CSCI 184 HW3

Due Date: May 31, 11:59pm 2024

All assignments MUST have your name, student ID, course name/number at the beginning of your documents.
Your homework MUST be submitted via Camino with the file format and name convention as follows:
For Question Answering part, you can either write by hand or type your answers, but please ensure your submission is **pdf** file with name "**HW#_Name.pdf**".
For programming questions, please upload your code and supporting files in "**HW#_Name.zip**".

If you have any questions, please don't hesitate to contact me :)

Q1) Perceptron Trees: To exploit the desirable properties of decision tree classifiers and perceptrons, Adam came up with a new algorithm called "perceptron trees", which combines features from both. Perceptron trees are similar to decision trees, however each leaf node is a perceptron, instead of a majority vote.
To create a perceptron tree, the first step is to follow a regular decision tree learning algorithm (such as ID3) and perform splitting on attributes until the specified maximum depth is reached. Once maximum depth has been reached, at each leaf node, a perceptron is trained on the remaining attributes which have not been used up in that branch. Classification of a new example is done via a similar procedure. The example is first passed through the decision tree based on its attribute values. When it reaches a leaf node, the final prediction is made by running the corresponding perceptron at that node.
Assume that you have a dataset with 6 binary attributes (A, B, C, D, E, F) and two output labels (-1 and 1). A perceptron tree of depth 2 on this dataset is given below. Weights of the perceptron are given in the leaf nodes. Assume bias=1 for each perceptron:

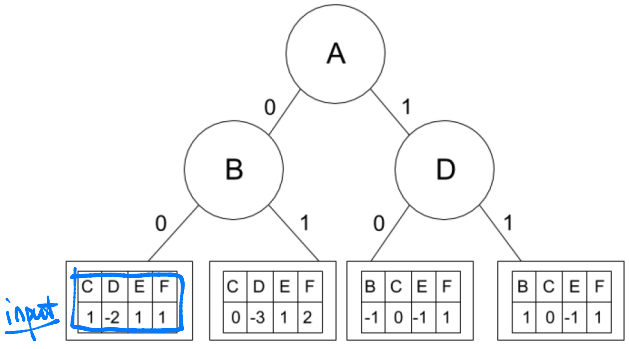


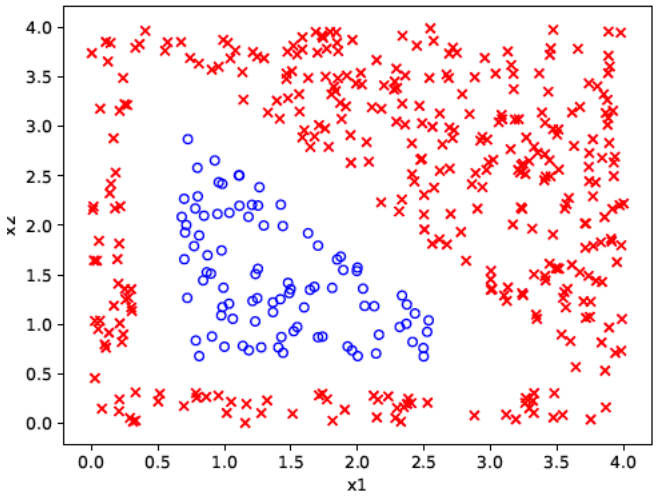
Figure 1: Perceptron Tree of max depth=2

Predict the output labels for the following two samples:

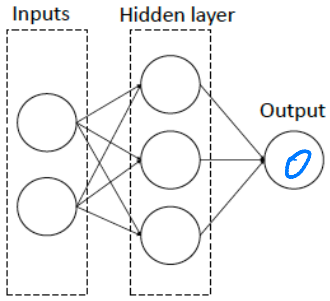
a) $x = (1, 1, 0, 1, 0, 1)$
b) $x = (0, 1, 0, 1, 0, 1)$

Q2)

Let $X = \{x^{(1)}, \cdots, x^{(m)}\}$ be a dataset of m samples with 2 features, i.e $x^{(i)} \in \mathbb{R}^2$. The samples are classified into 2 categories with labels $y^{(i)} \in \{0, 1\}$. A scatter plot of the dataset is shown in the following figure:



The examples in class 1 are marked as "×" and examples in class 0 are marked as "o". We want to perform binary classification using a simple neural network with the architecture shown in the following figure:



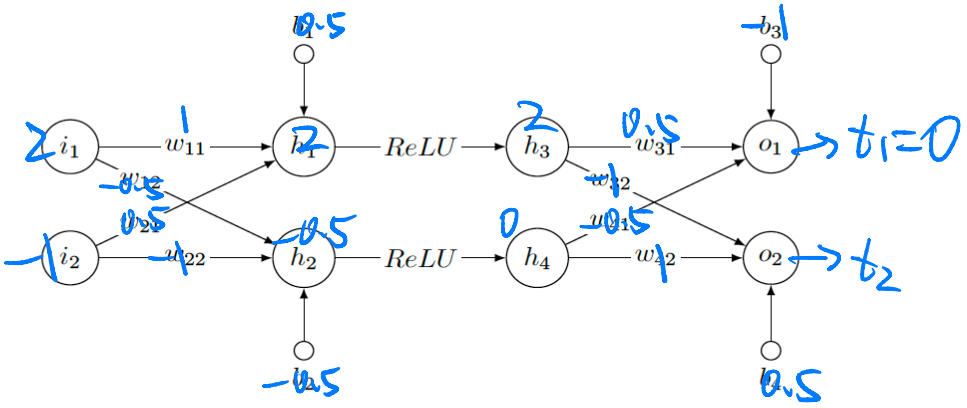
Denote the two features x_1 and x_2, the three neurons in the hidden layer h_1, h_2, and h_3, and the output neuron as o. Let the weight from x_i to h_j be $w_{i,j}^{[1]}$ for $i \in \{1, 2\}, j \in \{1, 2, 3\}$, and the weight from h_j to o be $w_j^{[2]}$. Finally, denote the intercept weight for h_j as $w_{0,j}^{[1]}$, and the intercept weight for o as $w_0^{[2]}$. For the loss function, we'll use average squared loss instead of the usual negative log-likelihood:

$$l = \frac{1}{m} \sum_{i=1}^{m} (o^{(i)} - y^{(i)})^2,$$

where $o^{(i)}$ is the result of the output neuron for example i.

Suppose we use the sigmoid function as the activation function for h_1, h_2, h_3, and o. What is the gradient descent update to $w_{1,2}^{[1]}$, assuming we use a learning rate of α? Your answer should be written in terms of $x^{(i)}, o^{(i)}, y^{(i)}$, and the weights.

Q3) Given the following neural network with fully connection layer and ReLU activations, including two input units (i_1, i_2), four hidden units (h_1, h_2) and (h_3, h_4). The output units are indicated as (o_1, o_2) and their targets are indicated as (t_1, t_2). The weights and bias of fully connected layer are called w and b with specific sub-descriptors.



The values of variables are given in the following table:

Variable	i_1	i_2	w_{11}	w_{12}	w_{21}	w_{22}	w_{31}	w_{32}	w_{41}	w_{42}	b_1	b_2	b_3	b_4	t_1	t_2
Value	2.0	-1.0	1.0	-0.5	0.5	-1.0	0.5	-1.0	-0.5	1.0	0.5	-0.5	-1.0	0.5	1.0	0.5

a) Compute the output (o_1, o_2) with the input (i_1, i_2) and network parameters as specified above. Write down all calculations, including intermediate layer results h_1, h_2, h_3, h_4.

b) Compute the mean squared error of the output (o_1, o_2) calculated above and the target (t_1, t_2).

c) Update the weight w_{21} using gradient descent with learning rate 0.1.

Xiaoming Huang
1604905

CSCI 184

HW3, Part 1

1. Using Sigmoid function

a) $x = (1, 1, 0, 1, 0, 1)$

$A \xrightarrow{1} D \xrightarrow{1}$	B	C	E	F
1	0	-1	1

$\Rightarrow z = 1 \cdot 1 + 0 + 0 + 1 \cdot 1 + 1$

$= 3$

$\Rightarrow g(z) = \frac{1}{1 + e^{-3}} = \underline{0.9526}$

$\Rightarrow \boxed{\hat{y} = 1}$

b) $x = (0, 1, 0, 1, 0, 1)$

$A \xrightarrow{0} B \xrightarrow{1}$ | C | D | E | F |
|---|---|---|---|
| 0 | -3 | 1 | 2 |

$\Rightarrow z = 0 + 1 \cdot (-3) + 0 + 1 \cdot 2 + 1$

$= 0$

$\Rightarrow g(z) = \frac{1}{2}$

$\Rightarrow \boxed{\hat{y} = 0}$,

since we have equal chances for both.

2. given $w_{12}^{[1]}$, Note for Sigmoid function, $\sigma'(z) = \sigma(z)(1-\sigma(z))$

As $h_2^{(i)} = \sigma\left(w_{02}^{[1]} + \sum_{j=1}^{2} w_{j2}^{[1]} x_j^{(i)}\right)$

$$\Rightarrow \frac{\partial h_2}{\partial w_{12}^{[1]}} = h_2^{(i)} \cdot (1-h_2^{(i)}) \cdot x_2^{(i)}$$

As $o^{(i)} = \sigma\left(w_0^{[2]} + \sum_{j=1}^{3} w_j^{[2]} h_j^{(i)}\right)$

$$\Rightarrow \frac{\partial o^{(i)}}{\partial h_2^{(i)}} = o^{(i)} \cdot (1-o^{(i)}) \cdot w_2^{[2]}$$

As $l = \frac{1}{m} \sum_{i=1}^{m} \left(o^{(i)} - y^{(i)}\right)^2$

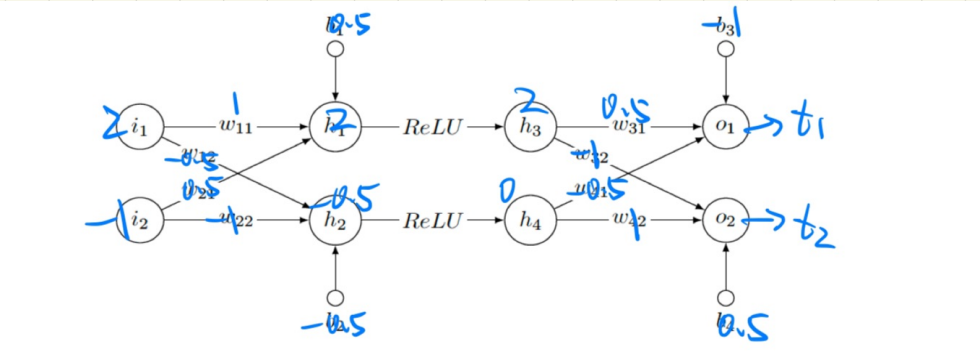
$$\Rightarrow \frac{\partial l}{\partial o} = \frac{1}{m} \sum_{i=1}^{m} 2\left(o^{(i)} - y^{(i)}\right)$$

the update for $w_{12}^{[1]}$ is

$$\Rightarrow w_{12}^{[1]} = w_{12}^{[1]} - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left[2\left(o^{(i)} - y^{(i)}\right) \cdot o^{(i)} \cdot (1-o^{(i)}) \cdot w_2^{[2]} \cdot h_2^{(i)} \cdot (1-h_2^{(i)}) \cdot x_2^{(i)} \right] \right]$$

3.

If assuming no activation function is used except for the 2 ReLU



a) $i_1 = 2$

$i_2 = -1$

$\Rightarrow h_1 = w_{11} \cdot i_1 + w_{21} \cdot i_2 + b_1$

$= (1)(2) + (0.5)(-1) + 0.5$

$= \boxed{2}$

$\Rightarrow h_2 = w_{12} \cdot i_1 + w_{22} \cdot i_2 + b_2$

$= (-0.5)(2) + (-1)(-1) - 0.5$

$= \boxed{-0.5}$

$\Rightarrow h_3 = \max(0, 2) = 2$

$\Rightarrow h_4 = \max(0, -0.5) =$

$\Rightarrow o_1 = w_{31} \cdot h_3 + w_{41} \cdot h_4 + b_3$

$= (0.5)(2) + (-0.5)(0) - 1$

$= \boxed{0}$

$\Rightarrow o_2 = w_{32} \cdot h_3 + w_{42} \cdot h_4 + b_4$

$= (-1)(2) + (1)(0) + 0.5$

$= \boxed{-1.5}$

so $(o_1, o_2) = (0, -1.5)$.

b) $MSE = \frac{1}{2}\left[(0-1)^2+(-1.5-0.5)^2\right] = \boxed{\frac{5}{2}}$

If consider o_1, o_2 as two seperate component,
we can apply regular scalar MSE, and take $n=2$.

Also if consider $\vec{o}$ and $\vec{t}$ as vector and take $\|\vec{o}-\vec{t}\|^2$,
since there are 2 component, It's also reasonable to average by 2.

c).

$$w_{21} = w_{21} - 0.1\left[(0-1)\cdot w_{31}\cdot 1\cdot(-1) + (-2)\cdot w_{32}\cdot 1\cdot(-1)\right]$$

$$= 0.5 - 0.1\left[(-1)\cdot 0.5\cdot 1\cdot(-1) + (-2)(-1)\cdot 1(-1)\right]$$

$$= \boxed{0.65}$$