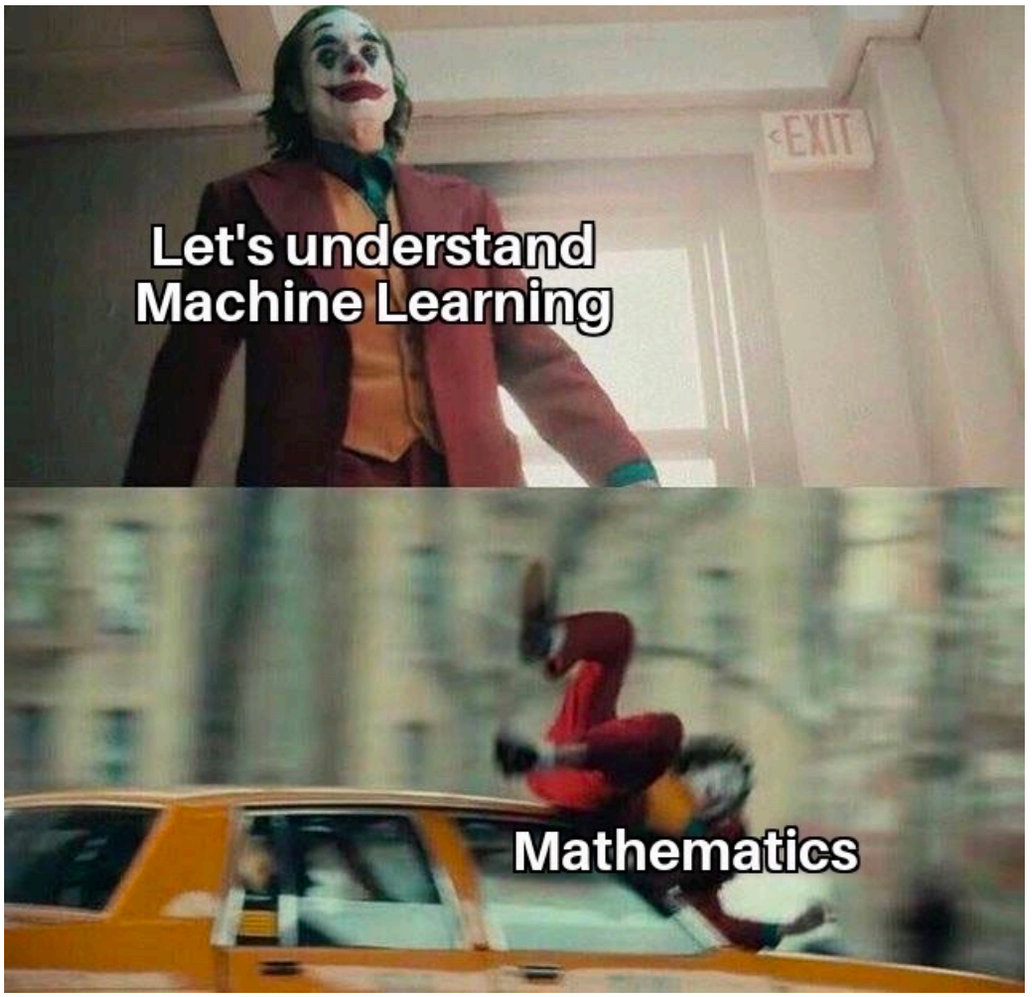
February 21, 2023

DSML : Math for ML.

Linear Algebra : Halfspaces and Distances.



Recap:

(a) **Vectors**: $\bar{x}, \bar{y} \in \mathbb{R}^d$ $\qquad \bar{x}^T = [x_1 \; x_2 \; \cdots \; x_d]$

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \qquad \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}$$

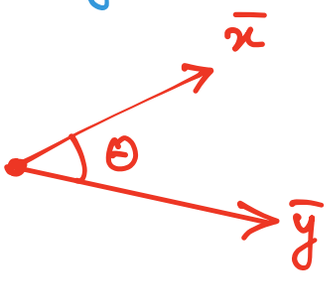
(b) **Norm**: $\bar{x} \in \mathbb{R}^d$ (length, magnitude)

$$\|\bar{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$$

(c) **Dot Product**: $\bar{x} \cdot \bar{y} = \bar{x}^T \bar{y}$

$$= \sum_{i=1}^{d} x_i \cdot y_i$$

(d) **Angle between two vectors**:

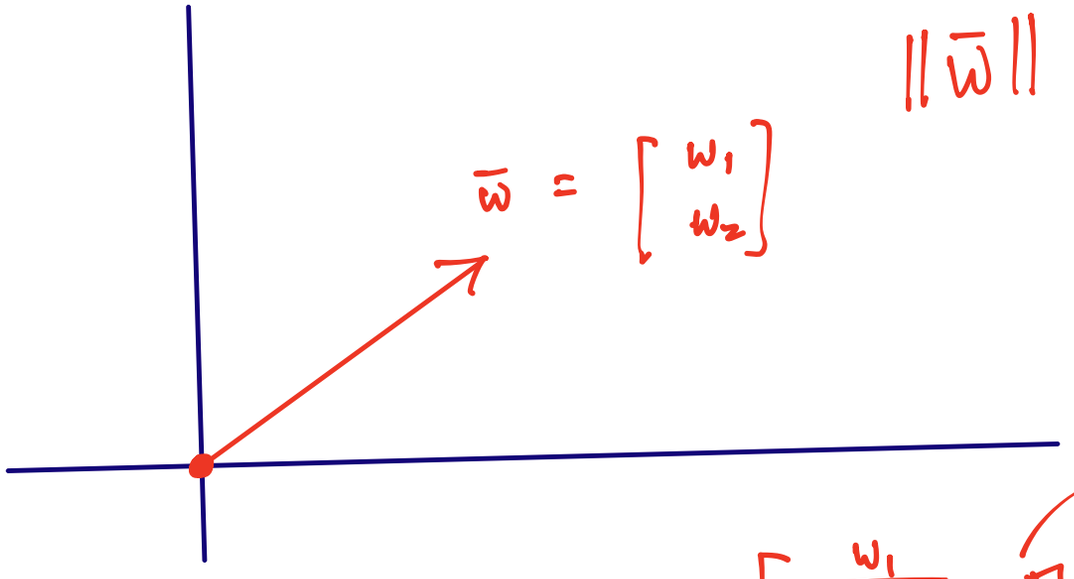


$$\cos\theta = \frac{\bar{x}^T \bar{y}}{\|\bar{x}\| \cdot \|\bar{y}\|}$$

Unit Vectors:

"Unit vectors represent the direction in which a vector is pointing towards"

* Vector with magnitude 1. → Unit vector.



$$\bar{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$\|\bar{w}\| = \sqrt{w_1^2 + w_2^2}$$

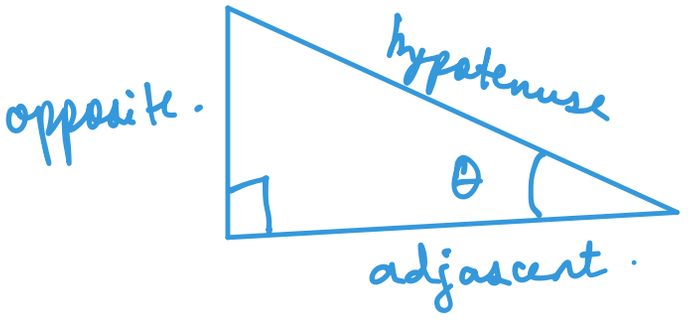
$$\frac{\bar{w}}{\|\bar{w}\|} = \hat{w} = \begin{bmatrix} \frac{w_1}{\sqrt{w_1^2+w_2^2}} \\ \frac{w_2}{\sqrt{w_1^2+w_2^2}} \end{bmatrix}$$

$$\|\hat{w}\| = \sqrt{\frac{w_1^2}{w_1^2+w_2^2} + \frac{w_2^2}{w_1^2+w_2^2}}$$

$$= \sqrt{\frac{w_1^2+w_2^2}{w_1^2+w_2^2}}$$

$$= \sqrt{1} = 1 \quad //$$

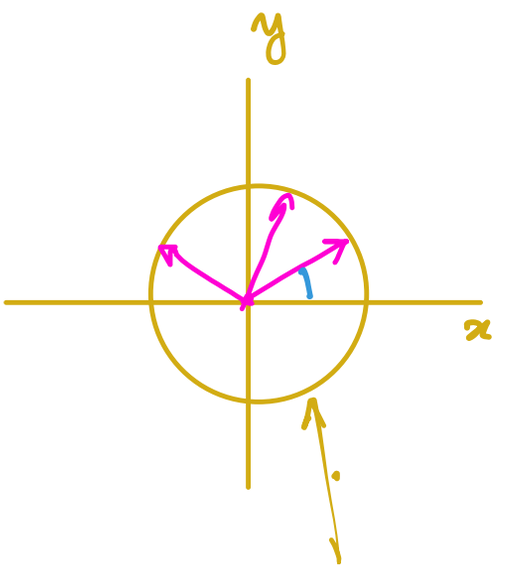
Basic Trigonometry



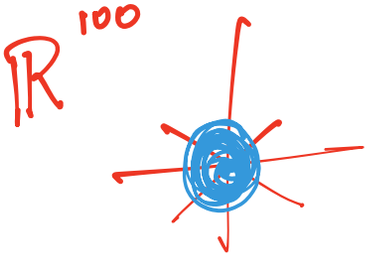
$$\sin(\theta) = \frac{\text{opposite}}{\text{hypotenuse}}$$

$$\cos(\theta) = \frac{\text{adjascent}}{\text{hypotenuse}}$$

$$\tan(\theta) = \frac{\text{opposite}}{\text{adjascent}}.$$

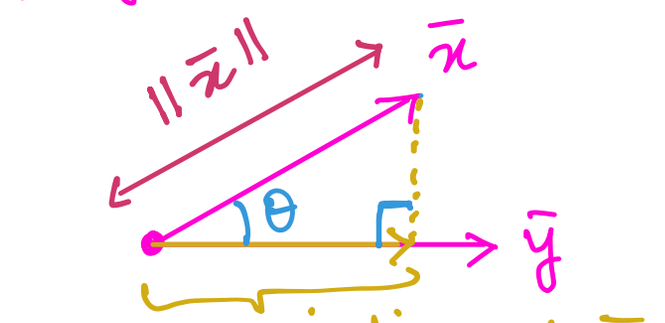


circle with radius 1.



Projection :

$\bar{x}^T \bar{y} = \bar{x} \cdot \bar{y}$



$$\cos(\theta) = \frac{\|\bar{p}\|}{\|\bar{x}\|}$$

$$\frac{\bar{x}^T \bar{y}}{\|\bar{x}\| \cdot \|\bar{y}\|} = \frac{\|\bar{p}\|}{\|\bar{x}\|}$$

$$\|\bar{p}\| = \cancel{\|\bar{x}\|} \cdot \frac{\bar{x}^T \bar{y}}{\cancel{\|\bar{x}\|} \cdot \|\bar{y}\|}$$

$$\|\bar{p}\| = \frac{\bar{x}^T \bar{y}}{\|\bar{y}\|} = \frac{\bar{x} \cdot \bar{y}}{\|\bar{y}\|}$$

$$= \bar{x}^T \left(\frac{\bar{y}}{\|\bar{y}\|} \right) = \bar{x}^T \hat{y} .$$

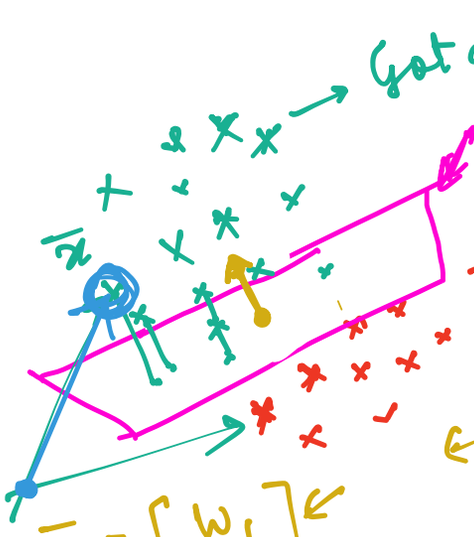
large dimensional Classification problems.

not useful

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
0	LP001002	Male	No	0	Graduate	No	5849	0.0	NaN	360.0	1.0	Urban
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban

$X = \{\bar{x}_i\}_{i=1}^{n}$

$\bar{x}_i \rightarrow$ feature vectors.



Find one (n-1)-dim hyperplane which can separate these.

So, I will optimize to find

$w_1x_1 + w_2x_2 + \dots + w_nx_n + w_0 = 0$

$\bar{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$

$\frac{\bar{w}}{\|\bar{w}\|}$

$\bar{x} = [\text{Unit vector}]$

$\bar{x} \in \mathbb{R}^n$

$\frac{\bar{x}}{\|\bar{x}\|} \rightarrow$ Direction in which $\bar{x}$ is pointing

Shifting 2-D line.

L : $w_1 x_1 + w_2 x_2 + w_0 = 0$

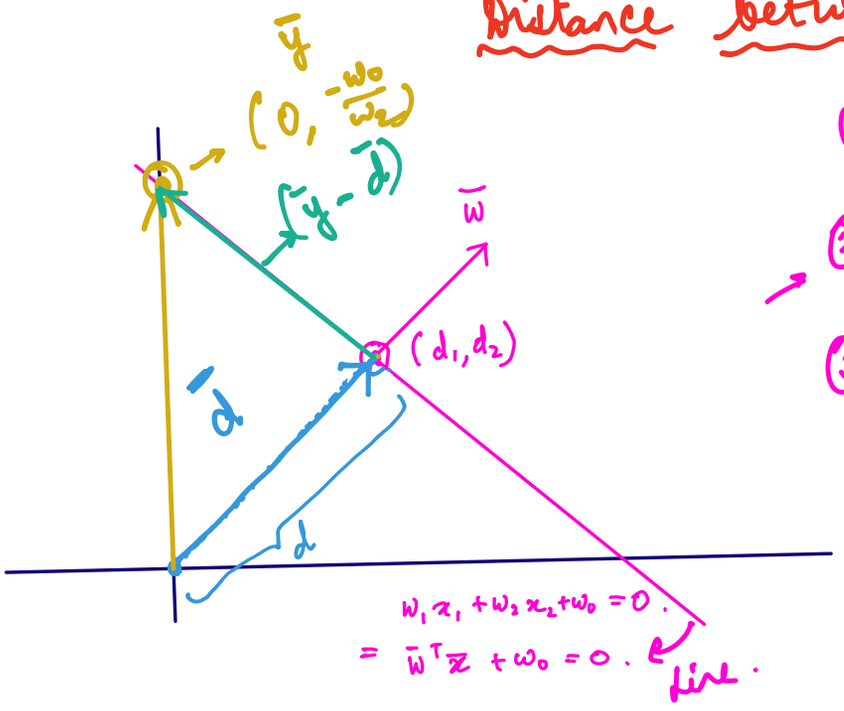
Q1] How to shift this line by 'a' units to the right?

$$w_1(x_1 - a) + w_2 x_2 + w_0 = 0$$

Q2] How to shift this line by 'b' units to the top?

$$w_1 x_1 + w_2(x_2 - b) + w_0 = 0$$

Distance between origin and a line



① $\|\bar{d}\| = \sqrt{d_1^2 + d_2^2}$

② $w_1 d_1 + w_2 d_2 + w_0 = 0.$

③ $\hat{w} = \frac{\bar{w}}{\|\bar{w}\|}$, $\bar{d} = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} k\frac{w_1}{\|\bar{w}\|} \\ k\frac{w_2}{\|\bar{w}\|} \end{bmatrix}$

Substitute ③ in ② →

$d_1 = \frac{k \cdot w_1}{\|\bar{w}\|}$, $d_2 = \frac{k\, w_2}{\|\bar{w}\|}$

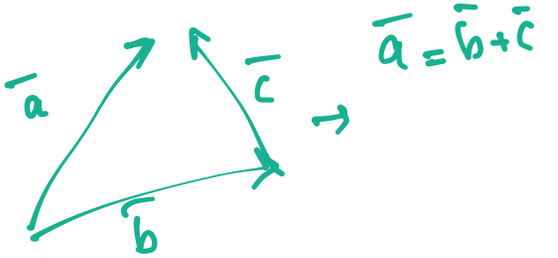
④ To get (d_1, d_2), substitute k in ③

$k \frac{w_1^2}{\|\bar{w}\|} + k \frac{w_2^2}{\|\bar{w}\|} + w_0 = 0.$

$\bar{d} = \begin{bmatrix} -\frac{w_0 w_1}{(w_1^2 + w_2^2)} \\ -\frac{w_0 w_2}{(w_1^2 + w_2^2)} \end{bmatrix}$

$k \left(\frac{w_1^2 + w_2^2}{\|\bar{w}\|} \right) = -w_0$

$k = \frac{-w_0 \times \|w\|}{(w_1^2 + w_2^2)}$



$$\bar{y} = \left(0, -\frac{w_0}{w_2}\right)$$

$$w_1 x_1 + w_2 x_2 + w_0 = 0.$$

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2}.$$

$$\bar{d} = \begin{bmatrix} \frac{-w_0 w_1}{(w_1^2 + w_2^2)} \\ \\ \frac{-w_0 w_2}{(w_1^2 + w_2^2)} \end{bmatrix}$$

$$d^T (\bar{y} - \bar{d}) = \frac{w_0 w_1}{(w_1^2 + w_2^2)} \times \frac{-w_0 w_1}{w_1^2 + w_2^2}$$

$$+ \left(\frac{-w_0}{w_2} + \frac{w_0 w_2}{(w_1^2 + w_2^2)}\right) \times \frac{-w_0 w_2}{w_1^2 + w_2^2}.$$

$$\left(\frac{-w_0 (w_1^2 + w_2^2) + w_0 w_2^2}{w_2 (w_1^2 + w_2^2)}\right) - \frac{w_0 w_2}{w_1^2 + w_2^2}$$

$$d^T(\bar{y} - \bar{d}) = \frac{-w_0^2 w_1^2}{(w_1^2 + w_2^2)^2} + \frac{w_0^2 (w_1^2 + w_2^2) - w_0^2 w_2^2}{(w_1^2 + w_2^2)^2}$$

$$= \frac{-w_0^2 w_1^2 + w_0^2 w_1^2 + w_0^2 w_2^2 - w_0^2 w_2^2}{(w_1^2 + w_2^2)^2}$$

$$= \frac{0}{(w_1^2 + w_2^2)^2} = 0$$

Distance between a line and the origin.

$$\bar{d} = \begin{bmatrix} \frac{-w_0 w_1}{(w_1^2 + w_2^2)} \\ \\ \frac{-w_0 w_2}{(w_1^2 + w_2^2)} \end{bmatrix}$$

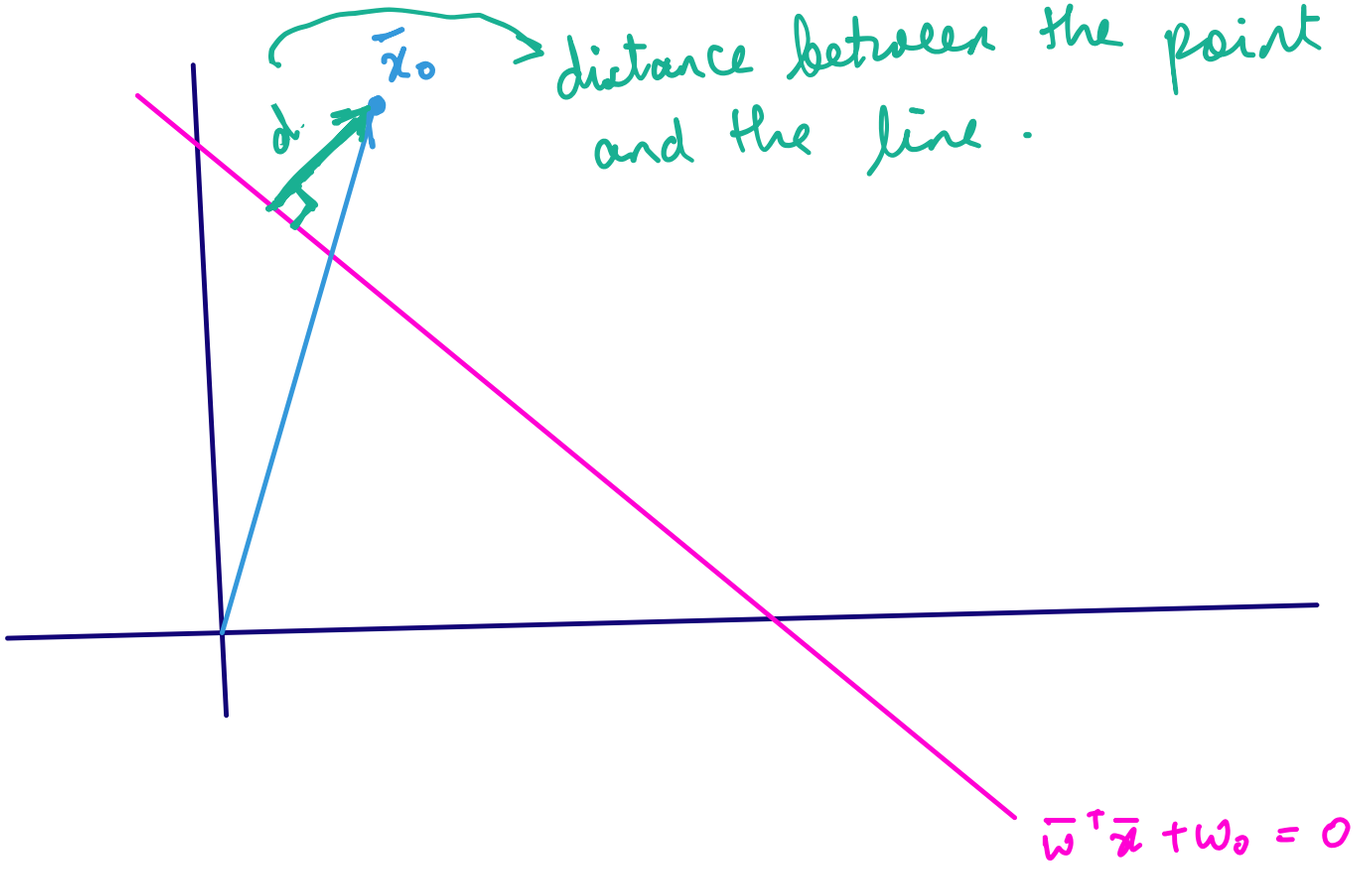
Hence, the distance between the origin and a line $\bar{w}^T \bar{x} + w_0 = 0$ is given by

$$d = \frac{w_0}{\|\bar{w}\|}$$

$$\|\bar{d}\| = \sqrt{\frac{w_0^2 w_1^2}{(w_1^2 + w_2^2)^2} + \frac{w_0^2 w_2^2}{(w_1^2 + w_2^2)^2}}$$

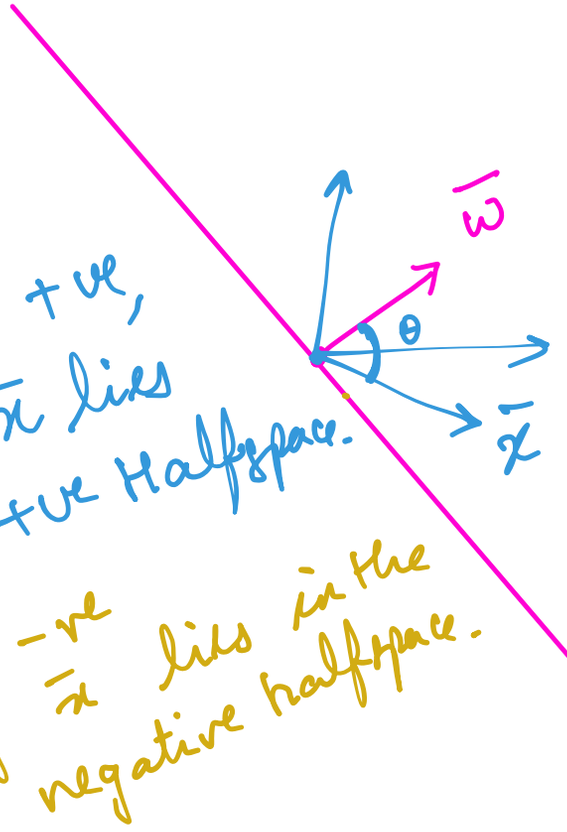
$$= \sqrt{\frac{w_0^2 \cancel{(w_1^2 + w_2^2)}}{(w_1^2 + w_2^2)^{\cancel{2}}}} = \frac{w_0}{\sqrt{w_1^2 + w_2^2}} = \frac{w_0}{\|\bar{w}\|}$$

Distance between a point and a line.



Final answer: $$d = \frac{\bar{w}^T \bar{x}_0 + w_0}{\|\bar{w}\|}$$

$\cos\theta = \dfrac{\bar{w}^T \bar{x}}{\|\bar{w}\| \cdot \|\bar{x}\|}$



$\bar{w}^T \bar{x} \rightarrow$ +ve, we say $\bar{x}$ lies in the +ve Halfspace.

$\bar{w}^T \bar{x} \rightarrow$ -ve, we say $\bar{x}$ lies in the negative halfspace.

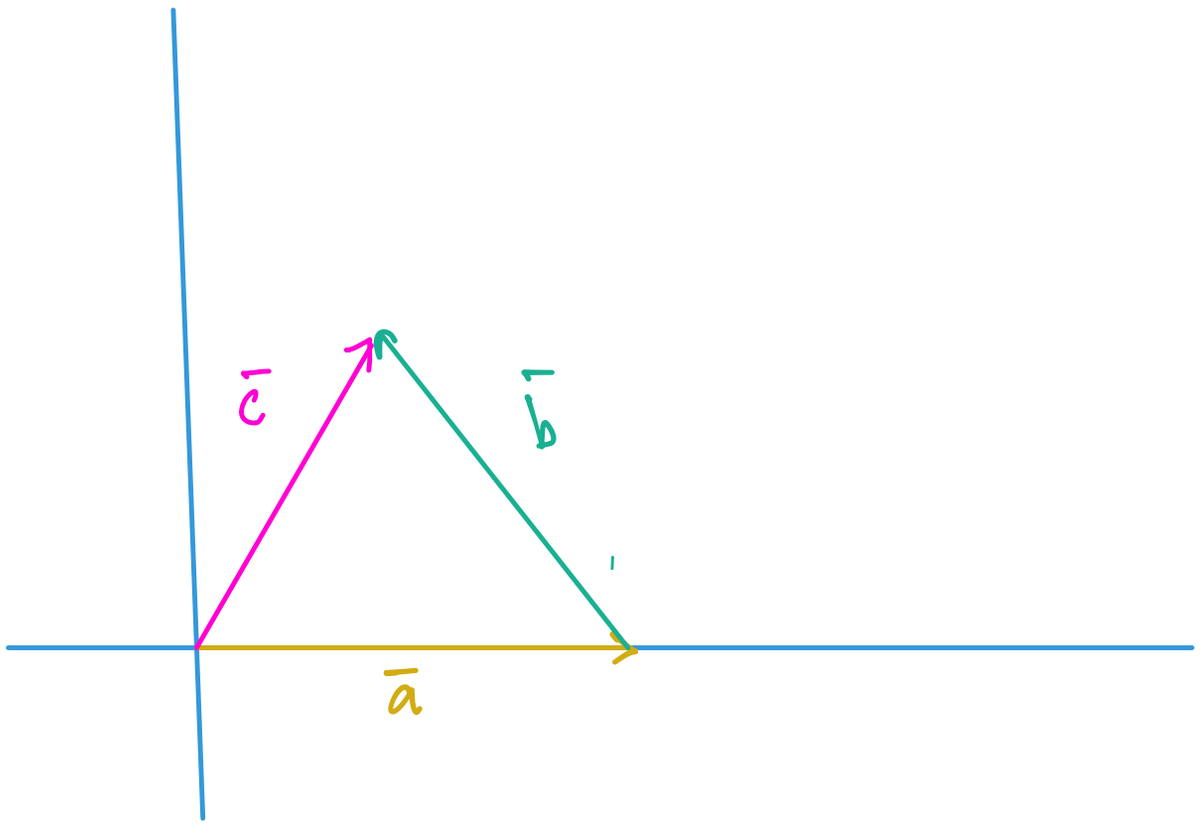
$\cos\theta \rightarrow$ +ve when $0 \le \theta \le 90°$, $270° \le \theta \le 360°$

$\rightarrow$ -ve when $90° \le \theta \le 270°$

$\text{sign}(\bar{w}^T \bar{x})$

$\rightarrow$ +ve when $0 \le \theta \le 90°$, $270° \le \theta \le 360°$

$\rightarrow$ -ve when $90° \le \theta \le 270°$



$$\bar{c} = \bar{a} + \bar{b}$$

$$\bar{b} = \bar{c} - \bar{a} .$$

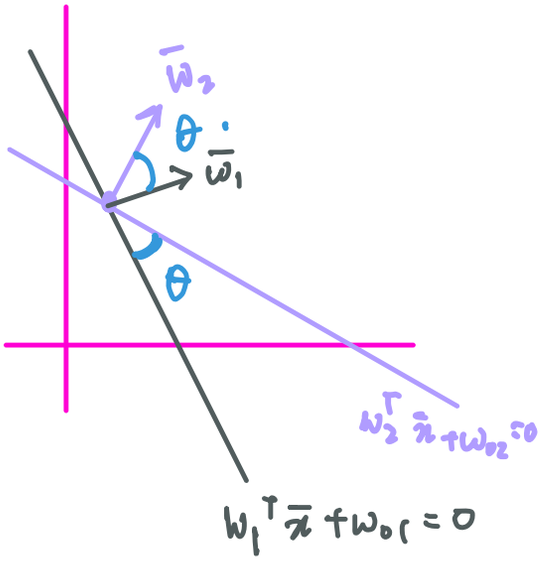
$$w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + w_0 = 0.$$

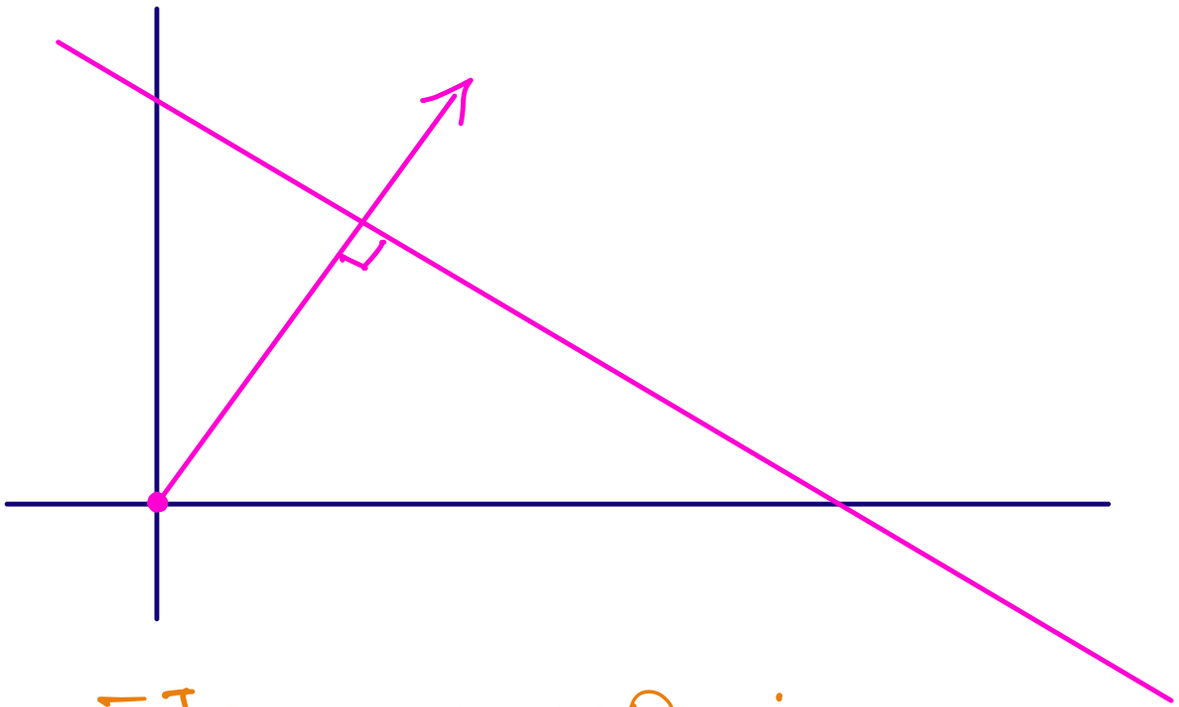
$$\Rightarrow \bar{w}^T \bar{x} + w_0 = 0$$

D_1 : $w_1^T \bar{x} + w_{01} = 0$

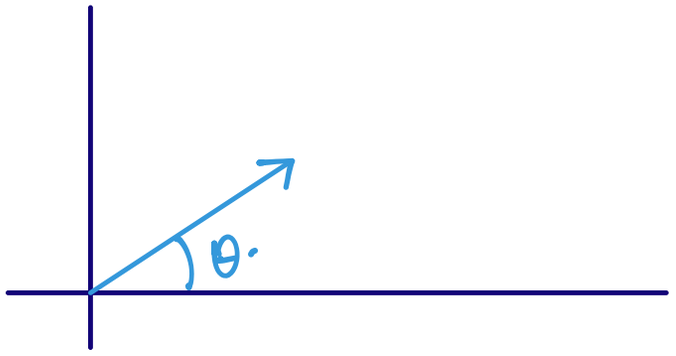
D_2 : $w_2^T \bar{x} + w_{02} = 0.$

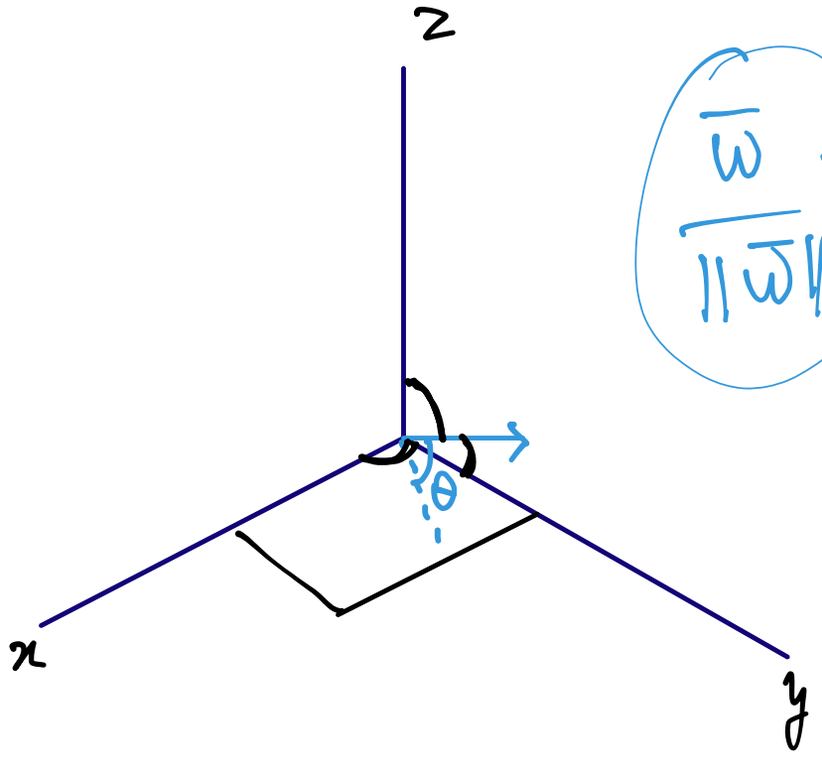
$$\cos(\theta) = \frac{\bar{w}_1^T \bar{w}_2}{\|\bar{w}_1\| \cdot \|\bar{w}_2\|}$$



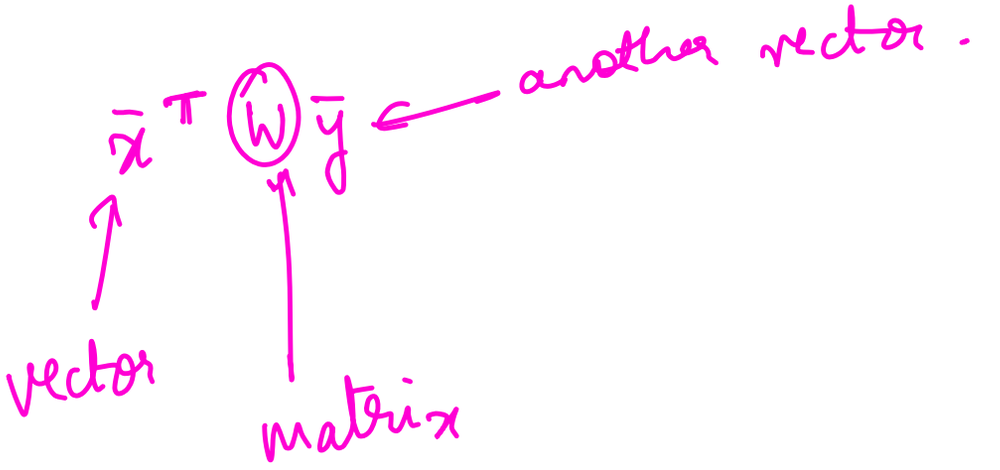


If $\bar{w}^T \bar{x} + w_0 = 0$ is a D-dim hyperplane, then, $\bar{w}$ is $\perp^r$ to all points lying on the d-Dim hyperplane.





$\frac{\bar{w}}{\|\bar{w}\|}$ → direction.

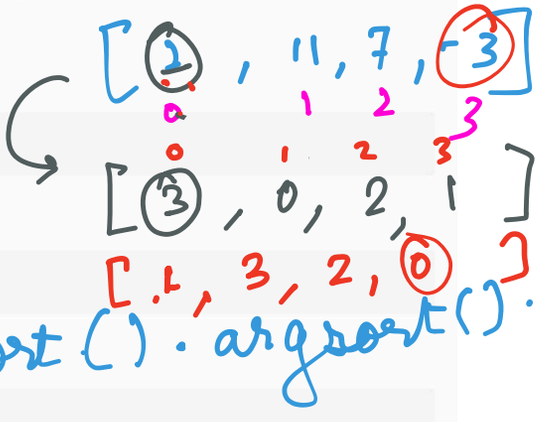


$$(\bar{x}^T W \bar{y})^T = \bar{y}^T W^T \bar{x}$$

$$\bar{x}^T \bar{y}$$

Problem Description:

Given two arrays, representing two continuous features from a certain dataset, your task is to calculate the **Spearman rank correlation coefficient** between these two arrays.



Input Format:

```
The input contains two arrays, passed to the function calculateSRCC()
```

Output Format:

```
Return the calculated value, rounded off up to 3 decimal places
```

Input Constraints:

```
1 <=  len(arr) <= 10000, for both the input arrays
-2 <= arr[i] <= 2
```

Sample Input:

```
-0.5722404243297754 1.8421092051183359 0.06532025415596275 -0.021812955390535325 0.2386780529341842
-0.04628825915644745 0.2073075701391909 1.0467470550072373 -0.31174784507404363 0.13821676157233972
```

Sample Output:

```
0.6
```

Output Explanation:

Using the Spearman correlation coefficient formula on both the arrays, we can easily calculate the value of the coefficient.

Code Constraints:

You aren't allowed to use scipy, stats, or any other libraries for inbuild correlation coefficient calculation

Note: Feel free to use NumPy to compute the helper functions.

Sun Pharmaceutical Industries claims that a person's IQ improves after they use the Donepezil drug.

To test this claim a trial was conducted considering **20** patients. An IQ test was conducted for these patients before giving the drug and an IQ test was conducted for the same set of patients after the drug the recorded results are shown below.

```
IQ_before=[101,124,89,57,135,98,69,105,114,106,97,121,93,116,102,71,88,108,144,99]

IQ_after=[113,127,89,70,127,104,69,127,115,99,104,120,95,129,106,71,94,112,154,96]
```

Perform an appropriate test to test the claim at **90%** confidence.

1 Spearman rank Correlation

2 Percentage

3 Customer tip

Q 1 Grocery delivery

Q 2 Drill sergeant

Q 3 Correlation

Q 4 Suitable for T-test

Q 5 Train & Test data

Q 6 IQ test

Q 7 Cloud computing company

Q 8 Dating website

Q 9 Coin Flip

Q 10 Prakruthi cafe

Needs to be fixed.

A drill sergeant claims that his battalion is able to complete the toughest obstacle course in **5 minutes or less.**

35 soldiers in the battalion are picked at random and asked to undergo the course, which they complete with an **average** time of **4.78 minutes** and a **standard deviation** of **1.8 minutes**.

To test the drill sergeant's claim, an alpha of **0.05** was selected, and the critical z-score was **-1.64**. The calculated z-score for the sample was **-0.12**. What conclusion can be drawn from this?

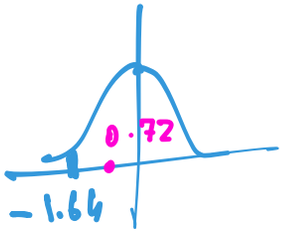
should be -0.72.

$H_0 : \mu = 5$

$H_a : \mu \leq 5$.

-1.64.

Less than -1.64.



$n = 35$.

$m_{obs} = 4.78$

$\sigma = 1.8$.

$$Z = \frac{4.78 - 5}{1.8/\sqrt{35}}$$

$=$

X Correct (marked) answer: Test conclusively shows alt. hypothesis is true. ↑ wrong ans.

A cloud computing company offers storage services and owns data centers in four geographical regions to reduce the latency of its services.

The company is carrying out a study to compare the load on each data center, measured by the number of million 'fetch' requests per day.

- H_0: The load is the same among all the regions.
- H_1: The load in atleast one of the regions is different from the others.

It was found that the population variances across the four regions were approximately the same, and the data in each group followed an approximately Normal distribution.

The mean square between the groups (i.e, **MSB**) is **180.23**, and the mean square within the groups (i.e, **MSW**) is **8.159**, Calculate the **F-test statistic** while the critical F-score was calculated as 3.49 at the alpha=0.05 significance level. What can you conclude about the test?

$$F_{stat} = \frac{MSB}{MSW} = \frac{180.23}{8.159}$$

$$= 22.08 \gg 3.49.$$

* Mistake,
correct answer: Reject H_0.

Problem Description:

A Chinese restaurant collects data on their customer's tips given to waiters. The dataset includes information about the bill, the tip paid, and customer details such as the sex of the customer, whether the customer smokes, and the day and time that the customer visited the restaurant.

We wish to study the relationship between the tips paid and customer details. Specifically, the restaurant claims that the **female customers tip more than the male customers**. Test the claim with the help of the given dataset at a **90%** confidence interval.

Note: Calculate the **p-value** with the help of obtained z-score and return the status as '**Reject**' or '**Fail to reject**'.

Input Description:

A data frame.

Output Description:

Status ('Reject' or 'Fail to reject')

* no issues here.

Sample Input:

	total_bill	tip	sex	smoker	day	time
0	16.99	1.01	Female	No	Sun	Dinner
1	10.34	1.66	Male	No	Sun	Dinner
2	21.01	3.50	Male	No	Sun	Dinner
3	23.68	3.31	Male	No	Sun	Dinner
4	24.59	3.61	Female	No	Sun	Dinner

Sample Output:

Reject

Sample Explanation:

From the given data frame, first, calculate the z-score using two sample z-test.

Based on the z-score calculate p_value and return the status as 'Reject' or 'Fail to reject'.