Concepts of Probabilities

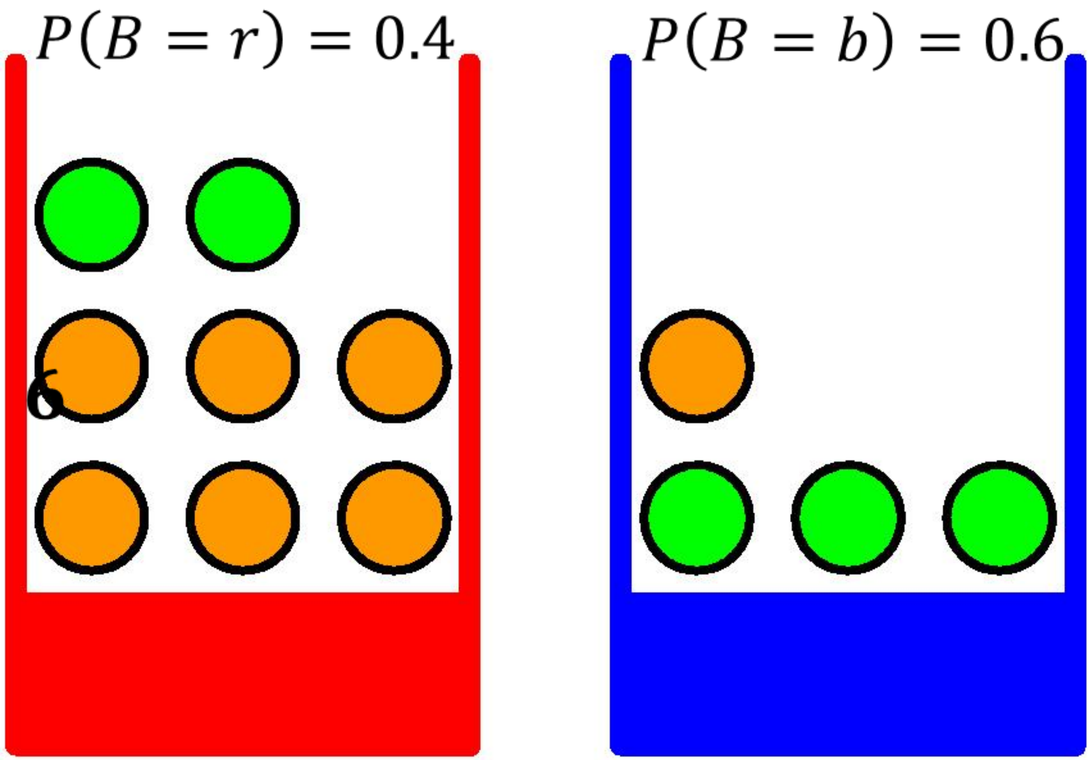
Outline

- Conditional, Joint, Marginal Probabilities
- Sum Rule and Product Rule
- Bayes’ Theorem
- Probability Distribution
  - Bernoulli Distribution
  - Normal/Gaussian Distribution
  - Central Limit Theory
- Monte Carlo Approximation

Introduction to Probability

A simple orienting example

$\boldsymbol{a = 2} \quad \boldsymbol{o = 6}$



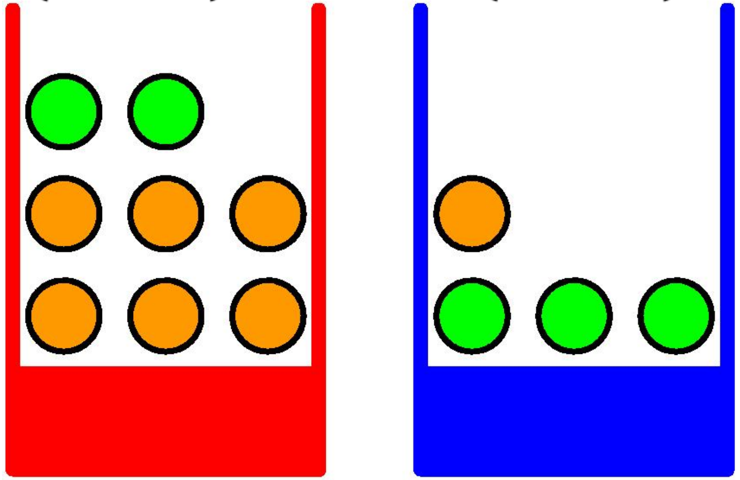
Random Variables:
B = {b,r} □ Baskets
F = {a,p} □ Fruits

$\boldsymbol{a = 3} \quad \boldsymbol{o = 1}$

1. What is the probability of picking a fruit is orange?
2. What is the probability that I will peak the fruit from red basket **given that** fruit was orange?

A simple orienting example

$P(B = r) = 0.4$ $P(B = b) = 0.6$



100 trials

Probability Table

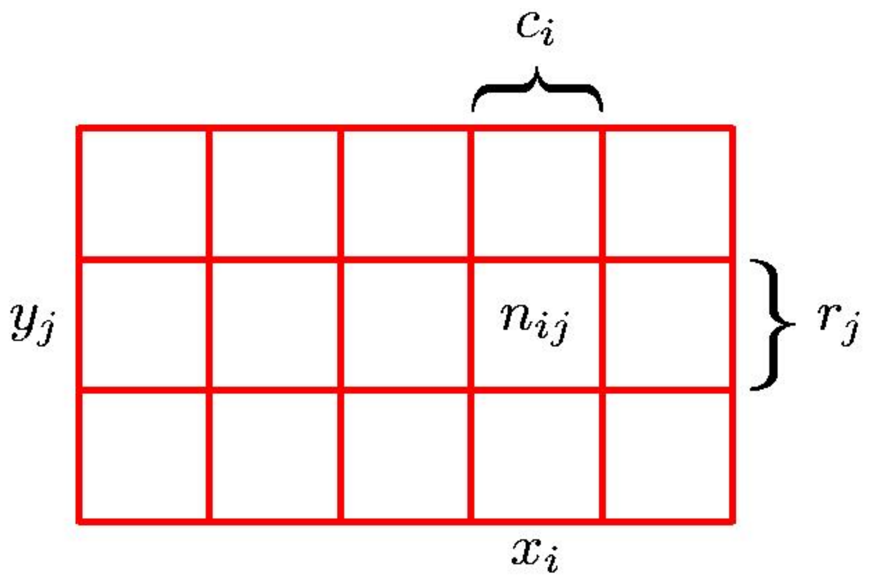
Distribution

	F=o	F=a	
B=r	r,o 30	r,a 10	
B=b	b,o 15	b,a 45	

$$= \frac{1}{4} * 60 = 15$$

Joint Probability (Discrete)

	F = o	F= a	Total
B = r	30	10	40
B = b	15	45	60
	45	55	



Joint probability

The probability that X will take the value x_i and Y will take the value y_j

$$P(X = x_i, Y = y_j)$$

Let the number of trials that $X = x_i \; and \; Y = y_j$ be n_{ij}

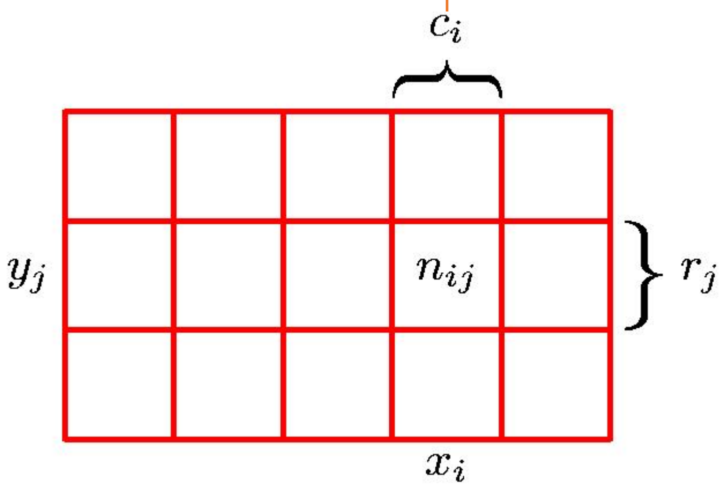
Then, $P\left(X = x_i, Y = y_j\right) = \frac{n_{ij}}{N}$

$$\boldsymbol{P(B = r, F = a) = \frac{10}{100} = 0.1}$$

Adapted from Dr Christopher Bishop's slides

Sum Rule

	F = o	F= a	Total
B = r	30	10	40
B = b	15	45	60
	45	55	



Let number of trials that $X = x_i$ be c_i

Then, $P(X = x_i) = \frac{c_i}{N}$ Marginal probability

$$P(F = o)$$
$$= P(F = o, B = r) + P(F = o, B = b)$$
$$= 0.3 + 0.15 = 0.45$$

$$c_i = \sum_j n_{ij}$$

$$\Rightarrow P(X = x_i) = \sum_j \frac{n_{ij}}{N}$$

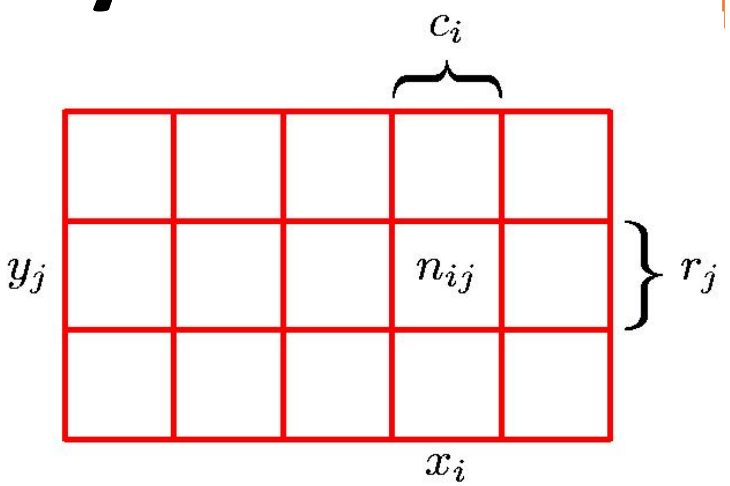
$$\boxed{\Rightarrow P(X = x_i) = \sum_j P(X = x_i, Y = y_j)}$$

Sum rule of probability

Adapted from Dr Christopher Bishop's slides

Conditional Probability

	F = o	F= a	Total
B = r	30	10	40
B = b	15	45	60
	45	55	



The probability that Y will take the value y_j **given that** X has taken the value x_i

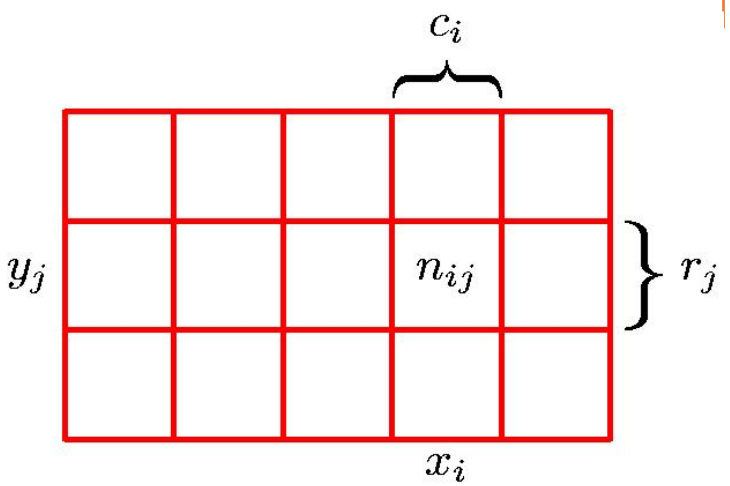
$$P(Y = y_j | X = x_i)$$

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Adapted from Dr Christopher Bishop's slides

Product Rule

	F = o	F= a	Total
B = r	30	10	40
B = b	15	45	60
	45	55	



$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

$$\Rightarrow \boxed{P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i)P(X = x_i)}$$

Product rule of probability

Joint probability Conditional probability Marginal probability

Adapted from Dr Christopher Bishop’s slides

Bayes' Theorem

Product Rule $P(X,Y) = P(Y|X)P(X)$

Similarly, $P(Y,X) = P(X|Y)P(Y)$

Since $P(X,Y) = P(Y,X)$ we obtain that

$$P(Y|X)P(X) = P(X|Y)P(Y)$$

So, $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$

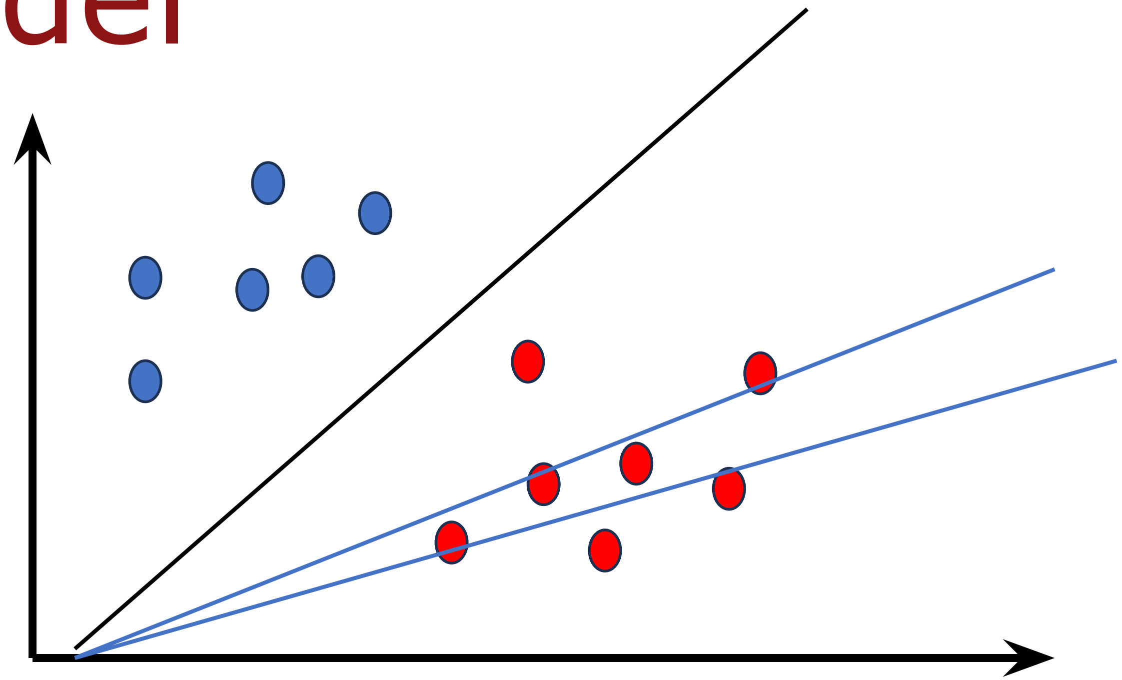
Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Bayesian Theorem

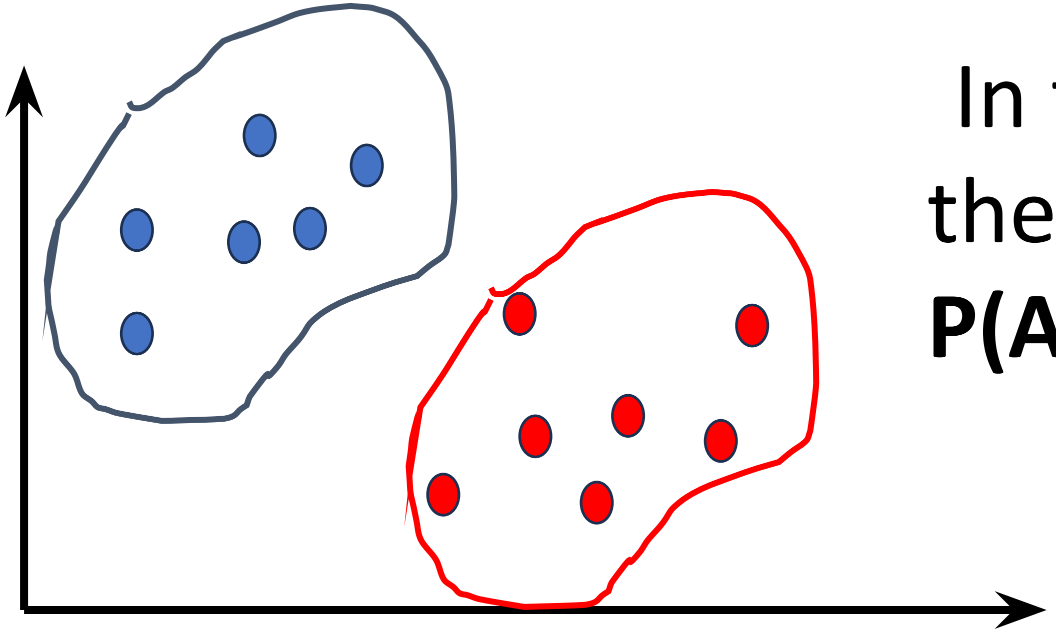
Generative Model

Discriminative vs Generative Model



A Discriminative model models the **decision boundary between the classes**.

Discriminative vs Generative Model



In final both of them is predicting the conditional probability **P(Animal | Features)**

A **Generative Model** explicitly models the actual distribution of each class

Discriminative vs Generative Model

A Generative Model learns the **joint probability distribution p(x,y).** It predicts the conditional probability with the help of **Bayes Theorem**.

P(Y|X) = P(X|Y)* P(Y) / P(X)

Posterior = (Likelihood * Prior) / Evidence

Discriminative model learns the **conditional probability distribution p(y|x)**.
Both of these models were generally used in **supervised learning** problems.

Discriminative vs Generative Model

Generative classifiers

- Assume some functional form for **P(Y), P(X|Y)**
- Estimate parameters of **P(X|Y), P(Y)** directly from training data
- Use Bayes rule to calculate **P(Y |X)**

Discriminative Classifiers

- Assume some functional form for **P(Y|X)**
- Estimate parameters of **P(Y|X)** directly from training data

Generative classifiers

- Naïve Bayes
- Bayesian networks
- Markov random fields
- Hidden Markov Models (HMM)

Discriminative Classifiers

- Logistic regression
- Scalar Vector Machine
- Traditional neural networks
- Nearest neighbour

Application of Bayesian classifiers

- Text-based classification such as spam or junk mail filtering, author identification, or topic categorization
- Medical diagnosis such as given the presence of a set of observed symptoms during a disease, identifying the probability of new patients having the disease
- Network security such as detecting illegal intrusion or anomaly in computer networks

Naïve Bayes Algorithm

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- **Naïve:** The occurrence of a certain feature is independent of the occurrence of other features.
- Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

Posterior = (Likelihood * Prior) / Evidence

Naïve Bayes Algorithm

$$P(Y=1|X) = P(X|Y=1) * P(Y=1) \cancel{/} P(X)$$

$$P(Y=0|X) = P(X|Y=0) * P(Y=0) \cancel{/} P(X)$$

$$\underset{k \in \{1, \ldots K\}}{argmax} P(C_k) \prod_{i=1}^{n} P(X_i | C_k)$$

By changing the class label, there is no change on denominator. So, we can eliminate calculation of Evidence (P(X)) in the Naïve Bayes.

$$P(Y|X) = P(X|Y) * P(Y)$$

One assumption in Naïve Bayes is all the features are independent and contribute same to the class label. So, considering this, above equation can be re-write as considering 4 features i.e X=[x1,x2,x3,x4];

$$P(Y|X) = P(X1|Y) * P(X2|Y) * P(X3|Y) * P(X4|Y) * P(Y)$$

- Class assignment is selected based on ***maximum a posteriori*** **(MAP)** rule

Example: Training Naïve Bayes Tennis Model

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

P(Play=Yes) = 9/14

P(Play=No) = 5/14

Create probability lookup tables based on training data

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

P(Y=Yes|(Sunny, Cool, High, Strong)) = 2/9 * 3/9 * 3/9 * 3/9 * 9/14 **=0.0053**

P(Y=No|(Sunny, Cool, High, Strong)) = 3/5 * 1/5 * 4/5 * 3/5 * 5/14 **=0.0206**

P(Y=Yes|(Sunny, Cool, High, Strong)) = P(Sunny|Y=Yes) *P(Cool|Yes)* P(High|Yes)* P(Strong |Yes)* P(Y)

Example: Training Naïve Bayes Tennis Model

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

P(Play=Yes) = 9/14

P(Play=No) = 5/14

Create probability lookup tables based on training data

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

P(Y=Yes|(D9)) = ?

P(Y=Yes|(D9)) = 2/9 * 3/9 * 6/9 * 6/9 * 9/14 **=0.021**

P(Y=No|(D9)) = 3/5 * 1/5 * 1/5 * 2/5 * 5/14 **=0.003**

Laplace Smoothing

- **Problem:** categories with no entries result in a value of "0" for conditional probability

- **Solution:** add "1" to numerator and denominator of empty categories

$$P(C|X) = \overset{0}{P(X_1|C)} * P(X_2|C) * P(C)$$

$$P(X_1|C) = \frac{1}{Count(C) + 1}$$

$$P(X_2|C) = \frac{Count(X_2 \,\&\, C) + 1}{Count(C) + m}$$

Types of Naïve Bayes

Naïve Bayes Model	Data Type
Bernoulli	Binary (T/F)
Multinomial	Discrete (e.g. count)
Gaussian	Continuous

Probability Distribution

Bernoulli, Gaussian (Normal), Central Limit Theory

Binary Variables (1)

- Coin flipping: heads=1, tails=0

$$p(x=1|\mu)=\mu$$

- Bernoulli Distribution

$$\begin{aligned}\mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu)\end{aligned}$$

Parameter Estimation (1)

- Maximum Likelihood for Bernoulli
- Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)
- $$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

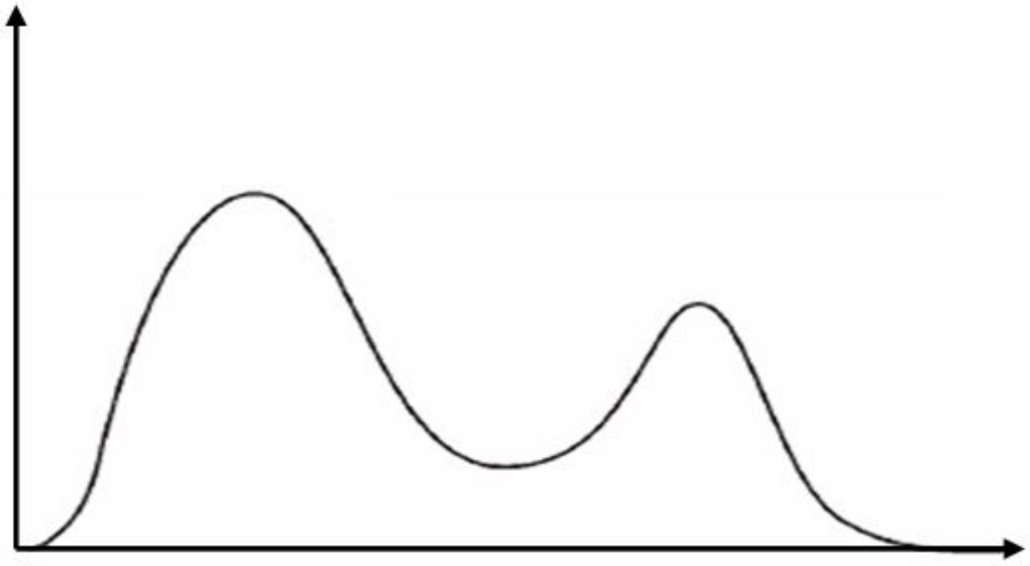
$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$

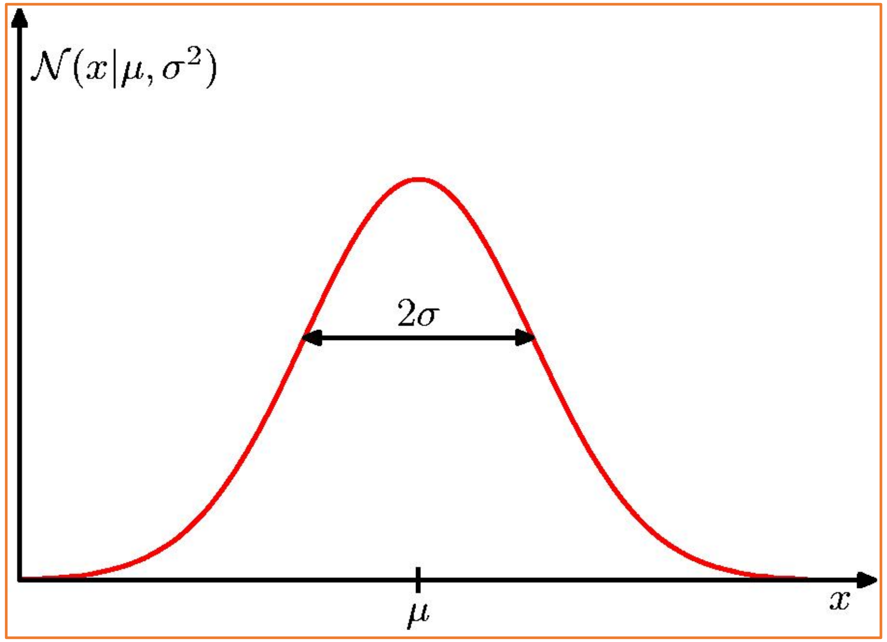
Bimodal Distribution

- The bimodal distribution occurs due to the combination of two groups that have different mean heights between them.

$$\boldsymbol{P(X=k)=(p^{k}.(1-p)^{1-k}).n_r^c}$$



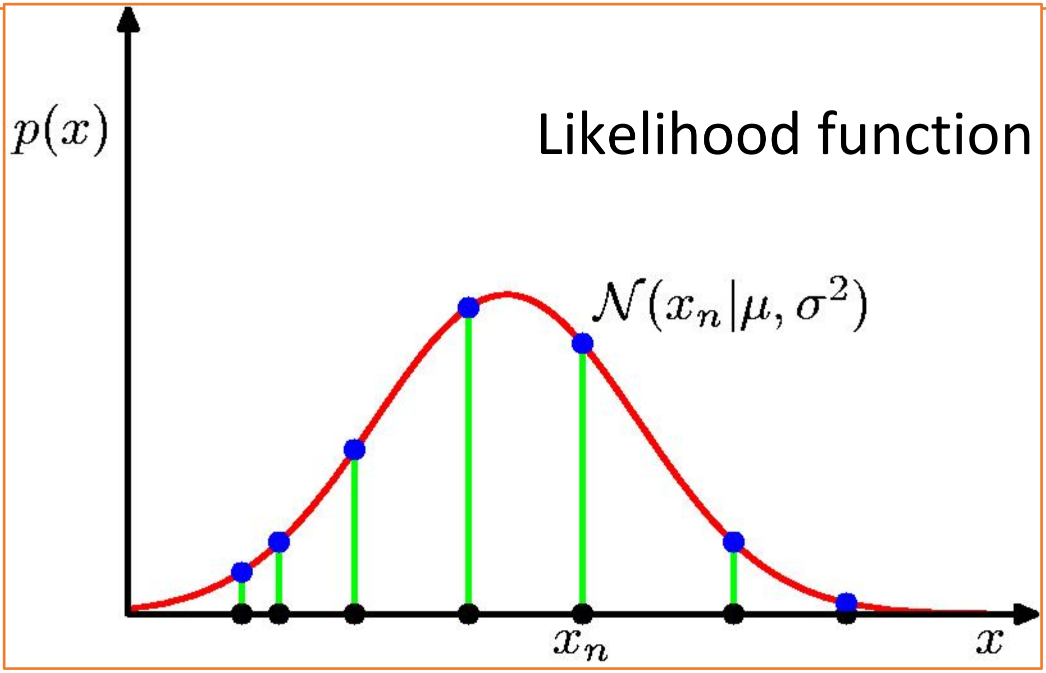
The Gaussian Distribution



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Central Limit Theory

- Central Limit Theorem is generally used to predict the characteristics of a population from a set of sample.
- It uses sampling distribution to generalize the samples and use to calculate approx mean, standard daviation and other important parameters.
- CLT states that if you have a population with **mean μ, sd σ,** and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be normally distributed.
- Example Self study : https://www.geeksforgeeks.org/central-limit-theorem/

Bayesian Network

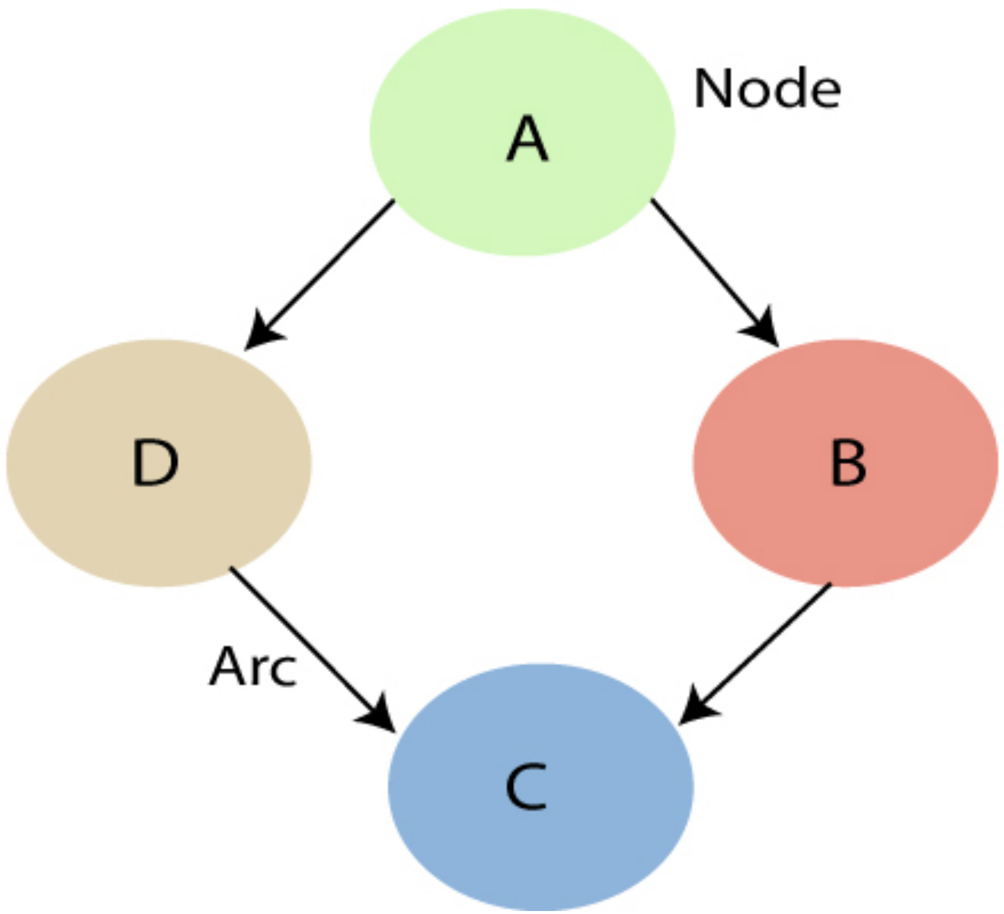
Supervised Machine Learning

Introduction

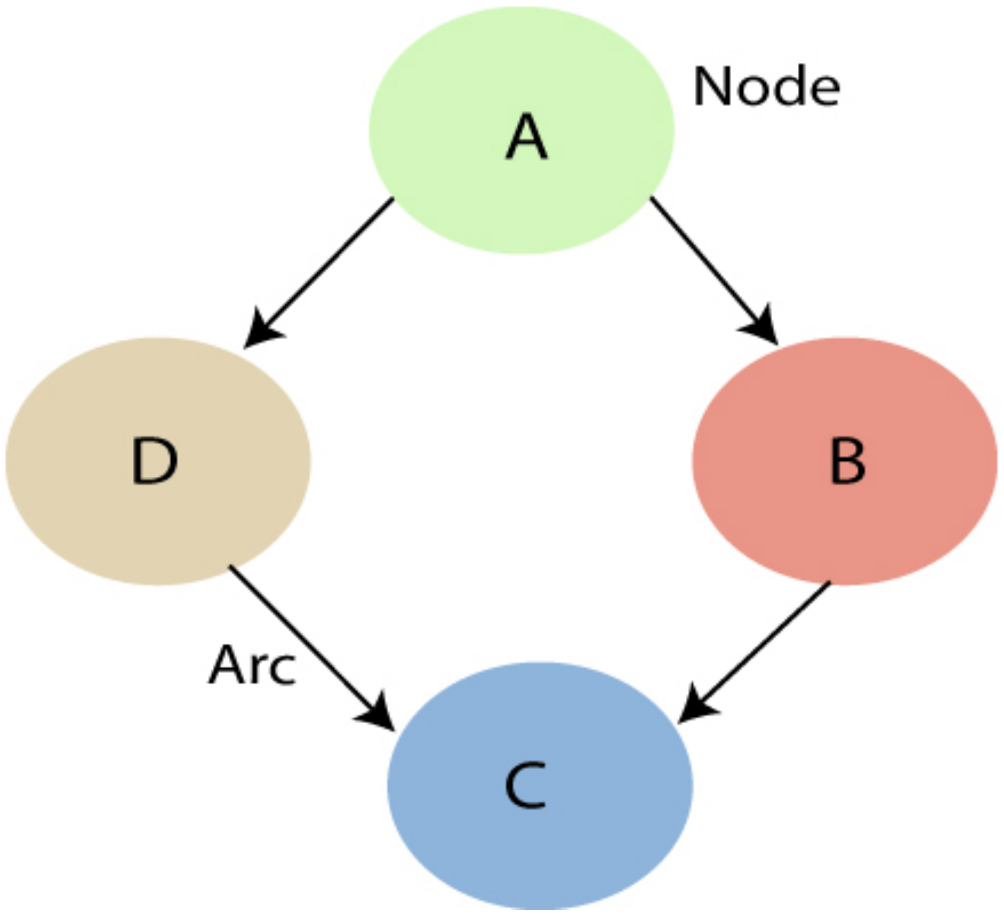
- Probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph.
- **Bayes network, belief network, decision network, or Bayesian model.**
- **Applications:**
  - **Prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction**, and **decision making under uncertainty**

Introduction

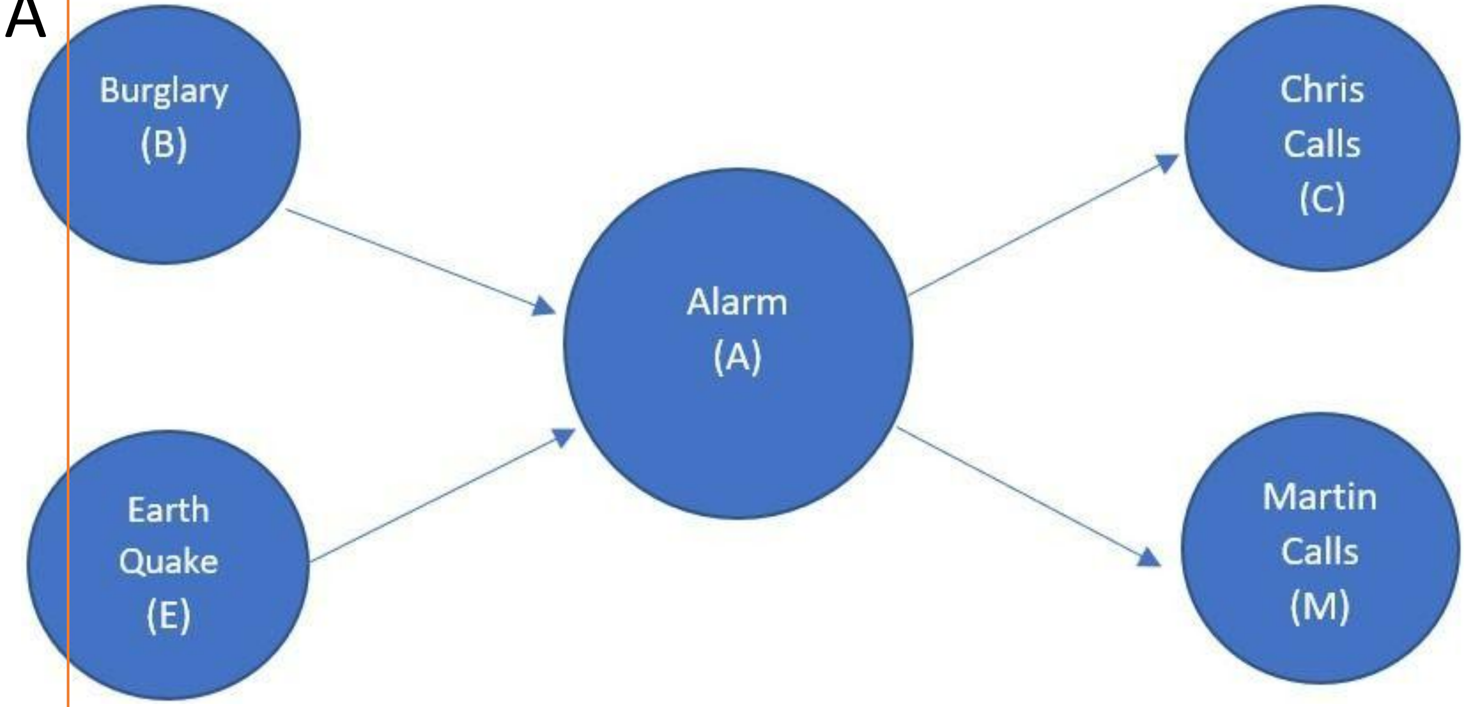
- Consists of two parts
  - **Directed Acyclic Graph**
  - **Table of conditional probabilities.**



- Each **node** corresponds to the random variables, and a variable can be **continuous** or **discrete**.
- **Arc or directed arrows** represent the causal relationship or conditional probabilities between random variables.
- Each node in the Bayesian network has condition probability distribution P(Xi |Parent(Xi)), which determines the effect of the parent on that node.
- **Bayesian network is based on Joint probability distribution and conditional probability.**

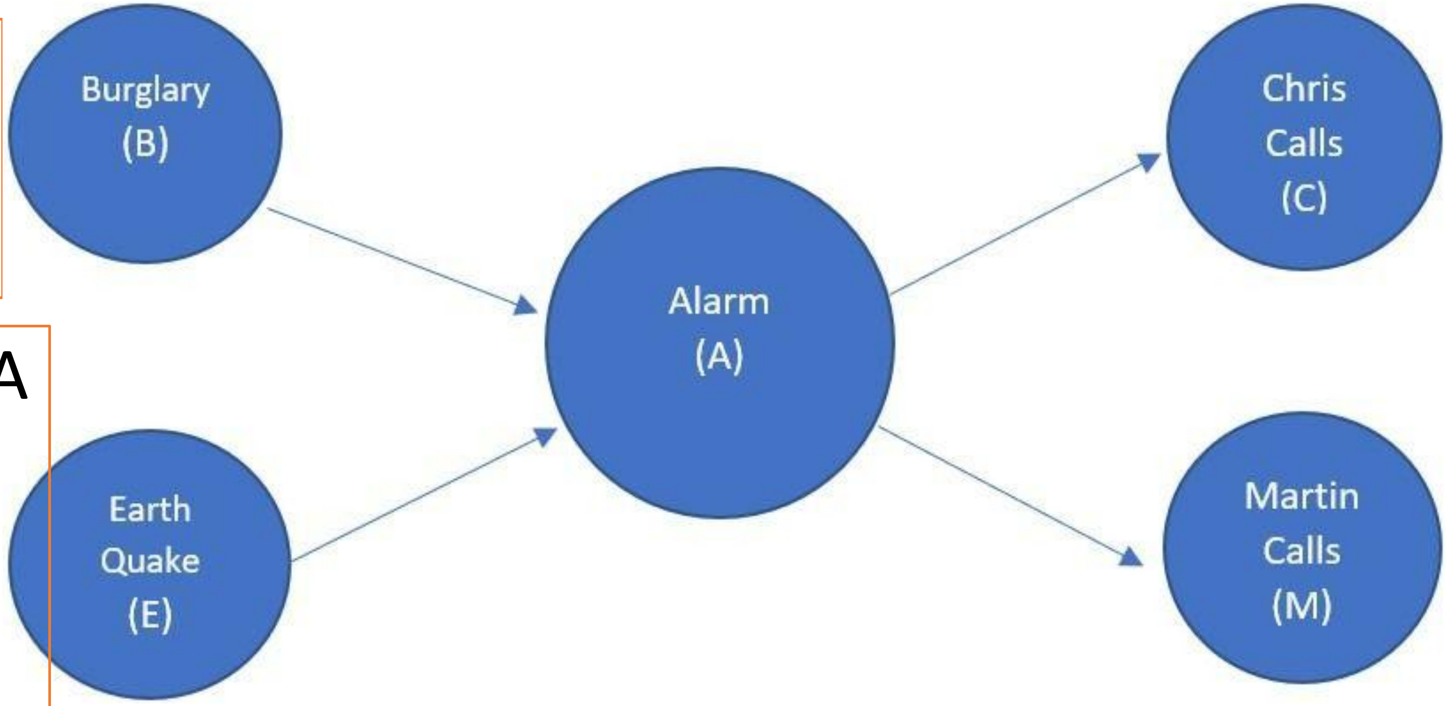


- P(B,E,A,C,M)=P(C|A)*P(M|A)*P(A|B,E)*P(B|E)*P(E)



Example

- P(B,E,A,C,M)=P(C|A)*P(M|A)*P(A|B,E)*P(B|E)*P(E)

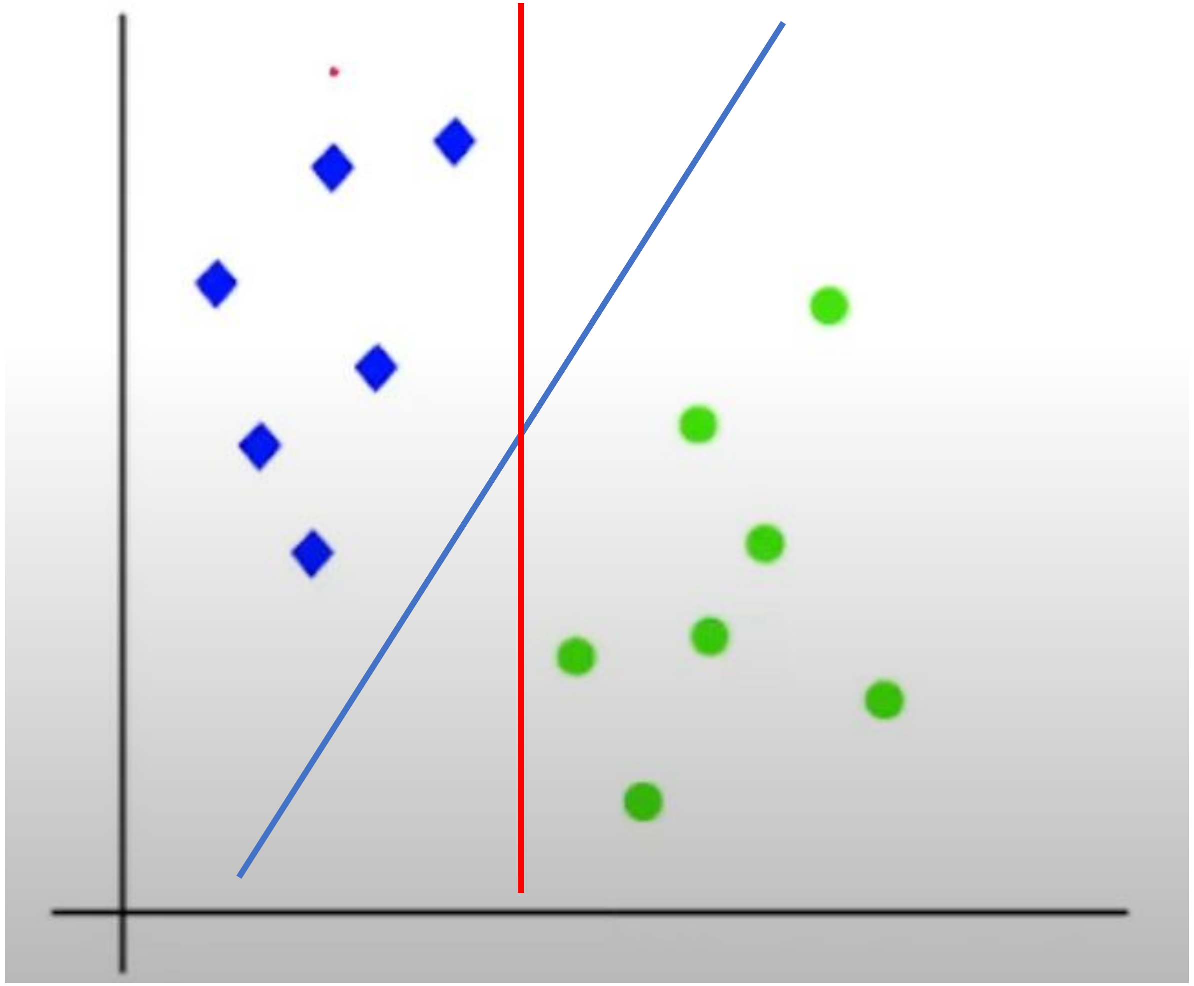


Support Vector Machine

Supervised Learning Model

Support Vector Machine (SVM)

- Used for both classification and regression
- Goal: To draw a best line/Decision boundary to separate class labels **Hyper plain**
-



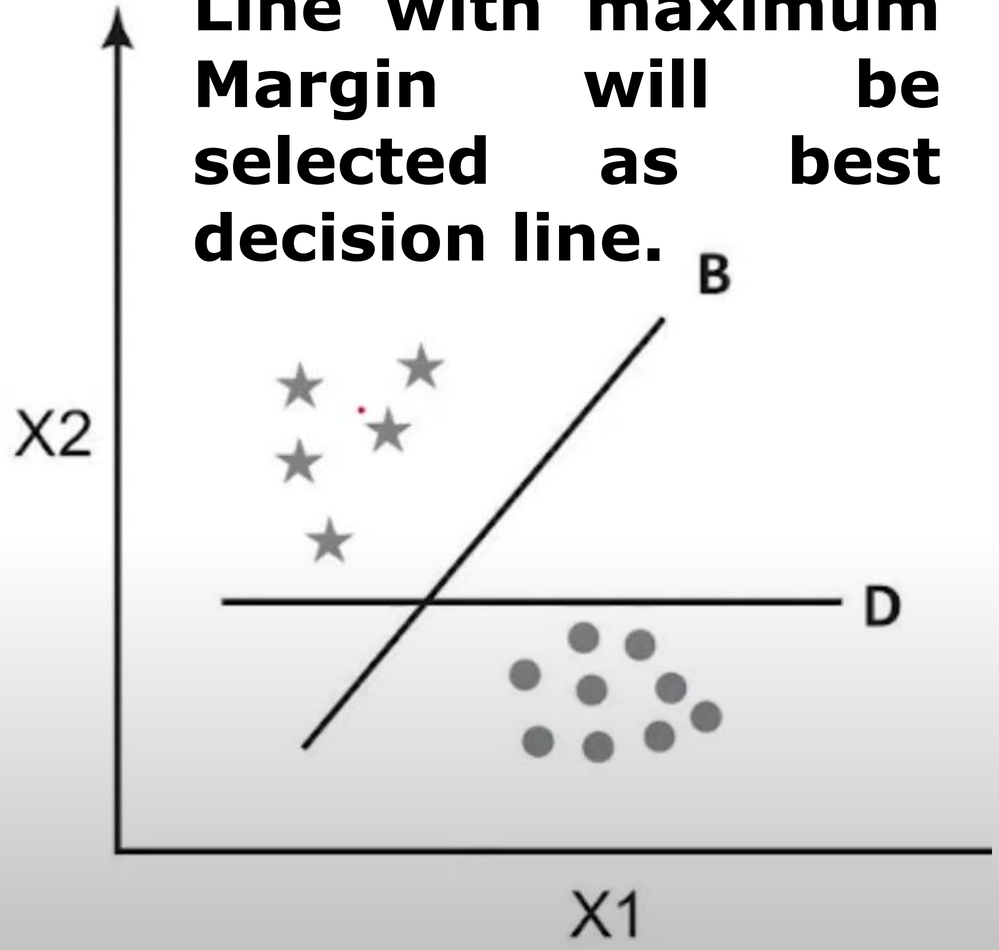
Nearest points to the decision boundary will be the **Support Vector** for that Respective boundary.

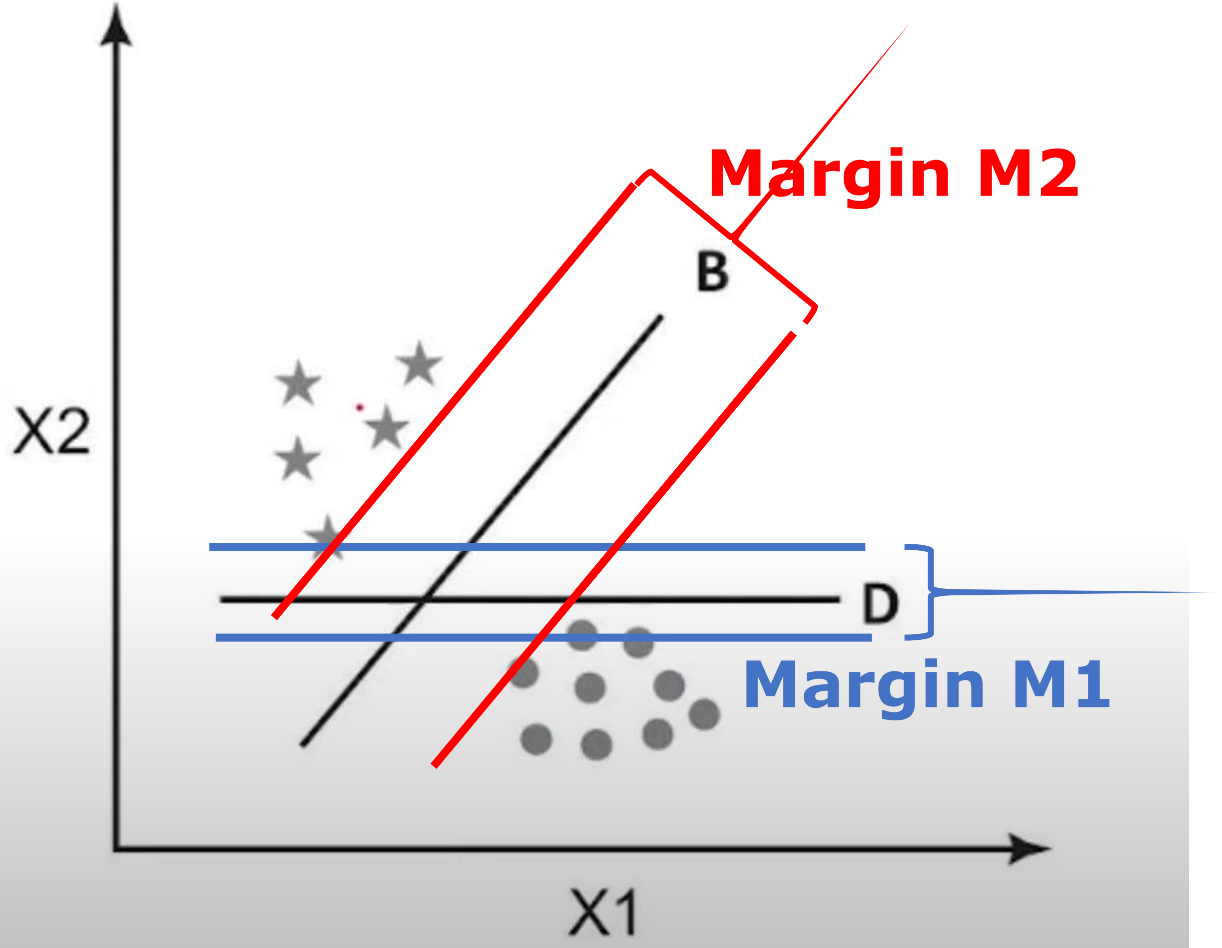
Two Types of SVM:

- **Linear** (Decision boundary is linearly separable)
- **Non-Linear** (Decision boundary is not linearly separable)

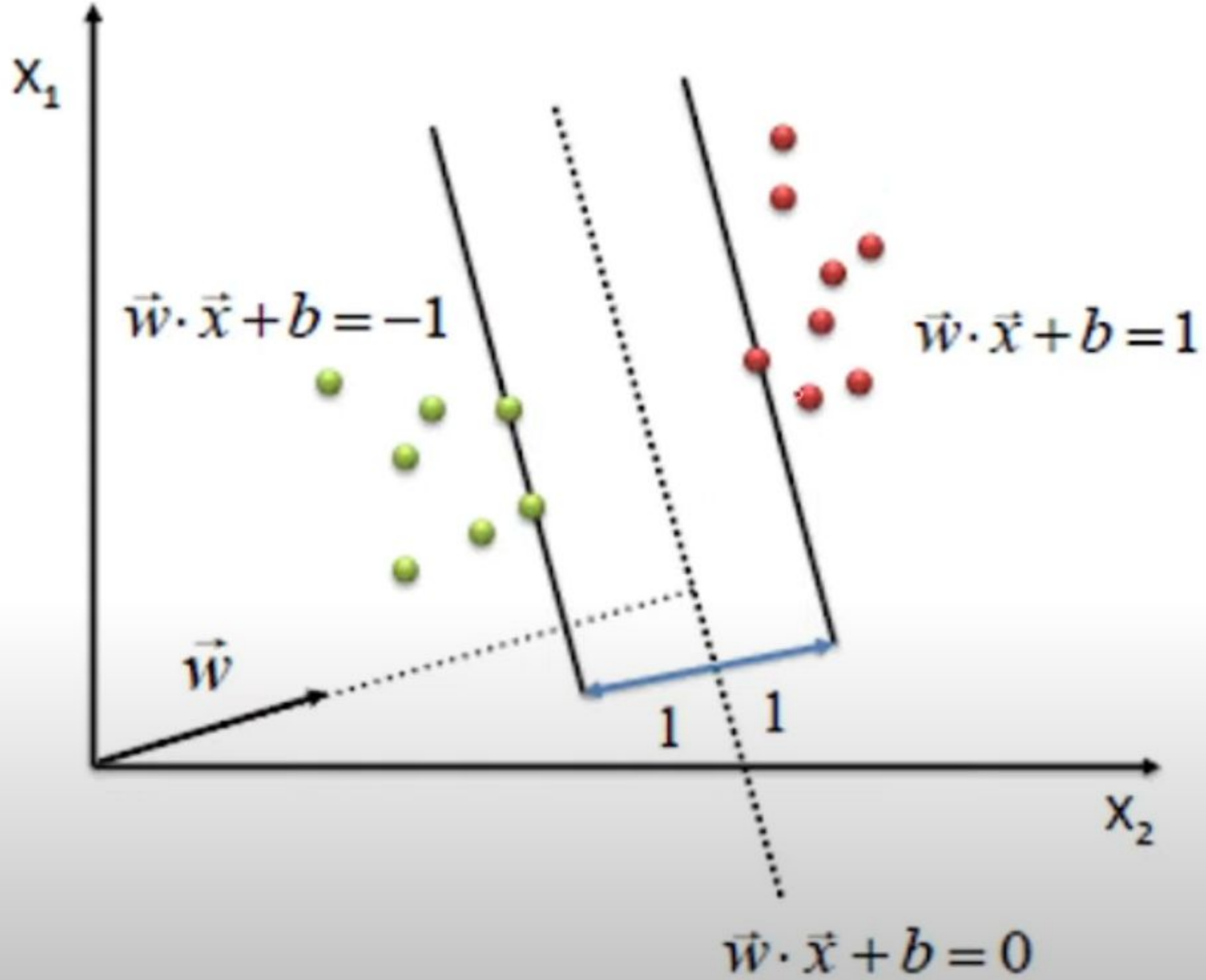
Linear SVM

Line with maximum Margin will be selected as best decision line.





SVM



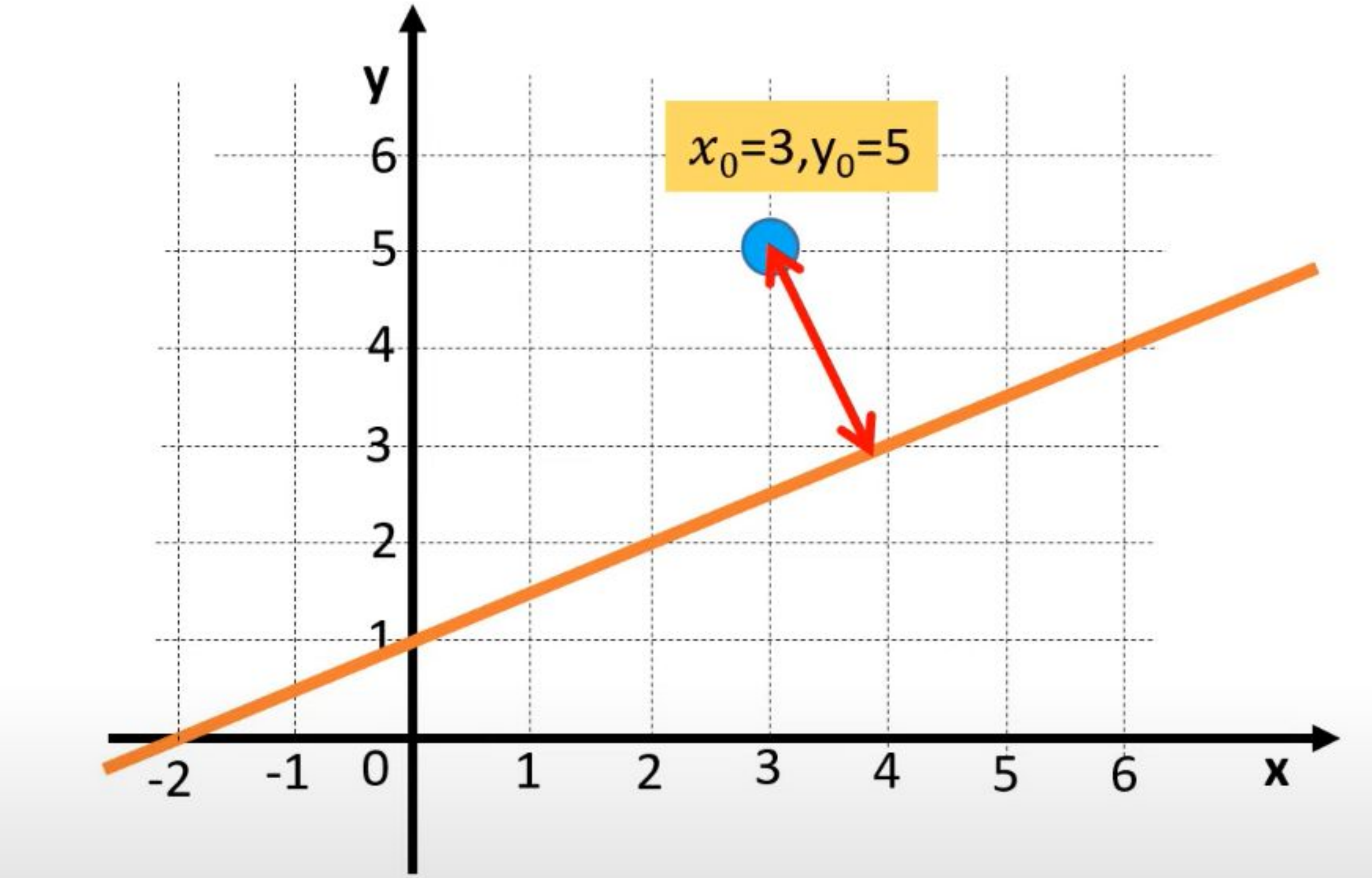
$$\boxed{\max \frac{2}{\|w\|}}$$

s.t.

$(w\cdot x+b)\geq 1, \forall x$ of class 1

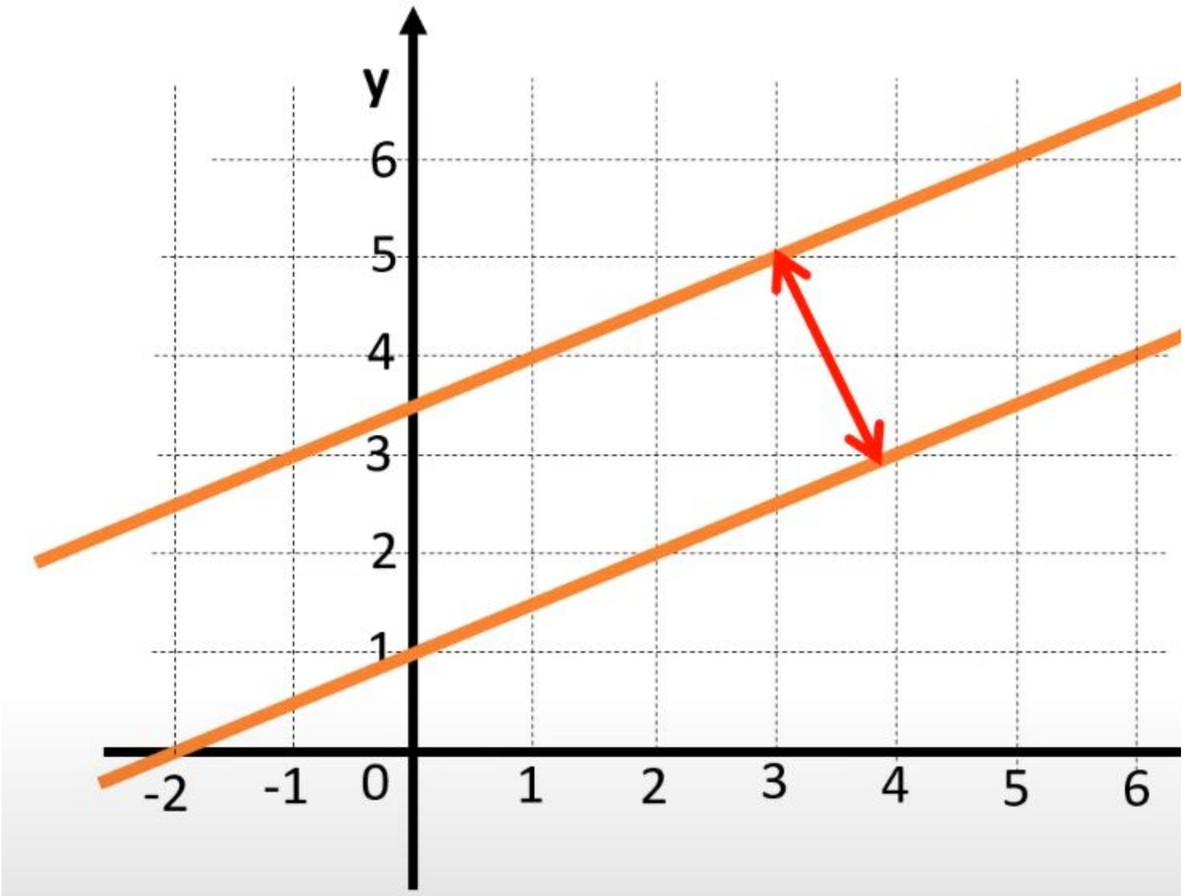
$(w\cdot x+b)\leq -1, \forall x$ of class 2

How to calculate Distance



$$d = \frac{|Ax_0 + By_o + C|}{\sqrt{A^2 + B^2}}$$

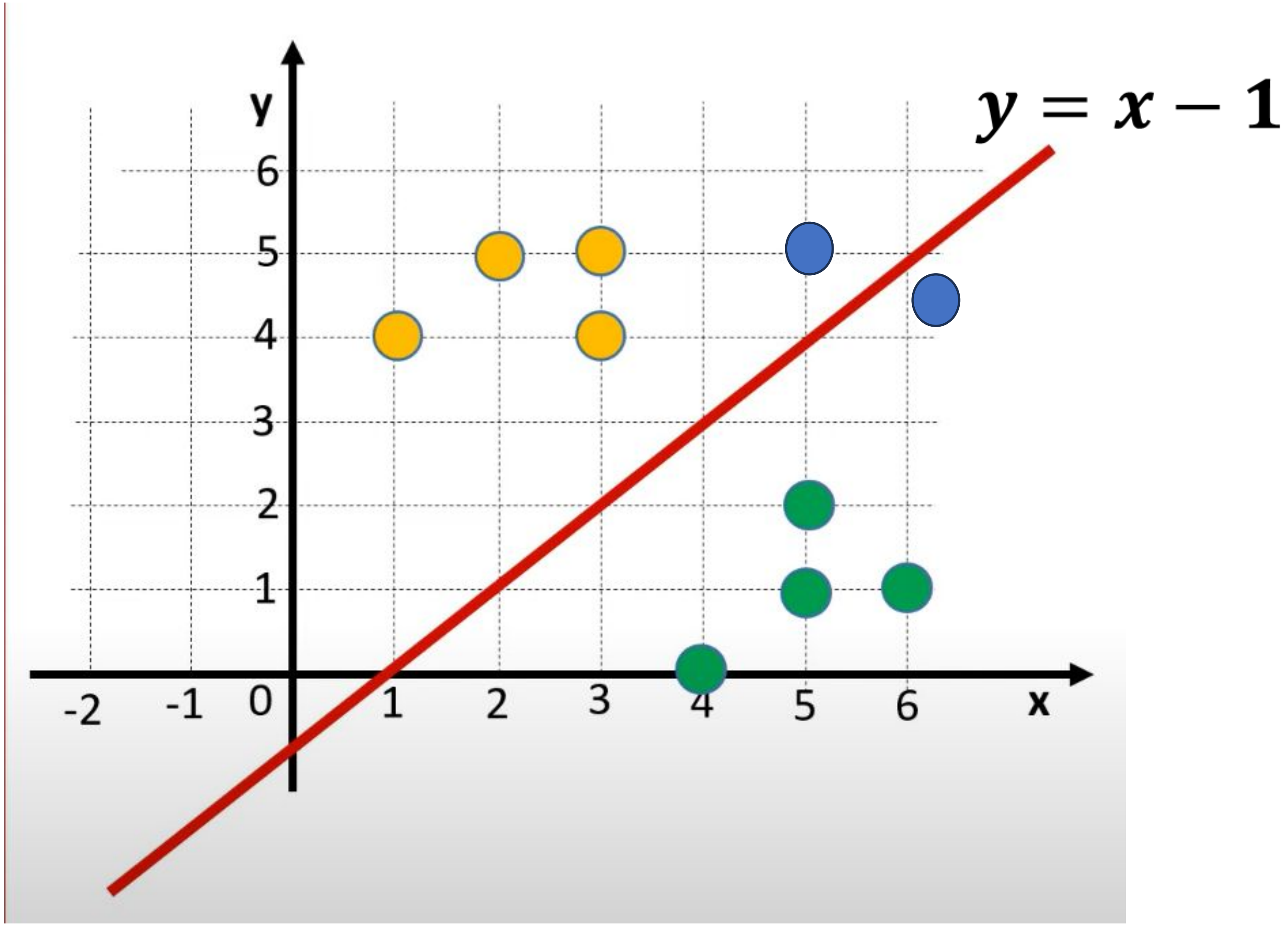
How to calculate Distance



$$d = \frac{|C_2 - C_1|}{\sqrt{A^2 + B^2}}$$

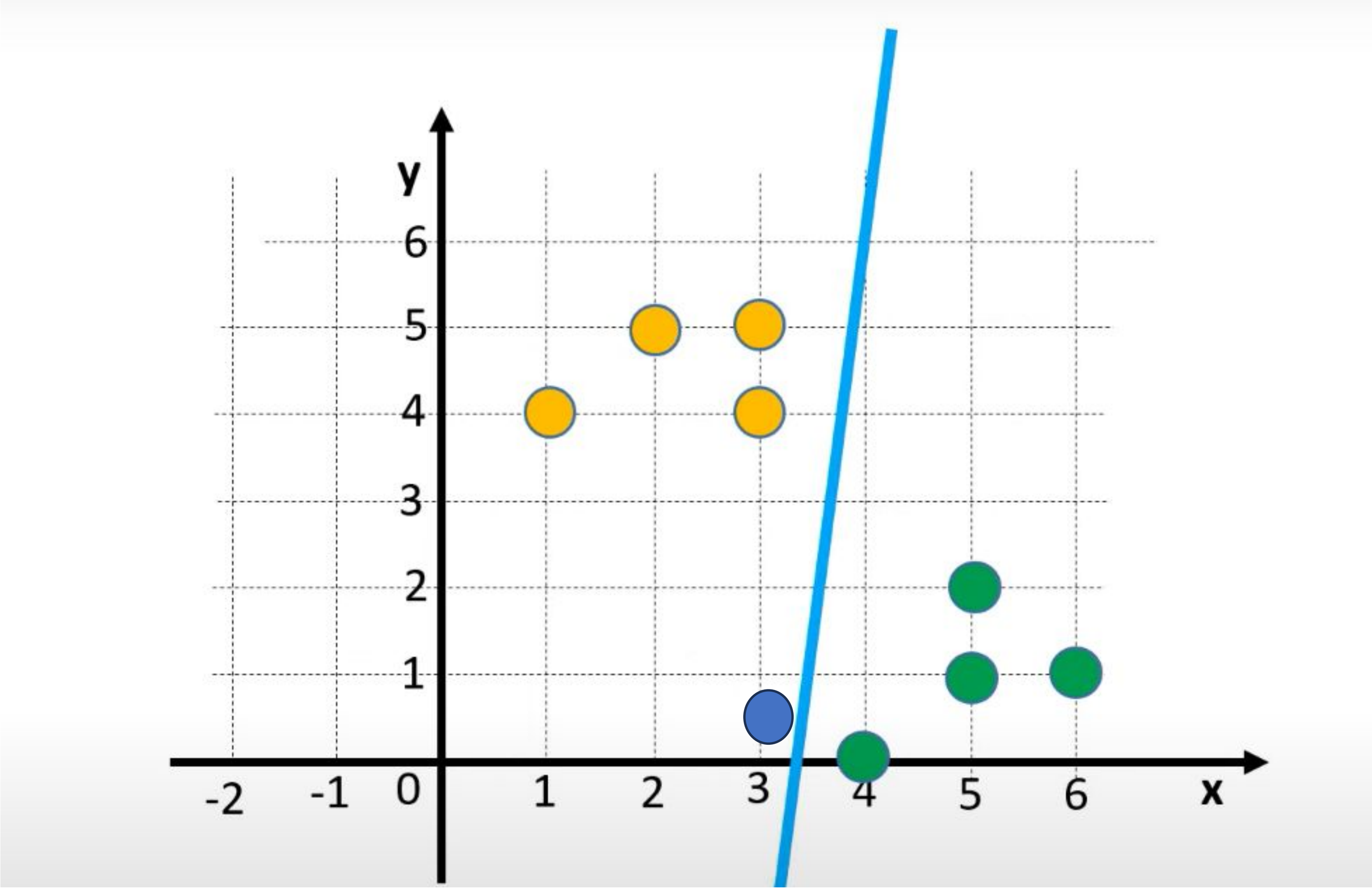
$$\boldsymbol{d} = \frac{\mathbf{2}}{[\![\boldsymbol{w}]\!]}$$

Example



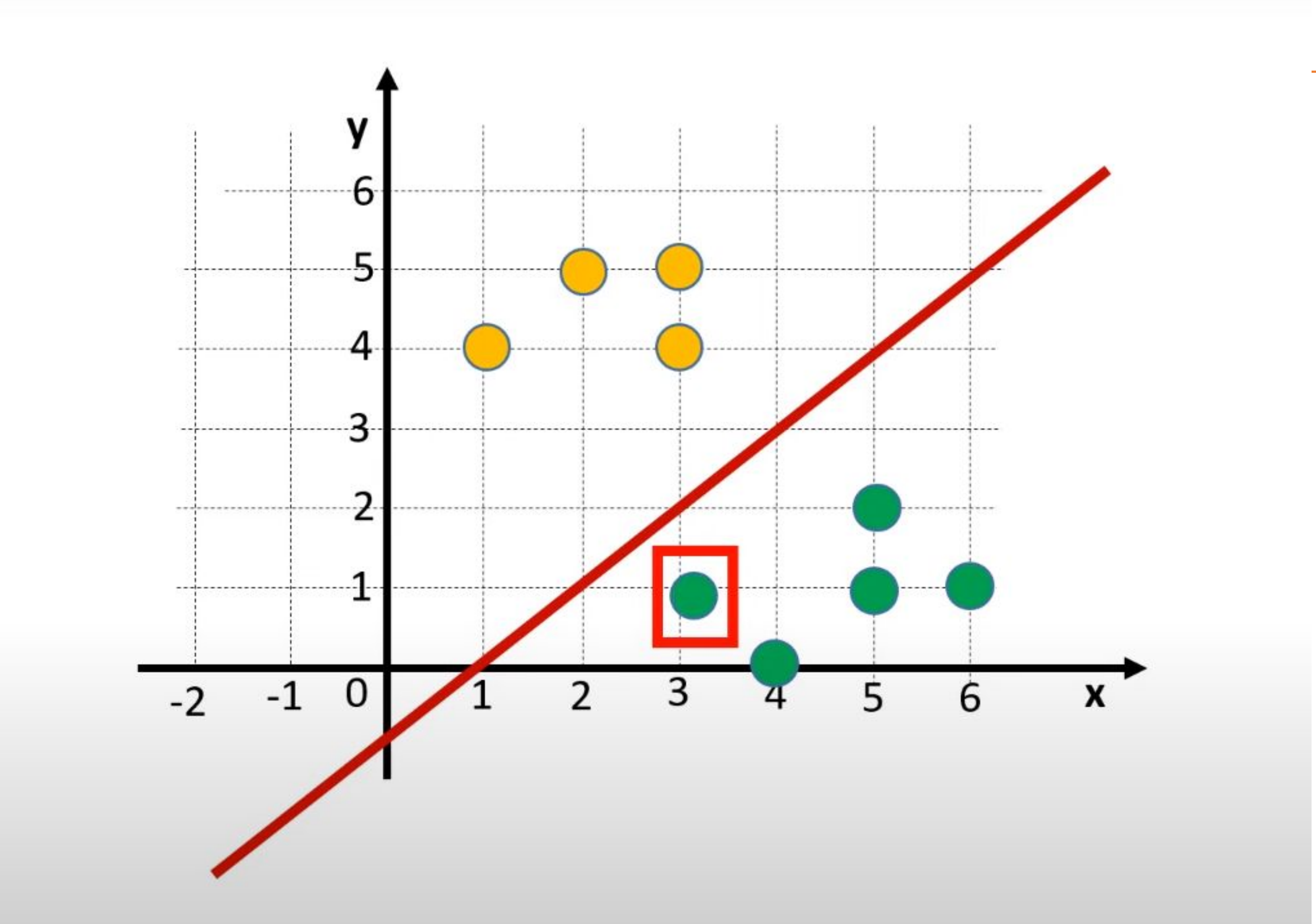
Group	x	y
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Example



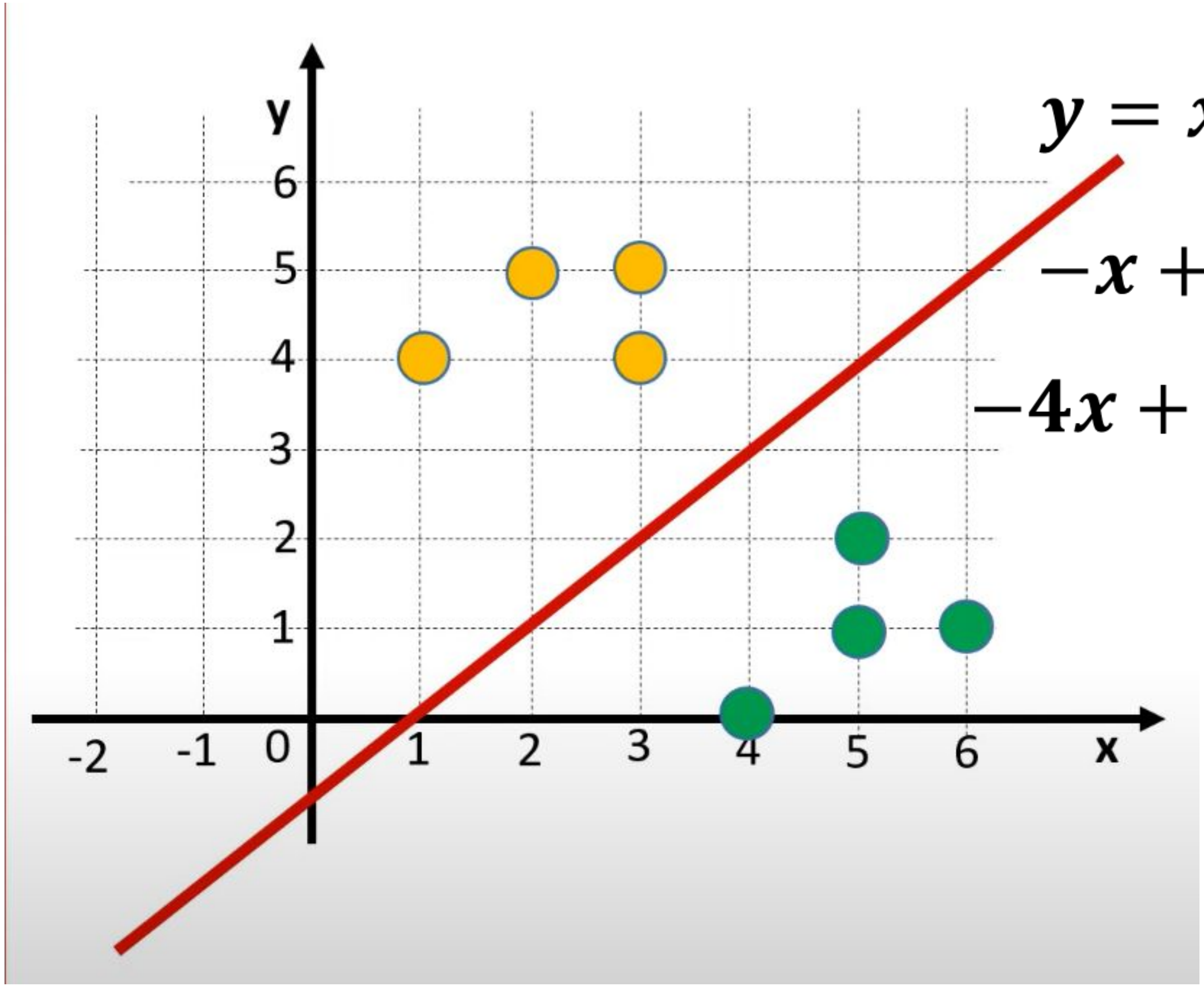
Group	x	y
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Example



Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Example



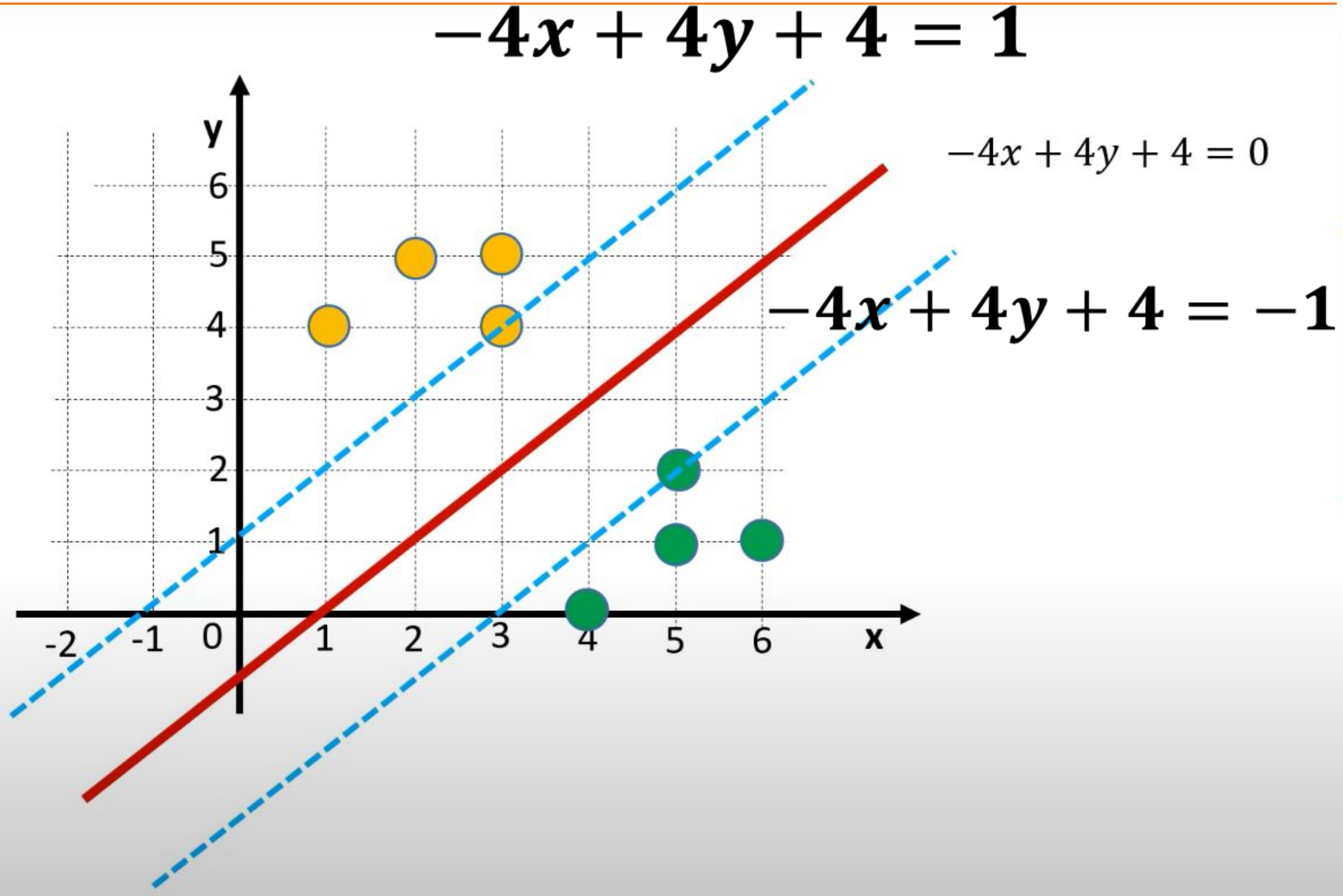
$$y = x - 1$$

$$-x + y + 1 = 0$$

$$-4x + 4y + 4 = 0$$

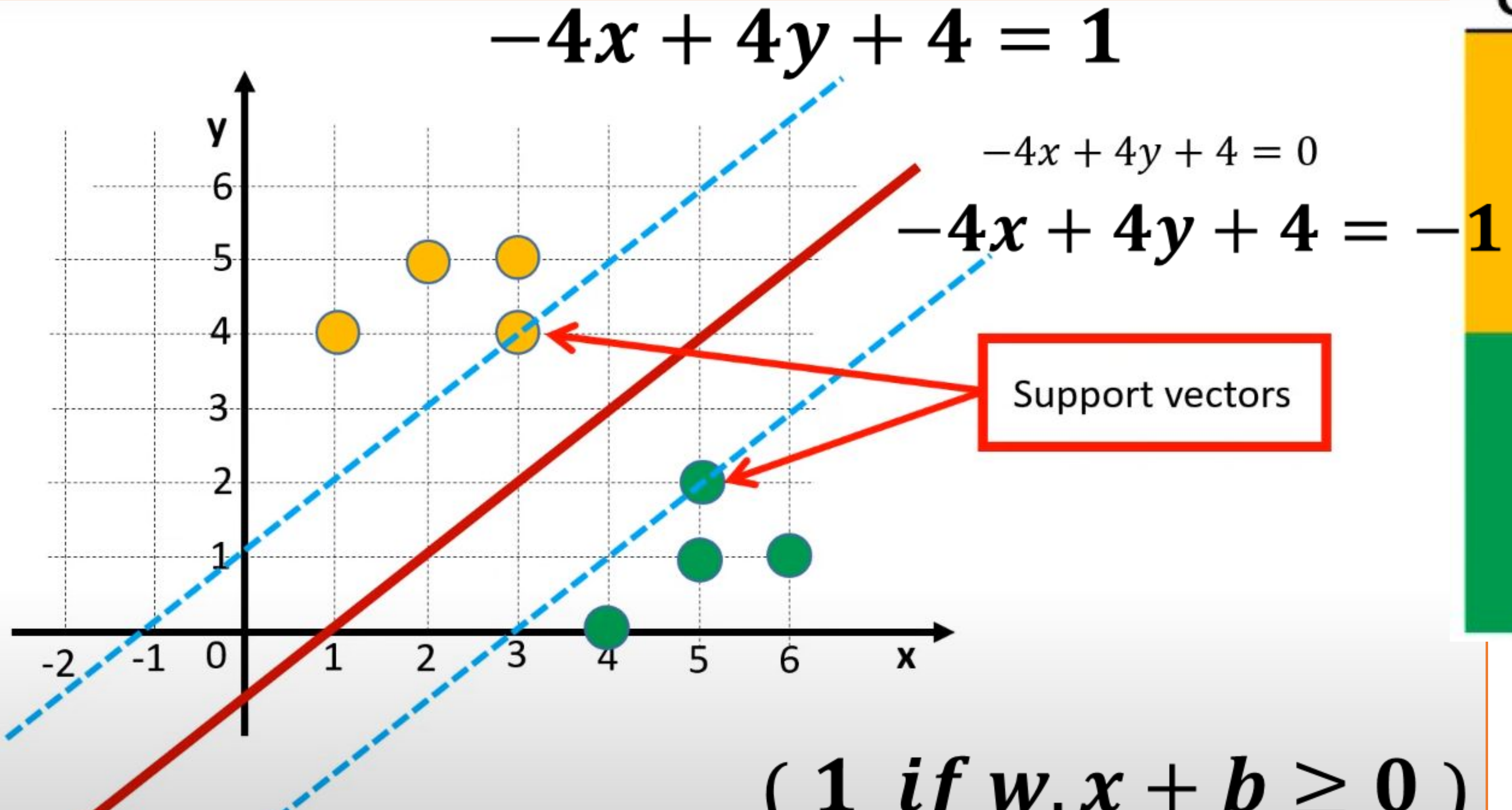
Group	*x*	*y*
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Example



Group	x	y
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Example: Try your self to plug the training data



$$y = \begin{cases} 1 \ if\ w.x + b \geq 0 \\ -1\ if\ w.x + b < 0 \end{cases}$$

Group	x	y
A	1	4
A	2	5
A	3	5
A	3	4
B	6	1
B	4	0
B	5	2
B	5	1

Summary of SVM

- Used for both classification and regression
- Goal: To draw a best line/Decision boundary to separate class labels **Hyper plain**
- When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line(if 2D).
- When the data is not linearly separable then we can use Non-Linear SVM, which means when the data points cannot be separated into 2 classes by using a straight line (if 2D) then we use some advanced techniques like kernel tricks to classify them.

Summary of SVM

- Important Terms
  - **Support Vectors:** These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.
  - **Margin:** it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin.
  - **Kernel**: Kernel is the mathematical function, which is used in SVM to map the original input data points into high-dimensional feature spaces, so, that the hyperplane can be easily found out even if the data points are not linearly separable in the original input space.