MACHINE LEARNING

Celebrating 20 Years

A Larsen & Toubro Group Company

Corporate Facts

Among the top **20** Global IT services companies

With an annual revenue of **$970Mn**

260+ clients

Including **50+** Fortune 500 companies

43 sales offices across **27** countries in US, EMEA & APAC

20,000+ employees

Working out of **23** delivery centers

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Types of Machine Learning Algorithms:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

CONTENT:

- REGRESSION (SIMPLE LINEAR REGRESSION, MULTIPLE LINEAR REGRESSION, POLYNOMIAL REGRESSION, SVR, DECISION TREE REGRESSION, RANDOM FOREST REGRESSION)
- ASSOCIATION RULE LEARNING (APRIORI, ECLAT)
- CLASSIFICATION (LOGISTIC REGRESSION, K-NN, SVN)

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

SIMPLE LINEAR REGRESSION

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Regressions

Simple Linear Regression

Constant

Coefficient

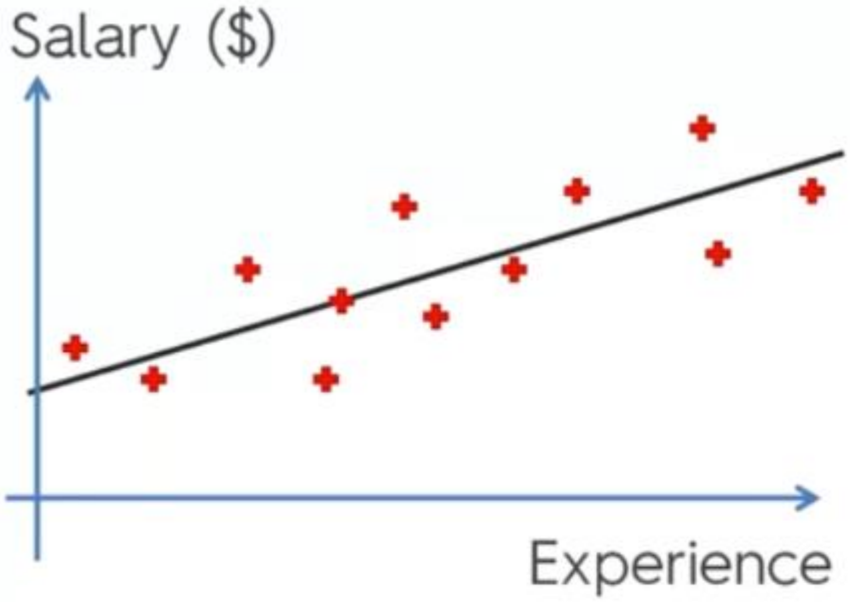
$$y = b_0 + b_1 * x_1$$

Dependent variable (DV)

Independent variable (IV)

Regressions

Simple Linear Regression:



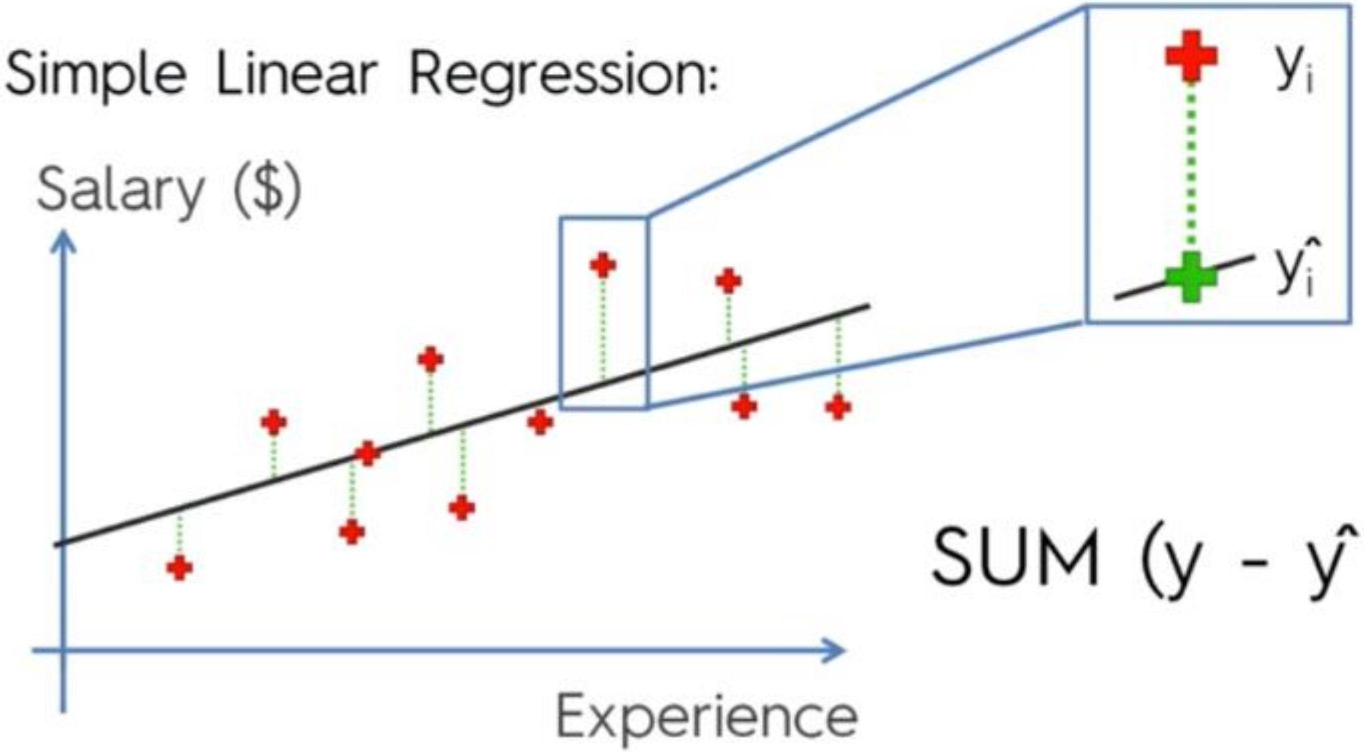
$$y = b_0 + b_1 * x$$

$$\downarrow$$

$$Salary = b_0 + b_1 * Experience$$

Ordinary Least Squares

Simple Linear Regression:



SUM $(y - \hat{y})^2$ -> min

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

MULTIPLE LINEAR REGRESSION

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Regressions

Simple Linear Regression

$$y = b_0 + b_1 {*} x_1$$

Multiple Linear Regression

Dependent variable (DV)   Independent variables (IVs)

$$y = b_0 + b_1 {*} x_1 + b_2 {*} x_2 + \ldots + b_n {*} x_n$$

Constant

Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Dummy Variables

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + b_3{}^*x_3 + \text{???}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Dummy Variables

Dummy Variables

Profit	R&D Spend	Admin	Marketing	~~State~~	New York	California
192,261.83	165,349.20	136,897.80	471,784.10	~~New York~~	1	0
191,792.06	162,597.70	151,377.59	443,898.53	~~California~~	0	1
191,050.39	153,441.51	101,145.55	407,934.54	~~California~~	0	1
182,901.99	144,372.41	118,671.85	383,199.62	~~New York~~	1	0
166,187.94	142,107.34	91,391.77	366,168.42	~~California~~	0	1

$$y = b_0 + b_1{}^*x_1 + b_2{}^*x_2 + b_3{}^*x_3 + ??? + b_4{}^*D_1$$

Dummy Variable Trap

Profit	R&D Spend	Admin	Marketing	State
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	California
191,050.39	153,441.51	101,145.55	407,934.54	California
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	California

Dummy Variables

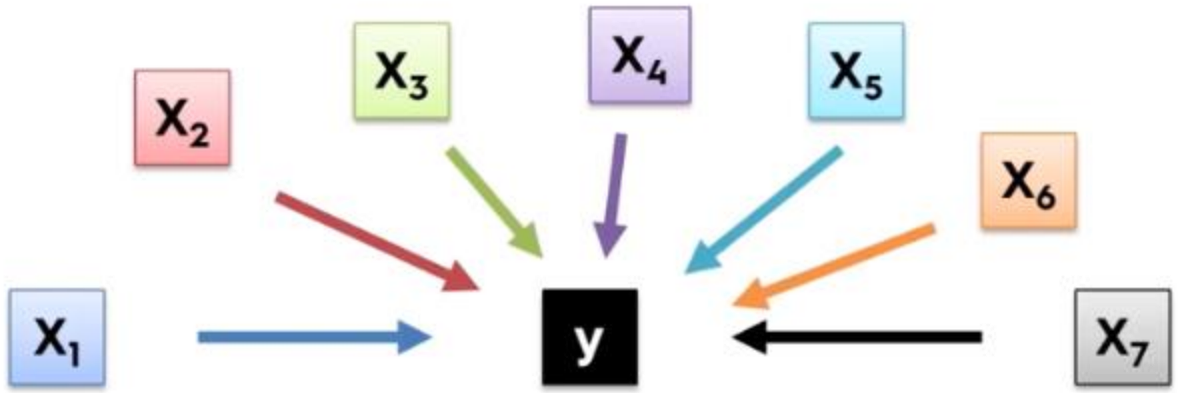
New York	California
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1{*}x_1 + b_2{*}x_2 + b_3{*}x_3 + b_4{*}D_1 + \cancel{b_5{*}D_2}$$

Always omit one dummy variable

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Building A Model



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Building A Model

5 methods of building models:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

Stepwise Regression

Building A Model

Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

STEP 2: Fit the full model with all possible predictors

STEP 3: Consider the predictor with the <u>highest</u> P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

Building A Model

Forward Selection

STEP 1: Select a significance level to enter the model (e.g. SL = 0.05)

STEP 2: Fit all simple regression models $\mathbf{y} \sim \mathbf{x_n}$ Select the one with the lowest P-value

STEP 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

STEP 4: Consider the predictor with the lowest P-value. If P < SL, go to STEP 3, otherwise go to FIN

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Building A Model

Forward Selection

STEP 1: Select a significance level to enter the model (e.g. SL = 0.05)

STEP 2: Fit all simple regression models **$y \sim x_n$** Select the one with the lowest P-value

STEP 3: Keep this variable and fit all possible models with one extra predictor added to the one(s) you already have

STEP 4: Consider the predictor with the <u>lowest</u> P-value. If ~~P < SL~~, go to STEP 3, otherwise go to FIN

FIN: Keep the previous model

Building A Model

Bidirectional Elimination

STEP 1: Select a significance level to enter and to stay in the model
e.g.: SLENTER = 0.05, SLSTAY = 0.05

STEP 2: Perform the next step of Forward Selection (new variables must have: P < SLENTER to enter)

STEP 3: Perform ALL steps of Backward Elimination (old variables must have P < SLSTAY to stay)

STEP 4: No new variables can enter and no old variables can exit

FIN: Your Model Is Ready

Activate Windows

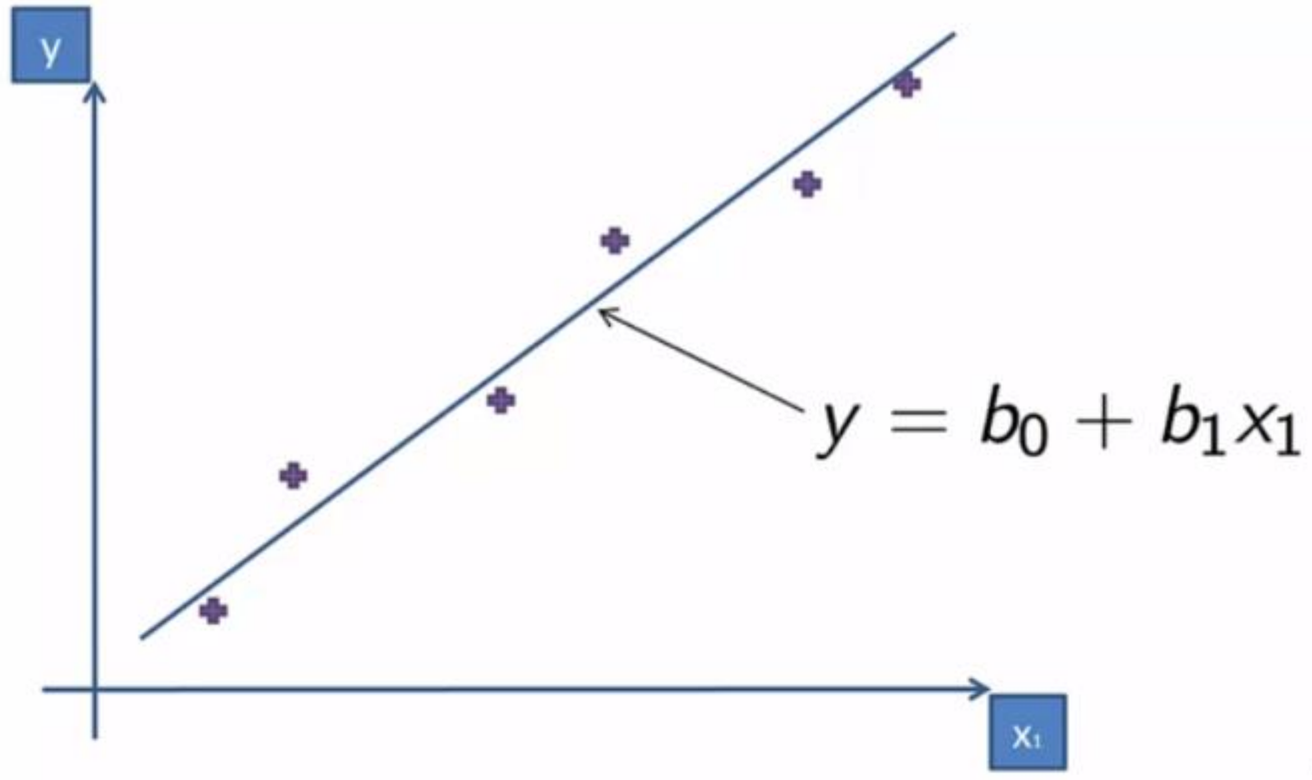
POLYNOMIAL REGRESSION

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

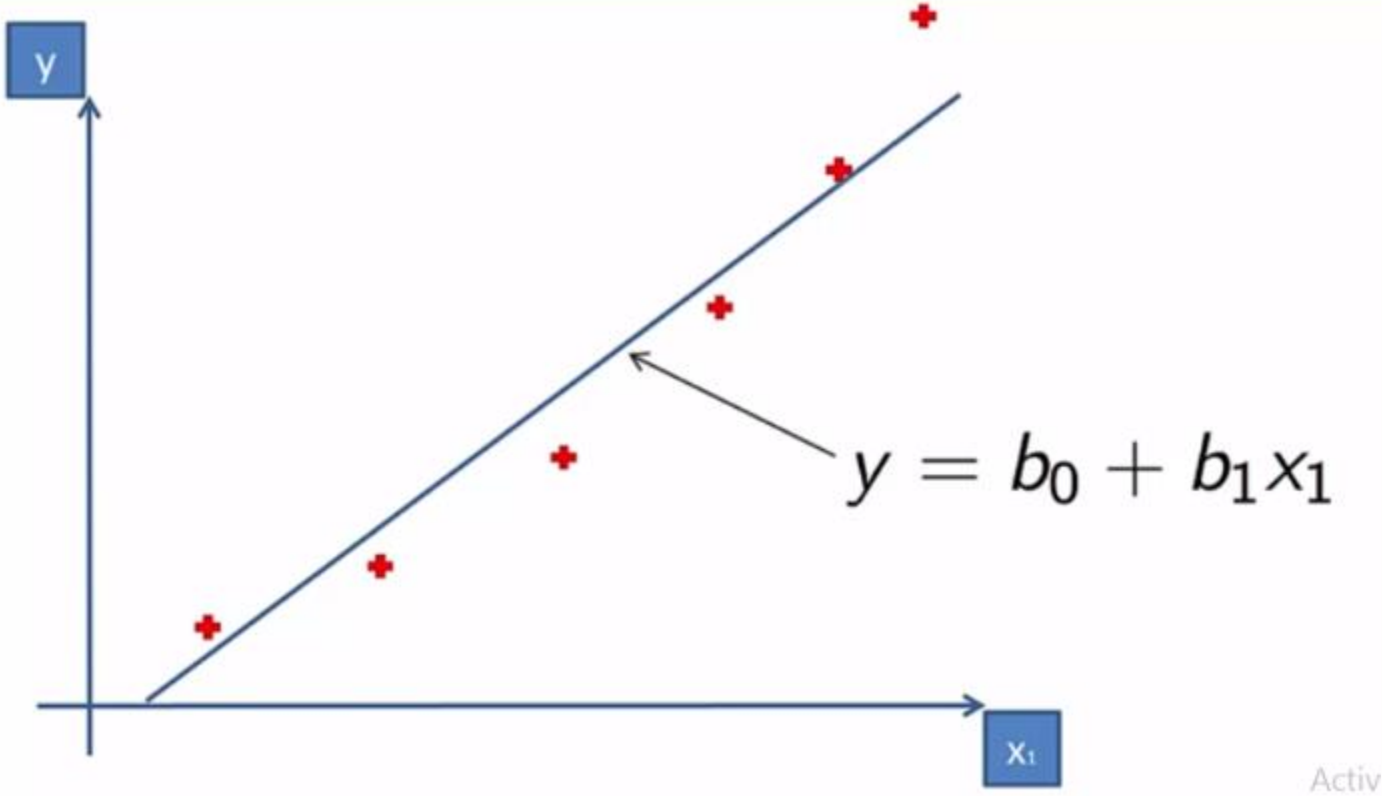
Regressions

Simple Linear Regression	$y = b_0 + b_1 x_1$
Multiple Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$
Polynomial Linear Regression	$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$

Simple Linear Regression

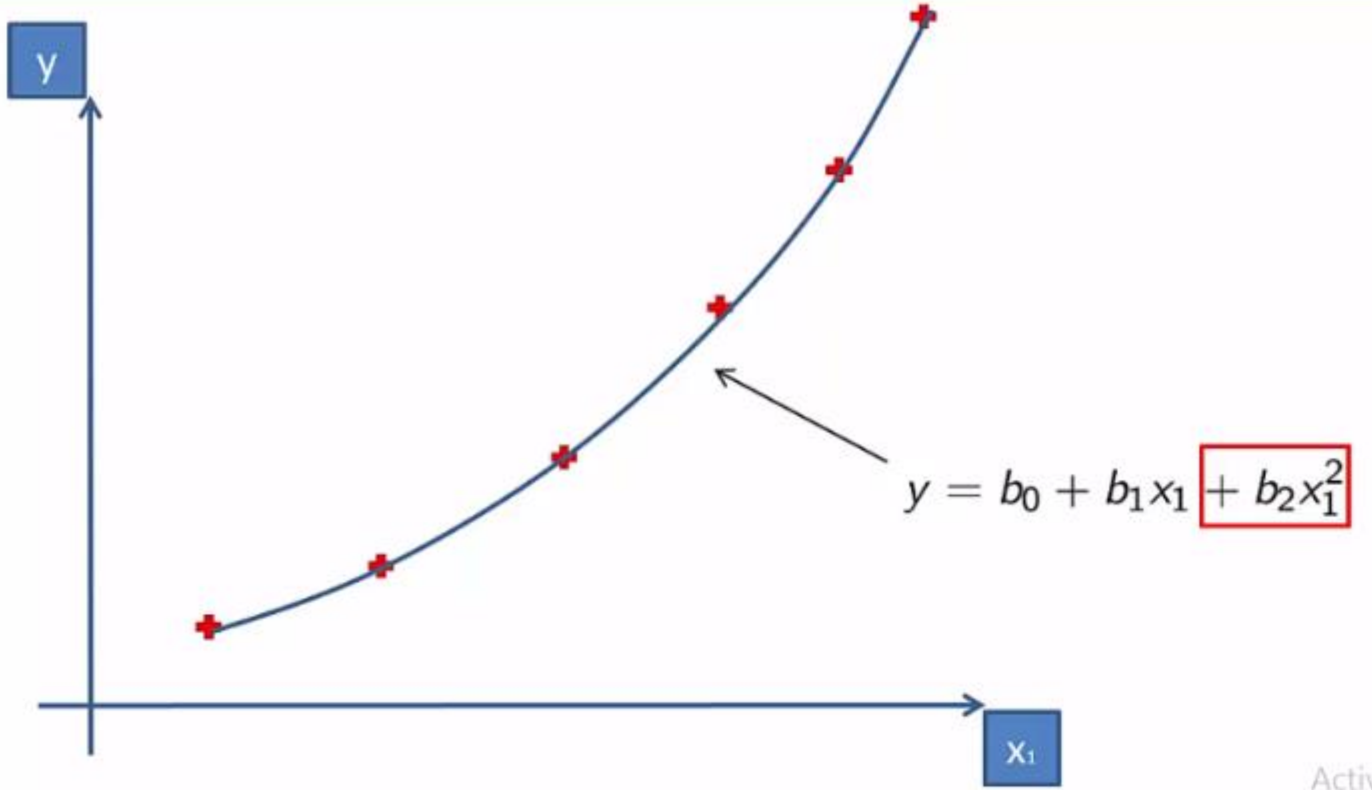


Simple Linear Regression



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Polinomial Regression

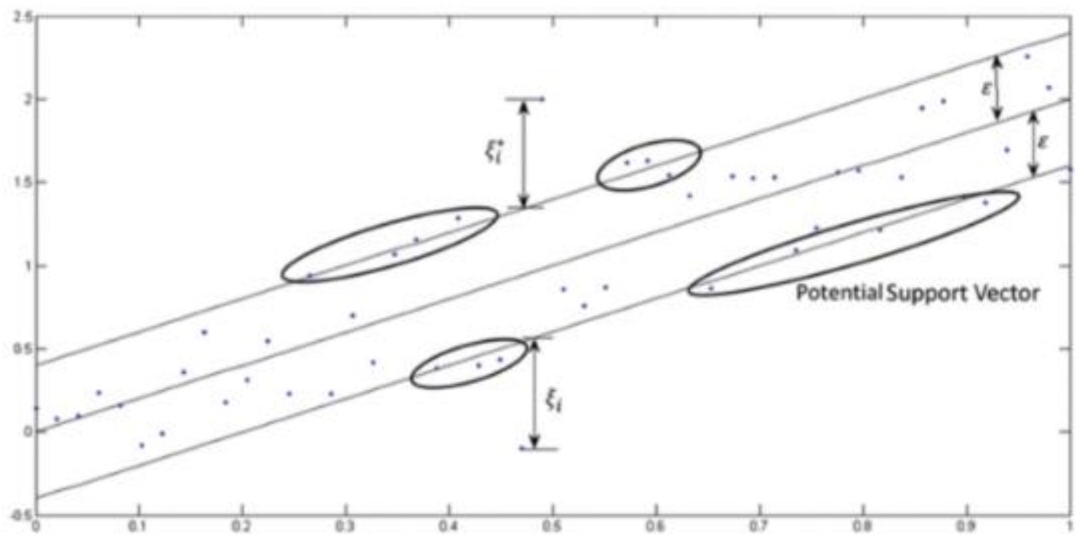


©Larsen & Toubro Infotech Ltd. Privileged and Confidential

SVR

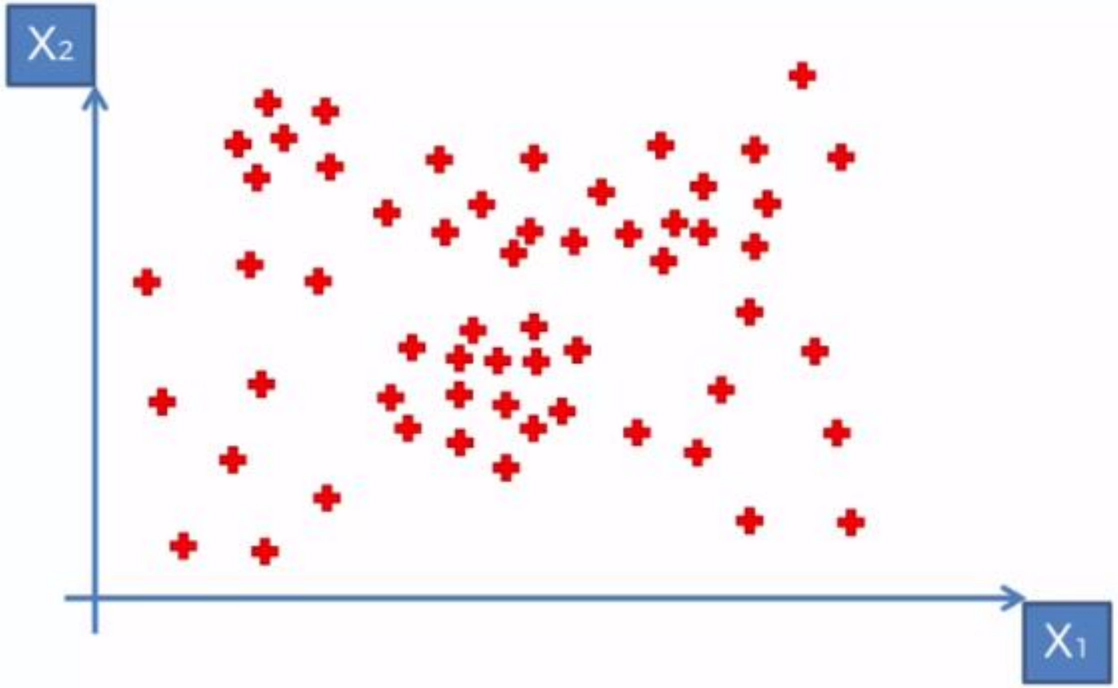
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Support Vector Regression - SVR



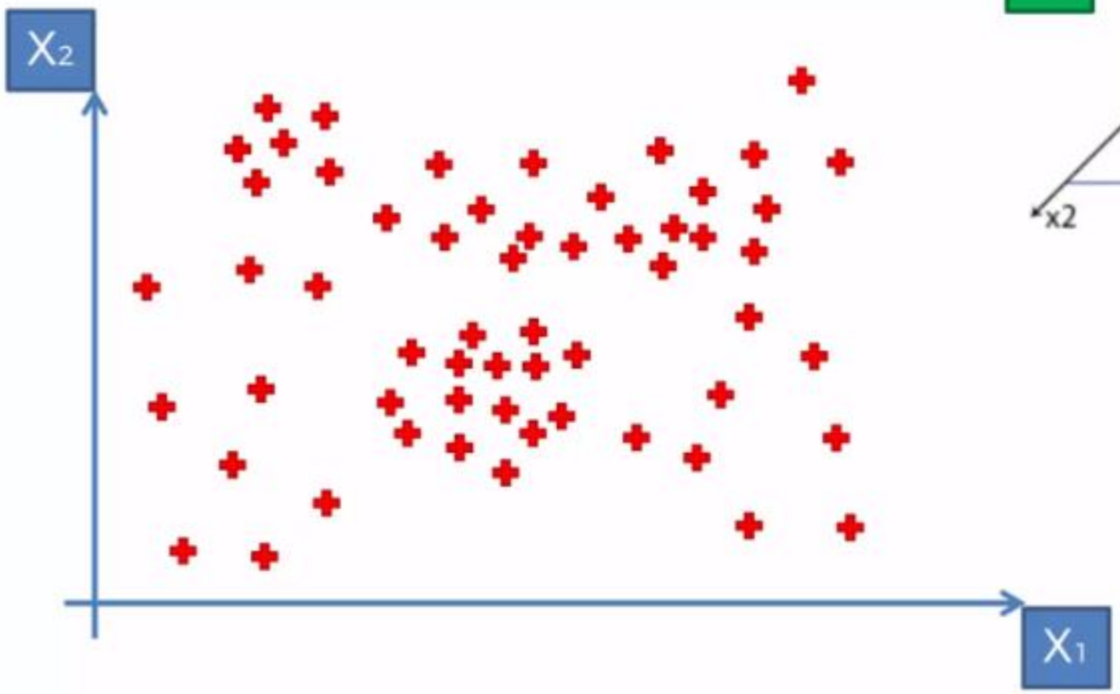
DECISION TREE REGRESSION

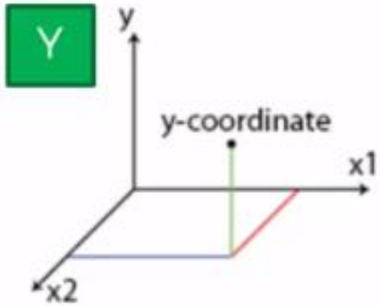
Decision Tree Intuition



Activate Windows

Decision Tree Intuition

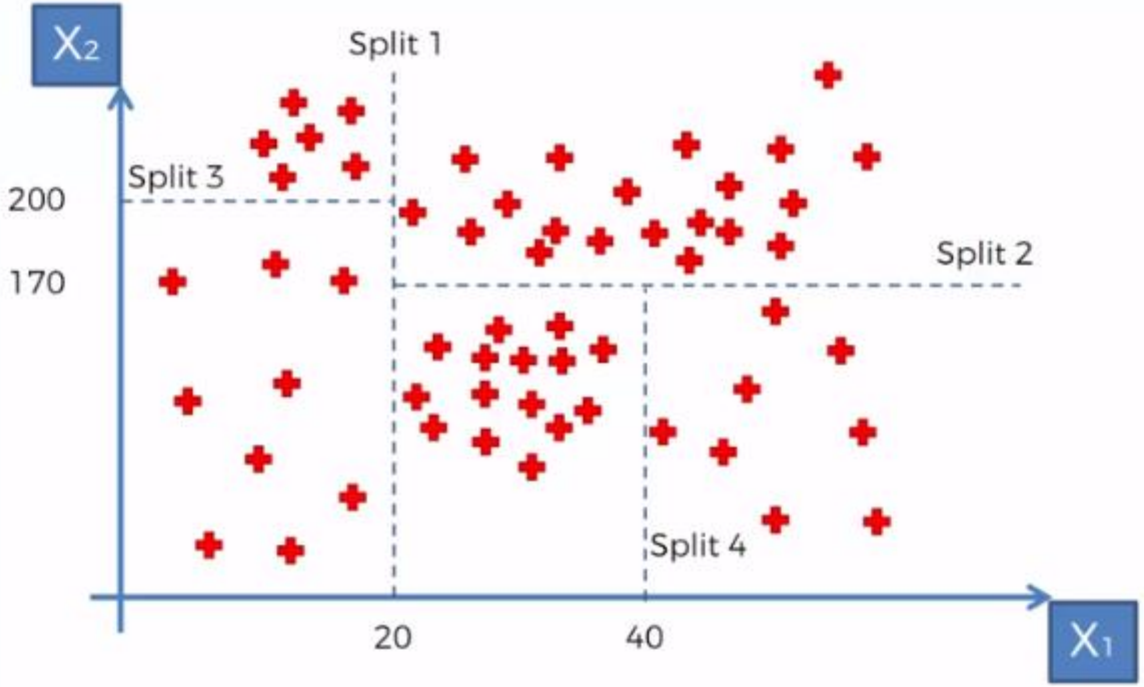




Activate Windows

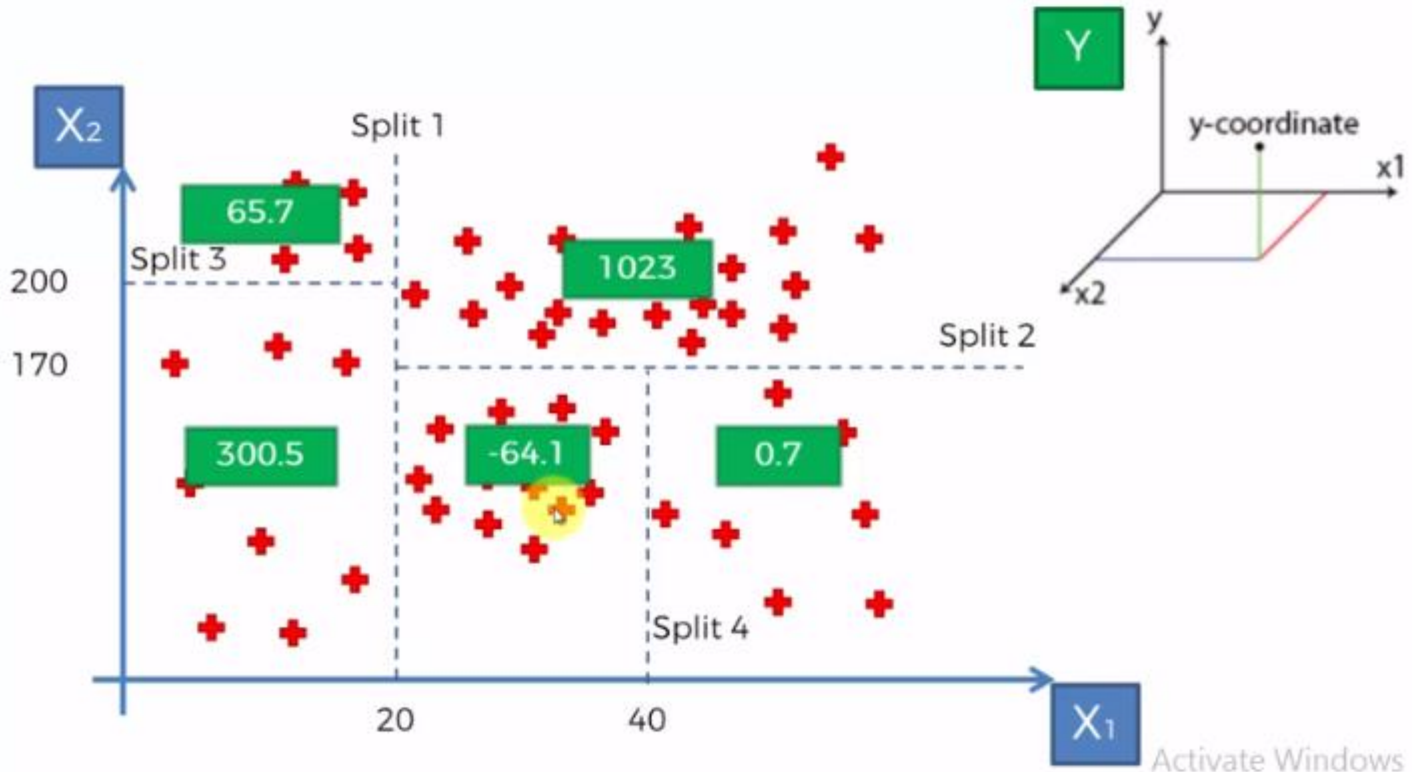
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Decision Tree Intuition

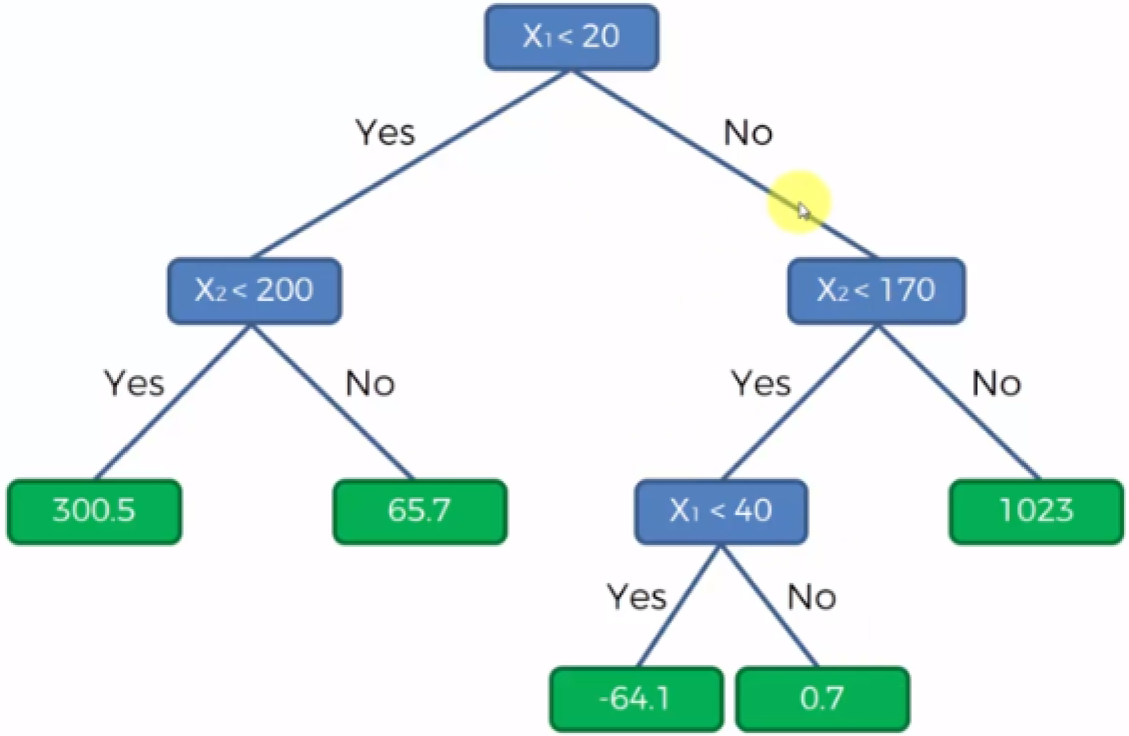


Activate Windows

Decision Tree Intuition



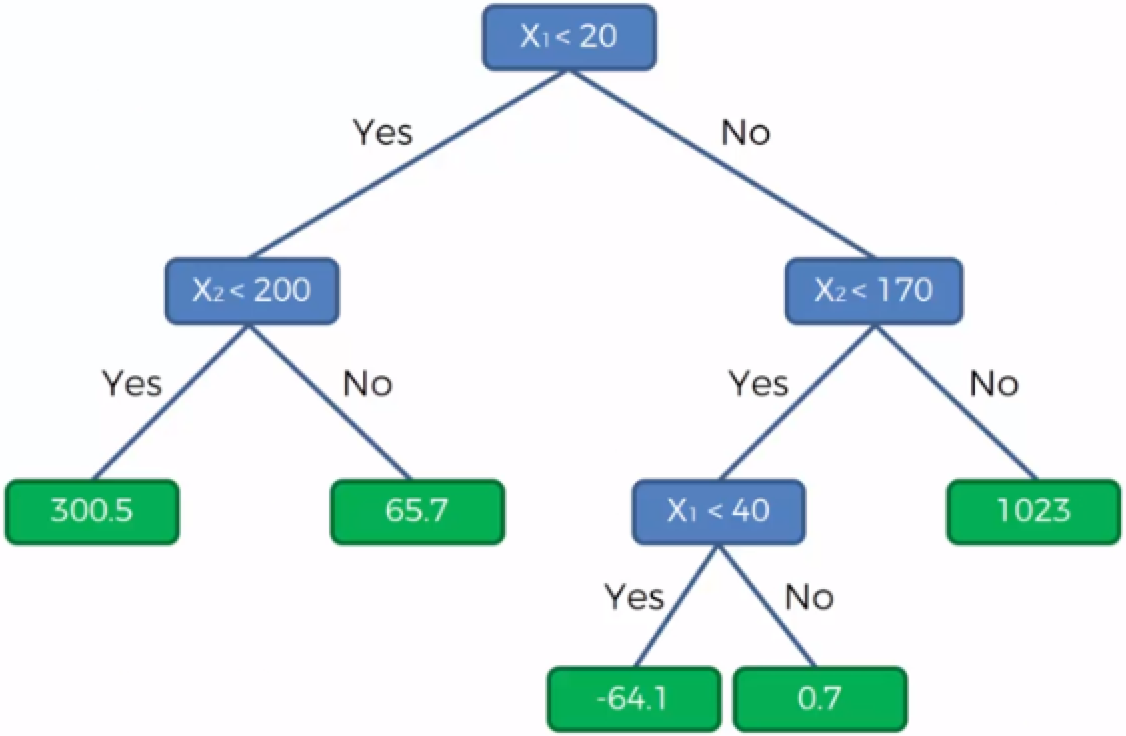
Decision Tree Intuition



Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Decision Tree Intuition



Activate Windows

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

RANDOM FOREST REGRESSION

Random Forest Intuition

STEP 1: Pick at random K data points from the Training set.

⇩

STEP 2: Build the Decision Tree associated to these K data points.

⇩

STEP 3: Choose the number Ntree of trees you want to build and repeat STEPS 1 & 2

⇩

STEP 4: For a new data point, make each one of your Ntree trees predict the value of Y to for the data point in question, and assign the new data point the average across all of the predicted Y values.

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Random Forest Intuition

Activate Windows

APRIORI

ARL - What is it all about ?

ARL - What is it all about ?

People who bought also bought …

ARL - Movie Recommendation

User ID	Movies liked
46578	Movie1, Movie2, Movie3, Movie4
98989	Movie1, Movie2
71527	Movie1, Movie2, Movie4
78981	Movie1, Movie2
89192	Movie2, Movie4
61557	Movie1, Movie3

Potential Rules:

Movie1 → Movie2

Movie2 → Movie4

Movie1 → Movie3

ARL - Market Basket Optimisation

Transaction ID	Products purchased
46578	Burgers, French Fries, Vegetables
98989	Burgers, French Fries, Ketchup
71527	Vegetables, Fruits
78981	Pasta, Fruits, Butter, Vegetables
89192	Burgers, Pasta, French Fries
61557	Fruits, Orange Juice, Vegetables
87923	Burgers, French Fries, Ketchup, Mayo

Potential Rules:

Burgers ⇒ French Fries

Vegetables ⇒ Fruits

Burgers, French Fries ⇒ Ketchup

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

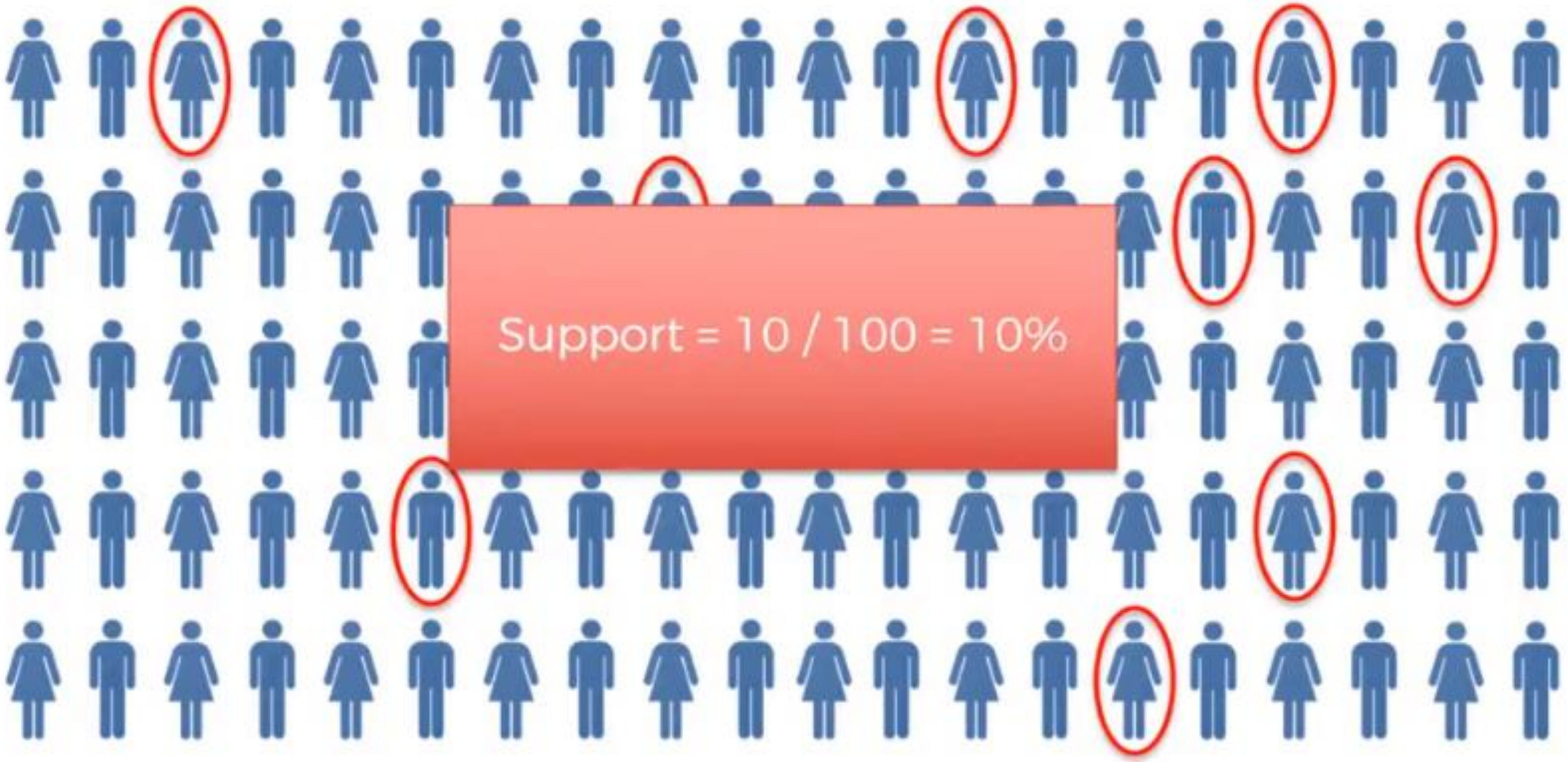
Apriori - Support

Movie Recommendation: $$\text{support}(\boldsymbol{M}) = \frac{\#\ \text{user watchlists containing } \boldsymbol{M}}{\#\ \text{user watchlists}}$$

Market Basket Optimisation: $$\text{support}(\boldsymbol{I}) = \frac{\#\ \text{transactions containing } \boldsymbol{I}}{\#\ \text{transactions}}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Apriori - Support



Apriori - Confidence

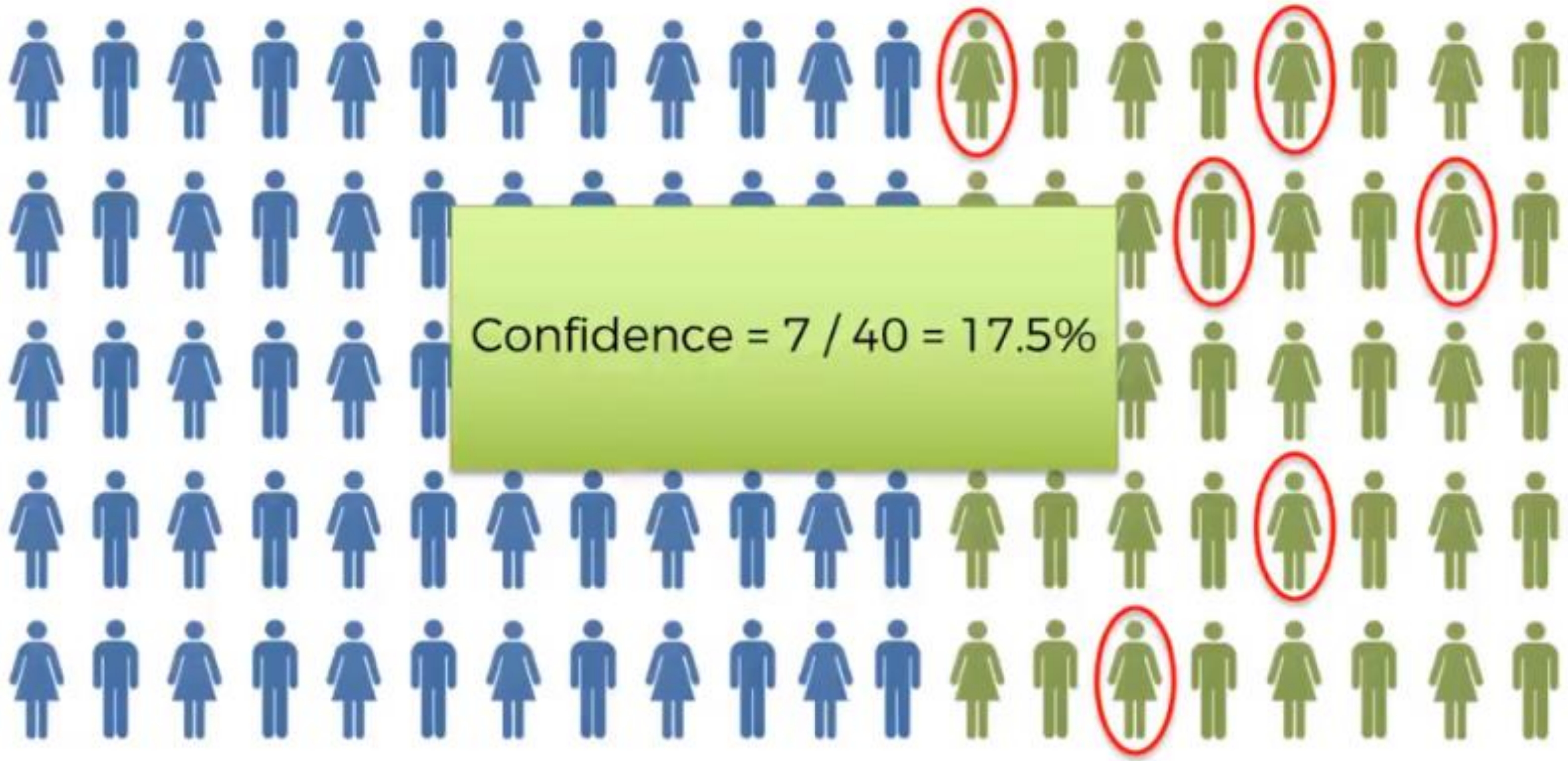
Movie Recommendation: $$\text{confidence}(M_1 \rightarrow M_2) = \frac{\#\text{ user watchlists containing } M_1 \text{ and } M_2}{\#\text{ user watchlists containing } M_1}$$

Market Basket Optimisation: $$\text{confidence}(I_1 \rightarrow I_2) = \frac{\#\text{ transactions containing } I_1 \text{ and } I_2}{\#\text{ transactions containing } I_1}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Apriori - Confidence

Apriori - Confidence



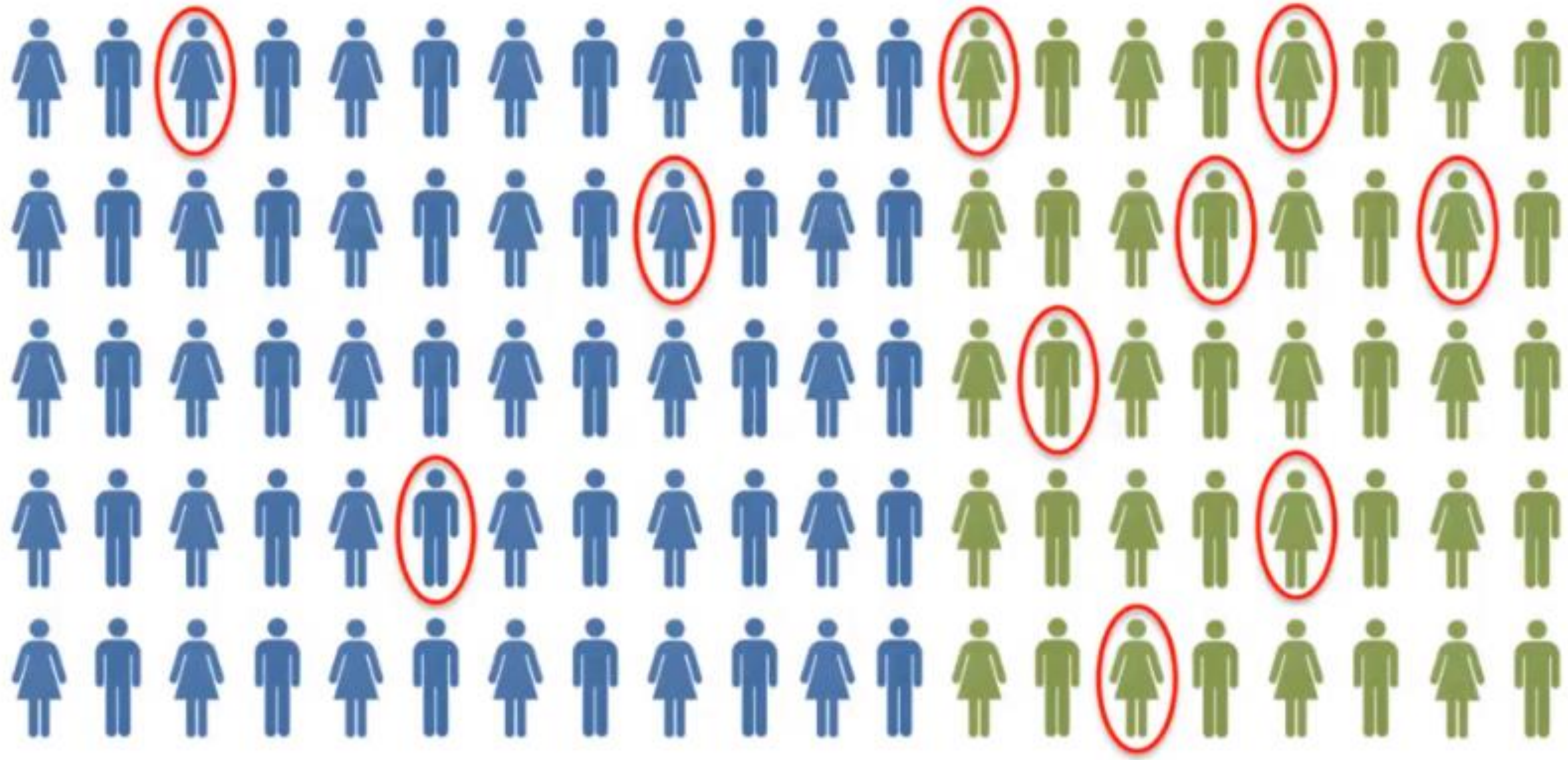
Apriori - Lift

Movie Recommendation: $$\text{lift}(M_1 \rightarrow M_2) = \frac{\text{confidence}(M_1 \rightarrow M_2)}{\text{support}(M_2)}$$

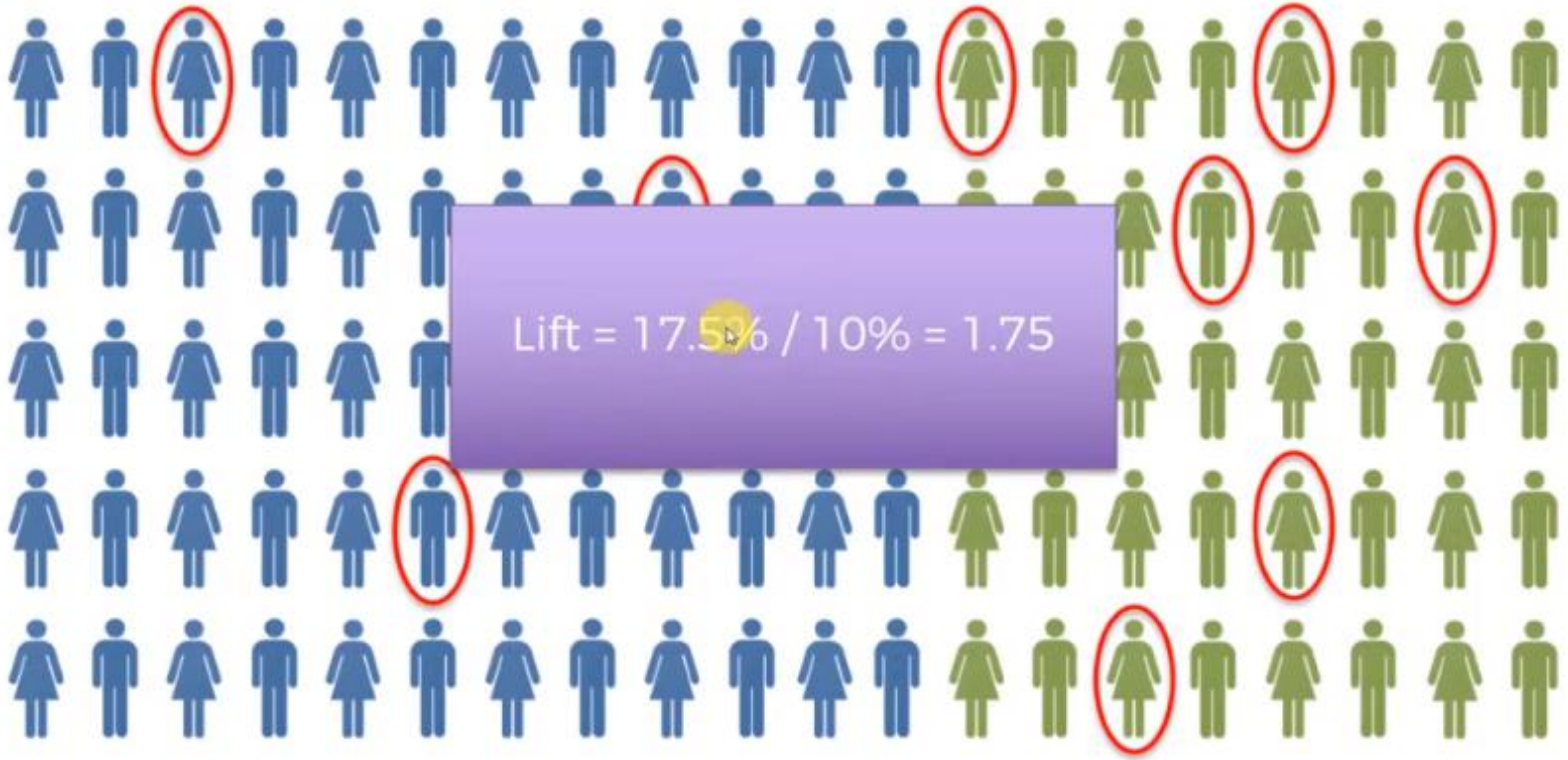
Market Basket Optimisation: $$\text{lift}(I_1 \rightarrow I_2) = \frac{\text{confidence}(I_1 \rightarrow I_2)}{\text{support}(I_2)}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Apriori - Lift



Apriori - Lift



Apriori - Algorithm

Step 1: Set a minimum support and confidence

↓

Step 2: Take all the subsets in transactions having higher support than minimum support

↓

Step 3: Take all the rules of these subsets having higher confidence than minimum confidence

↓

Step 4: Sort the rules by decreasing lift

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

ARL - Movie Recommendation

User ID	Movies liked
46578	Movie1, Movie2, Movie3, Movie4
98989	Movie1, Movie2
71527	Movie1, Movie2, Movie4
78981	Movie1, Movie2
89192	Movie2, Movie4
61557	Movie1, Movie3

Potential Rules:

Movie1 ⇒ Movie2

Movie2 ⇒ Movie4

Movie1 ⇒ Movie3

Activate Windows

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

ARL - Market Basket Optimisation

Transaction ID	Products purchased
46578	Burgers, French Fries, Vegetables
98989	Burgers, French Fries, Ketchup
71527	Vegetables, Fruits
78981	Pasta, Fruits, Butter, Vegetables
89192	Burgers, Pasta, French Fries
61557	Fruits, Orange Juice, Vegetables
87923	Burgers, French Fries, Ketchup, Mayo

Potential Rules:

Burgers ⇒ French Fries

Vegetables ⇒ Fruits

Burgers, French Fries ⇒ Ketchup

Activate Windows

ECLAT

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Eclat - Support

Movie Recommendation: $$\text{support}(\boldsymbol{M}) = \frac{\#\text{ user watchlists containing } \boldsymbol{M}}{\#\text{ user watchlists}}$$

Market Basket Optimisation: $$\text{support}(\boldsymbol{I}) = \frac{\#\text{ transactions containing } \boldsymbol{I}}{\#\text{ transactions}}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Eclat - Algorithm

Step 1: Set a minimum support

↓

Step 2: Take all the subsets in transactions having higher support than minimum support

↓

Step 3: Sort these subsets by decreasing support

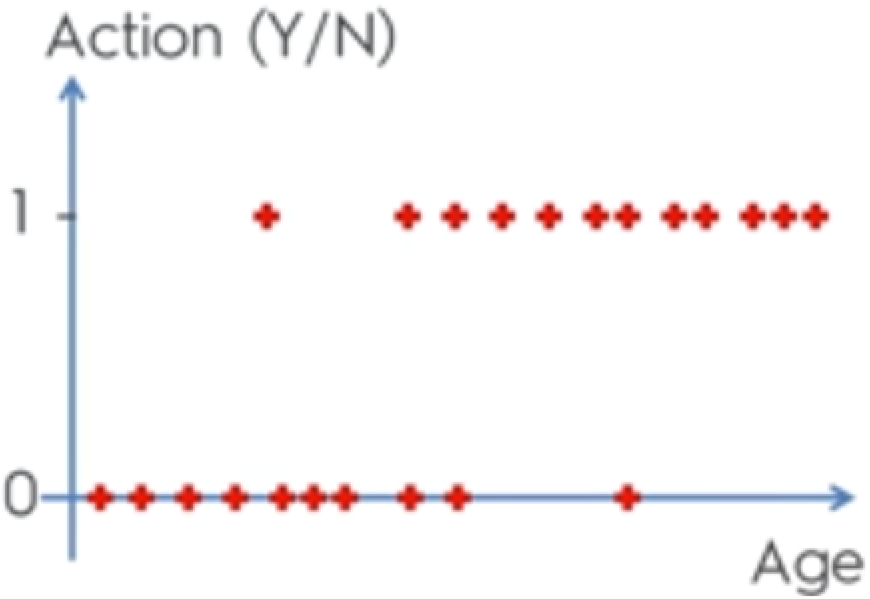
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

LOGISTIC REGRESSION

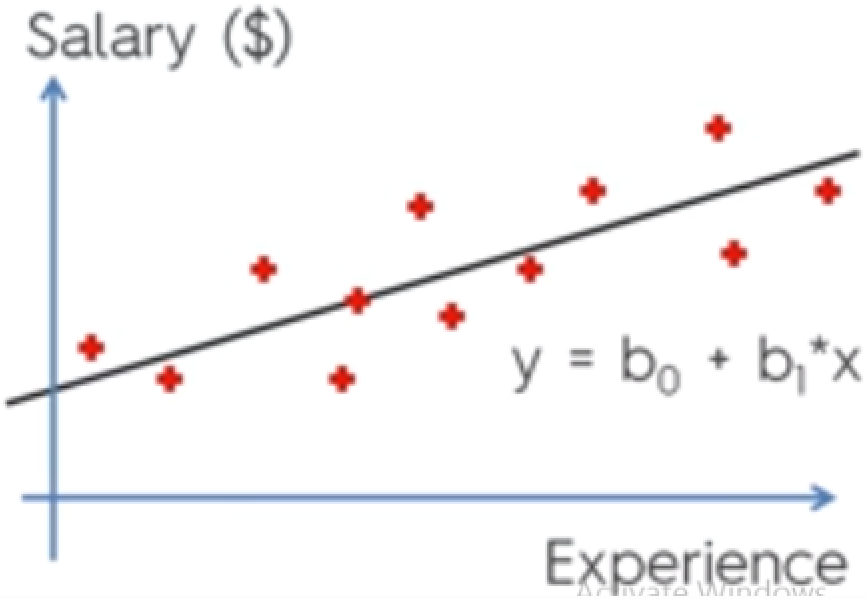
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Logistic Regression

This is new:

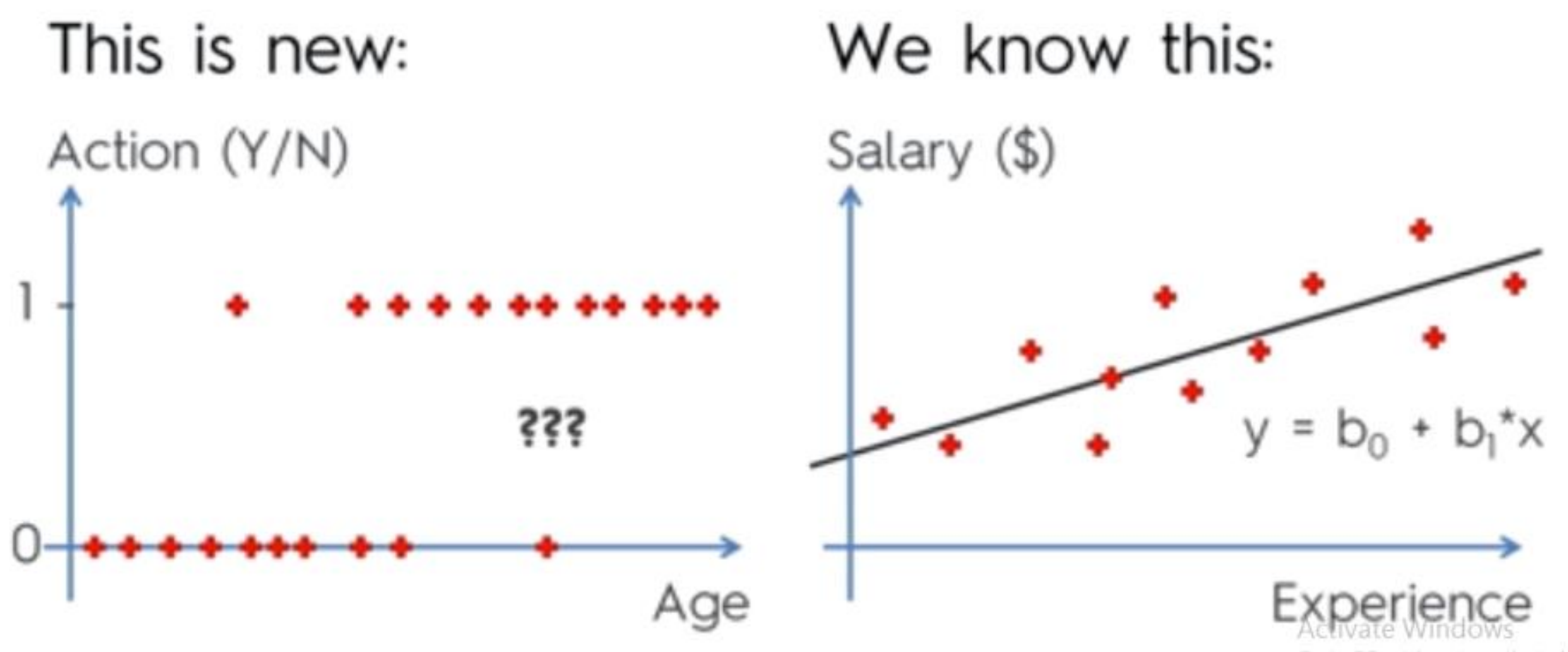


We know this:



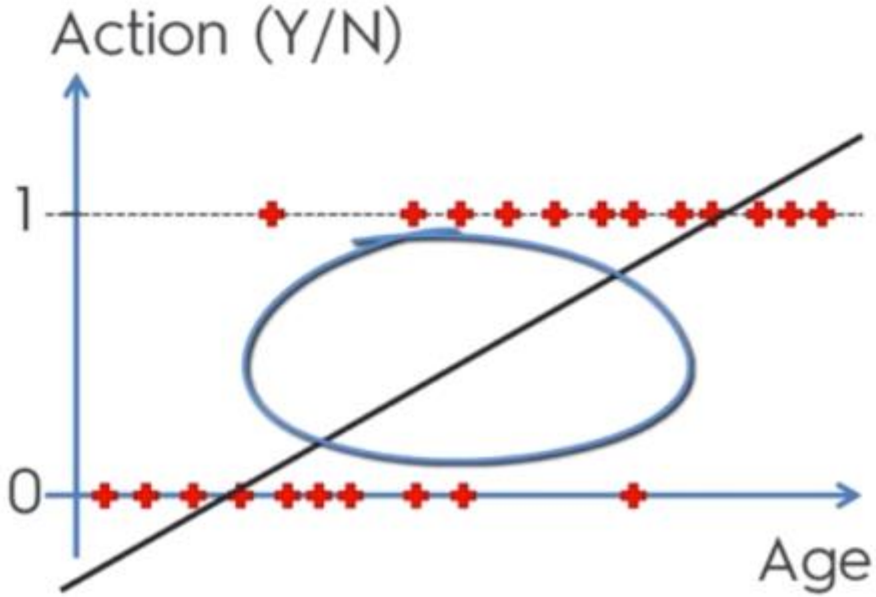
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Logistic Regression

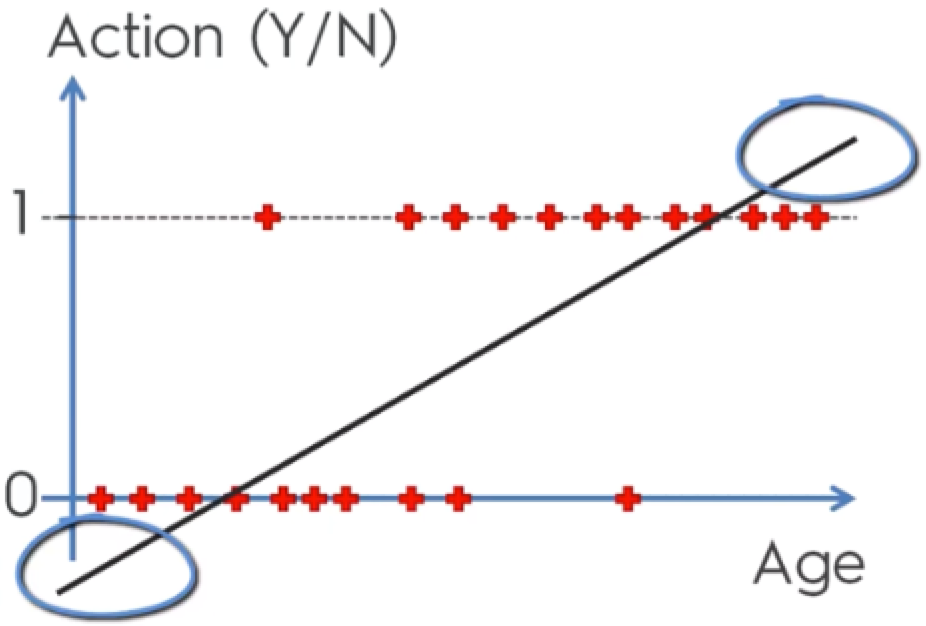


©Larsen & Toubro Infotech Ltd. Privileged and Confidential

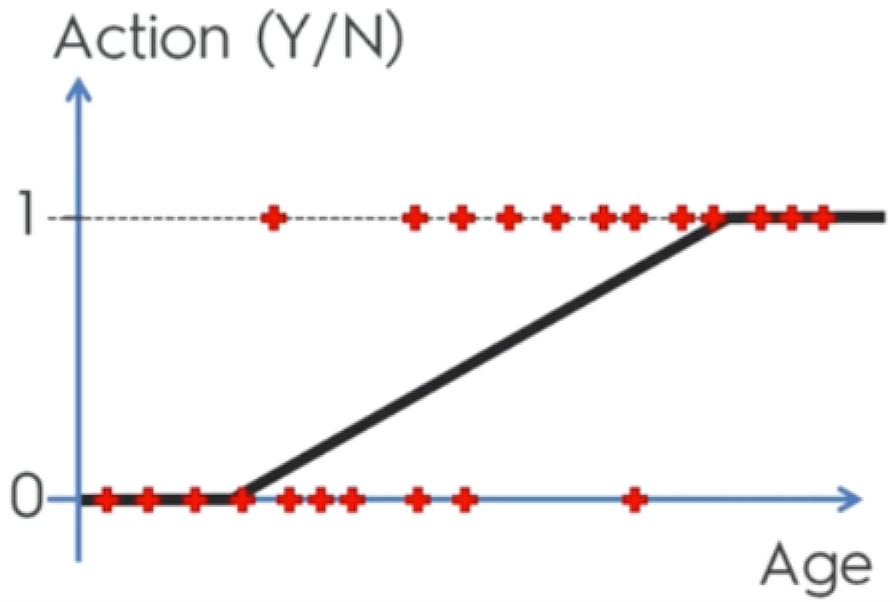
Logistic Regression



Logistic Regression



Logistic Regression



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

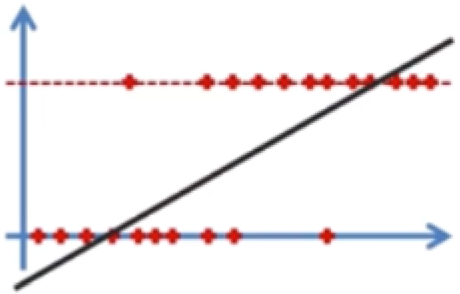
Logistic Regression

$$y = b_0 + b_1{}^{*}x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1{}^{*}x$$



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

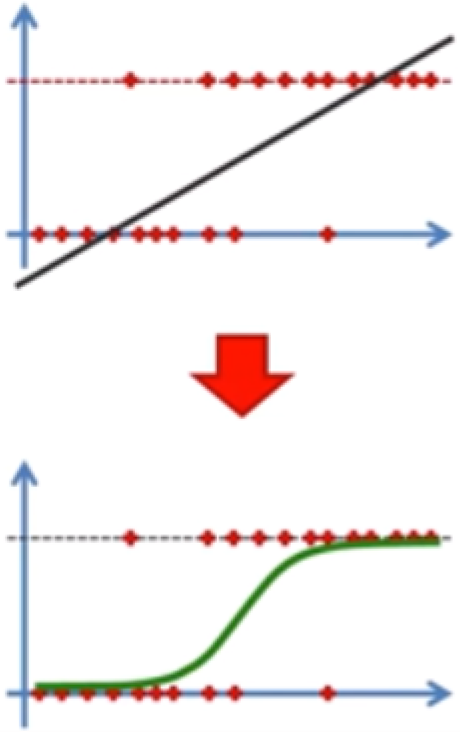
Logistic Regression

$$y = b_0 + b_1{}^*x$$

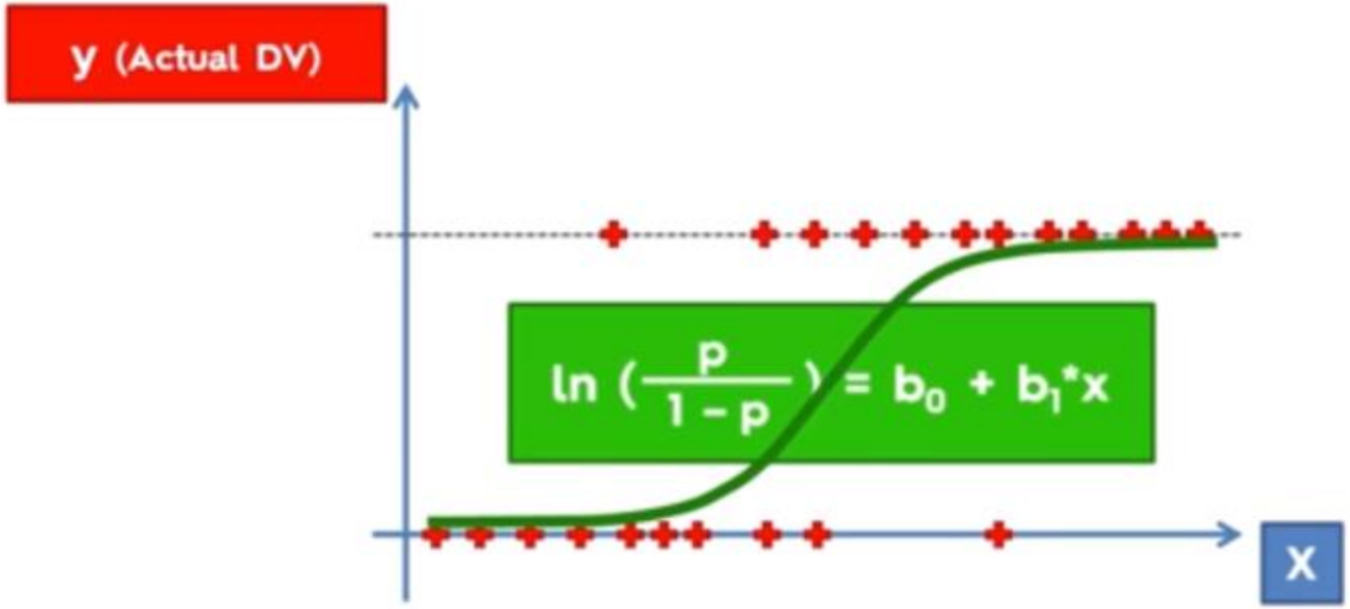
Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

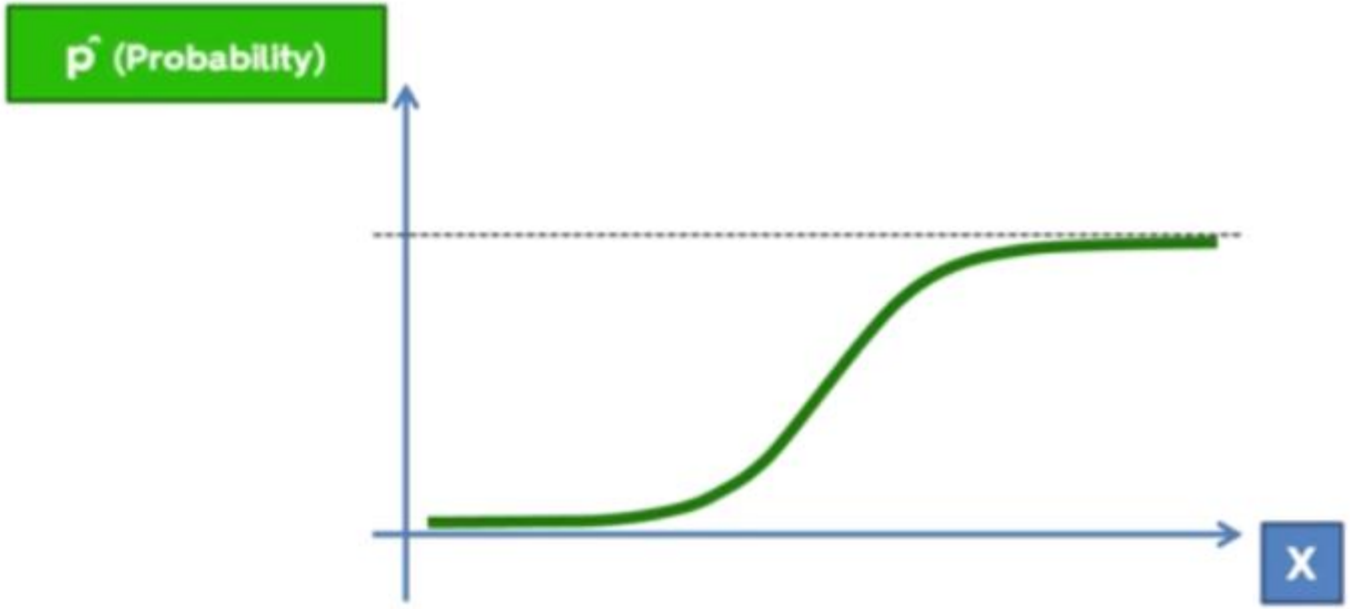
$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1{}^*x$$



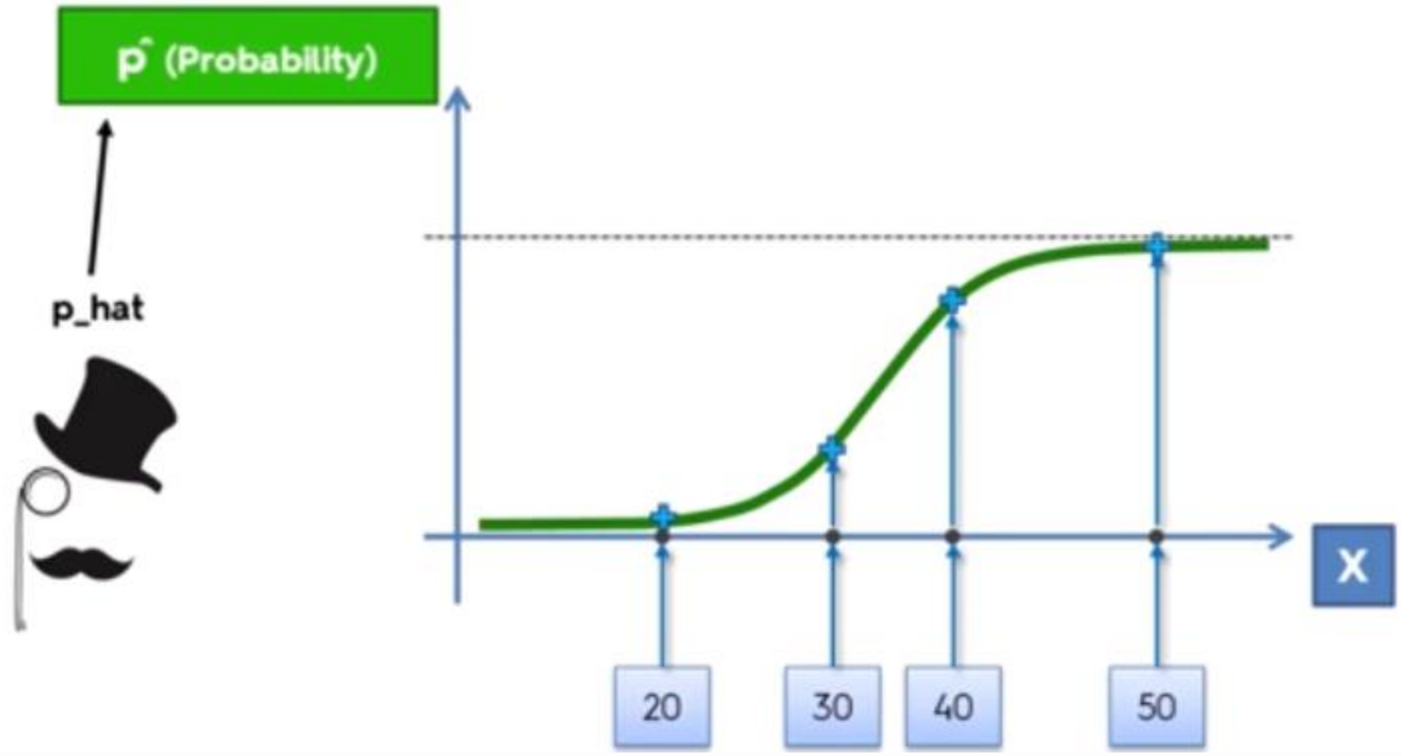
Logistic Regression



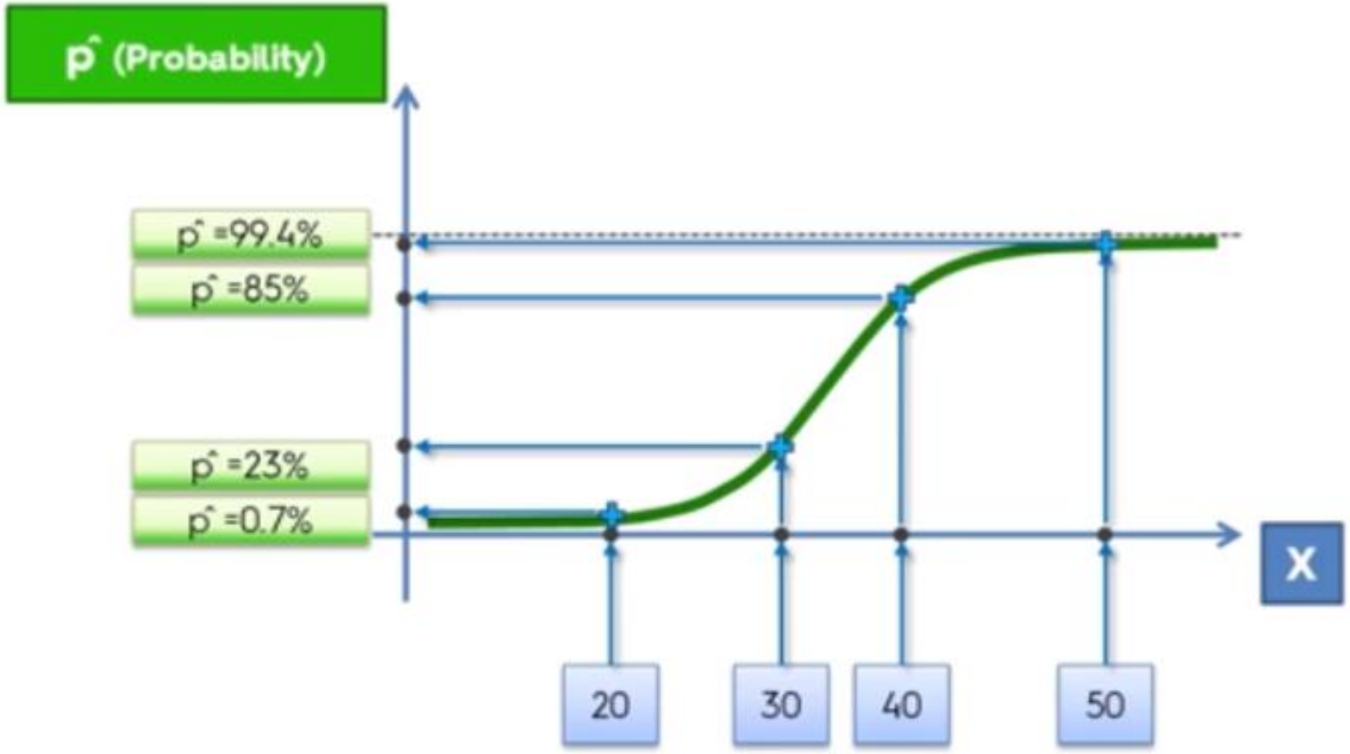
Logistic Regression



Logistic Regression

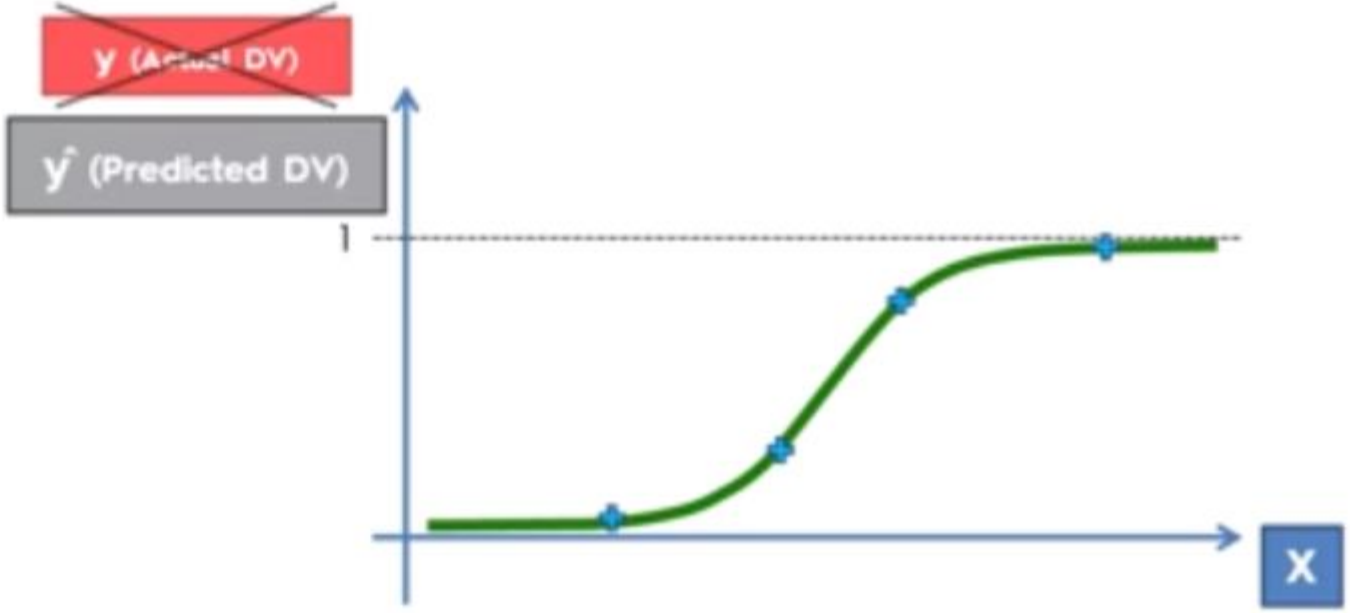


Logistic Regression

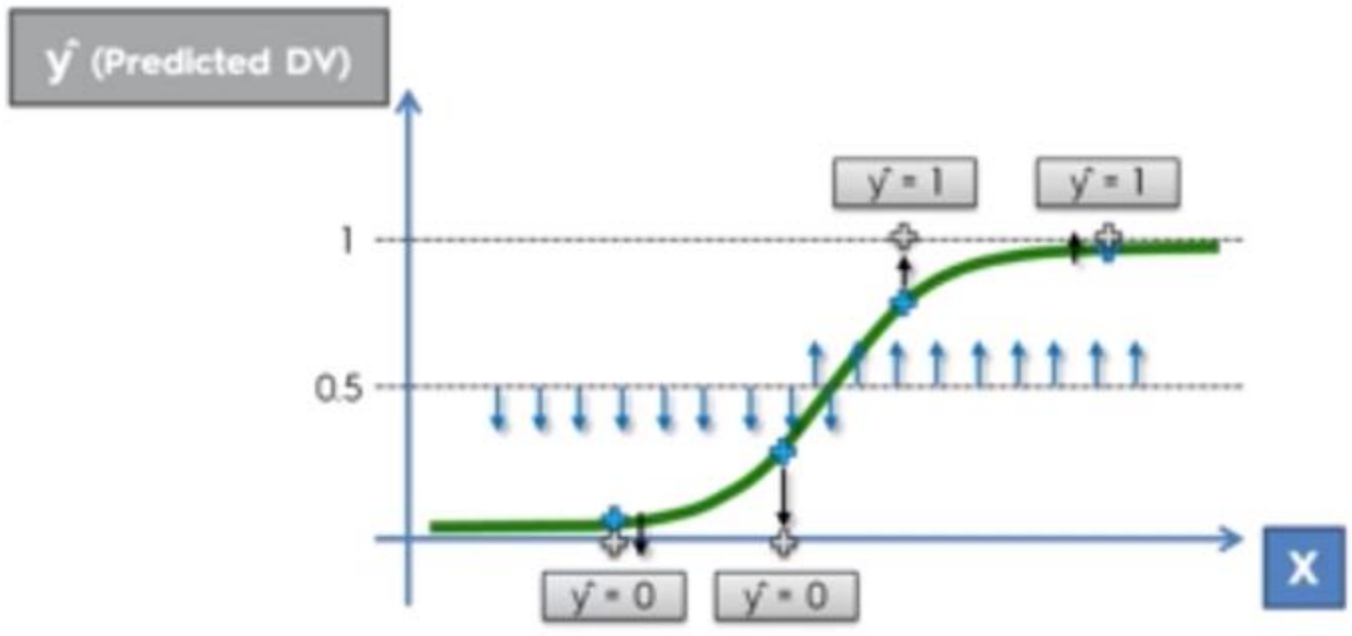


©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Logistic Regression

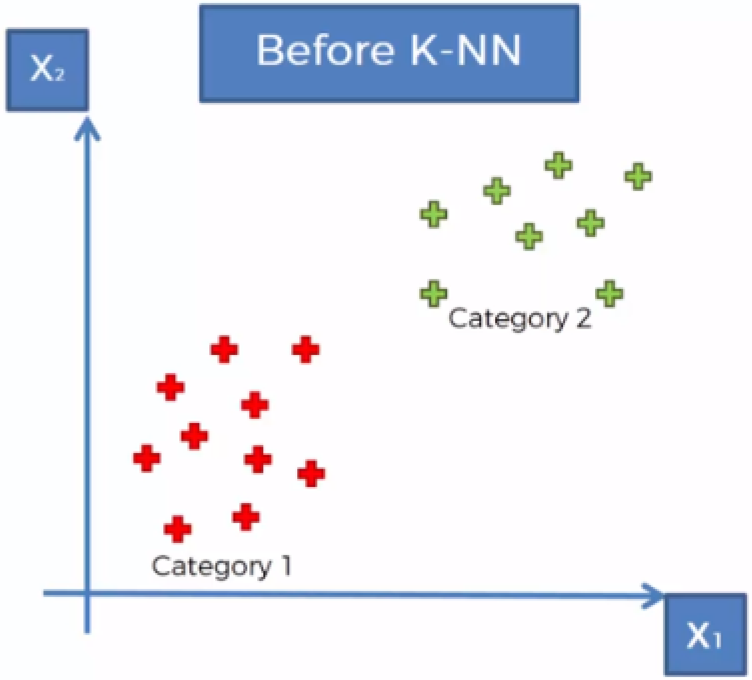


Logistic Regression

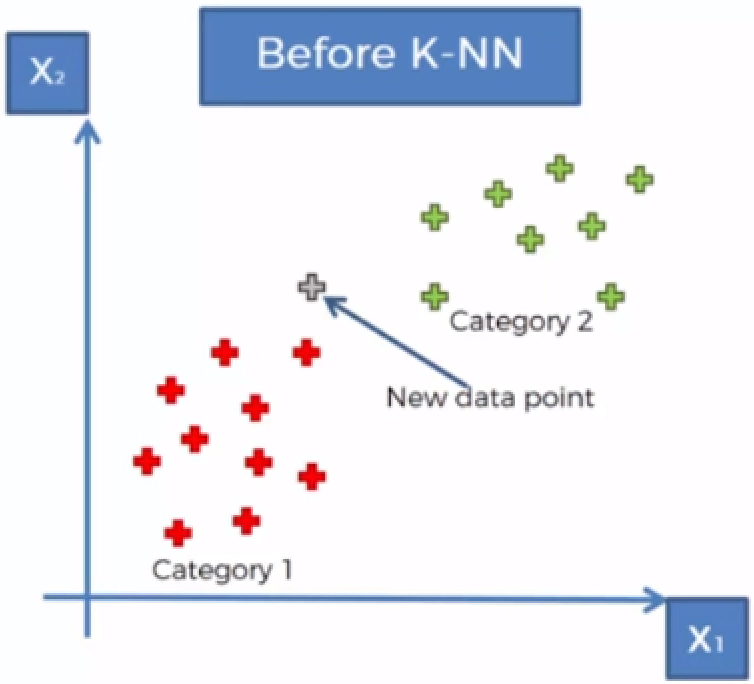


K-NN

What K-NN does for you

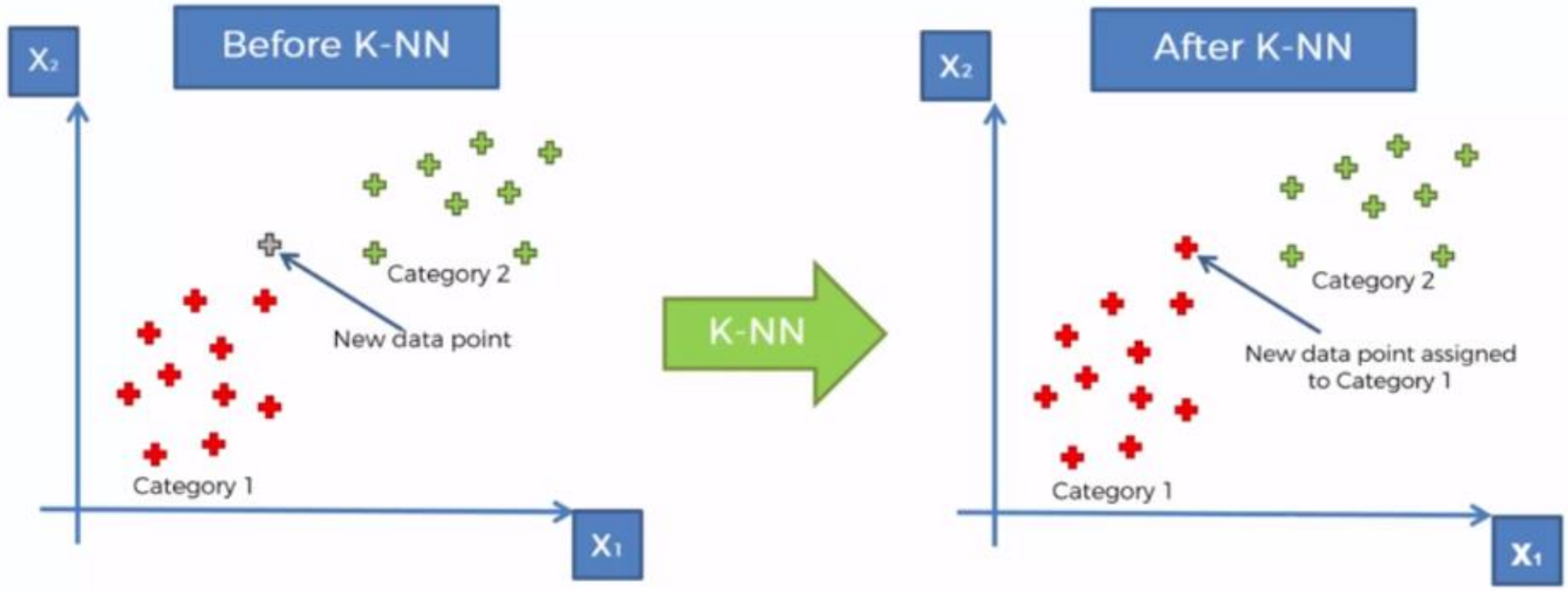


What K-NN does for you



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

What K-NN does for you



How did it do that ?

STEP 1: Choose the number K of neighbors

⇩

STEP 2: Take the K nearest neighbors of the new data point, according to the Euclidean distance

⇩

STEP 3: Among these K neighbors, count the number of data points in each category

⇩

STEP 4: Assign the new data point to the category where you counted the most neighbors

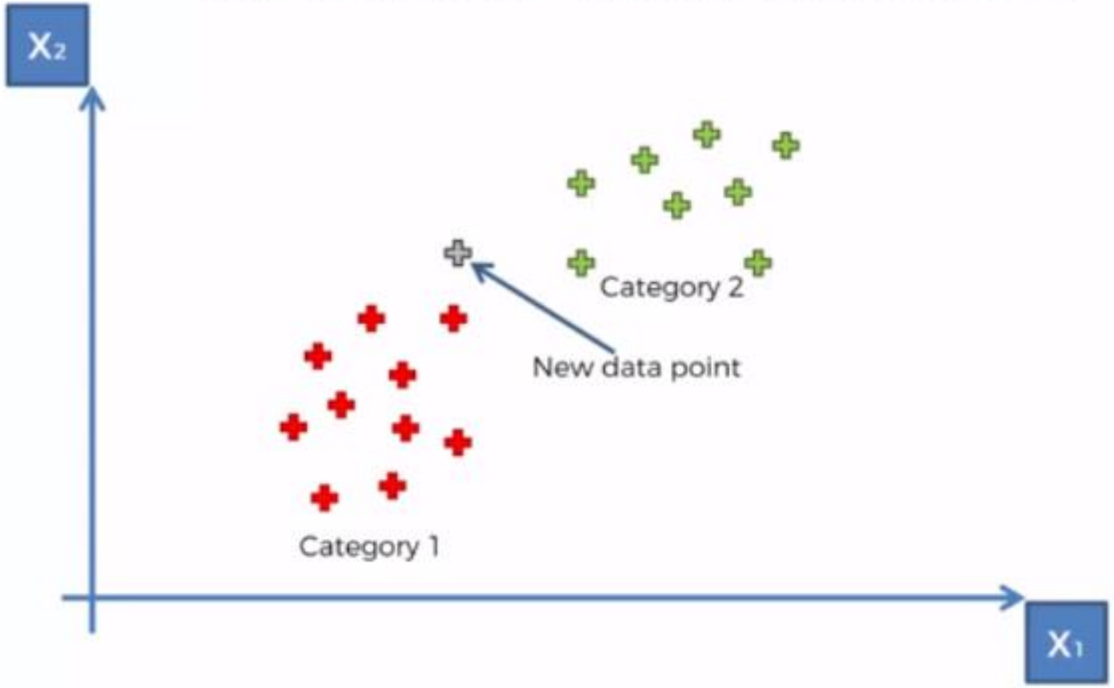
⇩

Your Model is Ready

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

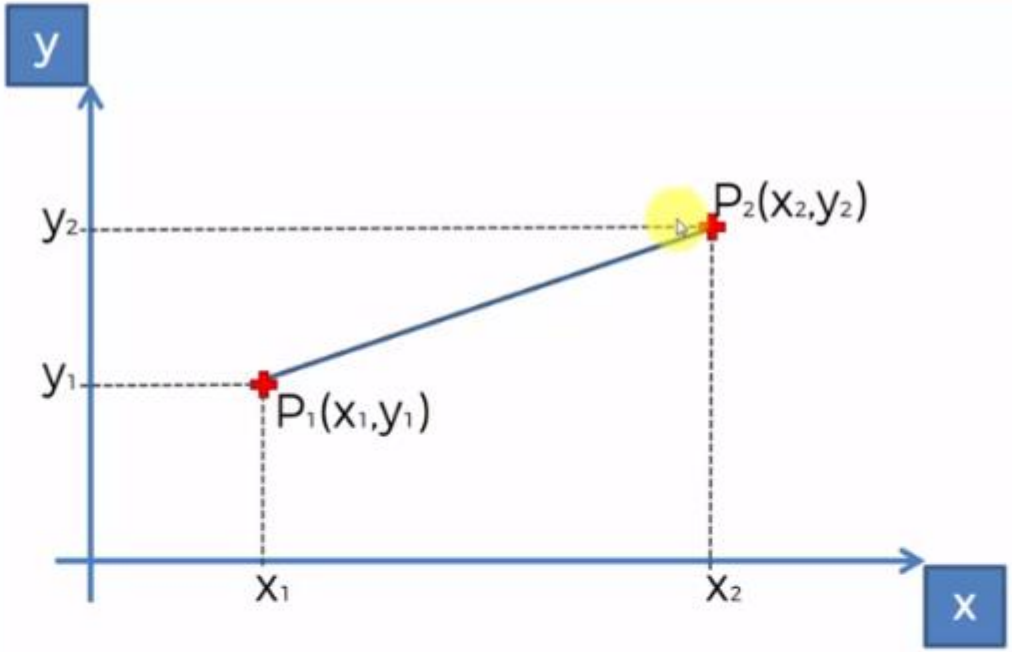
K-NN algorithm

STEP 1: Choose the number K of neighbors: K = 5



Activate Windows

Euclidean Distance



$$\text{Euclidean Distance between } P_1 \text{ and } P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

K-NN algorithm

STEP 2: Take the K = 5 nearest neighbors of the new data point, according to the Euclidean distance

Activate Windows

LTI

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

K-NN algorithm

STEP 3: Among these K neighbors, count the number of data points in each category

Activate Windows

K-NN algorithm

STEP 3: Among these K neighbors, count the number of data points in each category

Category 1: 3 neighbors

Category 2: 2 neighbors

Activate Windows

K-NN algorithm

STEP 4: Assign the new data point to the category where you counted the most neighbors

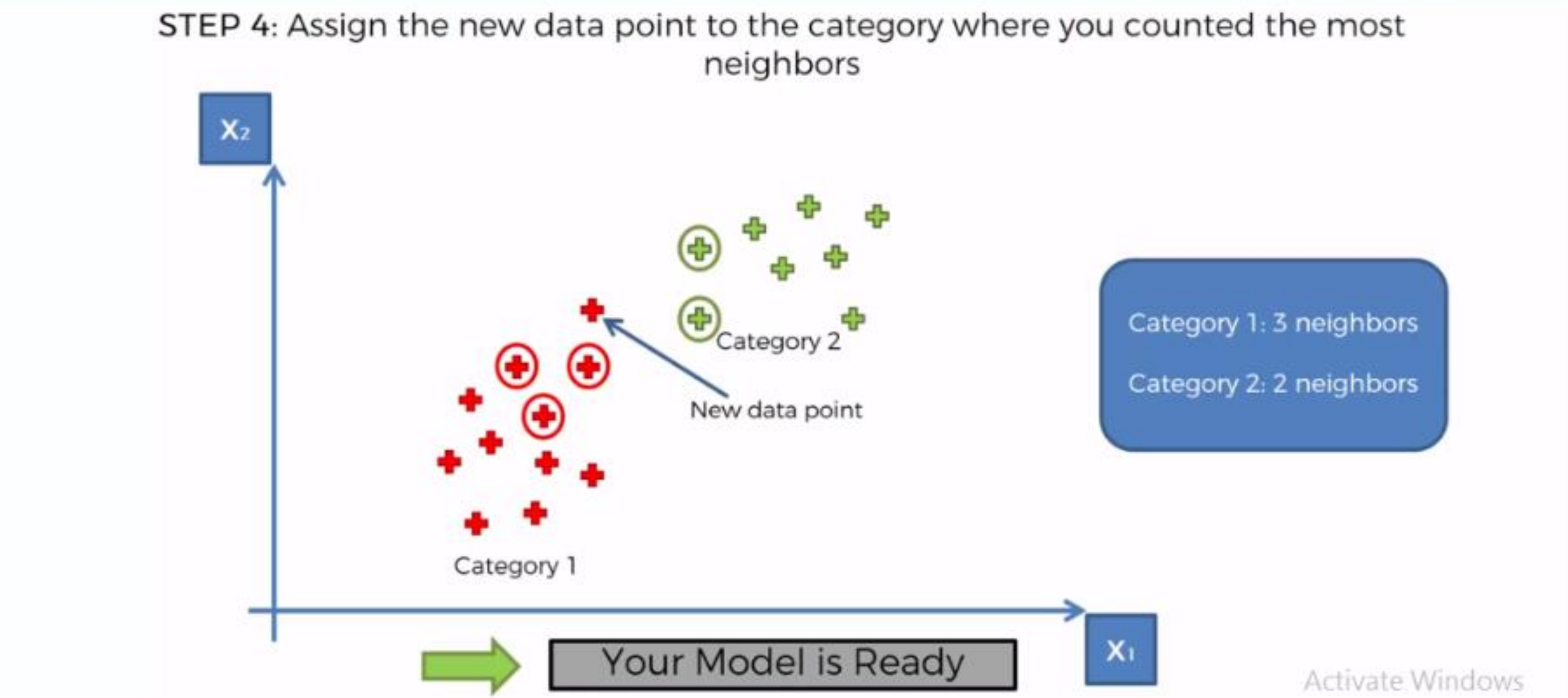
Category 1: 3 neighbors

Category 2: 2 neighbors

Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

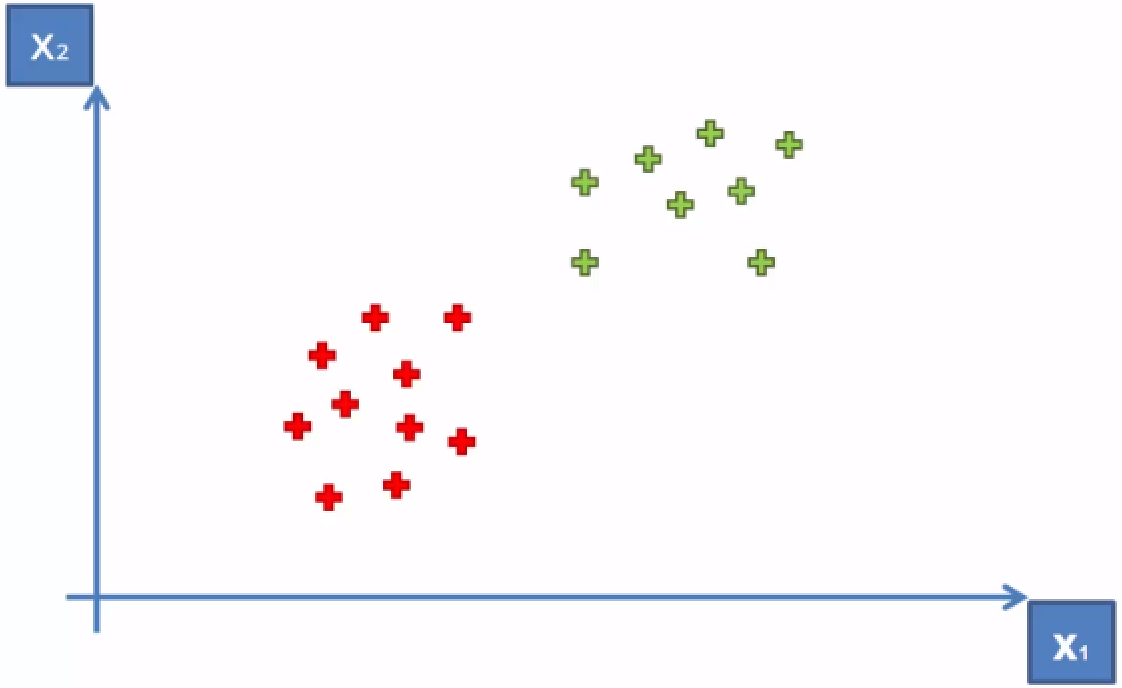
K-NN algorithm



SVM

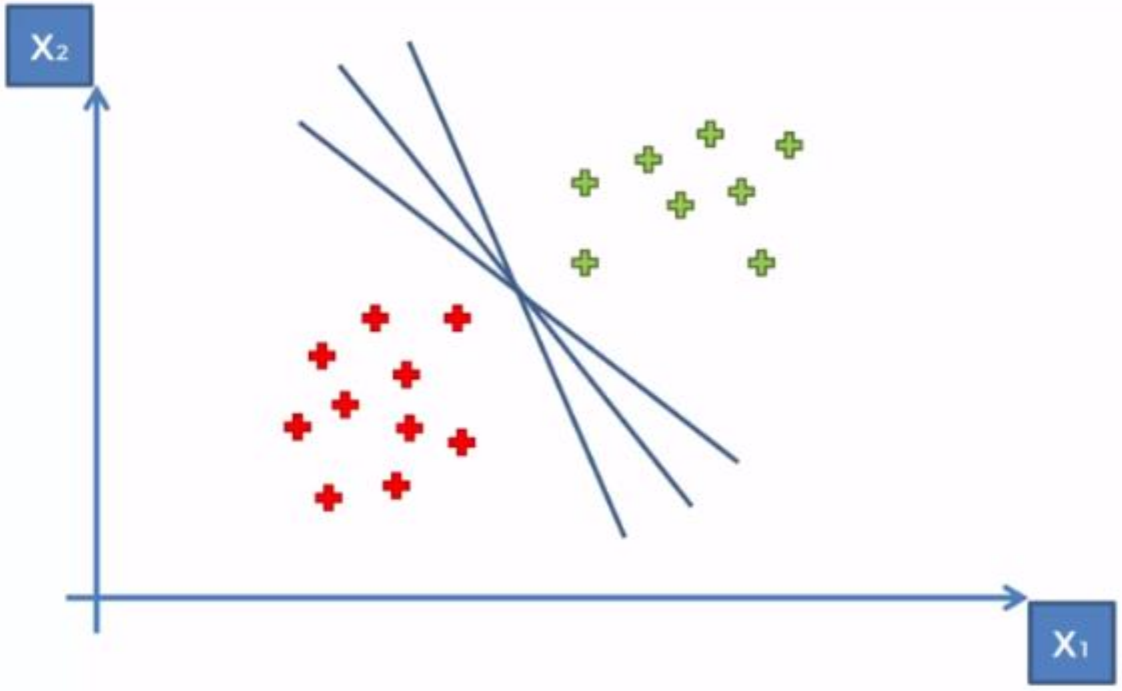
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

How to separate these points ?



Activate Windows

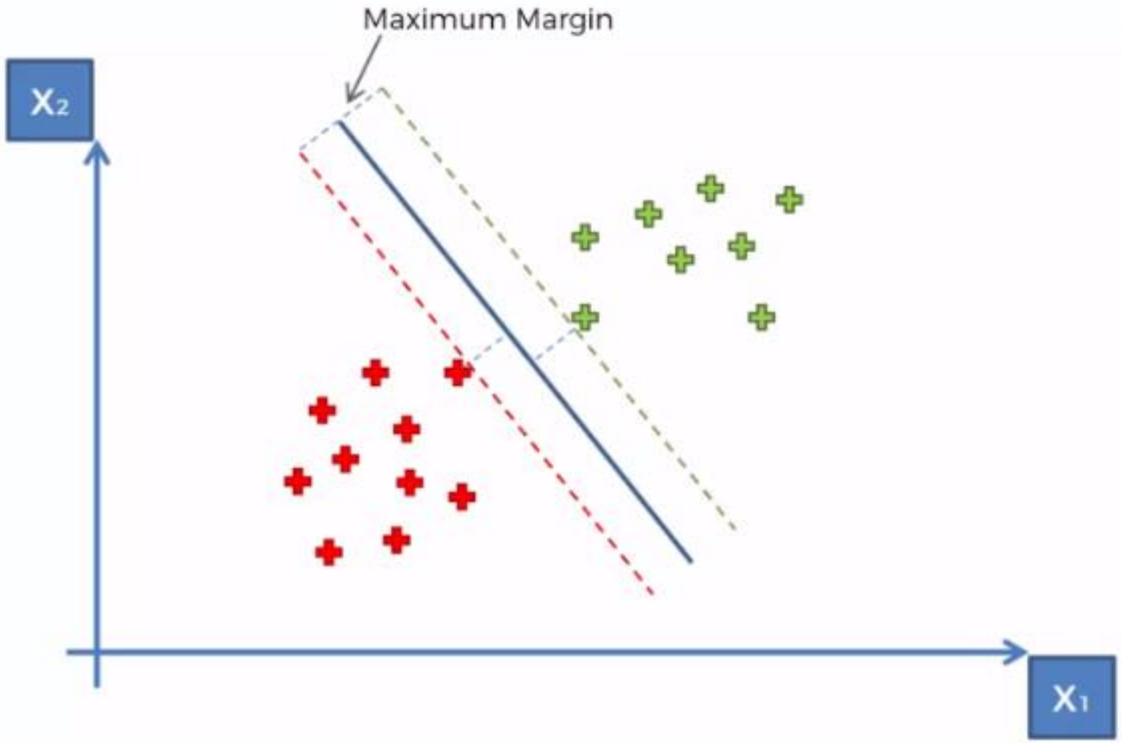
SVM



Activate Windows

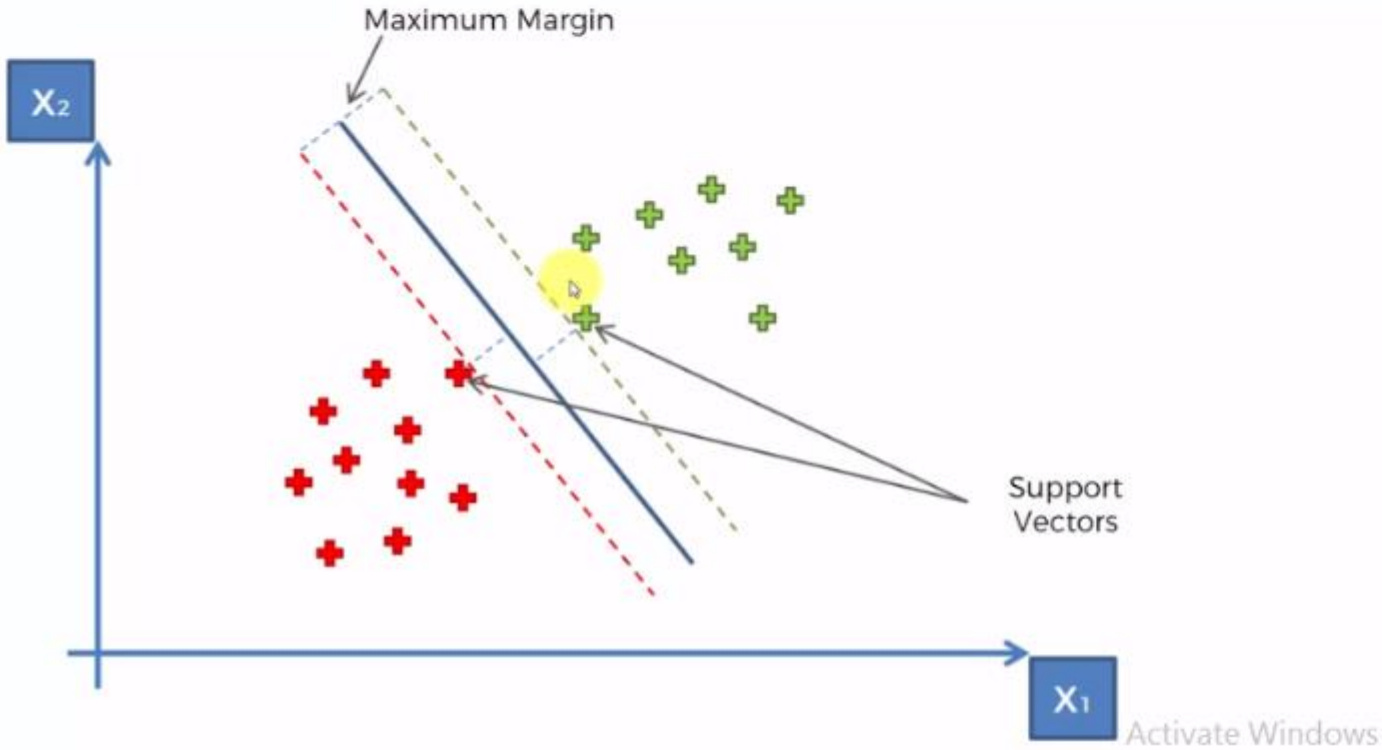
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Maximum Margin

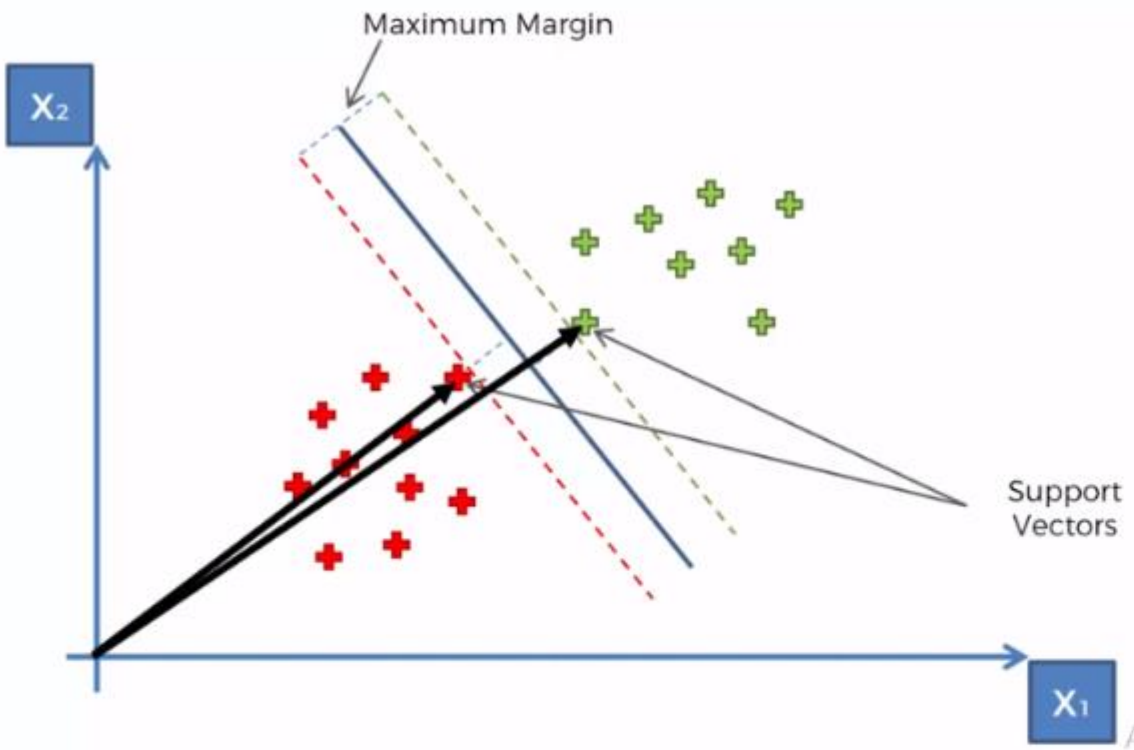


Activate Windows

Support Vectors



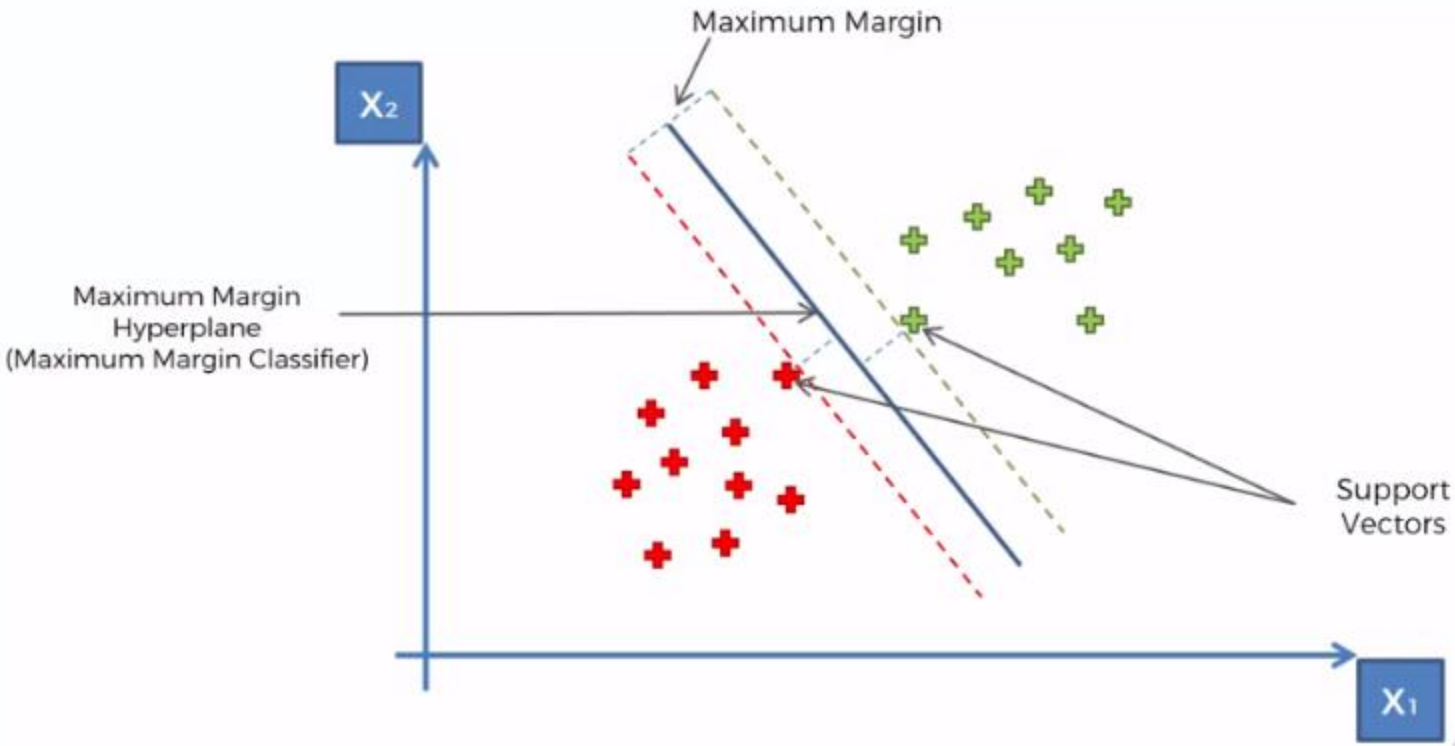
Support Vectors



Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

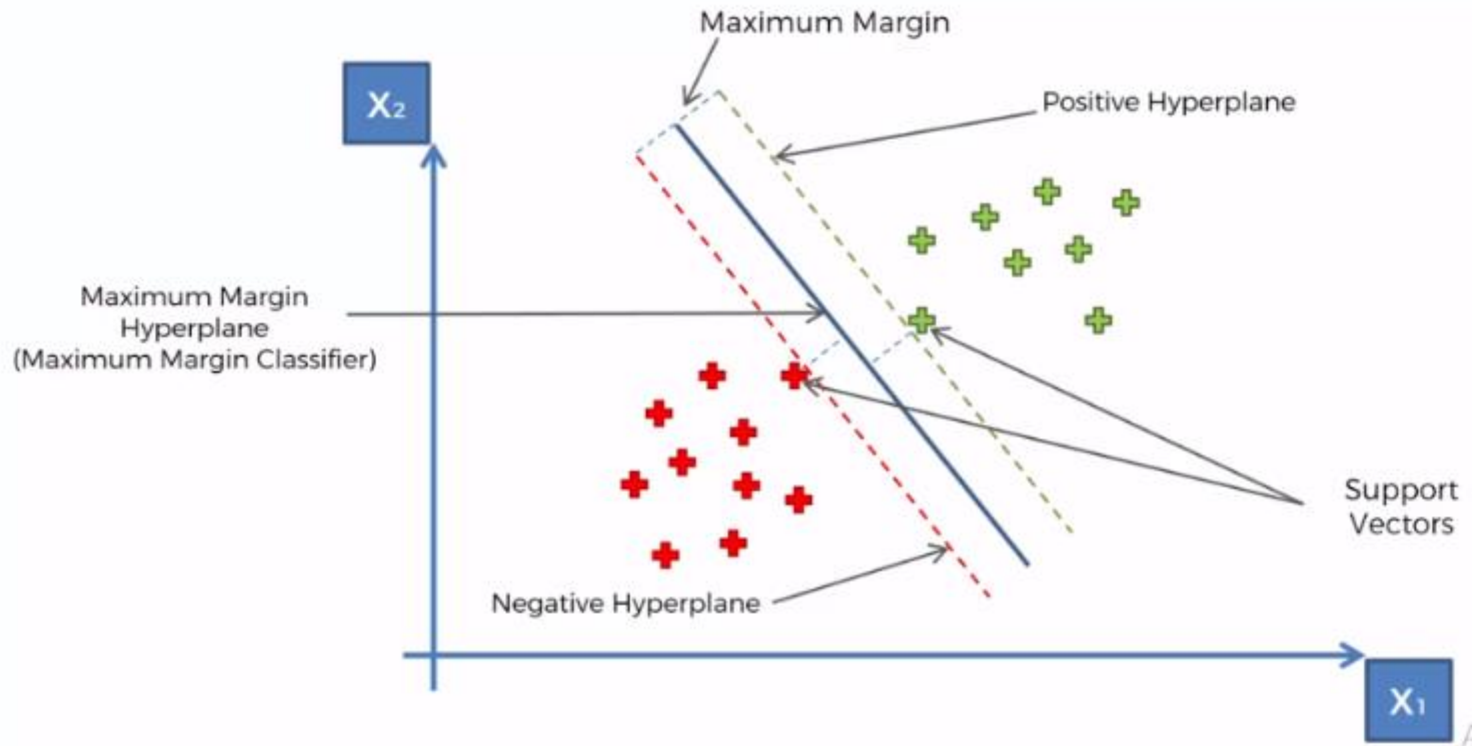
Hyperplanes



Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

Hyperplanes



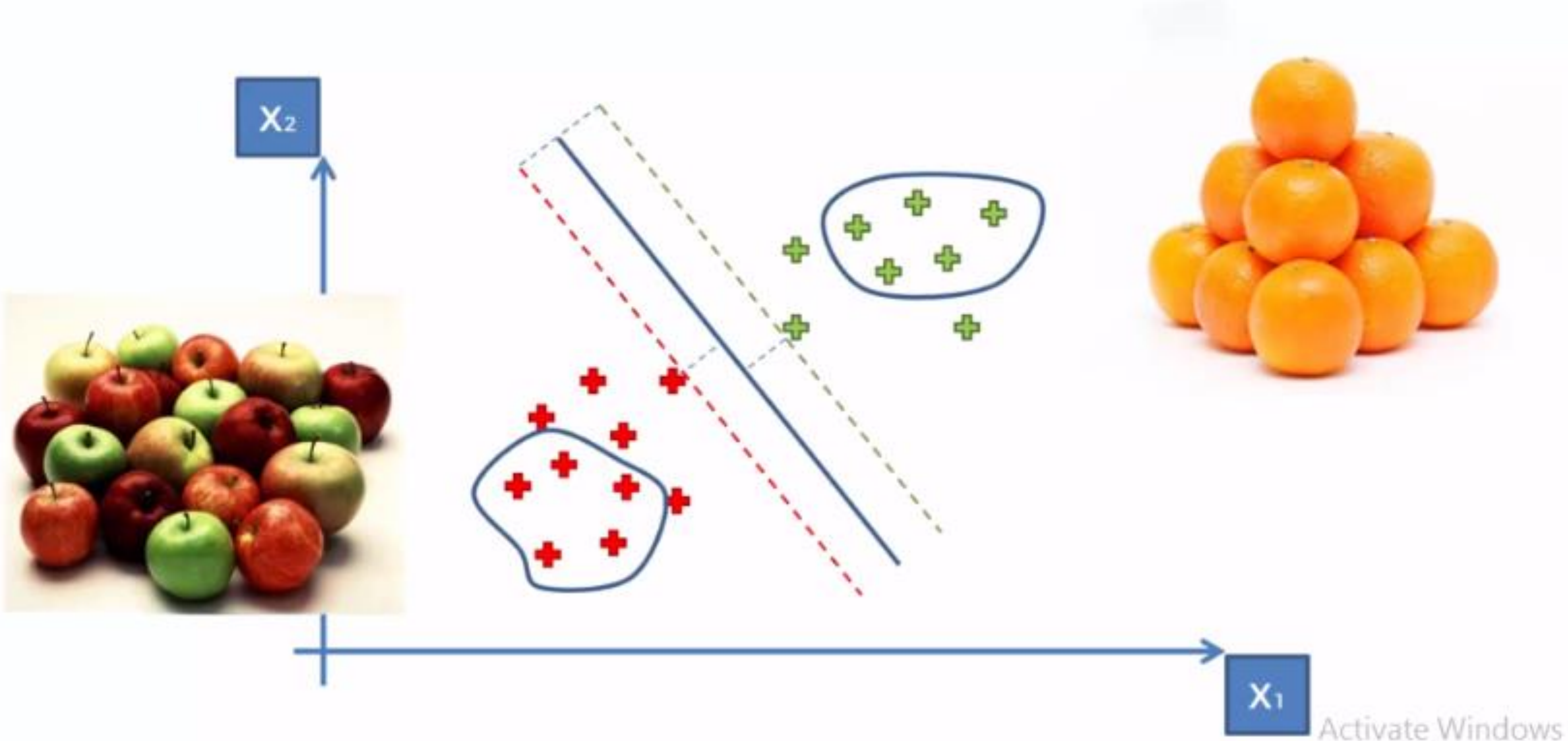
Activate Windows

What's So Special About SVMs?

What's So Special About SVMs?

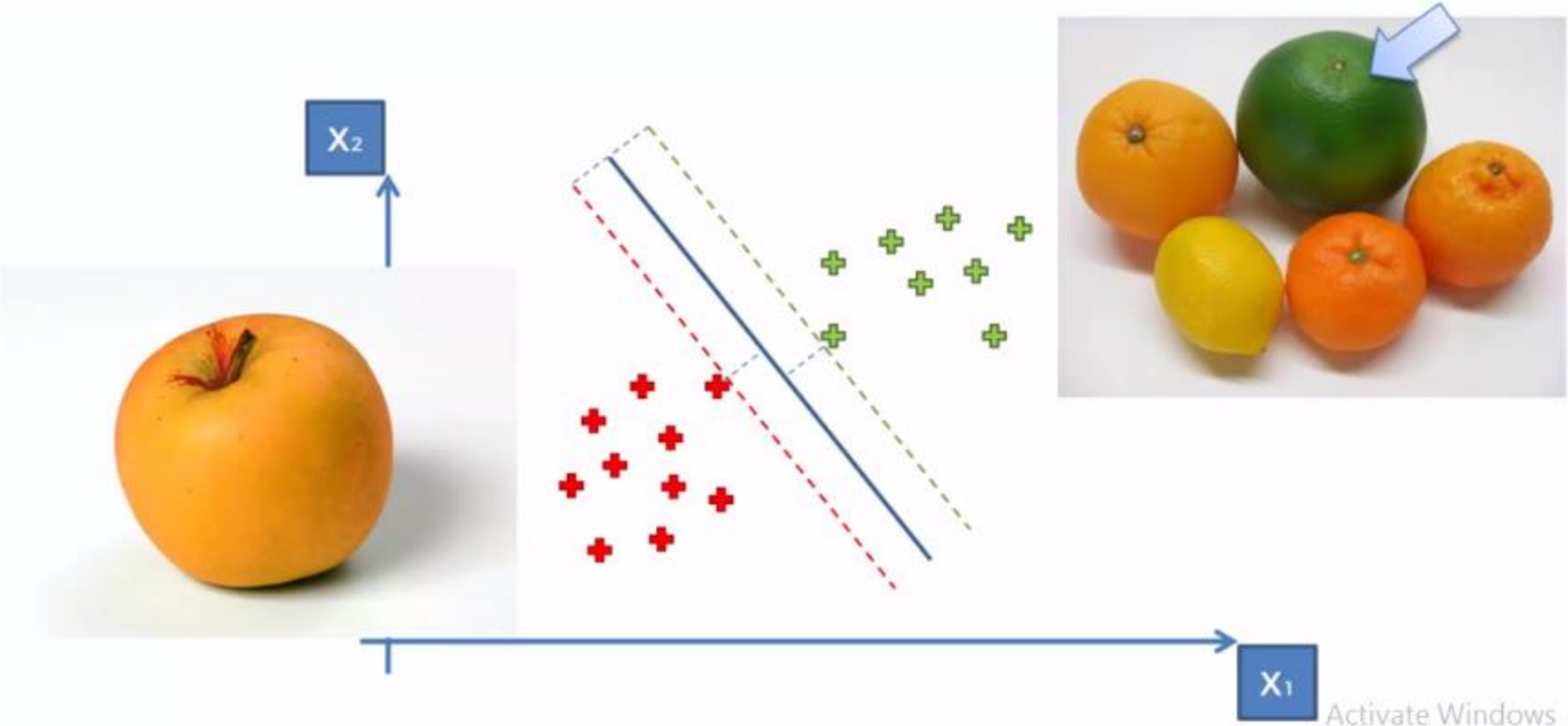
©Larsen & Toubro Infotech Ltd. Privileged and Confidential

What's So Special About SVMs?



©Larsen & Toubro Infotech Ltd. Privileged and Confidential

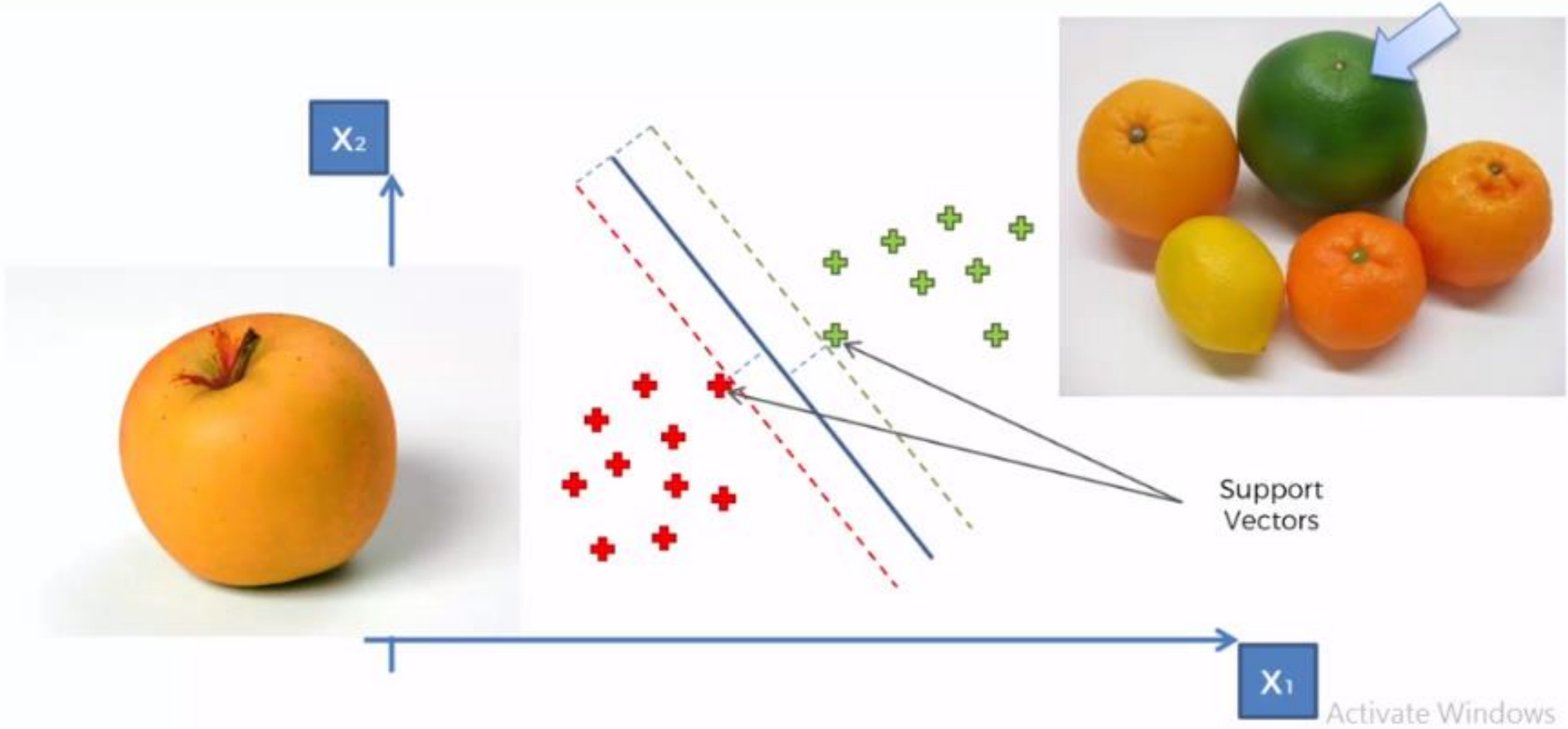
What's So Special About SVMs?



Activate Windows

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

What's So Special About SVMs?



SOURCE(S):

Machine Learning by Kirill Eremenko and Hadelin de Ponteves

©Larsen & Toubro Infotech Ltd. Privileged and Confidential

THANK YOU