1.

We have a dataset whose points have two features for binary classification: $\left\{\mathbf{x}_1 = \begin{bmatrix}0\\0\end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix}0\\1\end{bmatrix}\right\}$, with $y = 1$ and $\left\{\mathbf{x}_3 = \begin{bmatrix}2\\0\end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix}2\\2\end{bmatrix}\right\}$ with $y = -1$. Using the initial weights $\mathbf{w}^T = [0\ 0]$ and initial bias $b = 0$, run the standard perceptron algorithm ($\alpha = 0.5$) until it converges and find the weights and bias that perfectly classify the training set. Always present the data to the perceptron in the following order: $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$.

2.

Design a Perceptron with two inputs X_1 and X_2 that outputs the logical expression $Y = X_1 \wedge X_2$, i.e., determine its weights and bias w_1, w_2, b. Use a binary step function instead of the bipolar step function. (20 pts)

Note: you do not need to run the Perceptron rule to solve this problem. Just use your knowledge of plane geometry.

Index	X_1	X_2	$Y = X_1 \wedge X_2$
1	0	0	0
2	1	0	0
3	0	1	0
3	1	1	1

3.

Consider the following training data for a classification problem with three features:

X_1	X_2	X_3	Class (+, -)
1	1	1	+
0	1	1	-
1	0	1	-
0	0	1	-
1	1	0	-
0	1	0	-
1	0	0	-
0	0	0	-

Note that the rows represent the features for each data point. For example, the first row is equivalent to having the column vector $\mathbf{x}(1) = [1\ 1\ 1]^T$ as the first training instance with positive class label.

(a) Assume that the weights of a Perceptron classifier are $\mathbf{w}^{(1)} = [1\ 1\ 1]^T$ and its bias is $b^{(1)} = -2$ and its activation function is the usual bipolar step function (i.e the sign function). Ties are broken in favor of the positive class. What is the training error of this Perceptron on this dataset.

(b) Next, feed the output of the Perceptron in the previous section (2a) to another Perceptron layer with the weight matrix $\mathbf{W}^{(2)} = [-1\ \ -1\ \ -2]$ and the bias vector $\mathbf{b}^{(2)} = [0\ 1\ 2]^T$. Assume that the activation function for all the neurons in this layer is the so-called rectified linear unit (ReLu), which is characterized as $f^{(2)}(n) = \text{ReLu}(n) = max(0, n)$. Determine the output of this two-layer network (with parameters $\mathbf{w}^{(1)}, b^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(2)}$ and step and ReLu activation functions in the first and second layer, respectively.) when the input feature vector is $\mathbf{x}^* = [1\ 2\ 0]^T$.

4.

Assume that we have a neural network that classifies input vectors with three features into three classes. The weight matrix of both the first and second layer is the identity matrix (a matrix whose diagonal elements are 1's and off-diagonal elements are 0's). The elements of the bias vector of the first layer are all zeros and the elements of the bias vector of the second layer are $-\ln 2$'s. The activation function of the first layer is ReLU and the activation function of the second layer is linear. The second layer is followed by a softmax layer.

(a) Determine to what class the input vectors $\mathbf{x}_1^T = [\ln 2 \;\; \ln 6 \;\; -5]$ and $\mathbf{x}_2^T = [\ln 8 \;\; -2 \;\; \ln 2]$ are classified by calculating the three posterior probabilities $p_k(\mathbf{x}_i), k \in \{1, 2, 3\}$ for each of $\mathbf{x}_1$ and $\mathbf{x}_2$.

(b) If the total variation distance (see below) is used as a measure of confidence in classification, in an active learning setting, which of $\mathbf{x}_1$ and $\mathbf{x}_2$ must be labeled first by the oracle using the Uncertainty Sampling query selection strategy?

A measure of *certainty* or *confidence* for multi-class classification is the total variation distance between the posterior probability distribution yielded by a classifier with input $\mathbf{x}_i$ and the uniform distribution, defined as:

$$d_{TV}(\mathbf{x}_i) = \frac{1}{2} \sum_{k=1}^{K} \left| p_k(\mathbf{x}_i) - \frac{1}{K} \right|$$

where $p_k(\mathbf{x}_i)$ is the posterior probability of $\mathbf{x}_i$ being in class $k \in \{1, 2, \ldots, K\}$.

5.

Consider a dataset with 3 points in 1-D:

Index	X	Y
1	0	+
2	-1	-
3	+1	-

(a) Carefully sketch these three training points. Are the classes linearly separable?

(b) Consider mapping each point to 2-D using new feature vectors $\boldsymbol{\varphi}(x) = [u_1(x), u_2(x)]^T$, in which $u_1(x)$ and $u_2(x)$ are polynomial functions of x. Find a $\boldsymbol{\varphi}(x)$ such that data are linearly separable in the new feature space.

(c) The maximum margin classifier in the new space has the equation $\mathbf{w}^T\boldsymbol{\varphi}(x)+b = 0$. Find $\mathbf{w}$ and b. Determine the decision boundary in the original 1-D space and the class assigned to $x = \frac{1}{3}$ by the classifier.

(d) Find a two layer feedforward neural network whose input-output equation is exactly like the above maximum margin classifier, by determining $\mathbf{W}^{(1)}, \mathbf{b}^{(1)}$ and $\mathbf{W}^{(2)}, \mathbf{b}^{(2)}$, and $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$.

Hint: in this neural network, elements of $\mathbf{f}^{(1)}$ *are different polynomial functions.*

6.

Consider the following MLP:

Assume that:

$$\mathbf{W}_1 = \begin{bmatrix} 1 & 0 \\ 2 & -1 \\ 1 & 0 \end{bmatrix}$$

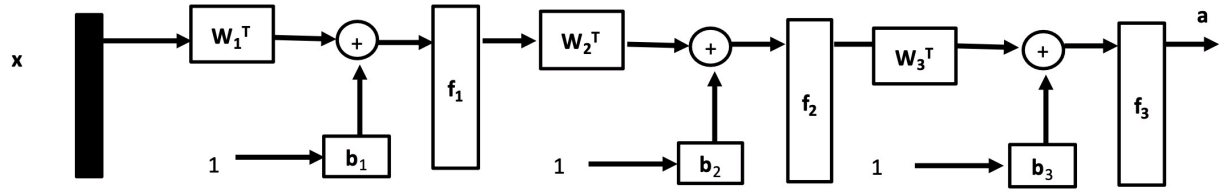
$$\mathbf{W}_2 = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix}$$

$$\mathbf{W}_3 = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

$$\mathbf{b}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \ \mathbf{b}_2 = \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \ \mathbf{b}_3 = \begin{bmatrix} 1 \end{bmatrix},$$

Also assume that the components of $\mathbf{f}_1$ are sign functions, the components of $\mathbf{f}_2$ are lines and the equation of each component of $\mathbf{f}_2$ is $f_2(n) = 0.2n$, and $f_3(n) = \tanh(2n)$. Answer the following questions: (30 pts)

(a) How many neurons are there in the first, second, and third layers, respectively?



(b) How many elements are there in the output $\mathbf{a}$ (i.e. what is the dimension of the vector $\mathbf{a}$)?

(c) Calculate the output of the network if the input is $\mathbf{x} = [1\ 0\ 1]^T$.