Exam No: _____

GANPAT UNIVERSITY

B. TECH.CSE (CBA/BDA) SEM- VII REGULAR EXAMINATION– NOV-DEC 2022

(2CSE702) MACHINE LEARNING

TIME: 3 HRS **TOTAL MARKS: 60**

SECTION-I

Q-1	**1**	How gradient descent can be used in linear regression.	**(01)**	**10**
		Answer: To find weights which minimizes error		
	2	What is the limitation of the k-means algorithm?	**(01)**	
		Answer: Outliers affect it very much		
	3	How does the value of learning rate affect the behavior of gradient descent?	**(01)**	
		Answer: If it is too high it causes fluctuation, if it is two low, it slows convergence		
	4	Calculate Euclidean distance between data points (13,20) and (12,16).	**(01)**	
		Answer:4.1		
	5	Explain a scenario which describes the importance of feature scaling in machine learning.	**(01)**	
		Answer: When attributes are in different range		
	6	Suppose you are given the set of data points, how you can classify them as Core point/Border point/Outlier using DBSCAN.	**(01)**	
		Answer: If a data point has m data points in its r neighborhood then it is a core point. If a datapoint is not core point but it is reachable from at least one core point then it is border point. if a point can't be either core or border point then it is outlier		
	7	Problem statement: Find whether a person has diabetes or not based on the age, weight,height and gender of the person. Which type of machine learning problem is this?	**(01)**	
		Answer: Classification		

8 In the problem statement given in Q7, which attributes are continuous attributes? **(01)**

Answer: age, weight, hight

9 List down names of algorithms which are based on bagging technique. **(01)**

Answer: Randomforest

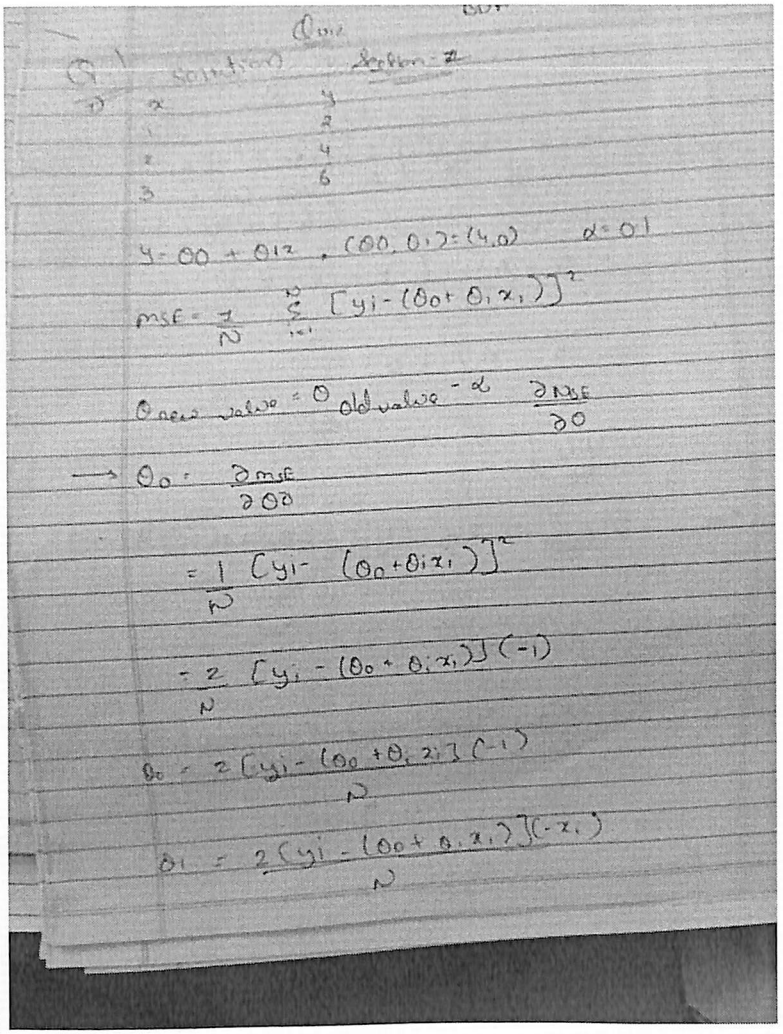
10 In the problem statement given in Q7, which tests should be used for feature selection. **(01)**

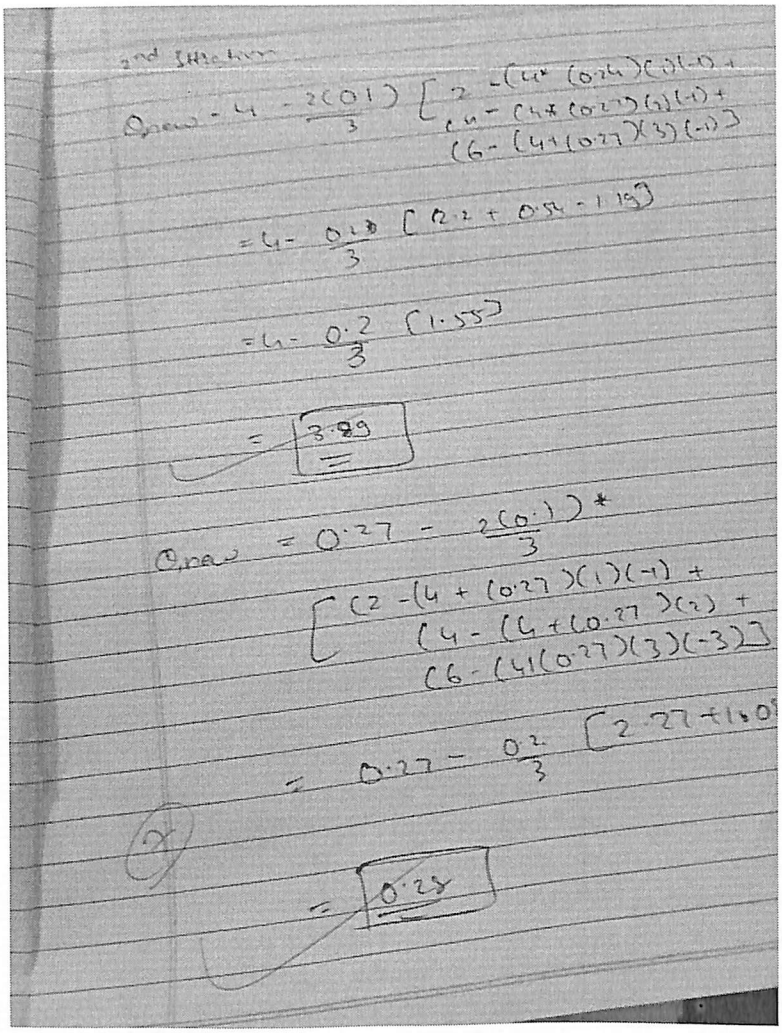
Answer: ANOVA, Chi square

Q-2 **(A)** Learn $y = \theta 0 + \theta 1x$ on the following dataset, using vanilla gradient descent where initially $(\theta 0, \theta 1)=(4, 0)$ and step-size, $\alpha = 0.1$. Find values of $\theta 0$ and $\theta 1$ for 2 iterations. **(5)** **10**

x	y
1	2
2	4
3	6

Answer:


Solution:
x
y
1
2
2
4
3
6
y = θ0 + θ1x , (θ0, θ1) = (4, 0) α = 0.1
MSE = 1/N Σ [yi - (θ0 + θ1xi)]^2
θ new value = θ old value - α ∂MSE/∂θ
θ0 = ∂MSE/∂θ0
= 1/N [yi - (θ0 + θ1xi)]^2
= 2/N [yi - (θ0 + θ1xi)] (-1)
θ0 = 2/N [yi - (θ0 + θ1xi)] (-1)
θ1 = 2/N [yi - (θ0 + θ1xi)] (-xi)



Q-2 (B) Find information gain of **Temperature** attribute of following dataset **(5)**

Temperature	Minutes played
High	20
Low	24
High	25
High	30
High	4
Low	22
Low	23
High	25
Low	15

Answer:
Entropy of dataset: 56.61
Entropy of low=variance of (24,22,23,15)=16.66
Entropy of high=variance of(20,25,30,4,25)=100.7
Information gain=56.61-(-5/9*100.7-4/9*16.66)=119.91

OR

Q-2 **(B)** Explain Hierarchical clustering in detail. "Hierarchical clustering is better than k-means clustering". Justify this statement **5**

Answer:
Strategies for hierarchical clustering generally fall into two types:
1. Hierarchical clustering algorithms build a hierarchy of clusters where each node is a cluster consisting of the clusters of its child nodes. Divisive: It is top-down, so you start with all observations in a large cluster and break it down into smaller pieces. Think about divisive as "dividing" the cluster.
2. Agglomerative: it is the opposite of divisive, so it is bottom-up, where each observation starts in its own cluster and pairs of clusters are merged together as they move up the hierarchy. Agglomeration means to amass or collect things, which is exactly what this does with the cluster. The Agglomerative approach is more popular among data scientists.

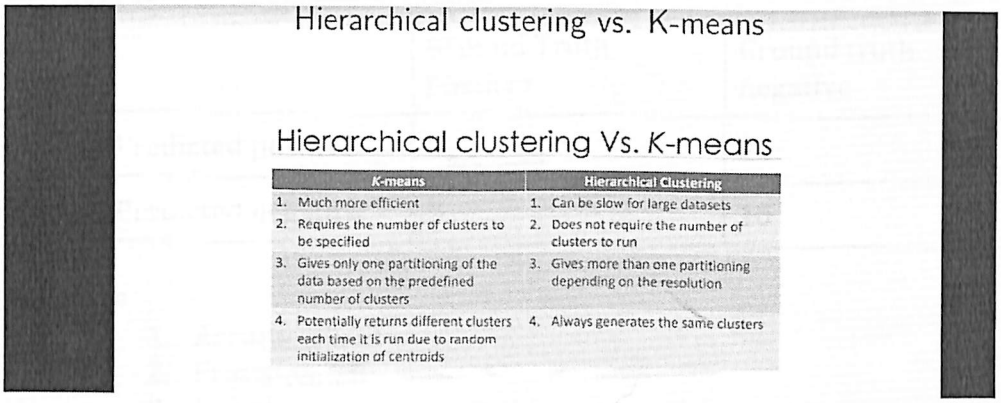
Divisive algorithm I

1. given a dataset (d1, d2, d3, ...dN) of size N at the top we have all data in one cluster
2. the cluster is split using a flat clustering method eg. K-Means etc
3. repeat
choose the best cluster among all the clusters to split split that cluster by the flat clustering algorithm untill each data is in its own singleton cluster

Divisive algorithm I

1. given a dataset (d1, d2, d3, ...dN) of size N at the top we have all data in one cluster
2. the cluster is split using a flat clustering method eg. K-Means etc
3. repeat

choose the best cluster among all the clusters to split split that cluster by the flat clustering algorithm untill each data is in its own singleton cluster



)-3 **Answer following questions (Attempt any TWO)** **10**

(A) Consider a clustering problem which has two clusters. Consider the data points with given initial membership values as following: **(5)**

Cluster	(2,3)	(5,7)	(4, 8)	(7, 9)
1)	0.3	0.7	0.4	0.1
2)	0.7	0.3	0.6	0.9

Calculate centroids for both the clusters using fuzzy c mean algorithm and find distance of each datapoint from both the centroids.

Answer:V11 = (0.82 *1 + 0.72 * 2 + 0.22 * 4 + 0.12 * 7) / ((0.82 + 0.72 + 0.22 + 0.12) = 1.568

V12 = (0.82 *3 + 0.72 * 5 + 0.22 * 8 + 0.12 * 9) / ((0.82 + 0.72 + 0.22 + 0.12) = 4.051

V11 = (0.22 *1 + 0.32 * 2 + 0.82 * 4 + 0.92 * 7) / ((0.22 + 0.32 + 0.82 + 0.92) = 5.35

V11 = (0.22 *3 + 0.32 * 5 + 0.82 * 8 + 0.92 * 9) / ((0.22 + 0.32 + 0.82 + 0.92) = 8.215

Centroids are: (1.568, 4.051) and (5.35, 8.215)

D11 = ((1 - 1.568)2 + (3 - 4.051)2)0.5 = 1.2

D12 = ((1 - 5.35)2 + (3 - 8.215)2)0.5 = 6.79

(B) Consider following confusion matrix **(5)**

	Ground Truth positive	Ground truth negative
Predicted positive	80	20
Predicted negative	5	10

Calculate

1. Accuracy:0.78
2. Precision:0.8
3. Recall:
4. F1 score:0.86
5. Write down your conclusion based on the value of above four.:

Answer:

1. Accuracy:0.78
2. Precision:0.8
3. Recall:0.94
4. F1 score:0.86
5. Write down your conclusion based on the value of above four.: Dataset is balanced

(C) Find clusters of data points (1,1), (2,1),(4,3),(5,4) with 2 clusters having (1,1) and (2,1) as initial centroid. Use k-means clustering algorithm to answer your question. Calculate for two iterations.

Answer:

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \\ \begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} & \begin{matrix} X \\ Y \end{matrix} \end{matrix}$$

$$\mathbf{G}^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\mathbf{c}_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3}\right) = \left(\tfrac{11}{3}, \tfrac{8}{3}\right)$$

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

$$\mathbf{G}^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{matrix} group-1 \\ group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

SECTION-II

Q-4 **Consider the scenario for Q1-Q5:** You are given the following information about a house: number of bedrooms, area(sqfeet), number of bathrooms, locality, ready possession or not. You are supposed to find the price of the house based on given details. Answer following questions for this scenario **10**

1 Which machine learning algorithms can be used to solve this problem? Justify your answer **(1)**

Answer: Regression models

2 How many neurons should be there in the input layer? **(1)**

Answer: 5

3 If we want to implement ANN for this problem, Which activation function should be used at the hidden layer? **(1)**

Answer: Sigmoid

4 If we want to implement ANN for this problem, Which activation function should be used at the output layer? **(1)**

Answer: Sigmoid

5 How many neurons should be there in the output layer? **(1)**
Answer:1

Consider the scenario for Q5-Q10: You are given the following information about a patient: Age, sex, weight, height, has alergy or not. You are supposed to find the suitable drug for the given patient. Answer following questions for this scenario.

6 Which machine learning algorithms can be used to solve this problem? Justify your answer **(1)**

Answer: Decision tree, KNN, ANN, Logistic regression

7 How many neurons should be there in the input layer? **(1)**

Answer:5

8 If we want to implement ANN for this problem, Which activation function should be used at the hidden layer? **(1)**

Answer: Sigmoid

9 If we want to implement ANN for this problem, Which activation function should be used at the output layer? **(1)**

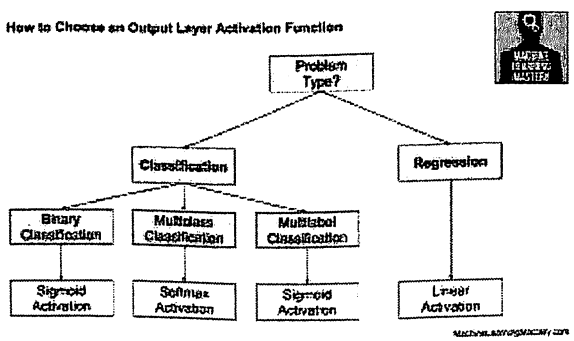
Answer: Sigmoid

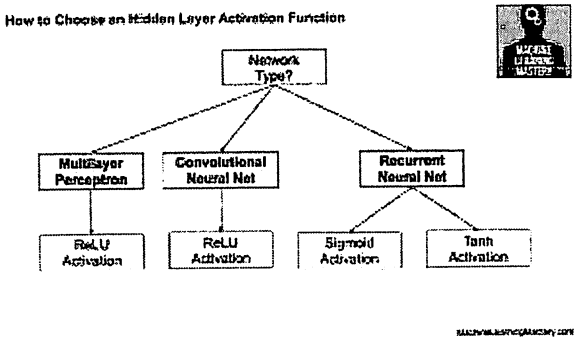
10 How many neurons should be there in the output layer? **(1)**

Answer: 1

Q-5 **(A)** List down the types of activation functions used in ANN. How can we decide which activation function should be used at the output layer based on the type of problem statement? **5** **10**

Answer: Sigmoid, Relu, Leakey relu, tanh, softmax





Q-5 (B) Consider following image matrix and filter matrix: **5**

image matrix:

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

Filter matrix:

1	-1	-1
-1	1	-1
-1	-1	1

Find output of applying convolutional layer(stride=1) with given filter and max pooling layer with window size 2×2.

Answer:

3	0
3	1

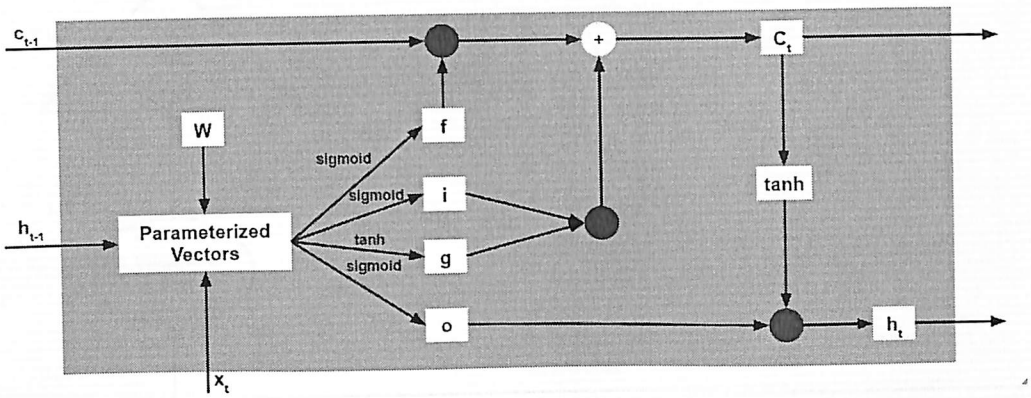
OR

Q-5 (B) "LSTM aims to provide a short-term memory for RNN that can last thousands of timesteps" Comment and discus its architecture in detail.: 5

Answer:
To solve the problem of Vanishing and Exploding Gradients in a Deep Recurrent Neural Network, many variations were developed. One of the most famous of them is the **Long Short Term Memory Network**(LSTM). In concept, an LSTM recurrent unit tries to "remember" all the past knowledge that the network is seen so far and to "forget" irrelevant data. This is done by introducing different activation function layers called "gates" for different purposes. Each LSTM recurrent unit also maintains a vector called the **Internal Cell State** which conceptually describes the information that was chosen to be retained by the previous LSTM recurrent unit. A Long Short Term Memory Network consists of four different gates for different purposes as described below:-

1. **Forget Gate(f):** It determines to what extent to forget the previous data.
2. **Input Gate(i):** It determines the extent of information be written onto the Internal Cell State.
3. **Input Modulation Gate(g):** It is often considered as a sub-part of the input gate and much literature on LSTM's does not even mention it and assume it is inside the Input gate. It is used to modulate the information that the Input gate will write onto the Internal State Cell by adding non-linearity to the information and making the information **Zero-mean**. This is done to reduce the learning time as Zero-mean input has faster convergence. Although this gate's actions are less important than the others and are often treated as a finesse-providing concept, it is good practice to include this gate in the structure of the LSTM unit.
4. **Output Gate(o):** It determines what output(next Hidden State) to generate from the current Internal Cell State.

The basic workflow of a Long Short Term Memory Network is similar to the workflow of a Recurrent Neural Network with the only difference being that the Internal Cell State is also passed forward along with the Hidden State



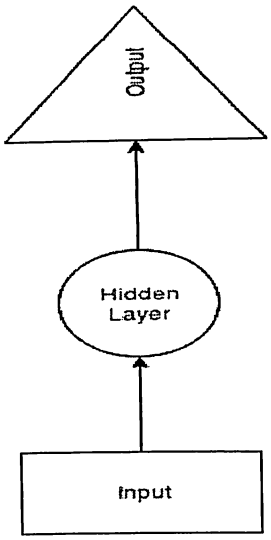
Q-6 **Answer following questions (Attempt any TWO)** **10**

(A) When we are using gradient descent as optimizer in ANN and sigmoid is used as activation function, which problem may occur. How we may solve that problem. **(5)**

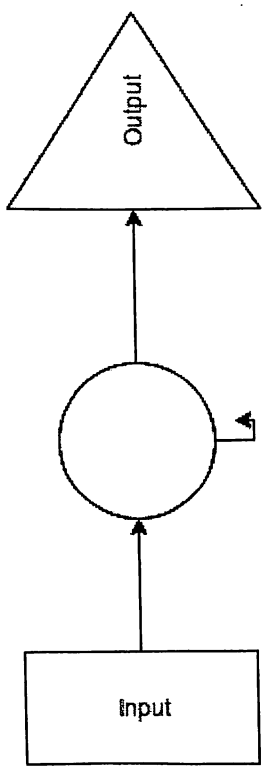
Answer: Exploding gradient descent occurs. It may be solved using other optimizers such as adagrade, adadelta

(B) Can you distinguish RNN from ANN and also tell how hidden layers are connected with each other in RNN and how the hidden state is updated and output is predicted at each neuron in the hidden layer. **(5)**

Answer:
RNN:

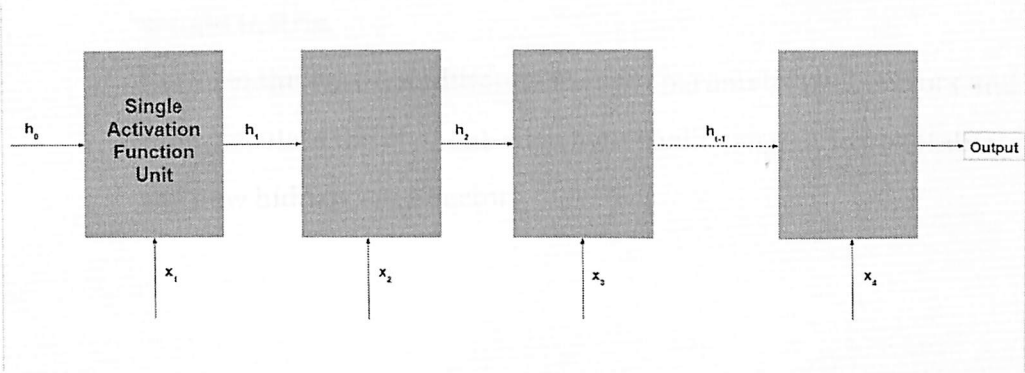


ANN:



Hidden state:
At each time step, the new hidden state is calculated using the recurrence relation as given above. This new generated hidden state is used to generate indeed a new hidden state and so on.

The basic work-flow of a Recurrent Neural Network is as follows:-



Note that is the initial hidden state of the network. Typically, it is a vector of zeros, but it can have other values also. One method is to encode the presumptions about the data into the initial hidden state of the network. For example, for a problem to determine the tone of a speech given by a renowned person, the person's past speeches' tones may be encoded into the initial hidden state. Another technique is to make the initial hidden state a trainable parameter. Although these techniques add little nuances to the network, initializing the hidden state vector to zeros is typically an effective choice.

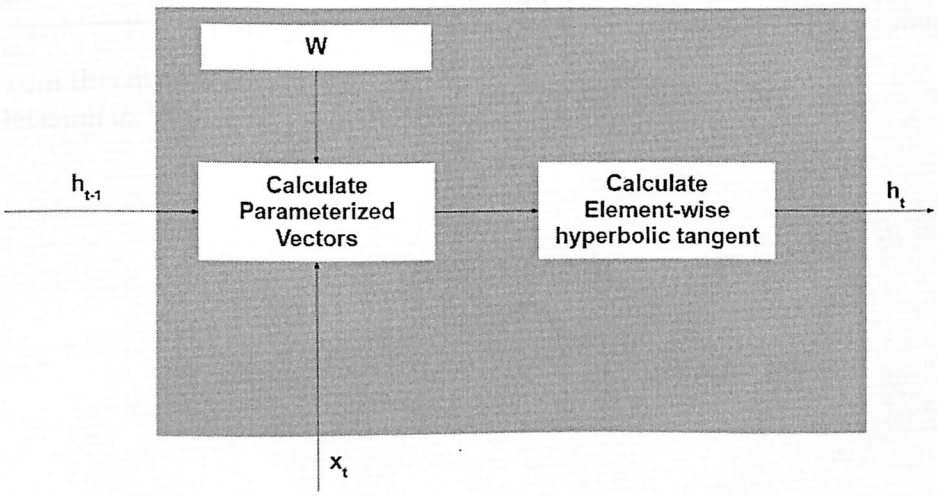
Working of each Recurrent Unit:

1. Take input the previously hidden state vector and the current input vector.

   Note that since the hidden state and current input are treated as vectors, each element in the vector is placed in a different dimension which is orthogonal to the other dimensions. Thus each element when multiplied by another element only gives a non-zero value when the elements involved are non-zero and the elements are in the same dimension.
2. Element-wise multiplies the hidden state vector by the hidden state weights and similarly performs the element-wise multiplication of the current input vector and the current input weights. This generates the parameterized hidden state vector and the current input vector.

Note that weights for different vectors are stored in the trainable weight matrix.

3. Perform the vector addition of the two parameterized vectors and then calculate the element-wise hyperbolic tangent to generate the new hidden state vector.



(C) Find answer for following dataset using KNN(k=3) **(5)**

ID	Height	Weight	Athlete
1	150	65	YES
2	160	70	YES
3	155	50	NO
4	142	65	YES
5	130	30	NO
6	145	60	?

Answer:

Distance from

Data point	Distance
1	7.01
2	18.02
3	14.14
4	5.83
5	33.54

From this distance:
Decision is: YES

END OF PAPER