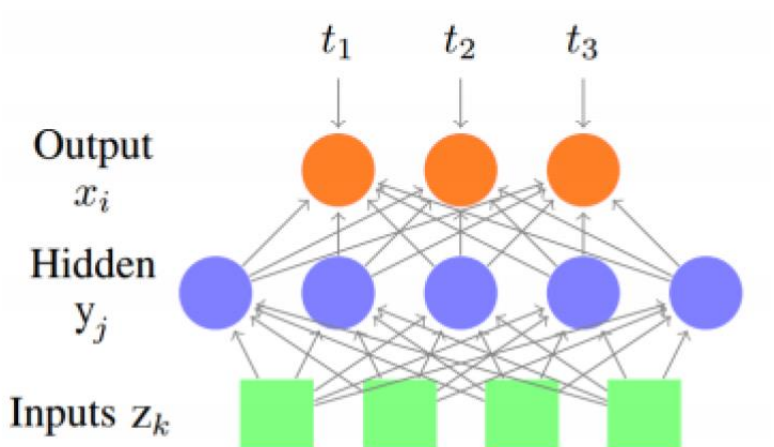
Problem 1 (15 points)

Implement the linear perceptron using stochastic gradient descent (SGD) or gradient descent (GD). Download the dataset "data3.mat". Use the whole data set as training, where each row consists of the feature vector x followed by the label $y \in \{-1, 1\}$ (last column). Show with figures the resulting linear decision boundary on the 2d x data. Show the evolution of binary classification error and the perceptron error with time (or number of iterations) from random initialization until convergence on a successful run (some random inits may not converge or may require many iterations). For GD, discuss the convergence behavior as you vary the step size (η).

Problem 2 (15 points)

Consider the following network, where x denotes output units, y denotes hidden units, and z denotes input units.



Consider:

a) [8 points] The cross-entropy error for a single example is:

$$E = -\sum_i (t_i \log(x_i) + (1 - t_i) \log(1 - x_i)),$$

where t is the target, and the logistic activation function for the output units is:

$$x_i = \frac{1}{1 + e^{-s_i}}, \quad \text{where} \quad s_i = \sum_j y_j w_{ji},$$

where w_{ji} denotes the weight of the edge between the j^{th} hidden unit and the i^{th} output unit. Assume hidden layers also use logistic activation function.

b) [7 points] The modified cross-entopy error for a single example is:

$$E = -\sum_i t_i \log(x_i)$$

and a softmax activation function for the output units is:

$$x_i = \frac{e^{s_i}}{\sum_{c=1}^{m} e^{s_c}},$$

where m is the number of outputs and the summation is taken over all outputs. Assume hidden layers use logistic activation function.

Derive the backpropagation updates in both cases (use the above error functions defined with respect to a single training example for your derivations rather than the sum of errors over the entire training dataset).

Problem 3 (10 points)

Consider the discrete distribution $\{p_k | k = 1, 2, \ldots, N\}$. The entropy of this distribution is given as $H = -\sum_{k=1}^{N} p_k \log p_k$. What is the distribution that maximizes this entropy? Show formal derivations using the method of Lagrange multipliers.

Problem 4 (10 points)

What is the VC dimension of axis-aligned squares? Justify your answer.