the following

$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} \left(y_i - g_{ideal}(\boldsymbol{x})\right)^2 \tag{11}$$

Remark: Know as least squares (Fitting the vertical offset!!!)

Generalize

If

- The dimensionality of each sample (data point) is d,
- You can extend each vector sample to be $\boldsymbol{x}^T = (\boldsymbol{1}, \boldsymbol{x'})$,
- We have:

$$\sum_{i=1}^{N} \left(y_i - \boldsymbol{x}^T \boldsymbol{w}\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 \tag{12}$$

Outline

DataLab
Data Science Community

What is $\boldsymbol{X}$

It is the Data Matrix

$$
\boldsymbol{X} = \begin{pmatrix} 1 & (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d \end{pmatrix} \tag{13}
$$

We know the following

$$
\frac{dx^T A x}{dx} = Ax + A^T x, \quad \frac{dAx}{dx} = A \tag{14}
$$

What is $\boldsymbol{X}$

It is the Data Matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & (\boldsymbol{x}_1)_1 & \cdots & (\boldsymbol{x}_1)_j & \cdots & (\boldsymbol{x}_1)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_i)_1 & & (\boldsymbol{x}_i)_j & & (\boldsymbol{x}_i)_d \\ \vdots & & & \vdots & & \vdots \\ 1 & (\boldsymbol{x}_N)_1 & \cdots & (\boldsymbol{x}_N)_j & \cdots & (\boldsymbol{x}_N)_d \end{pmatrix} \tag{13}$$

We know the following

$$\frac{d\boldsymbol{x}^T A\boldsymbol{x}}{d\boldsymbol{x}} = Ax + A^T x, \; \frac{dA\boldsymbol{x}}{d\boldsymbol{x}} = A \tag{14}$$

DataLab
Data Science Community

Note about other representations

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

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

We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^{T}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})=\boldsymbol{y}^{T}\boldsymbol{y}-\boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}-\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{w}+\boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \quad (16)$$

Making Possible to have by deriving with respect to w and assuming that $X^{T}X$ is invertible

$$\hat{w}=\left(X^{T}X\right)^{-1}X^{T}y \quad (17)$$

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

Thus, How we get the discriminant function?

Any Ideas?

We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})^{T}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w})=\boldsymbol{y}^{T}\boldsymbol{y}-\boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}-\boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{w}+\boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \quad (16)$$

Making Possible to have by deriving with respect to $\boldsymbol{w}$ and assuming that $\boldsymbol{X}^{T}\boldsymbol{X}$ is invertible

$$\hat{\boldsymbol{w}}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \quad (17)$$

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?

We can expand our quadratic formula!!!

Thus

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = \boldsymbol{y}^{T}\boldsymbol{y} - \boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{y}^{T}\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^{T}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{w} \quad (16)$$

Making Possible to have by deriving with respect to $\boldsymbol{w}$ and assuming that $\boldsymbol{X}^{T}\boldsymbol{X}$ is invertible

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \quad (17)$$

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!!!

$$g(\boldsymbol{x}) = \boldsymbol{x}^T \hat{\boldsymbol{w}} = \boldsymbol{x}^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \quad (18)$$

Pseudo-inverse of a Matrix

Definition

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

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

the pseudo inverse of A.

Reason

A^{+} inverts A on its image

What?

If $w \in image\,(A)$, then there is some $v \in \mathbb{R}^{n}$ such that $w = Av$. Hence:

$$A^{+}w = A^{+}Av = \left(A^{T}A\right)^{-1} A^{T}Av$$

Pseudo-inverse of a Matrix

Definition

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

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

the pseudo inverse of A.

Reason

A^{+} inverts A on its image

What?

If $w \in image\,(A)$, then there is some $v \in \mathbb{R}^{n}$ such that $w = Av$. Hence:

$$A^{+}w = A^{+}Av = \left(A^{T}A\right)^{-1} A^{T}Av$$

Pseudo-inverse of a Matrix

Definition

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

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

the pseudo inverse of A.

Reason

A^{+} inverts A on its image

What?

If $\boldsymbol{w} \in image\left(A\right)$, then there is some $\boldsymbol{v} \in \mathbb{R}^{n}$ such that $\boldsymbol{w} = A\boldsymbol{v}$. Hence:

$$A^{+}\boldsymbol{w} = A^{+}A\boldsymbol{v} = \left(A^{T}A\right)^{-1}A^{T}A\boldsymbol{v}$$

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N\times(d+1)}$
- $Image\left(X\right) = span\left\{X_1^{col}, ..., X_{d+1}^{col}\right\}$
- $\boldsymbol{x}_i \in \mathbb{R}^d$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X}_i^{col}, \boldsymbol{y} \in \mathbb{R}^N$

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N\times(d+1)}$
- $Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X_1^{col}}, ..., \boldsymbol{X_{d+1}^{col}}\right\}$
- $\boldsymbol{x}_i \in \mathbb{R}^d$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X}_i^{col}, \boldsymbol{y} \in \mathbb{R}^N$

Basically $\boldsymbol{y}$, the list of desired inputs the is being protected into

$$span\left\{\boldsymbol{X_1^{col}}, ..., \boldsymbol{X_{d+1}^{col}}\right\} \tag{19}$$

by the projection operator $X\left(X^T X\right)^{-1} X^T$.

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N \times (d+1)}$
- $Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, \ldots, \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$
- $\boldsymbol{x_i} \in \mathbb{R}^{d}$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{col}}, \boldsymbol{y} \in \mathbb{R}^{N}$

Basically $\boldsymbol{y}$, the list of desired inputs the is being protected into

$$span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, \ldots, \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\} \tag{19}$$

by the projection operator $\boldsymbol{X}\left(\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{T}$.

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N \times (d+1)}$
- $Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$
- $\boldsymbol{x_i} \in \mathbb{R}^d$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X}_{\boldsymbol{i}}^{\boldsymbol{col}}, \boldsymbol{y} \in \mathbb{R}^N$

Basically $\boldsymbol{y}$, the list of desired inputs the is being protected into

$$span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\} \tag{19}$$

by the projection operator $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$.

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N\times(d+1)}$
- $Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X_1^{col}}, ..., \boldsymbol{X_{d+1}^{col}}\right\}$
- $\boldsymbol{x_i} \in \mathbb{R}^{d}$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X_i^{col}}, \boldsymbol{y} \in \mathbb{R}^{N}$

Basically y, the list of desired inputs the is being protected into

$$span\left\{X_1^{col}, ..., X_{d+1}^{col}\right\} \tag{19}$$

by the projection operator $X\left(X^TX\right)^{-1}X^T$.

What lives where?

We have

- $\boldsymbol{X} \in \mathbb{R}^{N \times (d+1)}$
- $Image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X_1^{col}}, ..., \boldsymbol{X_{d+1}^{col}}\right\}$
- $\boldsymbol{x_i} \in \mathbb{R}^d$
- $\boldsymbol{w} \in \mathbb{R}^{d+1}$
- $\boldsymbol{X_i^{col}}, \boldsymbol{y} \in \mathbb{R}^N$

Basically $\boldsymbol{y}$, the list of desired inputs the is being protected into

$$span\left\{\boldsymbol{X_1^{col}}, ..., \boldsymbol{X_{d+1}^{col}}\right\} \tag{19}$$

by the projection operator $\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$.

Geometric Interpretation

We have

1. The image of the mapping $\boldsymbol{w}$ to $\boldsymbol{X}\boldsymbol{w}$ is a linear subspace of $\mathbb{R}^N$.
2. As $\boldsymbol{w}$ runs through all points $\mathbb{R}^{d+1}$, the function value $\boldsymbol{X}\boldsymbol{w}$ runs through all points in the image space $image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$.
3. Each $\boldsymbol{w}$ defines one point $\boldsymbol{X}\boldsymbol{w} = \sum_{j=0}^{d} w_j \boldsymbol{X}_j^{col}$.
4. $\hat{\boldsymbol{w}}$ is the point which minimizes the distance $d\left(\boldsymbol{y}, image\left(\boldsymbol{X}\right)\right)$.

Geometric Interpretation

We have

1. The image of the mapping $\boldsymbol{w}$ to $\boldsymbol{X}\boldsymbol{w}$ is a linear subspace of $\mathbb{R}^N$.
2. As $\boldsymbol{w}$ runs through all points $\mathbb{R}^{d+1}$, the function value $\boldsymbol{X}\boldsymbol{w}$ runs through all points in the image space $image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$.
3. Each $\boldsymbol{w}$ defines one point $\boldsymbol{X}\boldsymbol{w} = \sum_{j=0}^{d} w_j X_j^{col}$.
4. $\hat{w}$ is the point which minimizes the distance $d\left(y, image\left(\boldsymbol{X}\right)\right)$.

Geometric Interpretation

We have

1. The image of the mapping $\boldsymbol{w}$ to $\boldsymbol{X}\boldsymbol{w}$ is a linear subspace of $\mathbb{R}^N$.
2. As $\boldsymbol{w}$ runs through all points $\mathbb{R}^{d+1}$, the function value $\boldsymbol{X}\boldsymbol{w}$ runs through all points in the image space $image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$.
3. Each $\boldsymbol{w}$ defines one point $\boldsymbol{X}\boldsymbol{w} = \sum_{j=0}^{d} w_j \boldsymbol{X}_j^{col}$.
4. $\hat{\boldsymbol{w}}$ is the point which minimizes the distance $d\left(\boldsymbol{y}, image\left(\boldsymbol{X}\right)\right)$.

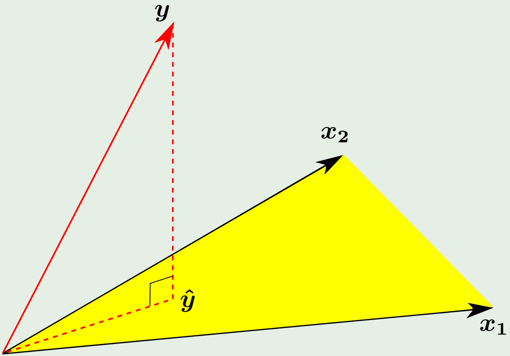
Geometric Interpretation

We have

1. The image of the mapping $\boldsymbol{w}$ to $\boldsymbol{X}\boldsymbol{w}$ is a linear subspace of $\mathbb{R}^N$.
2. As $\boldsymbol{w}$ runs through all points $\mathbb{R}^{d+1}$, the function value $\boldsymbol{X}\boldsymbol{w}$ runs through all points in the image space $image\left(\boldsymbol{X}\right) = span\left\{\boldsymbol{X}_{\boldsymbol{1}}^{\boldsymbol{col}}, ..., \boldsymbol{X}_{\boldsymbol{d+1}}^{\boldsymbol{col}}\right\}$.
3. Each $\boldsymbol{w}$ defines one point $\boldsymbol{X}\boldsymbol{w} = \sum_{j=0}^{d} w_j \boldsymbol{X}_j^{col}$.
4. $\hat{\boldsymbol{w}}$ is the point which minimizes the distance $d\left(\boldsymbol{y}, image\left(\boldsymbol{X}\right)\right)$.

Geometrically

Ahhhh!!!



Outline

DataLab
Data Science Community

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If X^TX almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
   1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
   1. It needs some form of feature selection.
   2. Possible some type of regularization

1. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
    1. It needs some form of feature selection.
    2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
    1. It needs some form of feature selection.
    2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
    1. It needs some form of feature selection.
    2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
   1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
   1. It needs some form of feature selection.
   2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^T\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
   1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
   1. It needs some form of feature selection.
   2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
    1. It needs some form of feature selection.
    2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Robustness

1. Least squares works only if $\boldsymbol{X}$ has full column rank, i.e. if $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ is invertible.
2. If $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ almost not invertible, least squares is numerically unstable.
    1. Statistical consequence: High variance of predictions.

Not suited for high-dimensional data

1. Modern problems: Many dimensions/features/predictors (possibly thousands).
2. Only a few of these may be important:
    1. It needs some form of feature selection.
    2. Possible some type of regularization

Why?

1. Treats all dimensions equally
2. Relevant dimensions are averaged with irrelevant ones

Issues with Least Squares

Problem with Outliers