Supervised Learning Techniques

Dr.S.Viswanadha Raju

Professor of CSE,JNTUH

svraju.jntu@gmail.com

Topics

1) Problem solving skills

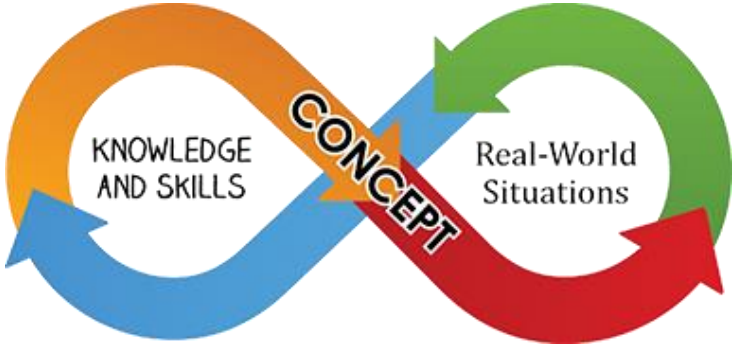
2) Machine Learning –Supervised Learning

3) Case study

4) Advantages AND Disadvantages

Problem Solving Skills -Conceptual Understanding

Conceptual Understanding Rather than rote learning and learning exam orientation



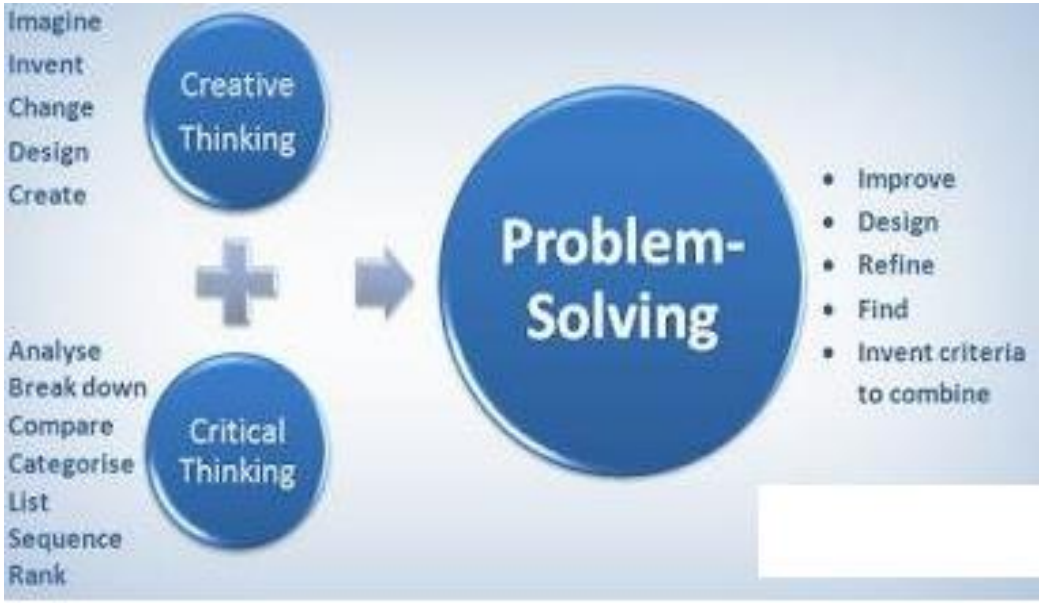
Lesson Type	Purpose	Lesson Objective
Conceptual Understanding	"I learned something new about math today."	"Understand that..."
Fluency/ Procedural Skills	"I learned and/or practiced a more efficient way to do math and I understand why it works."	"Practice or acquire skill in..."
Application	"I used my math knowledge today to solve a problem."	"Apply knowledge of ___ in order to..."

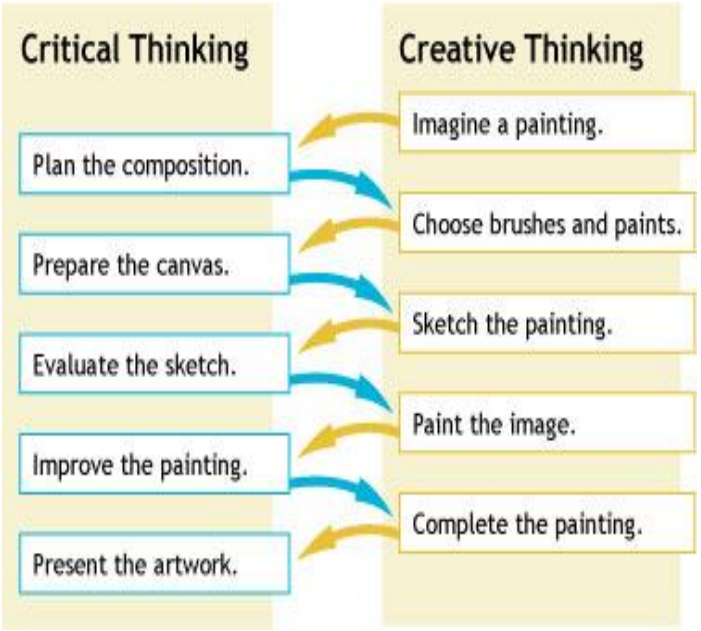
Example 2 : Suppose You are driving a car at the speed of

Example 3: Chicken curry : $y= w1*x1 + w2*x2 + ...+ w_n\ x_n + b$, **w1,w2,..... w_n** are weights of respective gradients , **x1,x2,.....x_n** are gradients ,**b**- flavor

Problem Solving Skills -Critical thinking and Creativity

Critical thinking and **Creativity** to encourage logical decision - making and innovation.





Supervised Learning:

It can apply what has been learned in the past to new data using **labeled examples** to predict future events.

Supervised Machine Learning

Input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

Suppose input data (x) that you can predict the output variables (Y) for that data.

.

Example a : 2, 4, 8, 16, … What is the next number?

$Y = 2^x$

- Ans: 32

Example b : 2, 4, … What is the next number?

- Ans: 8

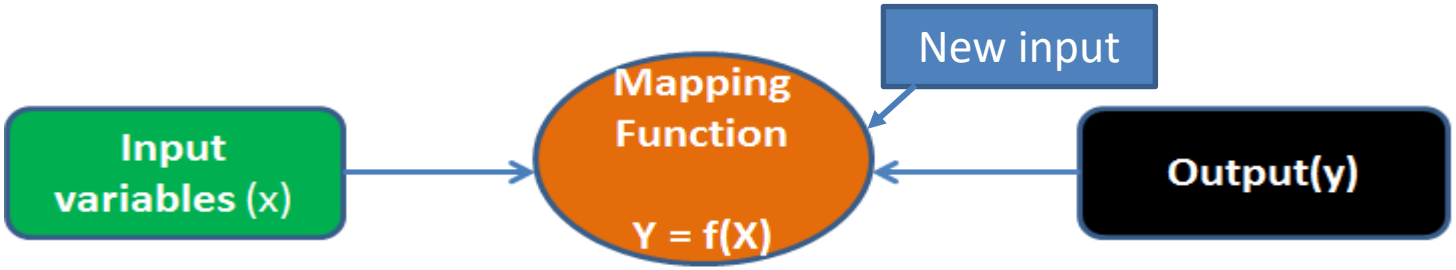
$Y = 2^x$

Example c : 2, 4,6,8 … What is the next number?

- Ans: 10

$Y = 2(x-1) + 2$

Supervised Learning



Goal : With the mapping function , when you have new input data (x) that you can predict the output variables (Y) for that data.

Example 1 :

- 2, 4, 8, 16, … What is the next number?

$y = 2^x$

Ans: 32

Supervised learning

Example 2 :

Text	Category
share your OTP	SPAM
Secure your OTP	NON SPAM
Give me CVV	SPAM
Share me PIN number	SPAM
Don't give your card details	NON SPAM

Consider the query string to classify as follows:

"Send me your OTP"=?? (SPAM / NON SPAM)

Unsupervised Machine Learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

Example 1

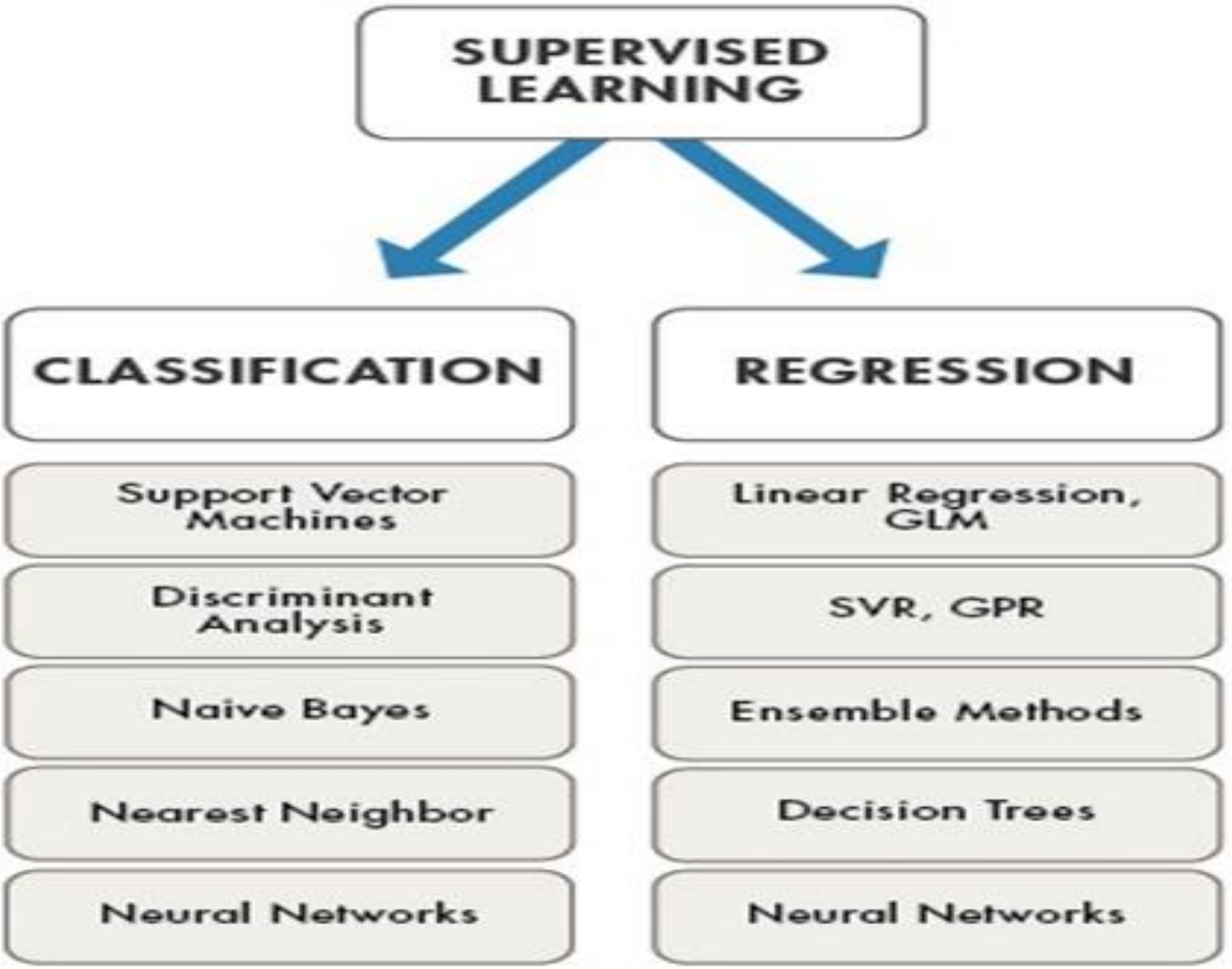
Text (input)	(Output) Category(SPAM /NON SPAM /etc.) ?
Share your OTP	
Secure your OTP	
Give me CVV	
Share me PIN number	
Don’t give your card details	

Example 2

Given a set of numbers {27,26,17,13,25,34,5,8,92}, partition into two sets: **Odd (27,17,13,25,5)** **and** **Even (26,34,8,92)**

- Why this? Why not single and two digit?
- Both mine and your solutions can be right?

Supervised learning



Dataset

Intensifying the Cyber Security using Naive Bayes Classifier

Text	Category
share your OTP	SPAM
Secure your OTP	NON SPAM
Give me CVV	SPAM
Share me PIN number	SPAM
Don’t give your card details	NON SPAM

Consider the query string to classify as follows:

“Send me your OTP”=?? (SPAM / NON SPAM)

Naive Bayes classifier

- It is simple and powerful algorithm for **classification** based on Bayes' Theorem with an assumption of independence among predictors.

- It is easy to build and particularly useful for very large data sets.

- It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Cont..

Finding probability of event B when event A is true

$$P(B / A) = \frac{P(A)}{P(A / B)\, P(B)}$$

P(A) and P(B) = The probability of A and B without any correspondence with each other.

P (B / A) = posterior probability [probability of event B after event A is true]

P(A|B) = The conditional probability that event A occurs, given that B has occurred.

Spam detection using Naive Bayes classifier

Naïve Bayes Algorithm:

- Consider and analyze the dataset
- Classify the dataset into different categories with the labels
- Apply the Naive Bayes classifier and train the machine on the dataset.
- Test the machine by giving a query with different categories .
- Highest probability value will decide the category of the query.

Case Study 1 : Dataset

Step 1: Consider the dataset

Text	Category
share your OTP	SPAM
Secure your OTP	NON SPAM
Give me CVV	SPAM
Share me PIN number	SPAM
Don't give your card details	NON SPAM

Step 2: Train the dataset and perform the text analysis results as follows

- Total no.of words in the dataset = 18
- Total no.of words in the **SPAM** category=10
- Total no.of words in the **NON SPAM** category=8
- Total no.of distinct words =12

Consider the query string to classify as follows:

"Send me your OTP"=?? (SPAM / NON SPAM)

Step 3: Find the probability of the given Query as follows

P(Send me your OTP) = P(Send)x P(me) x P(**your**) x P(**OTP**)

P(Send me your OTP/ SPAM)

= P(Send / SPAM) x P(me/ SPAM) x P(your/ SPAM) x P(OTP/ SPAM)

P(Send me your OTP/ NON SPAM)

= P(Send / NON SPAM) x P(me/ NON SPAM) x P(your/ NON SPAM)
x P(OTP/ NON SPAM)

Cont..

P(SPAM / Send me your OTP)

= (0/ 10) x (2/ 10) x (1/ 10) x (1/ 10)

= 0 x 2/10 x 1/10 x 1/10

= 0

If the probability of one of the words in the given data is zero.

This is known as the **'Zero Frequency'**. It can solve using the Laplace estimation or Smoothing. Laplace estimation is given by :

$$P(word) = \frac{(\text{Word count} + \alpha)}{(\text{Total number of words in a category} + \text{Total number of distinct words})}$$

Where $\alpha = 1$

Cont..

Step 4: **Probability of finding the query in the SPAM category**

$$P(\text{send}/\text{SPAM}) = \frac{0 + 1}{10+12} = 0.045$$

$$P(\text{ me}/\text{SPAM}) = \frac{2 + 1}{10+12} = 0.136$$

$$P(\text{your}/\text{SPAM}) = \frac{1 + 1}{10+12} = 0.090$$

$$P(\text{OTP}/\text{SPAM}) = \frac{1 + 1}{10+12} = 0.090$$

P(Send me your OTP/SPAM) = **0.0000495**

Probability of finding the query in the NON SPAM category

$$P(\text{send}/\text{NON SPAM}) = \frac{0 + 1}{8 + 12} = 0.050$$

$$P(\text{ me}/\text{NON SPAM}) = \frac{0 + 1}{8 + 12} = 0.050$$

$$P(\text{your}/\text{NON SPAM}) = \frac{2 + 1}{8 + 12} = 0.150$$

$$P(\text{OTP}/\text{NON SPAM}) = \frac{1 + 1}{8 + 12} = 0.100$$

$$P(\text{Send me your OTP}/\text{NON SPAM}) = \mathbf{0.0000375}$$

Cont..

Final probabilities over the query are

P(Send me your OTP/NON SPAM) = **0.0000375**

P(Send me your OTP/SPAM) = **0.0000495**

Step 5:From the above observations the query "**send me your OTP**" falls under the category of **SPAM** as it has the highest probability.

By the application of Naive Bayes classifier the spam detection can be easily detected and enhances cyber security

Advantages of Supervised Learning:

- To collect data or produce a data output from the previous experience
- To optimize performance criteria using experience
- To solve various types of real-world computation problems.
- To find out exactly how many classes are there before giving the data for training.

Disadvantages of Supervised Learning

- Decision boundary might be over trained if your training set which doesn't have examples that you want to have in a class
- It cannot cluster or classify data by discovering their features by its own, unlike unsupervised learning.
- In case of classification, if we give an input which is not from any of the classes in the training data, then the output may be a wrong class label.
- For example, let’s say you trained an image classifier with cats and dogs data. Then if you give the image of a giraffe, the output may be either cat or dog, which is not correct.

Thank you all

Dr.S.Viswanadha Raju

Professor of CSE,JNTUH

svraju.jntu@gmail.com