

Q1	Q2	Q3	Q4	Total

Student Name: _____

Number: _____

Yıldız Technical University

BLM4800–Introduction to Data Mining

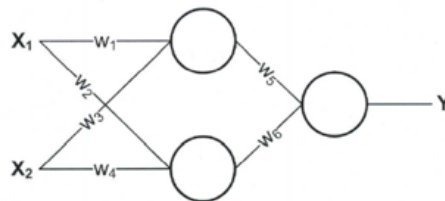
Final Exam - Spring 2022-2023

- **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 $X=(X_1, X_2)$. There are two activation function options;

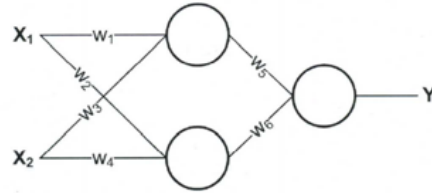
- **S:** sigmoid function $S(\alpha) = \text{sign}[\sigma(\alpha) - 0.5] = \text{sign}\left[\frac{1}{1+e^{(-\alpha)}} - 0.5\right]$
 - **L:** Linear function $L(\alpha) = c \cdot \alpha$
- In both cases $\alpha = \sum_i w_i X_i$

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



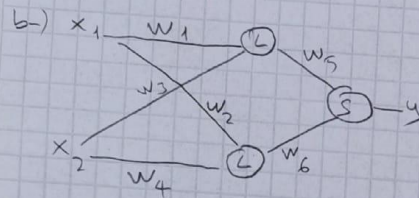
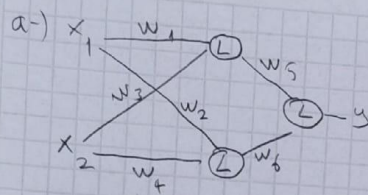
a-)In the field of neural networks, all activation functions inside the model must also be linear for the output to demonstrate linearity, a principle that denotes a straight proportionality between input and output. These activation functions are vital because they control how information moves through the network, basically changing a node's input signal into an output signal. As a result, the model's overall behavior is directly impacted by the linearity of these activation functions.

- b. [10 points]. To model this Artificial Neural Network $Y = \arg \max_y P(Y=y|X)$ as a binary logistic regression $P(Y=1|X) = \left(\frac{1}{1 + e^{(\beta_1 X_1 + \beta_2 X_2)}} \right)$, Write the appropriate activation functions in the gaps in the figure and explain your reasoning. (just write S or L inside the neurons).

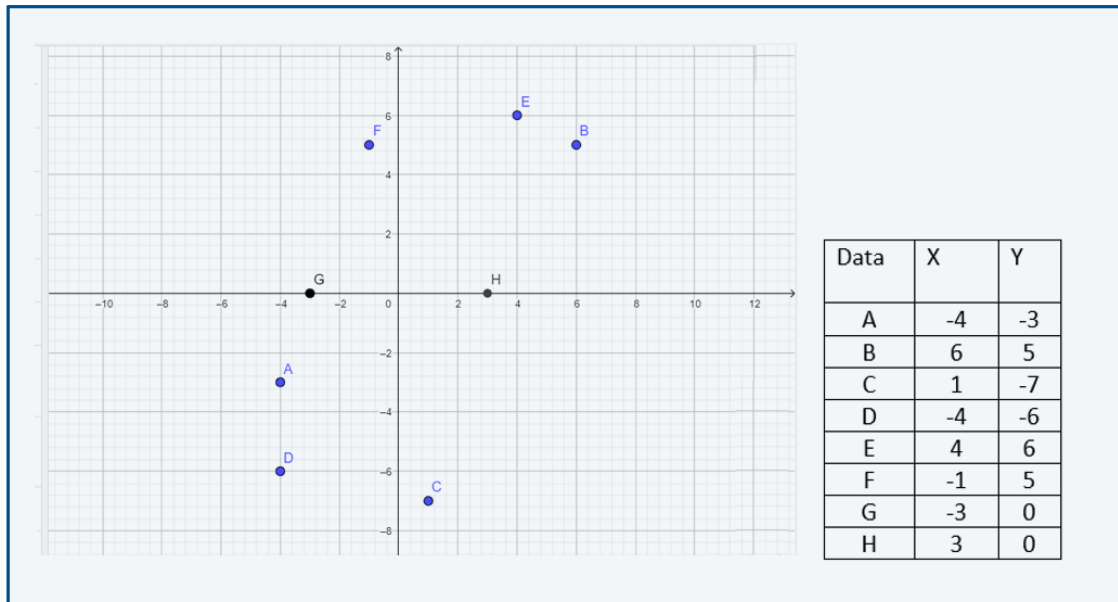


b-)The result strongly implies that a binary result is what is expected. This is made feasible by the neural network's integration of the sigmoid activation function (S(alpha)). Further, we may draw a parallel with the preceding explanation when we look at the input of the last layer, which follows a linear equation ($B_1 X_1 + B_2 X_2$). This suggests that the interior layers should have precisely linear activation functions (L) while designing them.

S!



M. Kaya Bulut 20011901



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 Euclidean distance.
- What happens to cluster assignments after ONE iteration?
 - 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		

S2

Old formula

a) First iteration completed

$$d(P, q) = d(q, P) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

cluster 1 \rightarrow (A, D, F, G)

cluster 2 \rightarrow (B, C, E, H)

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

b) Second iteration

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

$$\rightarrow ((-4 - 4 - 1 - 3)/4, (-3 - 6 + 5 + 0)/4) = (-3, -1)$$

$$\text{Centroid 2} = ((B_x + C_x + E_x + H_x)/4, (B_y + C_y + E_y + H_y)/4) \rightarrow$$

$$= ((6 + 1 + 4 + 3)/4, (5 - 7 + 6 + 9)/4) = (3.5, 1)$$

Second iteration completed

cluster 1 \rightarrow (A, C, D, G)

cluster 2 \rightarrow (B, E, F, H)

data	first iteration	second iteration
1	G	G
2	H	H
3	A	G
4	G	G
5	H	H
6	G	H
7	G	G
8	H	H

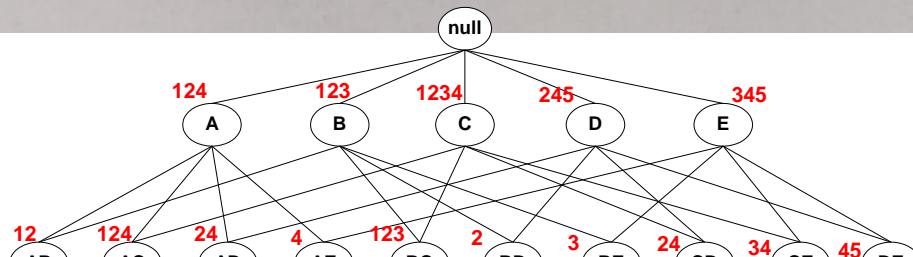
M. K. Agrawal

B. I. ut

20/11/2021

[Signature]

ID	Products
1	ABC
2	ABCD
3	BCE



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.

S3

a.

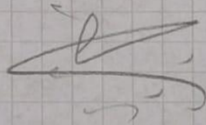
$\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{AB\}, \{AC\}, \{AD\}, \{BC\},$
 $\{CD\}, \{CE\}, \{DE\}, \{ABC\}, \{ACD\}$

b. $\{C\}, \{D\}, \{E\}, \{AC\}, \{BC\}, \{CE\}, \{DE\}, \{ABC\}, \{ACD\}$

c. $\{CE\}, \{DE\}, \{ABC\}, \{ACD\}$

M. Kaori Dkt

20211401



54

a. data point 3 \rightarrow (128, 82) first PC (0.59, 0.81)

$$\begin{array}{r} 128 \cdot 0.59 \\ \hline 75,52 \end{array} + \begin{array}{r} 82 \cdot 0.81 \\ \hline 66,42 \end{array} = 141,94$$

b. data point 5 \rightarrow (130, 84) second PC (-0.81, 0.59)

$$\begin{array}{r} 130 \cdot -0.81 \\ \hline -105,3 \end{array} + \begin{array}{r} 84 \cdot 0.59 \\ \hline 49,56 \end{array} = -55,74$$

M. Kargu Bulut 20011901

