CS124: Deep Learning - IIT Ropar

Nihar Ranjan Sahoo
Prime Minister Research Fellow, IIT Bombay

Instructors:
Prof. Mitesh M. Khapra, IIT Madras
Prof. Sudarshan Iyengar, IIT Ropar

NPTEL Problem Solving Session
Week-2
Date: 06/08/2022

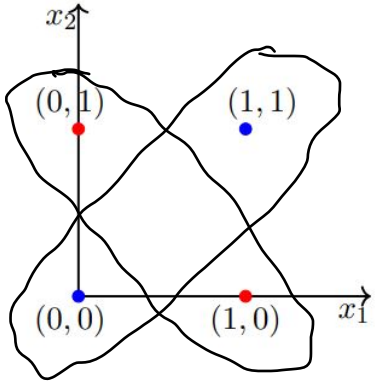
Outline:

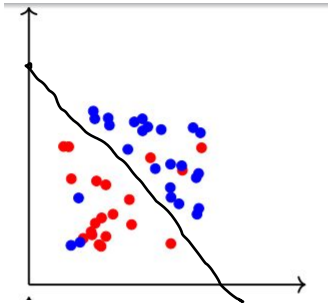
- Linearly Separable Boolean Functions
- Network of Perceptrons
- Sigmoid Neuron
- Supervised Machine Learning Setup
- Learning Parameters: Gradient descent
- Problem Solving

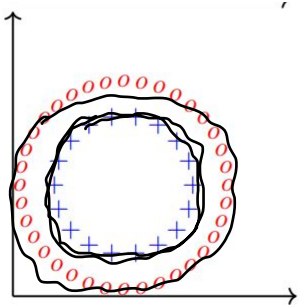
Linearly Separable Boolean Functions

x_1	x_2	XOR	
0	0	0	$w_0 + \sum_{i=1}^{2} w_i x_i < 0$
1	0	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
0	1	1	$w_0 + \sum_{i=1}^{2} w_i x_i \geq 0$
1	1	0	$w_0 + \sum_{i=1}^{2} w_i x_i < 0$

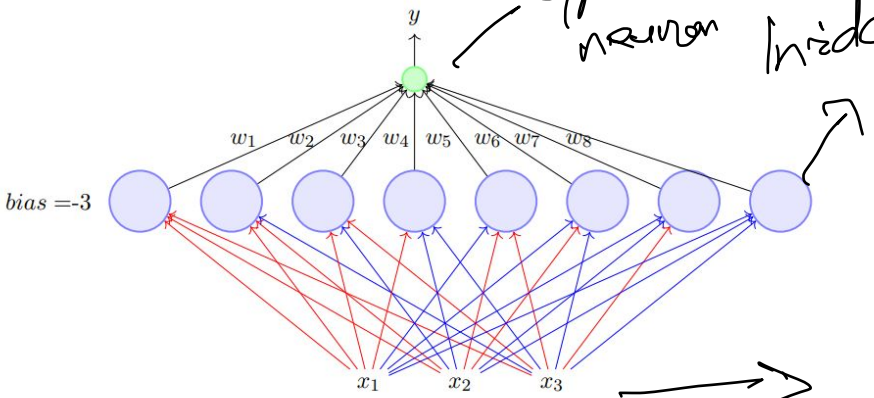
$$w_0 + w_1 \cdot 0 + w_2 \cdot 0 < 0 \implies w_0 < 0$$
$$w_0 + w_1 \cdot 0 + w_2 \cdot 1 \geq 0 \implies w_2 \geq -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 0 \geq 0 \implies w_1 \geq -w_0$$
$$w_0 + w_1 \cdot 1 + w_2 \cdot 1 < 0 \implies w_1 + w_2 < -w_0$$







Network of Perceptrons



Theorem

Any boolean function of n inputs can be represented exactly by a network of perceptrons containing 1 hidden layer with 2^n perceptrons and one output layer containing 1 perceptron

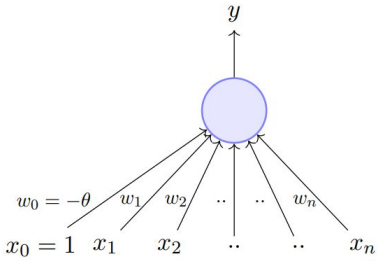
Proof (informal:) We just saw how to construct such a network

Note: A network of $2^n + 1$ perceptrons is not necessary but sufficient. For example, we already saw how to represent AND function with just 1 perceptron

Catch: As n increases the number of perceptrons in the hidden layers obviously increases exponentially

Sigmoid Neuron

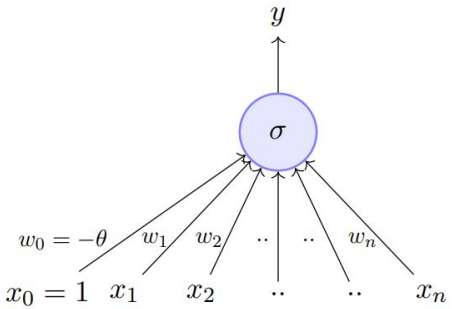
Perceptron



$$y = 1 \quad if \sum_{i=0}^{n} w_i * x_i \geq 0$$

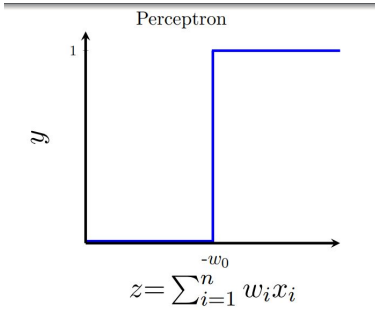
$$= 0 \quad if \sum_{i=0}^{n} w_i * x_i < 0$$

Sigmoid (logistic) Neuron

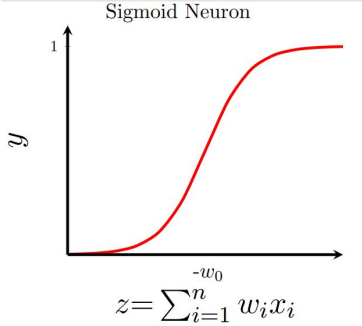


$$y = \frac{1}{1 + e^{-(\sum_{i=0}^{n} w_i x_i)}}$$

$\frac{1}{1+e^{-w^T x}}$



Not smooth, not continuous (at $w0$), **not differentiable**



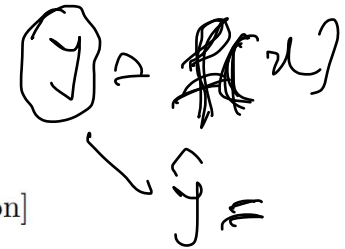
Smooth, continuous, **differentiable**

Supervised Machine Learning Setup

As an illustration, consider our movie example

- **Data:** $\{x_i = movie, y_i = like/dislike\}_{i=1}^{n}$
- **Model:** Our approximation of the relation between $\mathbf{x}$ and y (the probability of liking a movie).

$$\hat{y} = \frac{1}{1 + e^{-(\mathbf{w^T x})}}$$

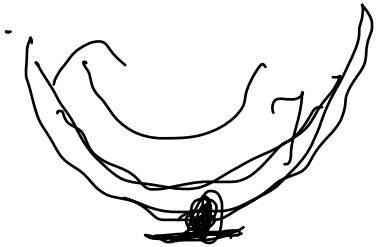


- **Parameter: w**
- **Learning algorithm:** Gradient Descent [we will see soon]
- **Objective/Loss/Error function:** One possibility is

$$\mathscr{L}(\mathbf{w}) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

The learning algorithm should aim to find a w which minimizes the above function (squared error between y and $\hat{y}$)

Gradient Descent



Gradient Descent Rule

- The direction u that we intend to move in should be at 180° w.r.t. the gradient
- In other words, move in a direction opposite to the gradient

Parameter Update Equations

$$w_{t+1} = w_t - \eta \nabla w_t$$
$$b_{t+1} = b_t - \eta \nabla b_t$$

$$where, \nabla w_t = \frac{\partial \mathscr{L}(w,b)}{\partial w}_{at\ w=w_t, b=b_t}, \nabla b = \frac{\partial \mathscr{L}(w,b)}{\partial b}_{at\ w=w_t, b=b_t}$$

Algorithm: gradient_descent()

$t \leftarrow 0$;
$max\_iterations \leftarrow 1000$;
while $t < max\_iterations$ **do**
  $w_{t+1} \leftarrow w_t - \eta \nabla w_t$;
  $b_{t+1} \leftarrow b_t - \eta \nabla b_t$;
  $t \leftarrow t + 1$;
end

Previous Iteration Problems

1) Squared error function is preferred over simple difference between the actual and predicted value of output. Identify the appropriate reason for this statement.

- square blows up the error
- In the second case, the positive and negative errors cancel out each other
- It is not differentiable
- The statement cannot be justified

2) Pick out the parameter that enables the model to learn from the given data for the given function:

$$\hat{y} = \frac{1}{1 + e^{-}(w^T x)}$$

- w
- x
- Both x and w
- None of the above

Previous Iteration Problems

3) Identify the function that cannot be used in a Machine learning model?

- $\hat{y} = w^T x$ ✓
- $\hat{y} = x^T W x$
- $\hat{y} = \frac{1}{1+e^{-x}}$ ✓ this can't be used
- $\hat{y} = \frac{1}{1+e^{wTx}}$

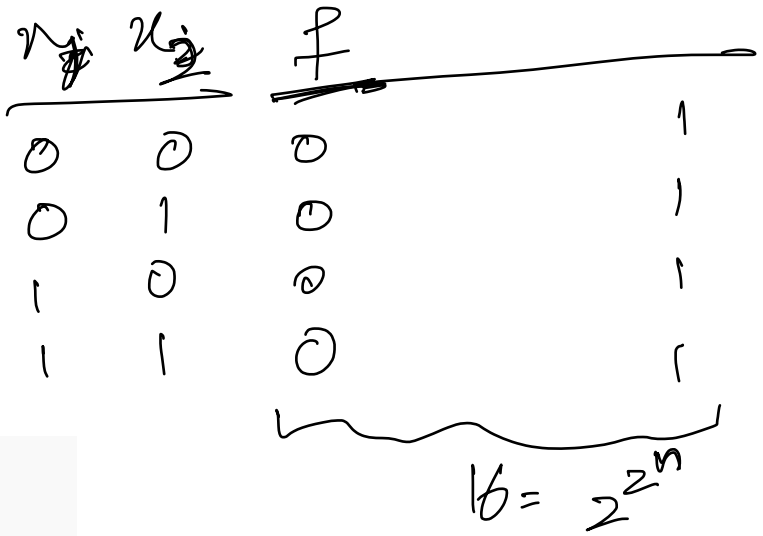
4) The condition for the new loss to be less than the current loss is $u^T \mathscr{L}(\theta) < 0$. What is the direction of u with respect t to the Gradient for which the decrease of loss is maximum?

- 45°
- 0°
- 90°
- 180° ✓

Previous Iteration Problems

5) Given that you have n inputs, how many boolean functions can be designed?

- 2^{2^n}
- 2^{n^2}
- $2^2 n$
- n^{2^2}



6) Which of the following statements is True for a Multilayer Perceptron model?

Statement I: Any boolean function can be represented by a single hidden layer

Statement II: The number of perceptrons for this hidden layer is n^2 where n is the number of classes

- Only I
- Only II
- Both I and II
- None

Previous Iteration Problems

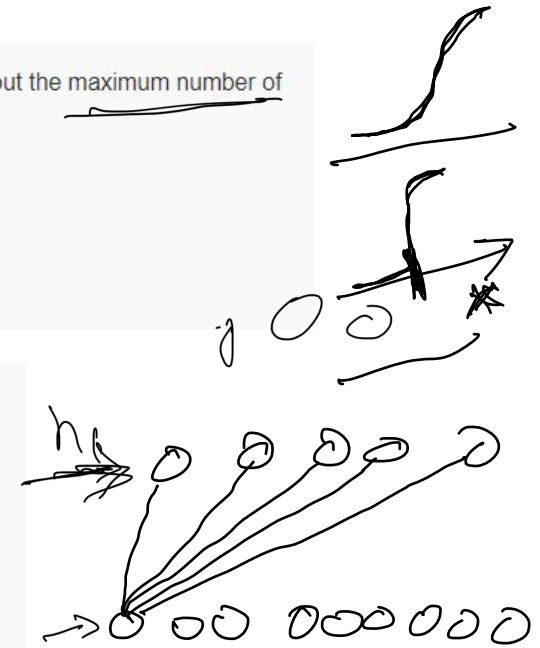
7) Suppose the number of neurons in the input layer is 9 and the consecutive hidden layer is 5 (In MLP). Pick out the maximum number of connections that can exist from the input layer to the hidden layer

- 14
- 45
- 18
- 10

8) Which of the following results when the value of b is changed in a sigmoid function given below?

$$\sigma(x) = \frac{1}{1 + e^{wx+b}}$$

- value of x at which transition occurs changes
- value of w changes
- the slope becomes steeper
- the transition point is found



Previous Iteration Problems

9) Consider a single perceptron that that takes a single input, whether the student has passed in the internal assessment. The perceptron predicts if the student can pass in the end semester exam, Now, if the bias is -0.5 and $w1$ is 1, what will be the prediction if the output value is 0.51 and 0.49?

- 1,0
- 0,1
- 1,1
- 0,0

10) Identify the statements that are True about learning algorithm.
Statement I. Maximizes objective function
Statement II. Learns the parameters from data

- I only
- II only
- I and II
- none