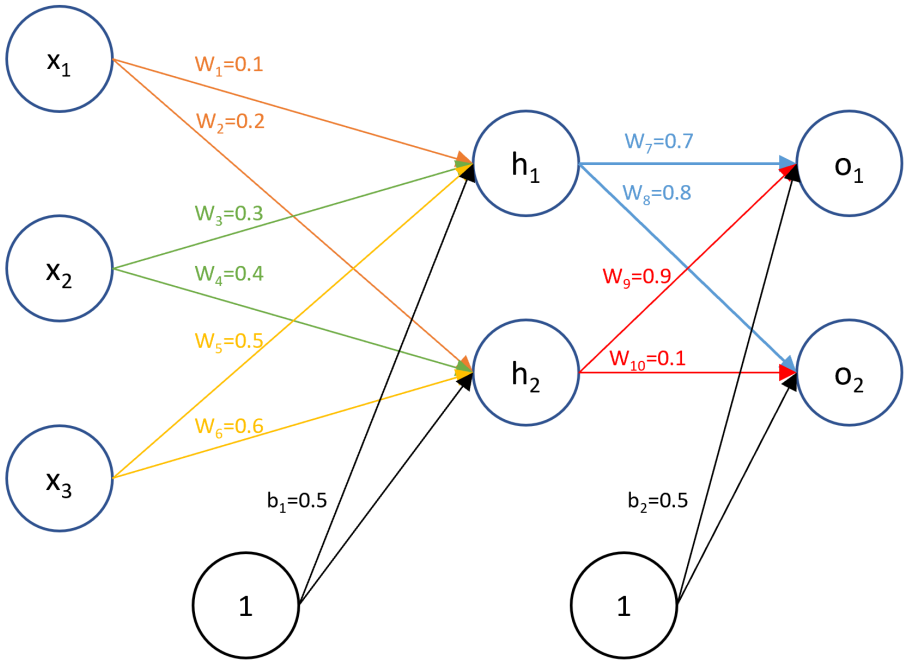
Candidates are required to answer in their own words as far as practicable.

1. Derive the gradient descent training rule assuming that the target function representation is: $O_d = w_0 + w_1 x_1 + w_2 x_2 + \ldots \ldots + w_n x_n$. Define explicitly the cost-function **E**, assuming that a set of training examples **D** is provided, where each training example **d**$\in$**D** is associated with the target output t_d. **[4 Marks]**
2. The following network is designed for binary classification with 3 input parameters; $x_1 = 1,\ x_2 = 4$ & $x_3 = 5$and target output as, $O_1 = 0.1$ & $O_2 = 0.05$. The network is trained using standard error back propagation, i.e. the square error should be minimized using gradient search. Now, forward Propagate through the network to get output values, backpropagate through the network to determine the error derivatives and update w_1. **[6 Marks]**



3. Why **weighted words** are important in Natural Language Processing. Describe **TF-IDF** with a suitable example. **[5 Marks]**
4. Draw a detailed computational graph of the sigmoid function, $y = 1/(1 + e^{-A^T x})$ where $A^T x = A_0 + A_1 x_1 + A_2 x_2$. **[4 Marks]**
5. From the given 3 by 3 confusion matrix, calculate the following:
   a) Accuracy & True Positive Rate, Positive Predictive Value
   b) F_1 Score & Support
   c) Micro-Average & Macro-Average
   d) Weighted average **[8 Marks]**

Confusion Matrix		Actual		
		A	**O**	**M**
Predicted	**A**	**7**	8	9
	O	1	**2**	3
	M	3	2	**1**

6. What is overfitting and why is it a problem?
   Give an example of a method to reduce the risk of overfitting.
   When is Ridge regression favorable over Lasso regression?
   Discuss. **[5 Marks]**

7. How is the best attribute selected in a decision tree? Select the root node from the sample data below:
[6 Marks]

Outlook	Temperature	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

8. Write short notes: **[4 * 3 = 12 Marks]**
    a. Vanishing gradient and exploding gradients
    b. Dimensionality reduction
    c. LSTM

9. Consider the logistic regression model $y = g(W^T x)$, trained using the binary cross entropy loss function where $g(z) = 1/(1 + e^{-z})$is the sigmoid function. Your friend proposes the modified logistic regression, using: $g(z) = e^{-z}/(1 + e^{-z})$. The model would still be trained using the binary cross entropy loss. How would the learnt model parameters, as well as the model predictions, differ from conventional logistic regression? Justify your answer mathematically. A purely graphical explanation is not sufficient. **[5 Marks]**

10. What does it mean to *pre-train* word embeddings and why is this useful? Can you think of an application where it may be useful to train "word embeddings," but the data is not text? Discuss. (That is, the embedded objects are not words.) **[5 Marks]**

"Think (design) globally; act (code) locally" → Anonymous .