Q1	Q2	Q3	Q4	Total

Student Name: ____________________ **Number:** ________________

Yıldız Technical University

BLM4800–Introduction to Data Mining

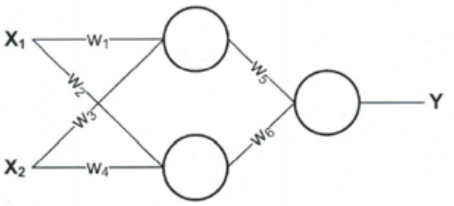
Final Exam - Spring 2022-2023

- **Duration:** 90 minutes
- **Exam information:**
    - Attempts to cheat in the exam will not be tolerated. If an attempt to cheat is discovered, it will be severely punished.
    - Read all the questions carefully before you start answering them.
    - The point value of each question is indicated next to the question.

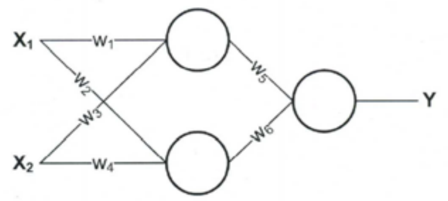
1. **[20 points]. The following two-layer Neural Network estimates the target variable Y using the weights of $w_2, \dots, w_6$ and activation functions on bivariate data such as *X=(X1, X2)*. There are two activation function options;**
    - **S:** sigmoid function $S(\alpha) = sign[\sigma(\alpha) - 0.5] = sign[\frac{1}{1+e^{(-\alpha)}} - 0.5]$
    - **L:** Lineer function $L(\alpha) = c.\alpha$

      In both cases $\alpha = \sum_i w_i X_i$

    a. [10 points]. To model a linear regression of this Artificial Neural Network as $\boldsymbol{Y = \beta_1 X_1 + \beta_2 X_2}$, write the appropriate activation functions in the blanks in the figure below and explain your reasoning (Write S or L inside the neurons).

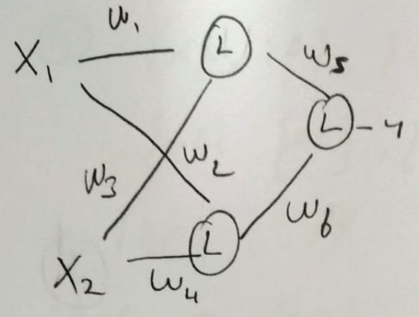


    b. [10 points]. To model this Artificial Neural Network $\boldsymbol{Y = arg\ max_y\ P(Y=y|X)}$ as a binary logistic regression $\boldsymbol{P(Y=1|X) = (\frac{1}{1+e^{(\beta 1X1 + \beta 2X2)}})}$, Write the appropriate activation functions in the gaps in the figure and explain your reasoning. (just write S or L inside the neurons).



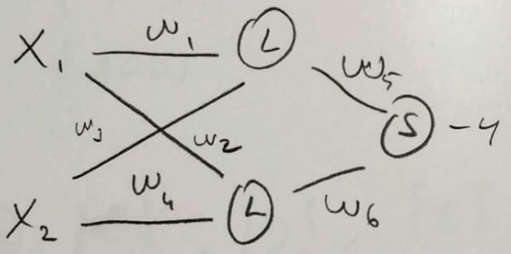
①

(a)



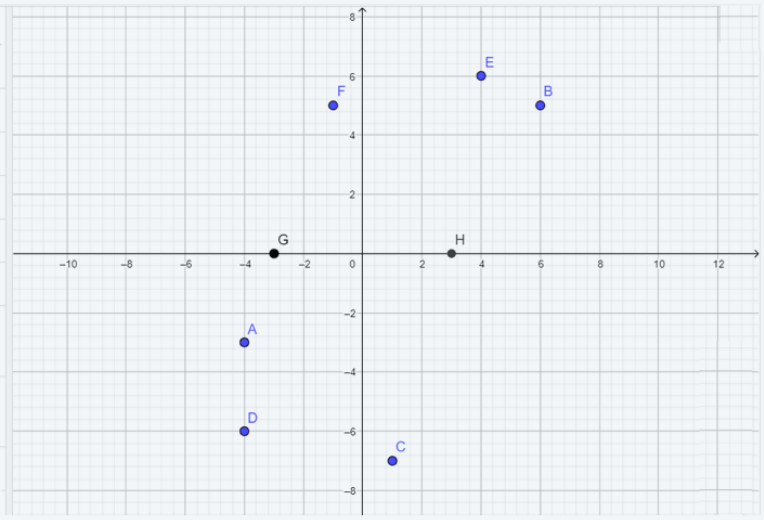
All of activation func in the model must be linear in order for the output of a NN to be linear

(b)



The provided information indicates that the desired outcome of the neural network is a binary output, achived through the use of the sigmoid activation func. The input of the last layer follows a linear func, implying that inner layers should also consist of entirely linear activatio

This suggests that network aims to linearly combine the input features to generate the desired binary outpu



Data	X	Y
A	-4	-3
B	6	5
C	1	-7
D	-4	-6
E	4	6
F	-1	5
G	-3	0
H	3	0

2. **[30 points]. Using the diagram above, answer the following questions in the context of cluster analysis.** Suppose you are given the G data point as the initial centroid of first cluster {cluster1: G point} and the H data point as the second cluster {cluster2: H point}. Simulate the K-Means algorithm for (#k=2) assuming it uses the Eucledian distance.
   a. What happens to cluster assignments after ONE iteration?
   b. What happens to cluster assignments after convergence? (Fill in the table below and show your calculations on the side)

Data	Cluster Assignment After First Iteration	Cluster Assignment After Second Iteration
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

②

(a-)

After first Loop

1. Cluster = A D F G
2. Cluster = B C E H

(b)

Second loop

$$1.\ \text{Centroid} = \left(\frac{(A_x + D_x + F_x + G_x)}{4}, \frac{(A_y + D_y + F_y + G_y}{4} \right)$$

$$= \frac{-4 -4 -1 -3}{4}, \frac{-3 -6 +5 +0}{4} = [-3, -1]$$

$$2.\ \text{Centroid} = \left(\frac{B_x + C_x + E_x + H_x}{4}, \frac{B_y + C_y + E_y + H_y}{4} \right)$$

$$= \left(\frac{6 + 1 + 3 + 4}{4}, \frac{(5 + 6 - 7)}{4} \right) = [3,5, 1]$$

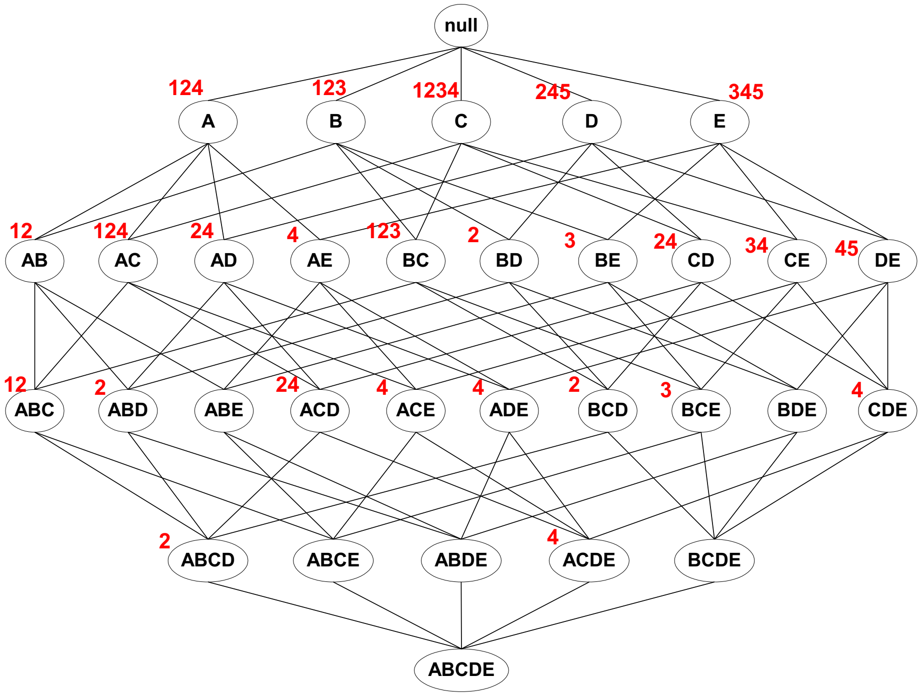
So

After Second Loop

1. Cluster = A C G D
2. Cluster = B F E H

Data	1.	2.
1	G	G
2	H	H
3	H	G
4	G	G
5	H	H
6	G	H
7	G	G
8	H	H

ID	Products
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



3. [26 points]. The diagram above shows the transaction number and the list of products sold within the transactions. Using this diagram, answer the following questions. Use the minimum support threshold value of 2 for these questions.

a. [6 points]. Using the diagram, write the list of all Frequent itemsets below.

b. [10 points]. *"An item set is a Closed Frequent itemset if none of its nearest supersets has the same support as the item set."* Using the diagram, write the Closed Frequent itemset list of all items below.

c. [10 points]. "*An item set is called a Maximal Frequent itemset if none of its immediate supersets are frequent.*" Using the diagram, write the Maximal Frequent itemset list of all items below.

③

ⓐ {A}, {B}, {C}, {D}, {E}

{AB}, {AC}, {AD}

{BC},

{CD}, {CE}

{DE}

{ABC}

{ACD}

ⓑ {C}, {D}, {E}

{AC}, {BC}, {CE}, {DE},

{ABC}, {ACD}

ⓒ {CE}, {DE}, {ABC}, {ACD}

Data	X	Y
1	126	78
2	128	80
3	128	82
4	130	82
5	130	84
6	132	86

4. **[24 points]. Using the data above, answer the following questions in the context of principal component analysis for data reduction. You want to reduce the data to a single size. Assume that the first Principal component is given as (0.59, 0.81). According to this;**
    a. [12 points]. What is the corresponding projection point on the First Principal component line for data point #3 (x=128, y=82)?

    b. [12 points]. Assume that the second Principal component is given as (-0.81, 0.59). What is the corresponding projection point on the second Principal component line for data point #5 (x=130, y=84)?

④ We have 2D (X,Y) and want to reduce to 1D

ⓐ by projecting it onto the 1. principal Component wich given (0.59, 0.81)

(x,y) — vector (u,v)

use formula dot product

1. PCA = dividing the PC vector by magnitute

Magnitute = $\sqrt{(0.59)^2 + (0.81)^2} \approx 1$

Product Dot Product = $(128 \times 0.59) + (82 \times 0,81)$

$= 75.52 + 66,42$

$= 141,94$

ⓑ Dot product = $(130 \times -0,81) + (84 \times 0,59)$

$= -105,3 + 49,56$

$= -55,74$