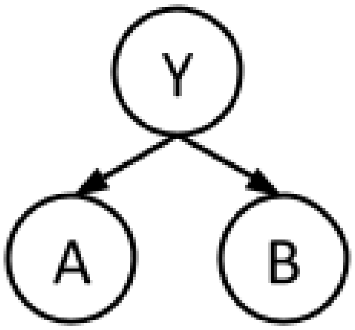
1 Naive Bayes

In this question, we will train a Naive Bayes classifier to predict class labels Y as a function of input features A and B. Y, A, and B are all binary variables, with domains 0 and 1. We are given 10 training points from which we will estimate our distribution.

A	1	1	1	1	0	1	0	1	1	1
B	1	0	0	1	1	1	1	0	1	1
Y	1	1	0	0	0	1	1	0	0	0



1. What are the maximum likelihood estimates for the tables $P(Y)$, $P(A|Y)$, and $P(B|Y)$?

Y	$P(Y)$
0	3/5
1	2/5

A	Y	$P(A\|Y)$
0	0	1/6
1	0	5/6
0	1	1/4
1	1	3/4

B	Y	$P(B\|Y)$
0	0	1/3
1	0	2/3
0	1	1/4
1	1	3/4

2. Consider a new data point $(A = 1,\ B = 1)$. What label would this classifier assign to this sample?

$$P(Y=0, A=1, B=1) = P(Y=0)\,P(A=1|Y=0)\,P(B=1|Y=0)$$
$$= (3/5)(5/6)(2/3)$$
$$= 1/3$$

$$P(Y=1, A=1, B=1) = 9/40$$

So the label is 0

3. Let's use Laplace Smoothing to smooth out our distribution. Compute the new distribution for $P(A|Y)$ given Laplace Smoothing with $k = 2$.

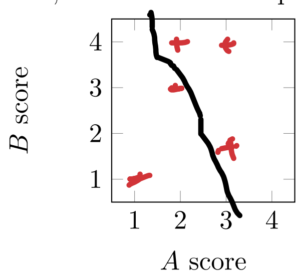
$$\frac{c(x,y)+k}{c(y)+k\,|x|}$$

A	Y	$P(A\|Y)$
0	0	3/10
1	0	7/10
0	1	3/8
1	1	5/8

2 Perceptron

You want to predict if movies will be profitable based on their screenplays. You hire two critics A and B to read a script you have and rate it on a scale of 1 to 4. The critics are not perfect; here are five data points including the critics' scores and the performance of the movie:

#	Movie Name	A	B	Profit?
1	Pellet Power	1	1	-
2	Ghosts!	3	2	+
3	Pac is Bac	2	4	+
4	Not a Pizza	3	4	+
5	Endless Maze	2	3	-



1. First, you would like to examine the linear separability of the data. Plot the data on the 2D plane above; label profitable movies with + and non-profitable movies with − and determine if the data are linearly separable. it is separable

2. Now you decide to use a perceptron to classify your data. Suppose you directly use the scores given above as features, together with a bias feature. That is $f_0 = 1$, $f_1 =$ score given by A and $f_2 =$ score given by B.

   Run one pass through the data with the perceptron algorithm, filling out the table below. Go through the data points in order, e.g. using data point #1 at step 1.

step	Weights	Score	Correct?
1	[-1, 0, 0]	$-1 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 = -1$	yes
2	[-1, 0, 0]	$-1 \cdot 1 + 0 \cdot 3 + 0 \cdot 2 = -1$	no
3	[0, 3, 2]	$0 \cdot 1 + 3 \cdot 2 + 2 \cdot 4 = 14$	yes
4	[0, 3, 2]	17	yes
5	[0, 3, 2]	12	no

add to wieght

− subtract

   Final weights: [-1, 1, -1] ← [0, 3, 2] − [1, 2, 3]

3. Have weights been learned that separate the data?

   current eq: −1 + A − B ← if ≥ 0 → +
   < 0 → −

   point 3 & 4 will be missed

4. More generally, irrespective of the training data, you want to know if your features are powerful enough to allow you to handle a range of scenarios. Circle the scenarios for which a perceptron using the features above can indeed perfectly classify movies which are profitable according to the given rules:

   [✓ (a)] Your reviewers are awesome: if the total of their scores is more than 8, then the movie will definitely be profitable, and otherwise it won't be. w = [8, 1, 1]

   (b) Your reviewers are art critics. Your movie will be profitable if and only if each reviewer gives either a score of 2 or a score of 3. No

   (c) Your reviewers have weird but different tastes. Your movie will be profitable if and only if both reviewers agree. No

3 Maximum Likelihood

A Geometric distribution is a probability distribution of the number X of Bernoulli trials needed to get one success. It depends on a parameter p, which is the probability of success for each individual Bernoulli trial. Think of it as the number of times you must flip a coin before flipping heads. The probability is given as follows:

$$P(X = k) = p(1 - p)^{k-1} \tag{1}$$

p is the parameter we wish to estimate.

We observe the following samples from a Geometric distribution: $x_1 = 5$, $x_2 = 8$, $x_3 = 3$, $x_4 = 5$, $x_5 = 7$. What is the maximum likelihood estimate for p?

$$L(p) = P(X=x_1)\,P(X=x_2)\,P(X=x_3)\,P(X=x_4)\,P(X=x_5)$$

$$= P(X=5)\,P(X=8)\,P(X=3)\,P(X=5)\,P(X=7)$$

$$= p^5 (1-p)^{23}$$

e.g. $P(X=5) = p\,(1-p)^4$

$P(X=8) = p\,(1-p)^7$

$$\log(L(p)) = 5\log(p) + 23\log(1-p)$$

max log likelihood -- take derivative and set to 0

$$\frac{5}{p} - \frac{23}{1-p} = 0$$

$$p = 5/28$$

1 Perceptron → Neural Nets

Instead of the standard perceptron algorithm, we decide to treat the perceptron as a single node neural network and update the weights using gradient-based optimization.

In lecture, we covered maximizing likelihood using gradient ascent. We can also choose to **minimize** a loss function that calculates the distance between a prediction and the correct label. The loss function for one data point is $Loss(y, y^*) = \frac{1}{2}(y - y^*)^2$, where y^* is the training label for a given point and y is the output of our single node network for that point.

We will compute a score $z = w_1x_1 + w_2x_2$, and then predict the output using an activation function g: $y = g(z)$.

1. Given a general activation function $g(z)$ and its derivative $g'(z)$, what is the derivative of the loss function with respect to w_1 in terms of $g, g', y^*, x_1, x_2, w_1$, and w_2?

   $$\frac{\partial Loss}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2}\left(g(w_1 x_1 + w_2 x_2) - y^* \right)^2$$

   if you derive the above your answer will be

   $$(g(w_1x_1 + w_2x_2) - y^*)\, g(w_1x_1 + w_2x_2)\, x_1$$

2. We wish to *minimize* the loss, so we will use gradient *descent* (not gradient ascent). What is the update equation for weight w_i given $\frac{\partial Loss}{\partial w_i}$ and learning rate α?

   $$w_i \leftarrow w_i - \alpha \frac{\partial\, loss}{\partial w_1}$$

3. For this question, the specific activation function that we will use is

   $$g(z) = 1 \text{ if } z \geq 0 \text{ , or } -1 \text{ if } z < 0$$

   Use gradient descent to update the weights for a single data point. With initial weights of $w_1 = 2$ and $w_2 = -2$, what are the updated weights after processing the data point $(x_1, x_2) = (-1, 2)$, $y^* = 1$?

   $g'(z) = 0 \rightarrow \frac{\partial\, loss}{\partial w_1}$ will be zero

   Weights are the same

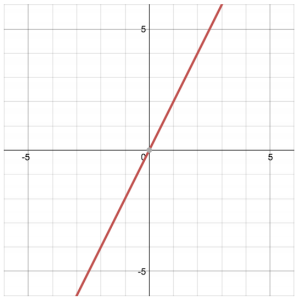
   $w_1 = 2$ $w_2 = -2$

4. What is the most critical problem with this gradient descent training process with that activation function?

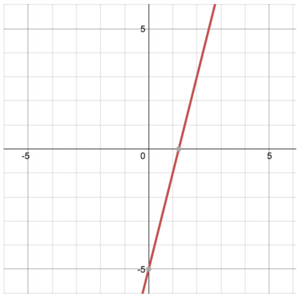
   The gradient of activation is zero no update will be made to weights

3 Neural Network Representations

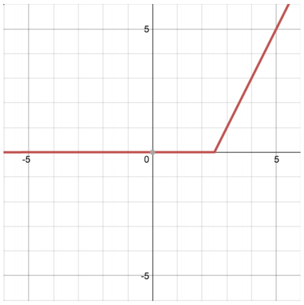
You are given a number of functions (a-h) of a single variable, x, which are graphed below. The computation graphs on the following pages will start off simple and get more complex, building up to neural networks. For each computation graph, indicate which of the functions below they are able to represent.



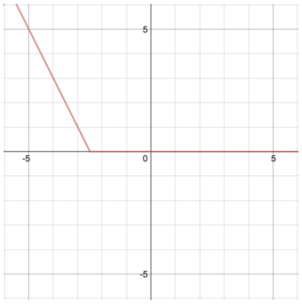
(a) $2x$



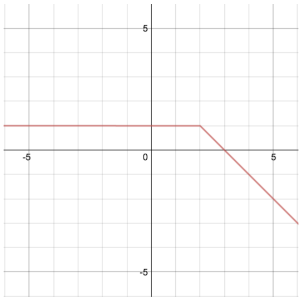
(b) $4x - 5$



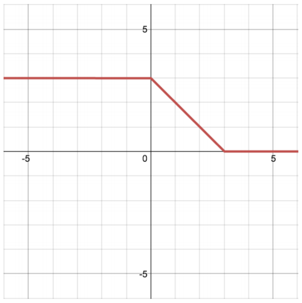
(c) $\begin{cases} 2x - 5 & x \geq 2.5 \\ 0 & x < 2.5 \end{cases}$



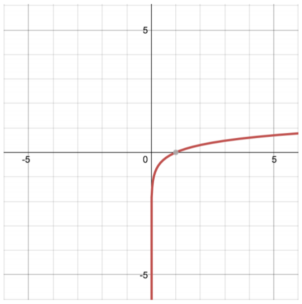
(d) $\begin{cases} -2x - 5 & x \leq -2.5 \\ 0 & x > -2.5 \end{cases}$



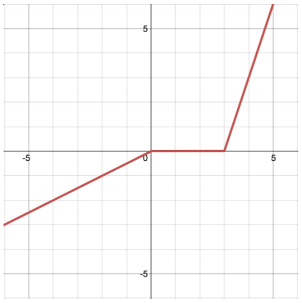
(e) $\begin{cases} -x + 3 & x \geq 2 \\ 1 & x < 2 \end{cases}$



(f) $\begin{cases} 3 & x \leq 0 \\ 3 - x & 0 < x \leq 3 \\ 0 & x > 3 \end{cases}$

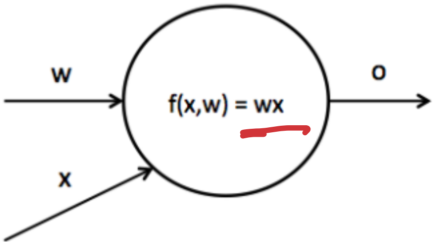


(g) $\log(x)$



(h) $\begin{cases} 0.5x & x \leq 0 \\ 0 & 0 < x \leq 3 \\ 3x - 9 & x > 3 \end{cases}$

1. Consider the following computation graph, computing a linear transformation with scalar input x, weight w, and output o, such that $o = wx$. Which of the funcions can be represented by this graph? For the options which can, write out the appropriate value of w.

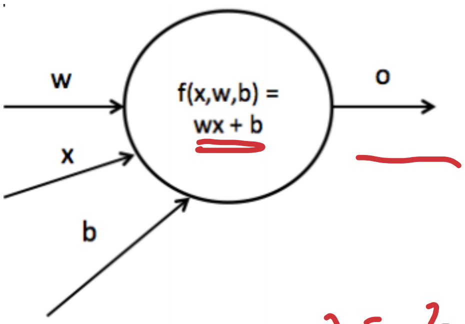


linear eq. → either a or b

no bias so it must be $\underline{a}$

a with w=2

2. Now we introduce a bias term b into the graph, such that $o = wx + b$ (this is known as an *affine* function). Which of the functions can be represented by this network? For the options which can, write out an appropriate value of w, b.

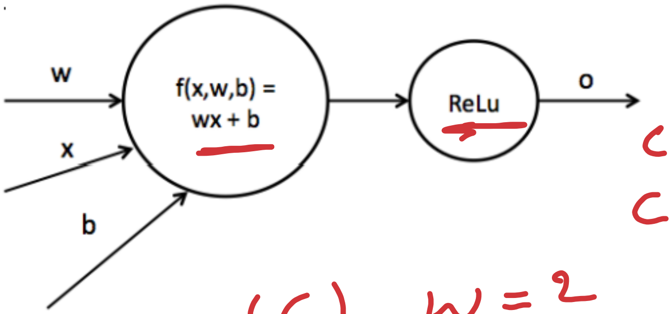


— linear → a or b

(a) $w = 2$ $b = 0$

(b) $w = 4$ $b = -5$

3. We can introduce a non-linearity into the network as indicated below. We use the ReLU non-linearity, which has the form $ReLU(x) = \max(0, x)$. Now which of the functions can be represented by this neural network with weight w and bias b? For the options which can, write out an appropriate value of w, b.



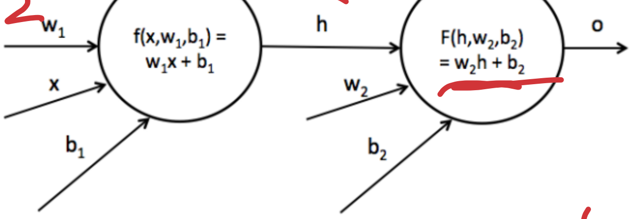
can produce c or d

(c) $w = 2$ $b = -5$

(d) $w = -2$ $b = -5$

4. Now we consider neural networks with multiple affine transformations, as indicated below. We now have two sets of weights and biases w_1, b_1 and w_2, b_2. We denote the result of the first transformation h such that $h = w_1x + b_1$, and $o = w_2h + b_2$. Which of the functions can be represented by this network? For the options which can, write out appropriate values of w_1, w_2, b_1, b_2.

d, $(w_1 x + b_1) + b_2$

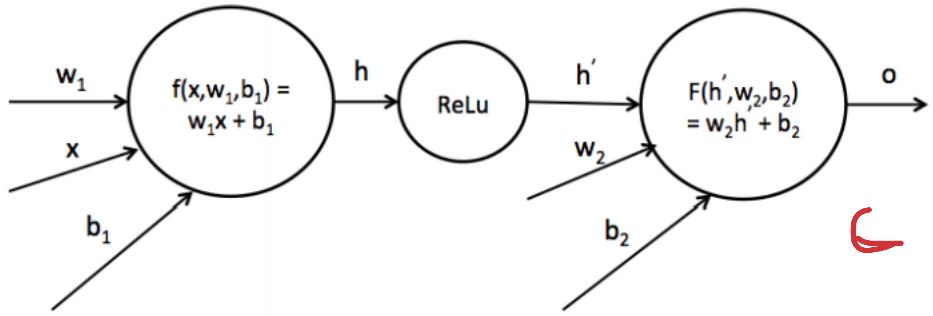


two linear → linear → a or b

(a) $w_1 = 2$ $w_2 = 1$ $b_1 = 0$ $b_2 = 0$

(b) $w_1 = 4$ $w_2 = 1$ $b_1 = 0$ $b_2 = -5$

5. Next we add a ReLU non-linearity to the network after the first affine transformation, creating a hidden layer. Which of the functions can be represented by this network? For the options which can, write out appropriate values of w_1, w_2, b_1, b_2.



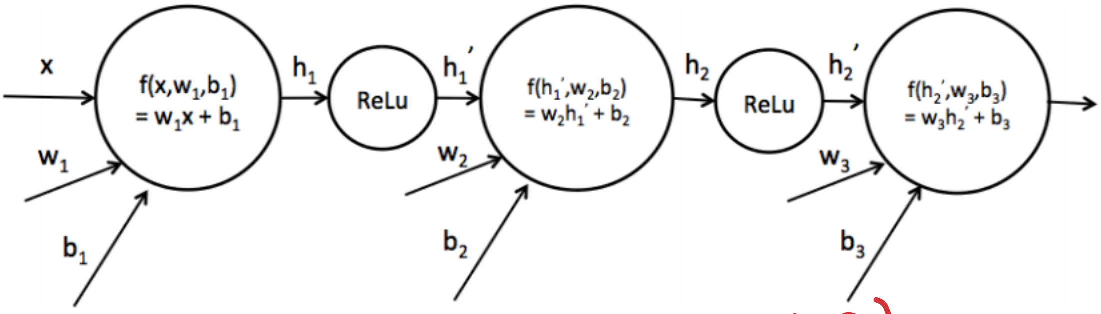
c, d, e

(c) $w_1 = 2$ $b_1 = -5$ $w_2 = 1$ $b_2 = 0$

(d) $w_1 = -2$ $b_1 = -5$ $w_2 = 1$ $b_2 = 0$

(e) $w_1 = 1$ $b_1 = -2$ $w_2 = -1$ $b_2 = 1$

6. Now we add another hidden layer to the network, as indicated below. Which of the functions can be represented by this network?


x
w_1
b_1
$f(x,w_1,b_1) = w_1x + b_1$
h_1
ReLu
h_1'
w_2
b_2
$f(h_1',w_2,b_2) = w_2h_1' + b_2$
h_2
ReLu
h_2'
w_3
b_3
$f(h_2',w_3,b_3) = w_3h_2' + b_3$

(c) (d) (e) (f)

Same as 5 except now f can be done