Data Science and Artificial Intelligence

Machine Learning

Support Vector Machines

Lecture No. 1

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic — Naive Bayes

Topic — Smoothing

Topic — Laplace smoothing

Topic

Topic

Topics to be Covered

Topic — Discriminative & generative learning

Topic — SVM

Topic — Naive Bayes

Topic — Advantage & disadvantage.

Topic


FIND
a way
NOT AN
excuse

REVISION

Summary of the last class

Solution to
Zero prob problem →

$\underbrace{\text{Old value} + \alpha}$

Old value + $K\alpha$

↓

K = No of values a dimension can take

REVISION

Summary of the last class

In case of Cont dimension ⇒ we use Gaussian PDF.

Q: Consider a classification problem with 10 classes $y \in \{1,2,\ldots,10\}$, and two binary features $x1, x2 \in \{0,1\}$.

Suppose:

$P(Y=1) = P(Y=2) = P(Y=3) - - - = \frac{1}{10}$

$p(Y=y)=1/10$,

$p(x1=1 \mid Y=y)=y/10$,

$p(x2=1 \mid Y=y)=y/540$

Which class will naïve Bayes classifier produce on a test item with $(x1=0, x2=1)$?

A. 1

B. 3

C. 5

D. 8

E. 10

$\underbrace{P(Y) \cdot P(x_1=0/Y=y) \cdot P(x_2=1/Y=y)}$

$P(Y) \to \frac{1}{10}$

$P(x_2=1/Y=y) \to \frac{y}{540}$

$\Rightarrow$ So we have to find Y value for which this is max

Since $x^1 = (0,1)$

So $P(x' = 0 / Y = y) = 1 - P(x' = 1 / Y = y)$

$= 1 - \frac{y}{10}$

$\underline{\underline{\text{So}}}$ $\max\left(\frac{1}{10}\left(1 - \frac{y}{10}\right)\left(y/540\right)\right)$ $\Rightarrow$ So we want y values that maximize this term.

$\frac{d}{dy}\frac{1}{10}\left(\frac{y}{540} - \frac{y^2}{5400}\right) = 0$

$\frac{1}{10}\left(\frac{1}{540} - \frac{2y}{5400}\right) = 0$

$\Rightarrow y = 10/2 \Rightarrow 5$

1. What type of algorithm is Naive Bayes used for in machine learning?

a. Classification

b. Regression

c. Clustering

d. Reinforcement learning

3. What is the "naive" assumption in Naive Bayes?

a. It assumes that all features are equally important.

b. It assumes that features are independent of each other.

c. It assumes that the dataset is small.

d. It assumes that features are dependent on each other.

6. In a binary classification problem, if the probability of an event occurring in Class A is 0.8 and in Class B is 0.~~3~~ 2, what is the odds ratio in favor of Class A?

a. 0.375

b. 1.5

c. ~~2.67~~ 4.0.

d. 3.33

odds Ratio in favour of class A ⇒ $\dfrac{\text{Probab that point belong to class A}}{\text{Probab that point donot belong to class A}}$

$\Rightarrow \dfrac{.8}{.2} = 4$.

1) Reduce overfitting 2) Reduce Complexity

9. In the context of Naive Bayes, what is Laplace smoothing (additive smoothing) used for? → $\alpha=1$

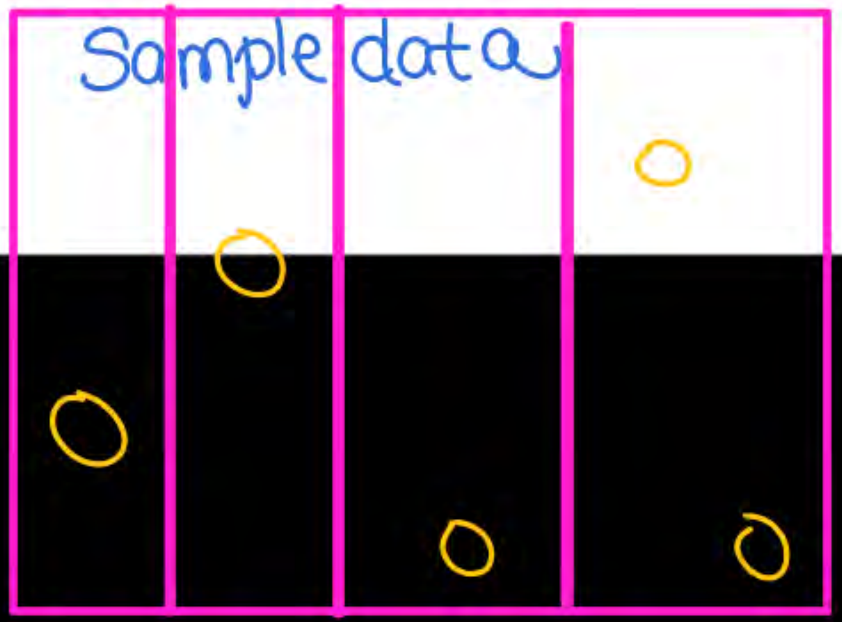
a. Reducing the impact of rare features

missing data

b. ~~Increasing the model's complexity~~

c. ~~Decreasing the training time~~

d. ~~Ignoring missing data~~



13. In a binary classification problem, a Naive Bayes classifier correctly classifies 85% of Class A instances and 90% of Class B instances. If the prior probabilities are P(Class A) = 0.4 and P(Class B) = 0.6, what is the overall accuracy of the classifier?

a. 0.48

b. 0.87

c. 0.90

d. 0.84

Total probability ⇒

$P(Acc) \Rightarrow P(Acc|A)P(A) + P(Acc|B)P(B)$

$\Rightarrow .85 \times .4 + .9 \times .6$

$\Rightarrow .88$

Advantage

Training time is less as compared to KNN, DT

non linear in nature

Work with both Categorical & Continuous dimension

Accuracy is better than other algo.

no Normalisation of data
No feature selection

Can handle missing data.

Disadvantages

assume that all dimension are independent

{ equal importance to all dimension

Sensitive to outliers

more training data

Zero probab Problem.

Cannot work for regression

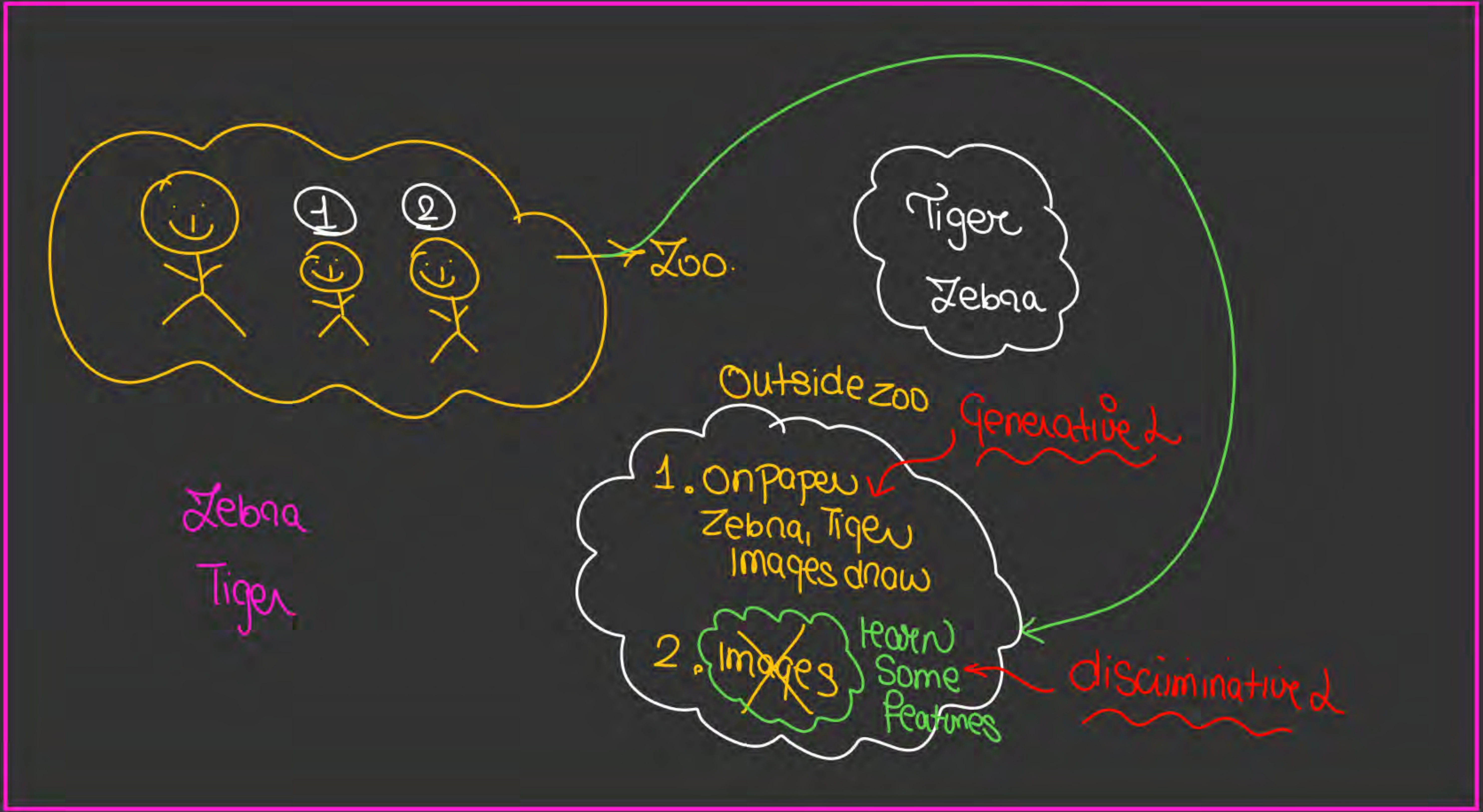
Naïve Bayes Classifier

Advantages of Naïve Bayes Classifier:

- Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
- It can be used for Binary as well as Multi-class Classifications.
- It performs well in Multi-class predictions as compared to the other Algorithms.
- It is the most popular choice for **text classification problems.**

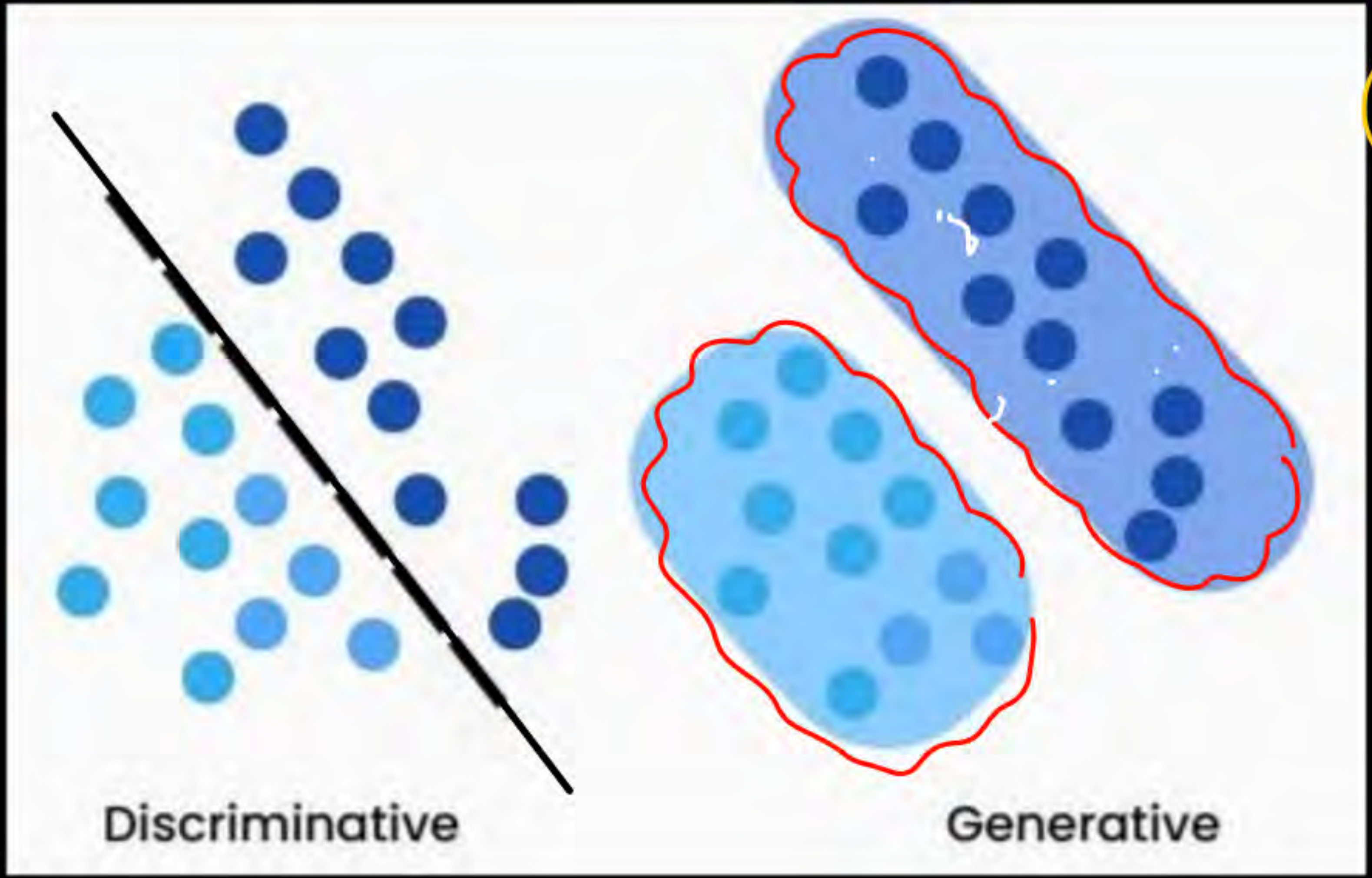
Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.



1
2
Zoo.
Tiger
Zebra
Outside zoo
1. On paper
Zebra, Tiger
Images draw
Generative
2. ~~Images~~
learn
Some
features
discriminative
Zebra
Tiger

Discriminative vs. Generative Learning



Naive Bayes
Bayes

Both the algorithm find the distribution of data

thus they can generate the new points

Story...

A father has two kids, Kid A and Kid B. Kid A has a special character whereas he can learn everything in depth. Kid B have a special character whereas he can only learn the differences between what he saw.

One fine day, The father takes two of his kids (Kid A and Kid B) to a zoo. This zoo is a very small one and has only two kinds of animals say a lion and an elephant. After they came out of the zoo, the father showed them an animal and asked both of them "**is this animal a lion or an elephant**?"

The Kid A, the kid suddenly draw the image of lion and elephant in a piece of paper based on what he saw inside the zoo. He compared both the images with the animal standing before and answered based on the **closest match** of image & animal, he answered: "The animal is Lion".

The Kid B knows only the differences, based on **different properties learned**, he answered: "The animal is a Lion".

Here, we can see both of them is finding the kind of animal, but the way of learning and the way of finding answer is entirely different. In Machine Learning, We generally call Kid A as a Generative Model & Kid B as a Discriminative Model.

Generative

- data distribution
- accuracy high
- Very large data needed
- Naive Bayes
  Bayes
- more effected by outlier

Disc.

- we only need a Classification line
- accuracy lower
- Very less data
- SVM, LR, KNN
  DT
- generally these are used, bcoz they need less Computation, less memory etc

Discriminative vs. Generative Learning

Let's consider an example.
Imagine yourself as a language classification system.

There are two ways you can classify languages.

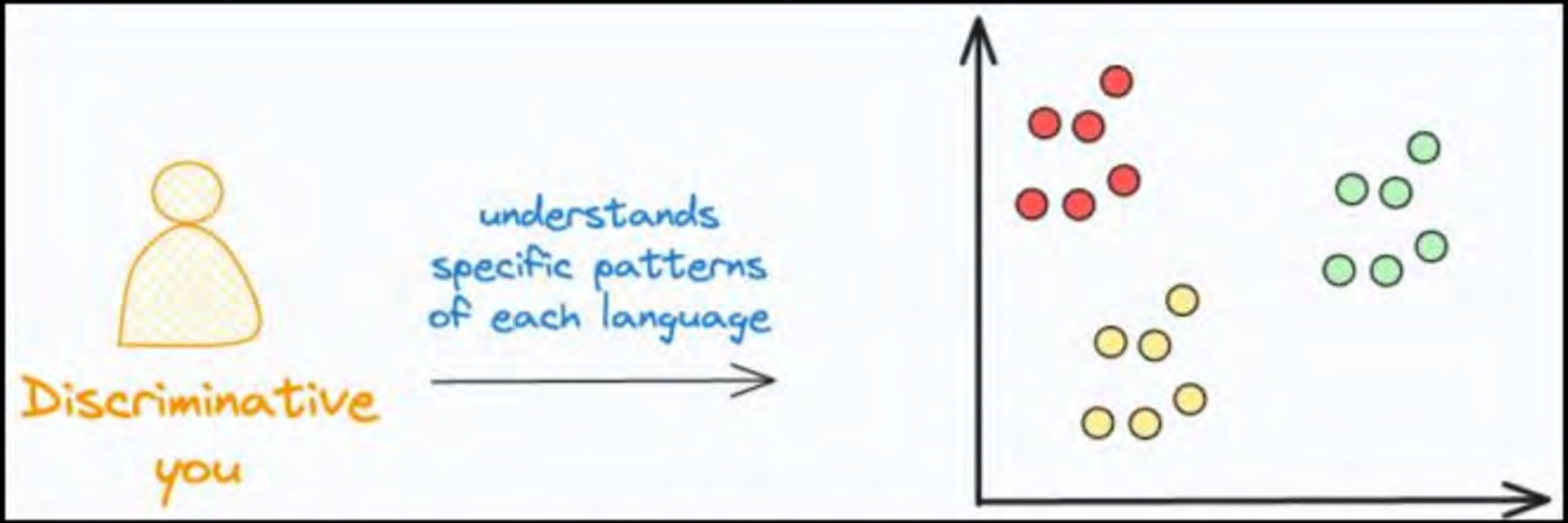
- ❑ Learn every language and then classify a new language based on acquired knowledge.
- ❑ Understand some distinctive patterns in each language without truly learning the language. Once done, classify a new language.

Can you figure out which of the above is generative and which one is discriminative?

Discriminative vs. Generative Learning

The second approach is a **discriminative approach**. This is because you only learned specific distinctive patterns of each language. It is like:

- If so and so words appear, it is likely "Langauge A."
- If this specific set of words appear, it is likely "Langauge B." and so on.



In other words, you learned the conditional distribution P(Language|Words).

Discriminative vs. Generative Learning

- ☐ Also, the above description might persuade you that generative models are more generally useful, but it is not true.
- ☐ This is because generative models have their own modeling complications.
- ☐ For instance, typically, generative models require more data than discriminative models.
- ☐ Relate it to the language classification example again.
- ☐ Imagine the amount of data you would need to learn all languages (generative approach) vs. the amount of data you would need to understand some distinctive patterns (discriminative approach).
- ☐ Typically, discriminative models outperform generative models in classification tasks.

Discriminative vs. Generative Learning

- In General, A Discriminative model models the **decision boundary between the classes.**
- A Generative Model explicitly models the **actual distribution of each class.**
- In final both of them is predicting the conditional probability P(Animal | Features). But Both models learn different probabilities.
- A Generative Model learns the **joint probability distribution p(x,y)**. It predicts the conditional probability with the help of **Bayes Theorem.**
- A Discriminative model learns the **conditional probability distribution p(y|x)**. Both of these models were generally used in **supervised learning** problems.

Discriminative Learning

- The discriminative model learn the boundaries between classes or labels in a dataset.
- Discriminative models focus on modelling the decision boundary between classes in a classification problem. The goal is to learn a function that maps inputs to binary outputs, indicating the class label of the input.
- Maximum likelihood estimation is often used to estimate the parameters of the discriminative model, such as the coefficients of a logistic regression model or the weights of a neural network.
- Discriminative models (just as in the literal meaning) separate classes. But these models are not capable of generating new data points. Therefore, the ultimate objective of discriminative models is to separate one class from another.
- If we have some outliers present in the dataset, discriminative models work better compared to generative models i.e., discriminative models are more robust to outliers.
- But overall the accuracy of discriminative model is less than the generative models.

Generative and Descriptive Learning

- Examples of Discriminative Models
  - Logistic regression
  - Support vector machines(SVMs)
  - Traditional neural networks
  - Nearest neighbor
  - Conditional Random Fields (CRFs)
  - Decision Trees and Random Forest
- Outliers have little to no effect on these models. They are a better choice than generative models, but this leads to misclassification problems which can be a major drawback.

Bayesian Decision Theory

Generative Learning

- Generative models are machine learning models that learn to generate new data samples similar to the training data they were trained on. They capture the underlying distribution of the data and can produce novel instances.
- So, the Generative approach focuses on the distribution of individual classes in a dataset, and the learning algorithms tend to model the underlying patterns or distribution of the data points (e.g., gaussian). These models use the concept of joint probability and create instances where a given feature (x) or input and the desired output or label (y) exist simultaneously.
- These models use probability estimates and likelihood to model data points and differentiate between different class labels present in a dataset. Unlike discriminative models, these models can also generate new data points.
- However, they also have a major drawback – If there is a presence of outliers in the dataset, then it affects these types of models to a significant extent.

Generative and Descriptive Learning

- Generative model
- As the name suggests, generative models can be used to generate new data points. These models are usually used in unsupervised machine learning problems.
- Generative models go in-depth to model the actual data distribution and learn the different data points, rather than model just the decision boundary between classes.
- These models are prone to outliers, which is their only drawback when compared to discriminative models. The mathematics behind generative models is quite intuitive too. The method is not direct like in the case of discriminative models. To calculate $P(Y|X)$, they first estimate the prior probability $P(Y)$ and the likelihood probability $P(X|Y)$ from the data provided.

- Discriminative models

↓

- Here we start with model parameters and find best values of parameter for which accuracy on $\underbrace{\text{training data}}_{\text{loss fxn}}$ is low

- So Discriminative model loss fxn $\approx$ MLE.

- So they also try to find PDF.

Logistic Reg → MLE

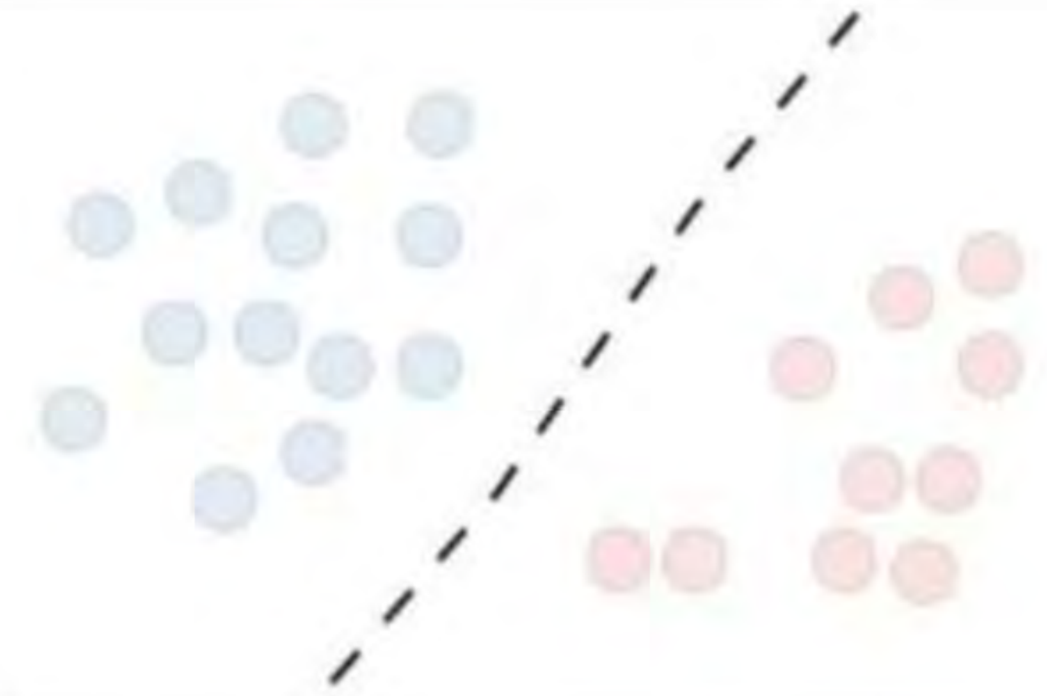
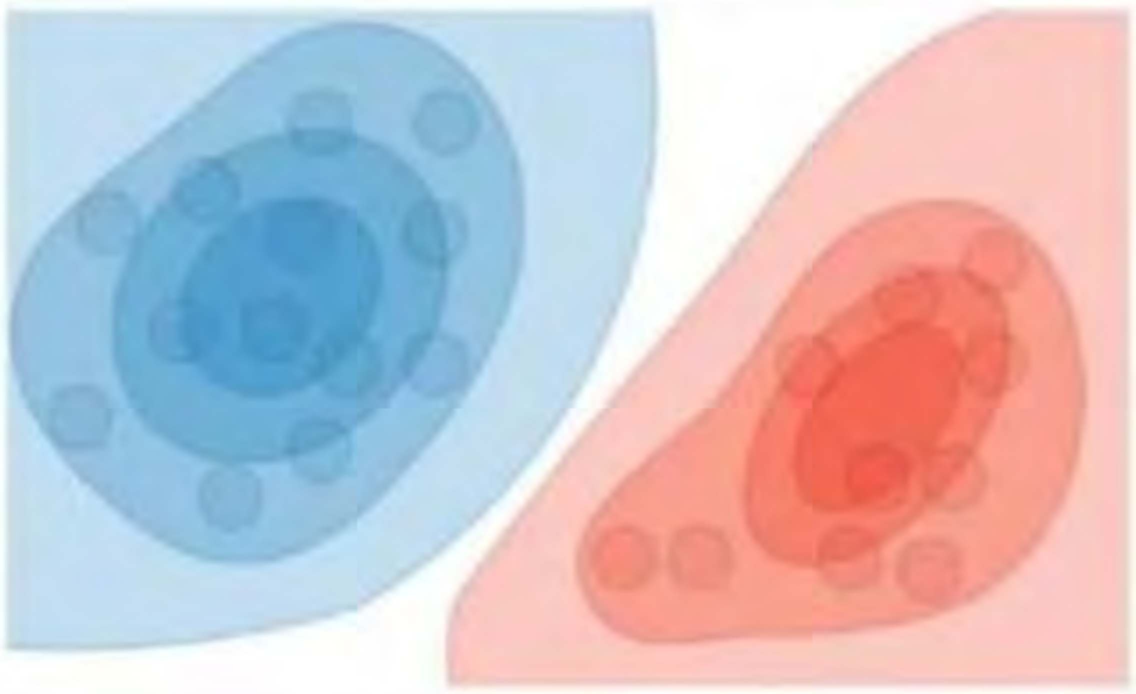
Linear Reg → OLS

OLS ↓ MLE

Both Gen/Disc model stay to find $P(Y|x)$ ⇒ discumative model do this by loss fxn

Gen model do this Using the Bayes theorem and Joint PDF $P(x, y)$

Generative and Descriptive Learning

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Given a discrete K-class dataset containing N points, where sample points are described using D features with each feature capable of taking V values, how many parameters need to be estimated for Naïve Bayes Classifier?

(A)	$V^D K$	(C)	VDK
(B)	K^{V^D}	(D)	$K(V + D)$

$((VDK) + K)$

Q1-1: Which of the following about Naive Bayes is incorrect?

- A Attributes can be nominal or numeric
- B Attributes are equally important
- C Attributes are statistically dependent of one another given the class value
- D Attributes are statistically independent of one another given the class value
- E All of above

Q1-2: Consider a classification problem with two binary features, $x_1, x_2 \in \{0,1\}$. Suppose $P(Y = y) = 1/32$, $P(x_1 = 1 | Y = y) = y/46$, $P(x_2 = 1 | Y = y) = y/62$. Which class will naive Bayes classifier produce on a test item with $x_1 = 1$ and $x_2 = 0$?

- A 16
- B 26
- C 31
- D 32

$P_Y \, P(x_1 = 1 / Y) \, P(x_2 = 0 | Y)$

$\frac{1}{32} \left(y/46 \right) \left(1 - y/62 \right)$

$\frac{d}{dy} \frac{1}{32} \left(\frac{y}{46} - y^2/46 \times 62 \right) \Rightarrow y = 31$

Q1-3: Consider the following dataset showing the result whether a person has passed or failed the exam based on various factors. Suppose the factors are independent to each other. We want to classify a new instance with Confident=Yes, Studied=Yes, and Sick=No.

(done)

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

- A Pass
- B Fail

12. Identify the parametric machine learning algorithm.

a) CNN (Convolutional neural network)

b) KNN (K-Nearest Neighbours) → not parametric

c) Naïve Bayes

d) SVM (Support vector machines)

Impact of outlier

- The outliers can Impact PDF is, $P(x_i/C_j)$ the class Conditioned PDF become distort.

Impact of missing data

- if we have some missing value in training data then we can see that it will not effect PDF of dimension

	Rain	Wind	Hum	Temp	
1				80	C_1
2				85	C_2
3				$\square$	
4				90	C_1
5				92	C_3
6				82	

- Test point

R	W		Temp

$$\Rightarrow \max P_{c_i}\, P(R|c_i)\, P(W|c_i)\, P(H|c_i)\, P(T|c_i)$$

Skip

So in testing if any point is missing then skip that dimension in analysis

- Time complexity ⇒ we need to find (ADM + M) number of Probab

  $\rightarrow O(ADm + m)$

- Space Complexity ⇒ we need to store these parameters

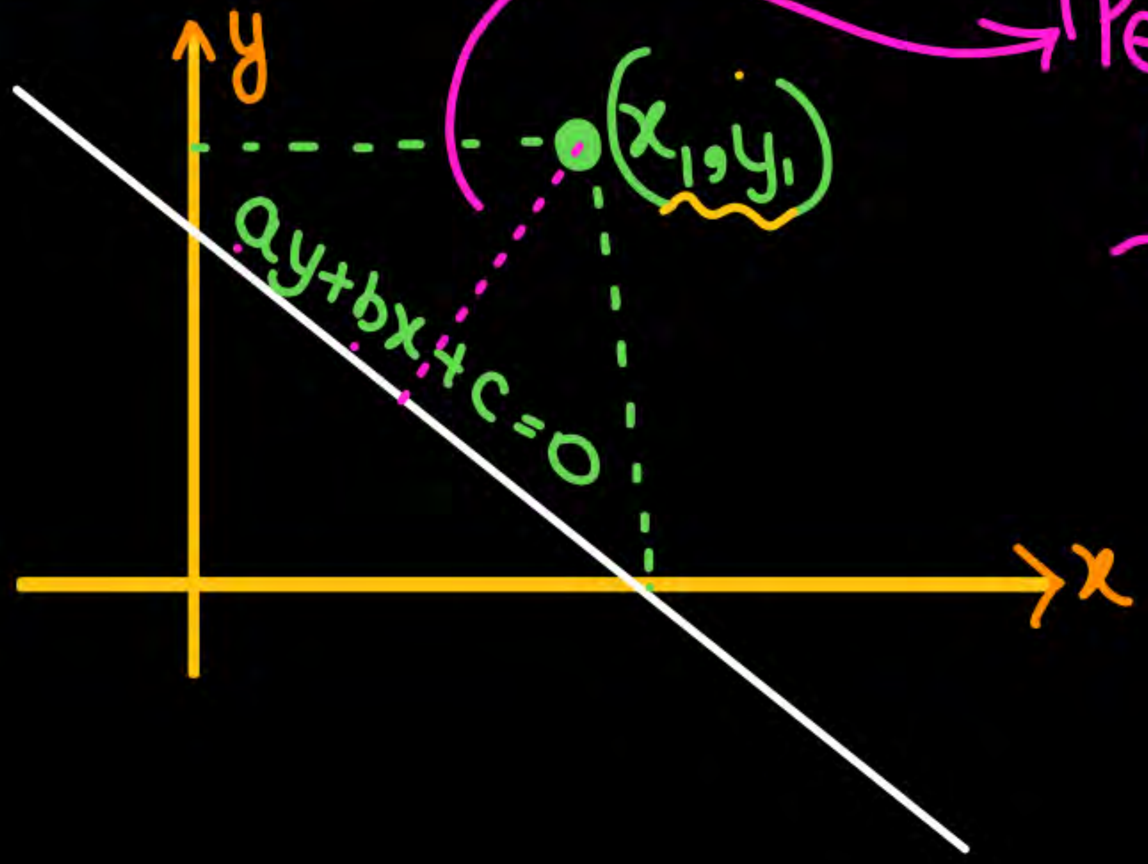
  $O(ADm + m)$

SVMs

Support Vector Machine

Distance of a point from a hyperplane...

Lets see some geometry...



Perpendicular distance of a point from a line

$$\frac{ay_1+bx_1+c}{\sqrt{a^2+b^2}}$$

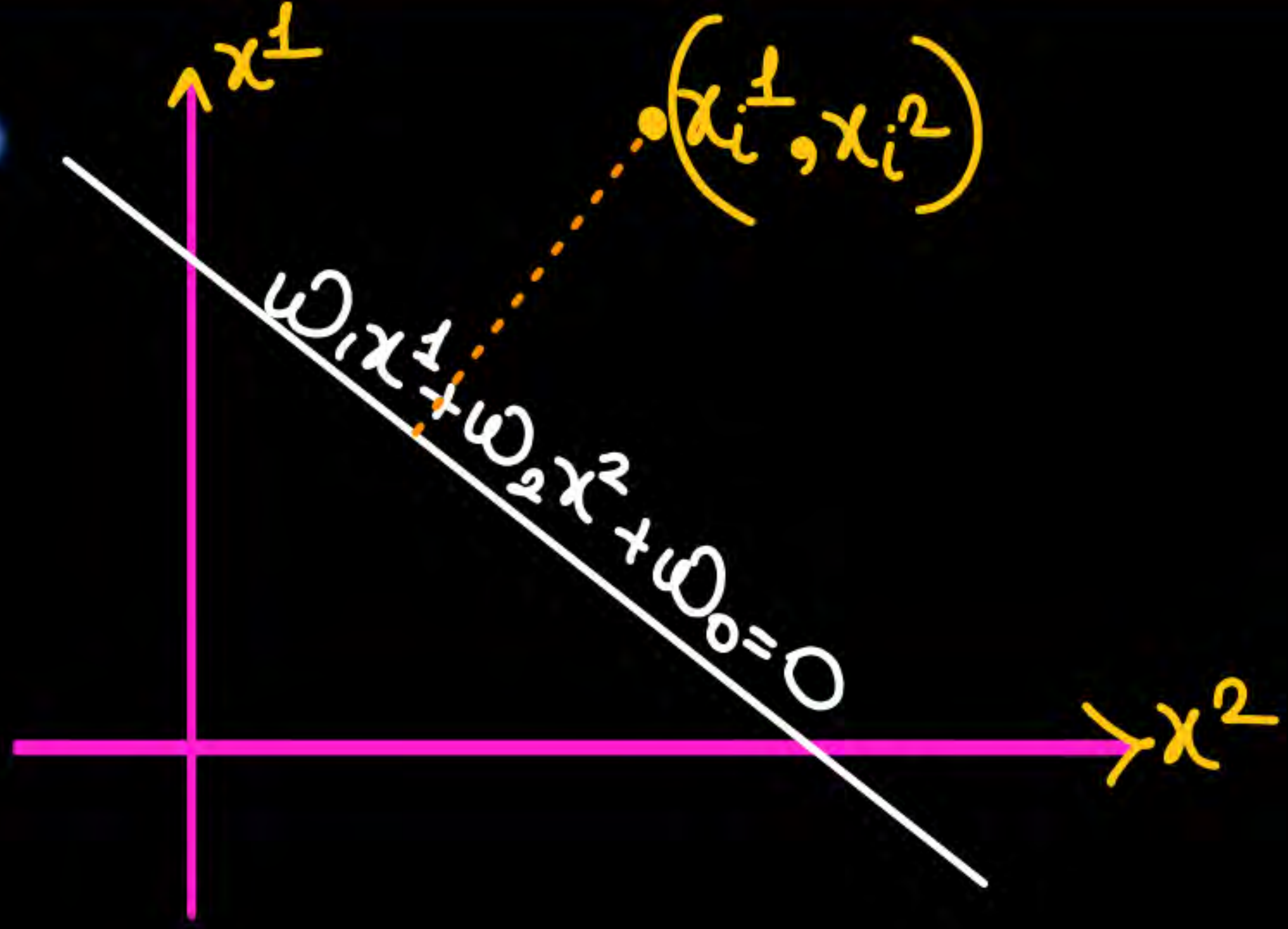
SVMs

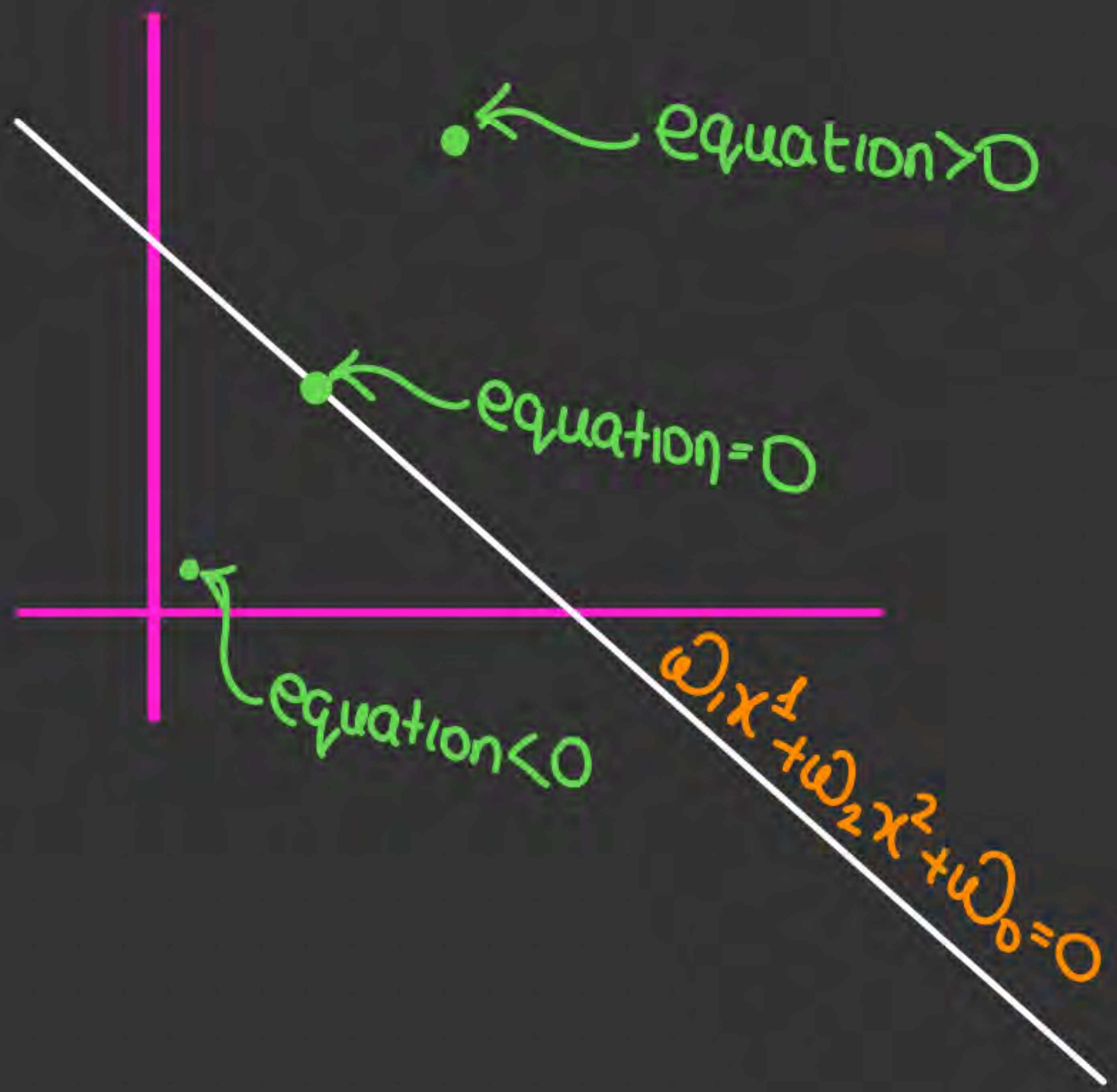
Support Vector Machine

Distance of a point from a hyperplane...

distance of point from line

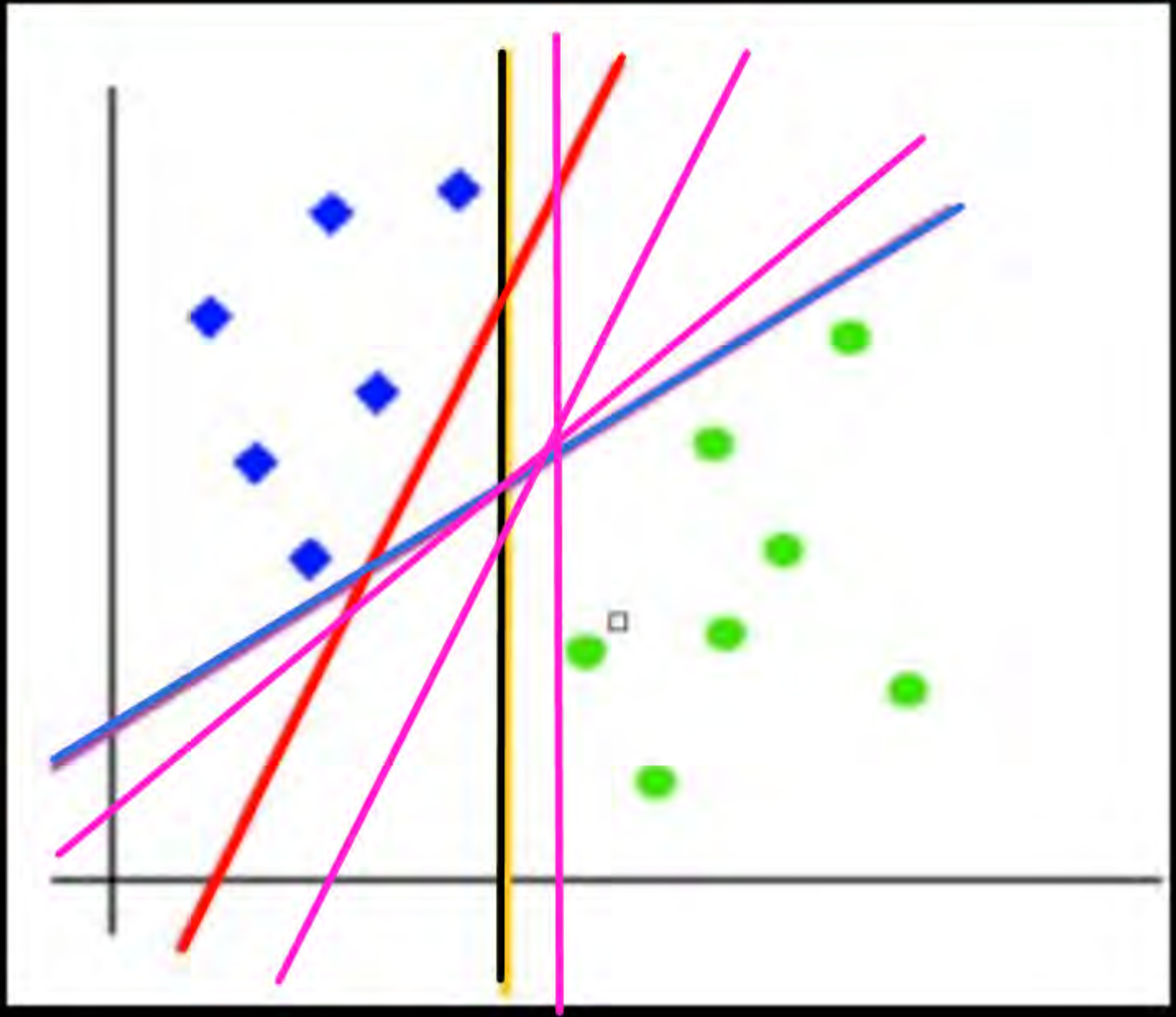
$$= \frac{\omega_1 x_i^1 + \omega_2 x_i^2 + \omega_0}{\sqrt{\omega_1^2 + \omega_2^2}}$$




equation>0
equation=0
equation<0
$\omega_1 x^1 + \omega_2 x^2 + \omega_0 = 0$

SVMs

Support Vector Machine



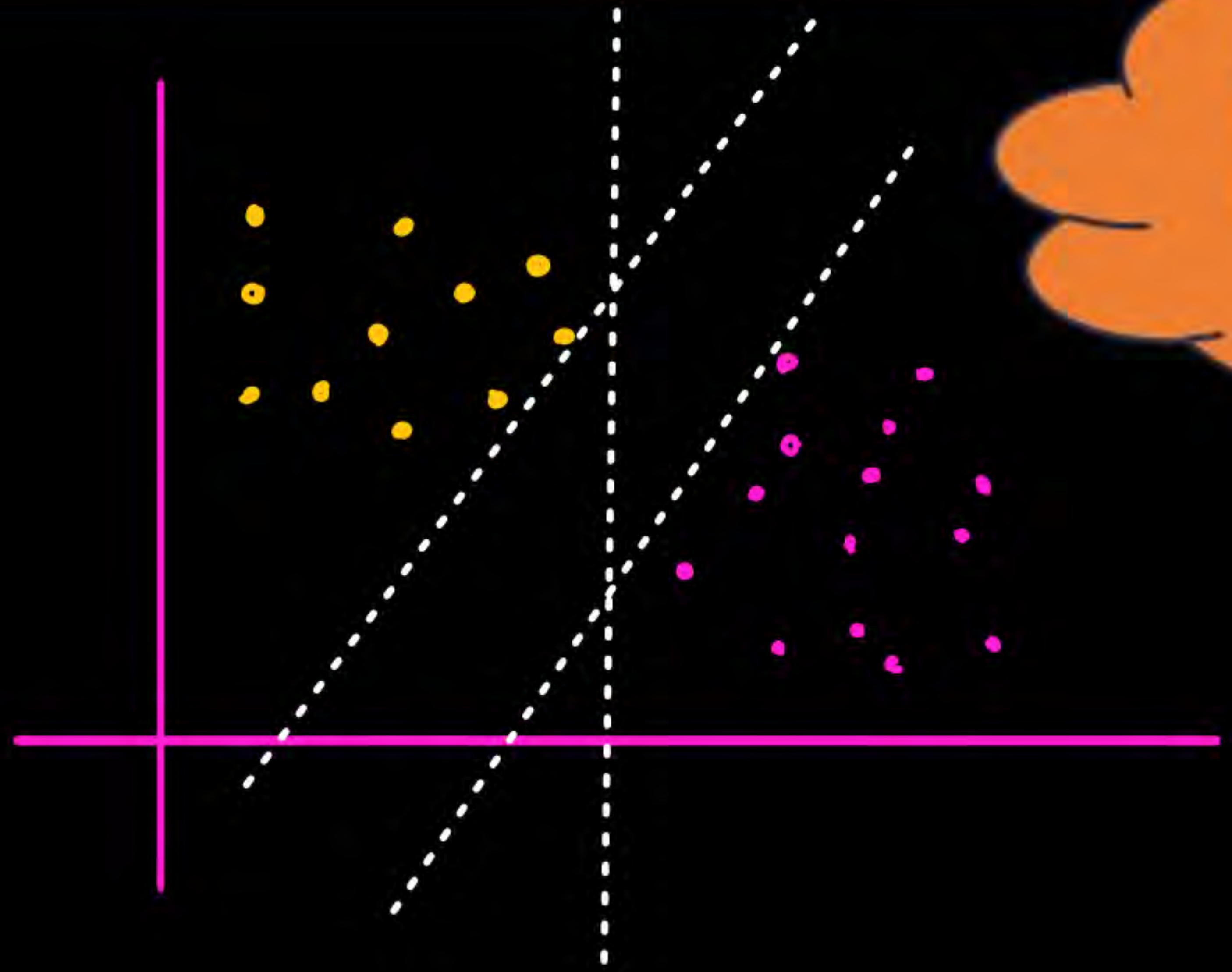
Why we need SVMs ⇒

Lets see the case of classification

* So as we can see, we can create many classifier, all giving zero error training
* But to get the best classifier we use svm

SVMs

Support Vector Machine

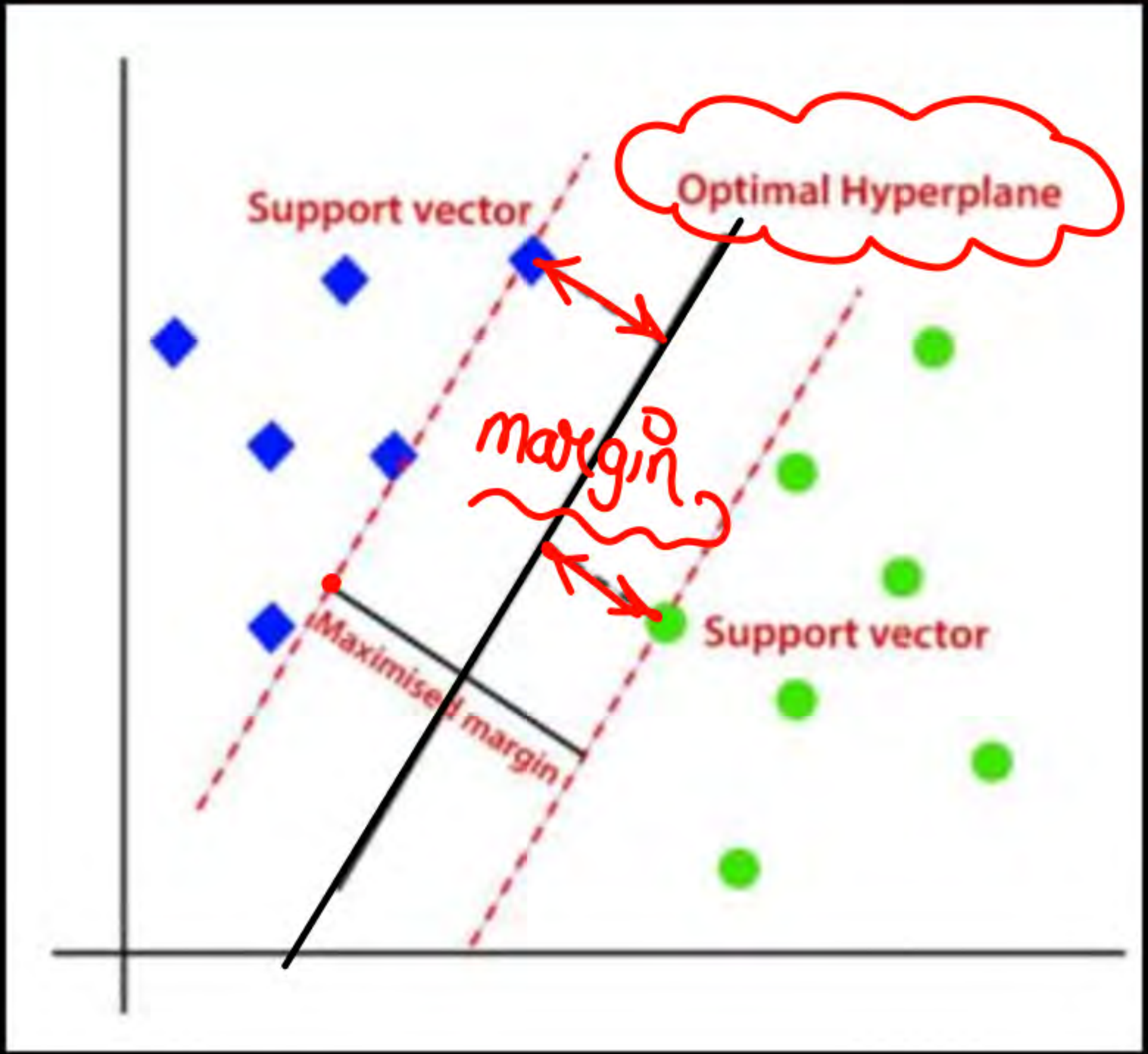


Why we need SVMs
Lets see the case of classification

Best classifier by svm is such that it maintain a margin from both classes distance

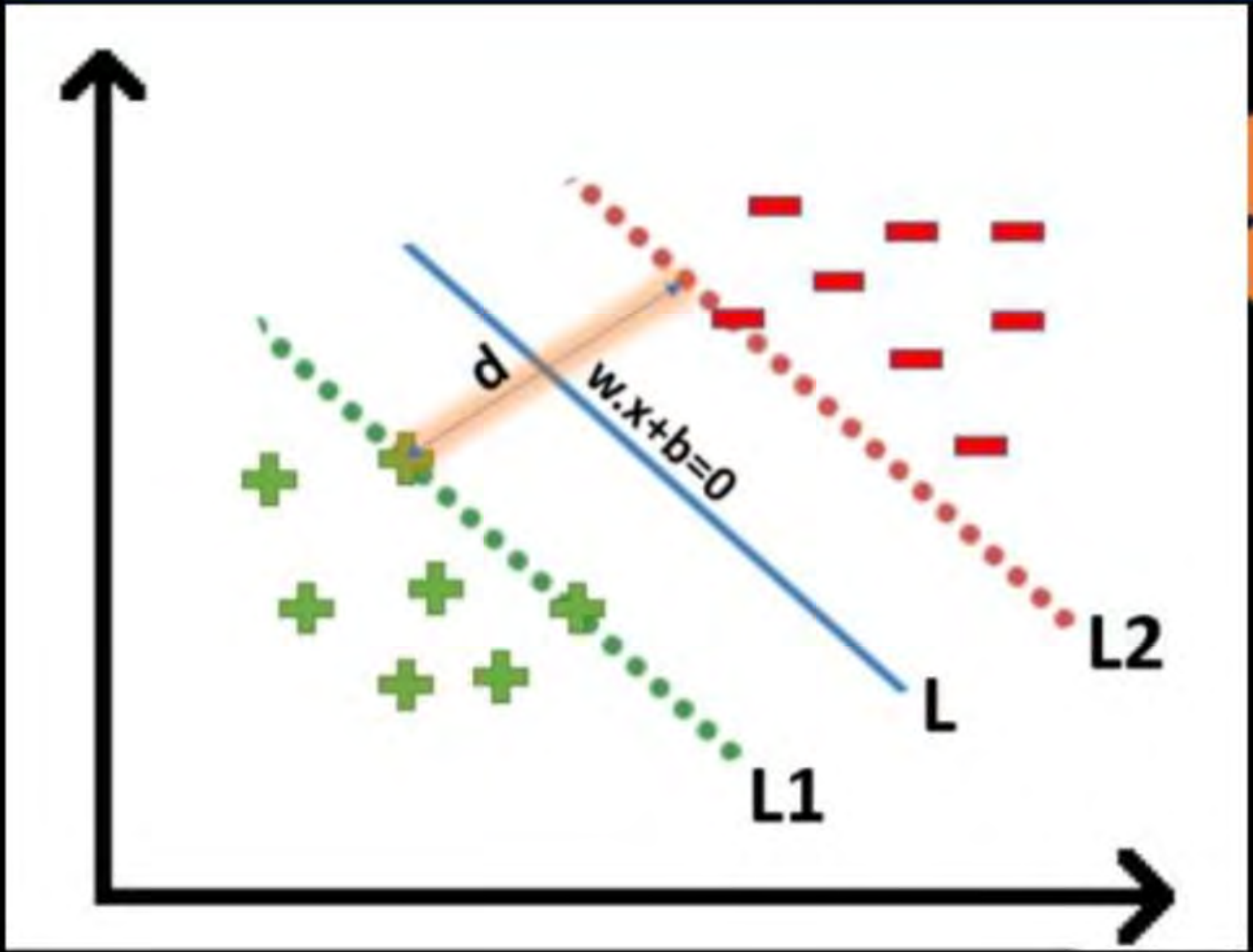
SVMs

Support Vector Machine



Why we need SVMs
Lets see the case of classification

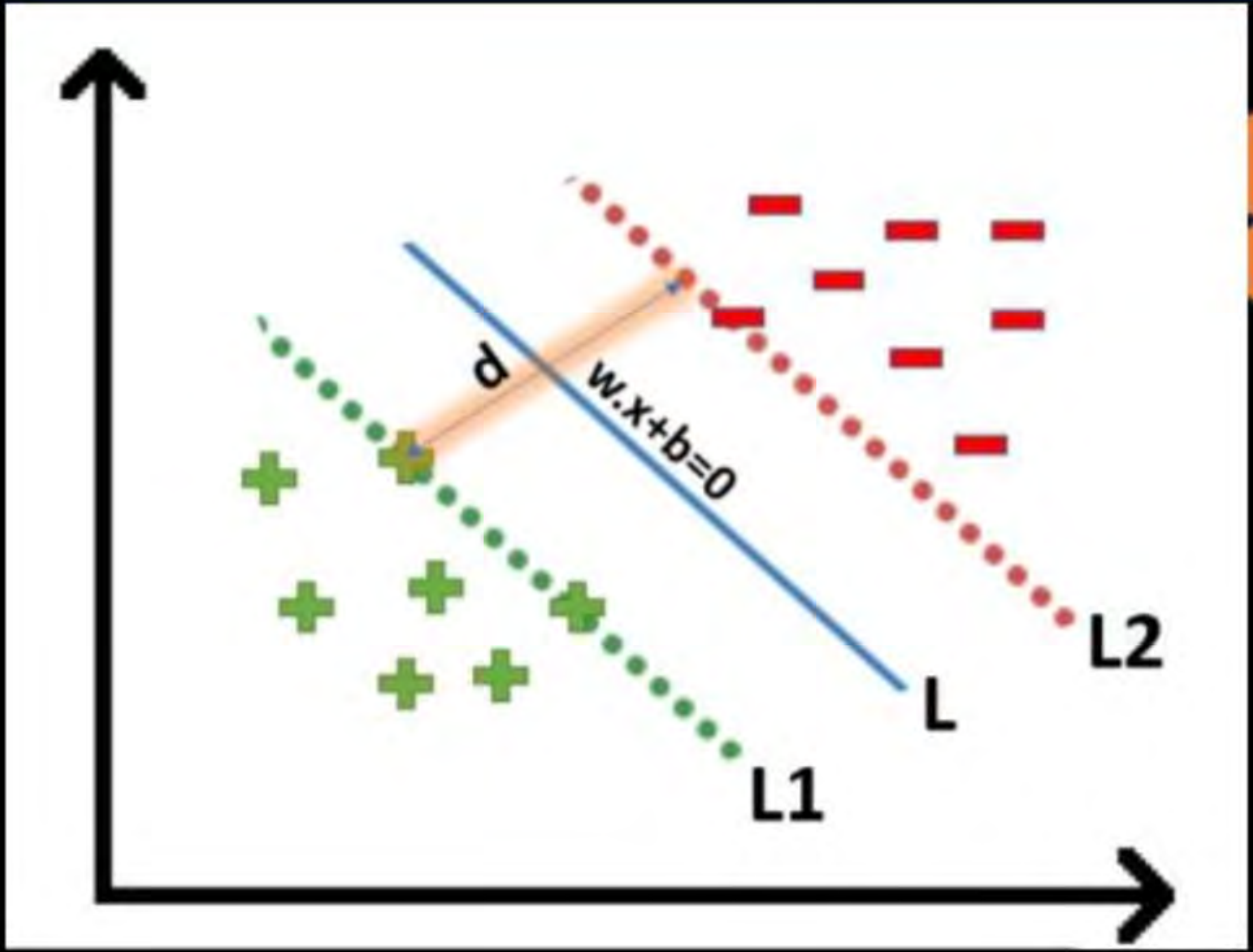
Support Vector Machine



So in SVM we want the classifier which has max separation from both the classes..

SVMs

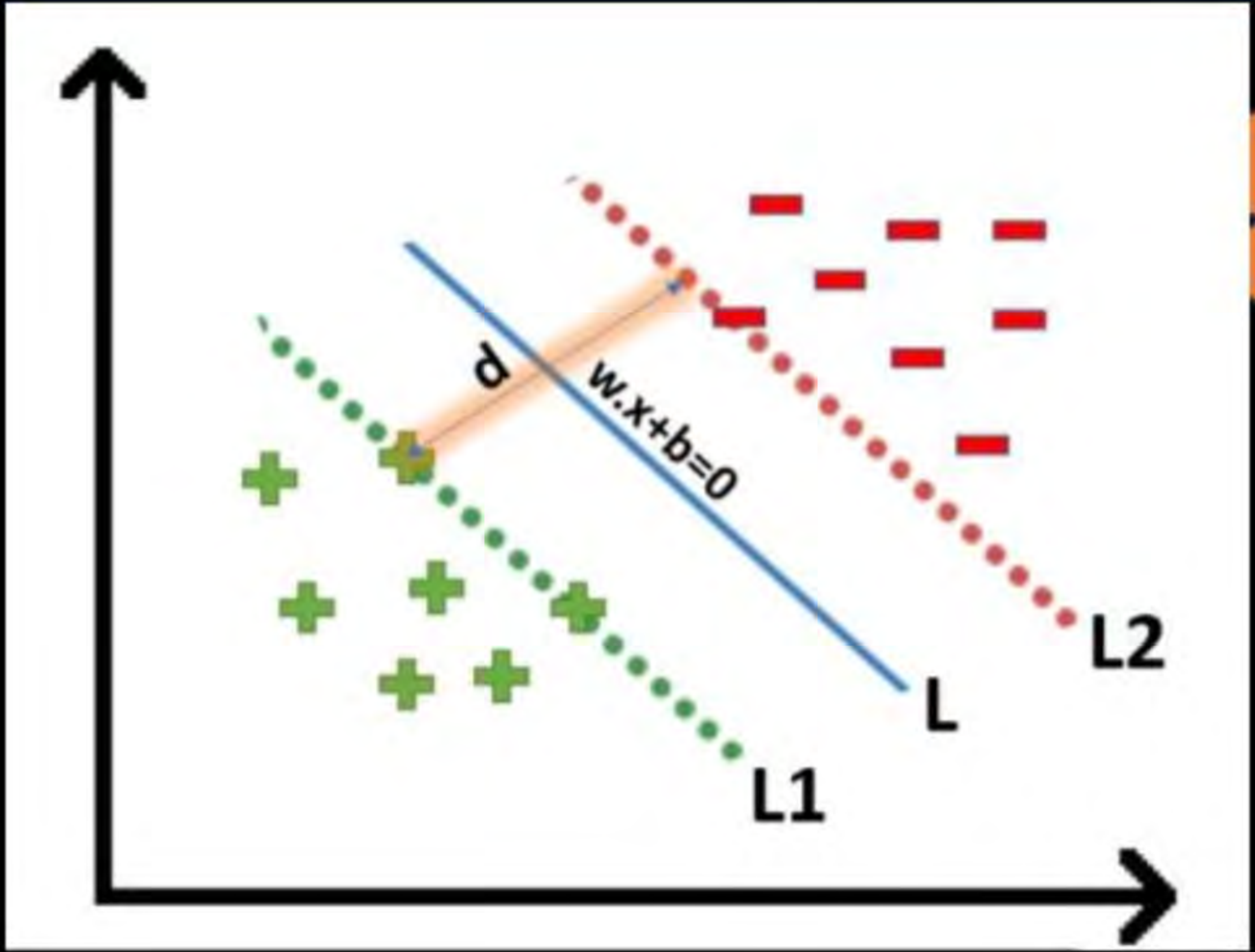
Support Vector Machine



The points above the line have …..
And below the line have ….

SVMs

Support Vector Machine



The equation of Line L, L1 and L2 will be ….

THANK - YOU