Q1	Q2	Q3	Q4	Total

Student Name: ______________________ **Number: __________________**

Yıldız Technical University

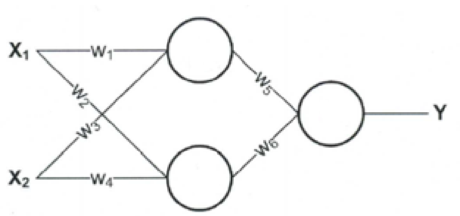
BLM4800–Introduction to Data Mining

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.

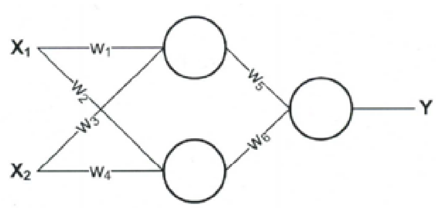
1. **[20 points]. The following two-layer Neural Network estimates the target variable Y using the weights of $w_2, \ldots, w_6$ and activation functions on bivariate data such as *X=(X1, X2)*. There are two activation function options;**
   - **S:** sigmoid function $S(\alpha) = sign[\sigma(\alpha) - 0.5] = sign[\frac{1}{1+e^{(-\alpha)}} - 0.5]$
   - **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 $\boldsymbol{Y} = \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2}$, 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.

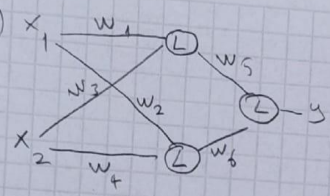
b. [10 points]. To model this Artificial Neural Network ***Y =arg max$_y$ P(Y=y|X)*** as a binary logistic regression ***P(Y=1|X)=*** $\left(\frac{1}{1+e^{(\beta 1X1\,+\,\beta 2X2\,)}}\right)$***,*** Write the appropriate activation functions in the gaps in the figure and explain your reasoning. (just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



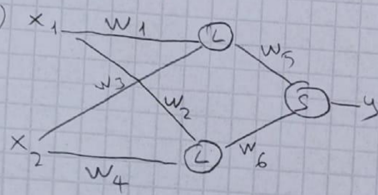
b-)The result strongly implies that a binary result is what is expected. This is made feasible by the neural network's integration of the sigmoid activation function (S(alpha)). Further, we may draw a parallel with the preceding explanation when we look at the input of the last layer, which follows a linear equation (B1 X1 + B2 X2). This suggests that the interior layers should have precisely linear activation functions (L) while designing them.

S1

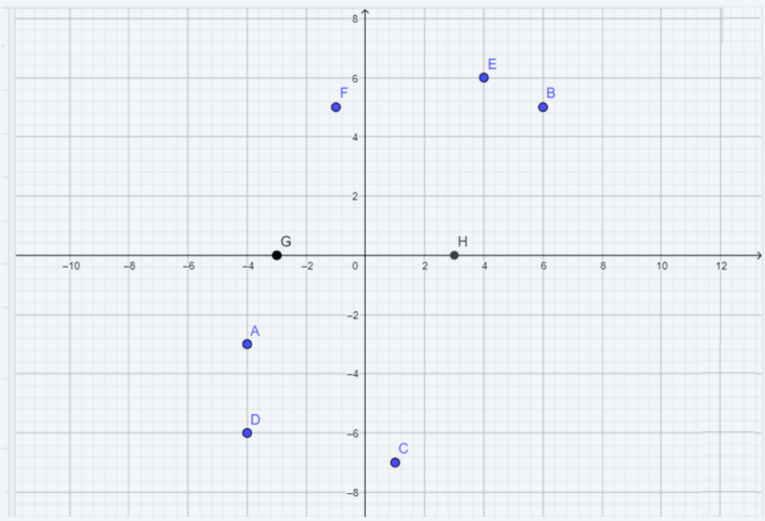
a-)



b-)



M. Kaya Bulut 20011401



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		

S2

Öklid formula

a) First iteration completed $d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 \; (q_n-p_n)^2}$

$\downarrow$

cluster 1 $\longrightarrow$ (A, D, F, G) cluster 2 $\rightarrow$ (B, C, E, H)

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

b) Second iteration

Centroid 1 = $((Ax+Dx+Fx+G_x)/4, (A_y+D_y+F_y+G_y)/4)$

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

Centroid 2 = $((Bx+Cx+Ex+Hx)/4, (B_y+C_y+E_y+H_y)/4) \rightarrow$

$= ((6+1+4+3)/4, (5-7+6+0)/4) = (3,5, 1)$

Second iteration completed

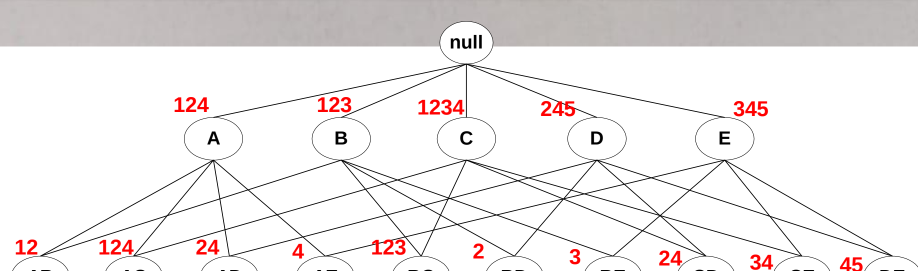
cluster 1 $\rightarrow$ (A, C, D, G) cluster 2 $\rightarrow$ (B, E, F, H)

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

M. Kagra Bulut

20011901

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.

We write all the elements whose number of bags is higher than 2

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. Kagan Bulut 22211401

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)?

   To reduce the dimensionality of the provided data, principle Component Analysis (PCA) allows one to transfer data points onto the principle components, which are effectively vectors. The axes that capture the most variety in the data are the major components.

   This particular situation involves projecting a 2D dataset (X, Y) onto the first main component, represented by the vector (0.59, 0.81), in order to reduce it to a 1D set.

   The mathematical procedure known as "dot product" can be used to project a point (x, y) onto a vector (u, v).

   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)?

   The projection of a data point onto the second principal component may also be determined using the dot product formula, much as the preceding approach.

   We start by determining the second main component's unit vector. We may accept the vector (-0.81, 0.59) as a unit vector since its magnitude is substantially close to 1 (rounding mistakes apart).

We may calculate the projection of this data point onto the second main component by calculating the dot product of the data point (130, 84) and the unit vector (-0.81, 0.59).

S4

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

$$\underbrace{128 * 0.59}_{75,52} + \underbrace{82 * 0.81}_{66,42} = 141,94$$

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

$$\underbrace{130 * -0.81}_{-105,3} + \underbrace{84 * 0.59}_{49,56} = -55,74$$

M. Kayra Bulut 20011901