Universidad Nacional de Colombia
Departarmento de Matemáticas

Matemáticas para el Aprendizaje de Máquina

Juan Antonio Rodríguez

2024

Problems Section 3

1. Problem 3.1 What kind of mapping from input to output would be created if the activation function in equation 3.1 was linear so that $a[z] = \phi_0 + \phi_1 z$? What kind of mapping would be created if the activation function was removed, so $a[z] = z$?

   (a) $a[z] = \phi_0 + \phi_1 z$. To give a proper answer we must replace $a[z]$ in the equation 3.1 with $\beta_0 + \beta_1 z$. Note that we have changed here the variables ϕ_i to β_i to avoid confusion with the ϕ_i that equation 3.1 has.

$$
\begin{aligned}
y &= \phi_0 + \phi_1 * (\beta_0 + \beta_1 * (\theta_{10} + \theta_{11}x)) + \phi_2 * (\beta_0 + \beta_1 * (\theta_{20} + \theta_{21}x)) \\
&+ \phi_3 * (\beta_0 + \beta_1 * (\theta_{30} + \theta_{31}x)) \\
&= \phi_0 + \phi_1 * (\beta_0 + \beta_1 * \theta_{10} + \beta_1 * \theta_{11}x) + \phi_2 * (\beta_0 + \beta_1 * \theta_{20} + \beta_1 * \theta_{21}x)) \\
&+ \phi_3 * (\beta_0 + \beta_1 * \theta_{30} + \beta_1 * \theta_{31}x)) \\
&= \phi_0 + (\phi_1 * \beta_0) + (\phi_1 * \beta_1 * \theta_{10}) + (\phi_1 * \beta_1 * \theta_{11})x + (\phi_2 * \beta_0) + (\phi_2 * \beta_1 * \theta_{20}) \\
&+ (\phi_2 * \beta_1 * \theta_{21})x + (\phi_3 * \beta_0) + (\phi_3 * \beta_1 * \theta_{30}) + (\phi_3 * \beta_1 * \theta_{31})x \\
&= [\phi_0 + (\phi_1 * \beta_0) + (\phi_1 * \beta_1 * \theta_{10}) + (\phi_2 * \beta_0) + (\phi_2 * \beta_1 * \theta_{20}) \\
&+ (\phi_3 * \beta_0) + (\phi_3 * \beta_1 * \theta_{30})] + [(\phi_1 * \beta_1 * \theta_{11}) + (\phi_2 * \beta_1 * \theta_{21}) + (\phi_3 * \beta_1 * \theta_{31})]x
\end{aligned}
$$

   If we let $\alpha_0 = \phi_0 + \phi_1 * \beta_0 + \phi_1 * \beta_1 * \theta_{10} + \phi_2 * \beta_0 + \phi_2 * \beta_1 * \theta_{20} + \phi_3 * \beta_0 + \phi_3 * \beta_1 * \theta_{30}$ and $\alpha_1 = \phi_1 * \beta_1 * \theta_{11} + \phi_2 * \beta_1 * \theta_{21} + \phi_3 * \beta_1 * \theta_{31}$, the equation above would end up being $y = \alpha_0 + \alpha_1 x$, which is the equation of a line (the one used in the regression model).

   (b) $a[z] = z$ We can use the same process we performed above.

$$
\begin{aligned}
y &= \phi_0 + \phi_1 * (\theta_{10} + \theta_{11}x) + \phi_2 * (\theta_{20} + \theta_{21}x) + \phi_3 * (\theta_{30} + \theta_{31}x) \\
y &= \phi_0 + \phi_1 * \theta_{10} + \phi_1 * \theta_{11}x + \phi_2 * \theta_{20} + \phi_2 * \theta_{21}x) + \phi_3 * \theta_{30} + \phi_3 * \theta_{31}x \\
y &= \phi_0 + \phi_1 * \theta_{10} + \phi_2 * \theta_{20} + \phi_3 * \theta_{30} + [\phi_1 * \theta_{11} + \phi_2 * \theta_{21} + \phi_3 * \theta_{31}]x
\end{aligned}
$$

   If we let $\alpha_0 = \phi_0 + \phi_1 * \theta_{10} + \phi_2 * \theta_{20} + \phi_3 * \theta_{30}$ and $\alpha_1 = \phi_1 * \theta_{11} + \phi_2 * \theta_{21} + \phi_3 * \theta_{31}$, the equation would end up being $y = \alpha_0 + \alpha_1 x$, which is again, the equation of a line (used in the regression model).

2. Problem 3.2 For each of the four linear regions in figure 3.3j, indicate which hidden units are inactive and which are active (i.e., which do and do not clip their inputs).

   Active hidden units are marked with an 'X':

Region	HU1	HU2	HU3
Region 1			X
Region 2	X		X
Region 3	X	X	X
Region 4	X	X	

3. Problem 3.3: Derive expressions for the positions of the "joints" in function in figure 3.3j in terms of the ten parameters ϕ and the input x. Derive expressions for the slopes of the four linear regions.

   For the "joints" it is not hard to notice that they appear in the result in the same $x-$coordinate as when the straight line in one of the hidden units crosses $y = 0$. With that idea in mind, the expressions for the joints can be obtained from the expression inside each hidden unit as follows:

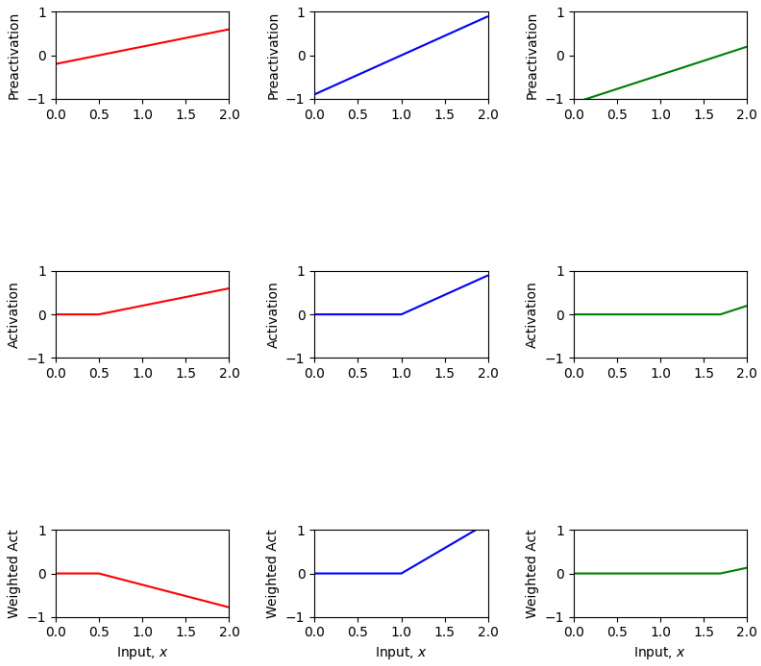
$$\begin{aligned} 0 &= \theta_{i0} + \theta_{i1}x \\ -\theta_{i0} &= \theta_{i1}x \\ x &= \frac{-\theta_{i0}}{\theta_{i1}} \end{aligned}$$

   This means, the result function have a joint when $x = \frac{-\theta_{i0}}{\theta_{i1}}$ for $i = 1, \ldots, n$ with n the number of hidden units.

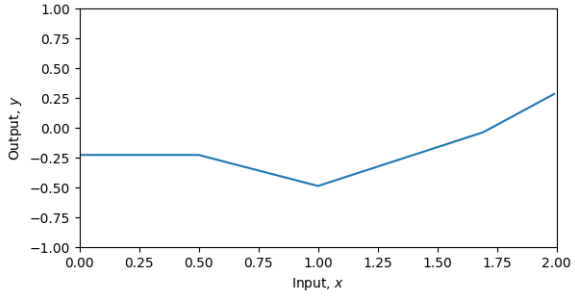
   For the slopes in each linear region, we have to take into consideration whether or not a hidden unit is active or not. In general, the slope of each linear region is the sum of, the slopes in the active hidden units (θ_{i1}) times the weight of each hidden unit (ϕ_i).

4. Problem 3.4 Draw a version of figure 3.3 where the y-intercept and slope of the third hidden unit have changed as in figure 3.14c. Assume that the remaining parameters remain the same.

   Pre-activation, activation and weighted-activation:



Result:



5. Problem 3.5 Prove that the following property holds for $\alpha \in \mathbb{R}^+$: $ReLU[\alpha \cdot z] = \alpha \cdot ReLU[z]$. This is known as the non-negative homogeneity property of the ReLU function.

   Proof. Let $z \in \mathbb{R}$, $\alpha \in \mathbb{R}^+$, then $ReLU[\alpha] = \alpha$. We have to consider two cases:

   (a) If $z \geq 0$, then $ReLU[z] = z$. Since $z \geq 0$, $\alpha \cdot z \geq 0$, therefore $ReLU[\alpha \cdot z] = \alpha \cdot z = \alpha \cdot ReLU[z]$.

   (b) If $z < 0$, then $ReLU[z] = 0$. Since $z < 0$, then $\alpha \cdot z < 0$, therefore $ReLU[\alpha \cdot z] = 0 = \alpha \cdot ReLU[z]$.

   We have proved that for $\alpha \in \mathbb{R}^+$, $ReLU[\alpha \cdot z] = \alpha \cdot ReLU[z]$. □

6. Problem 3.6 Following on from problem 3.5, what happens to the shallow network defined in equations 3.3 and 3.4 when we multiply the parameters θ_{10} and θ_{11} by a positive constant α and divide the slope ϕ_1 by the same parameter α? What happens if α is negative?

- If $\alpha > 0$: When we multiply both parameters θ_{10} and θ_{11} by α after factorization by common factor, we then can apply the result from problem 3.5, and get $\alpha * a[\theta_{10} + \theta_{11}x]$, since we have divided ϕ_1 by α, those α cancel each other out, and we get equation 3.1 again. Therefore, the shallow network defined in 3.3 and 3.4 remain unchanged.
- If $\alpha < 0$: We cannot use result 3.5. Therefore the shallow network is altered. The hidden unit 1, would shift its y-intercept changing its sign as well as its slope, which also changes its sign. Due to the changes in the sign, the activated region of hidden unit 1 would be different from the "original" one. All of that to say, it would result in a different neural network.

7. Problem 3.7 Consider fitting the model in equation 3.1 using a least squares loss function. Does this loss function have a unique minimum? i.e., is there a single "best" set of parameters?

   When thinking of the least squares loss function as a cuadratic function, one can assume that the curve this function creates is somewhat similar to a parabola (in $\mathbb{R}^{\nVdash}$) or a "cicular cup" (in $\mathbb{R}^{\nVdash}$) which both have a unique global minimum. Extending this idea into $\mathbb{R}^{\nVdash}$ one can conclude that the function is still going to have a unique minimum. However, depending on each problem and training dataset this global minimum may not be reached but a local minimim.

8. Problem 3.8 Consider replacing the ReLU activation function with (i) the Heaviside step function $heaviside[z]$, (ii) the hyperbolic tangent function $tanh[z]$, and (iii) the rectangular func- tion $rect[z]$, where:

$$heavside[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases} \qquad rect[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases}$$

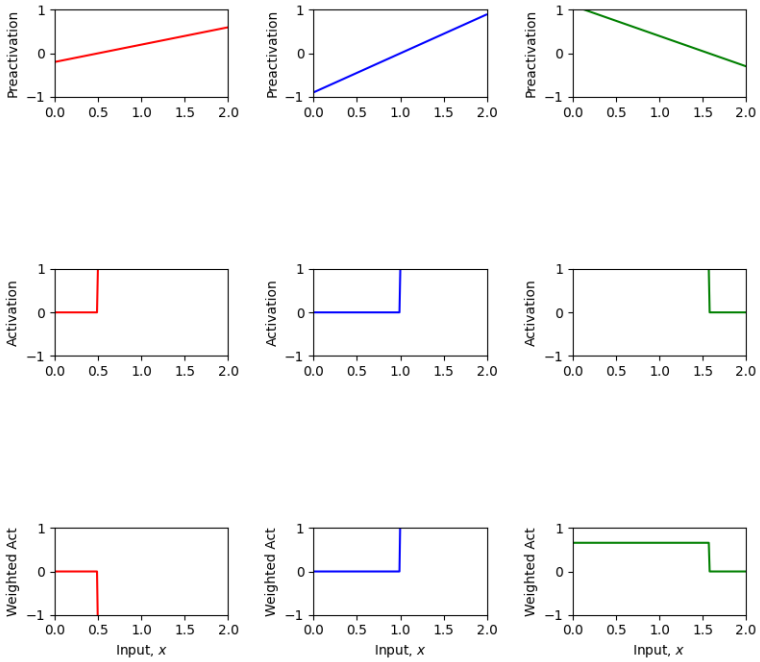
   Redraw a version of figure 3.3 for each of these functions. The original parameters were:

$$\begin{aligned} \phi &= \{\phi_0, \phi_1, \phi_2, \phi_3, \theta_{10}, \theta_{11}, \theta_{20}, \theta_{21}, \theta_{30}, \theta_{31}\} \\ &= \{-0.23, -1.3, 1.3, 0.66, -0.2, 0.4, -0.9, 0.9, 1.1, -0.7\} \end{aligned}$$

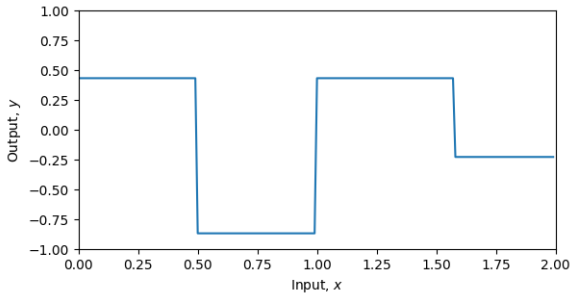
   Provide an informal description of the family of functions that can be created by neural networks with one input, three hidden units, and one output for each activation function.

   - Using $heavside[z]$

     Pre-activation, activation and weighted-activation:

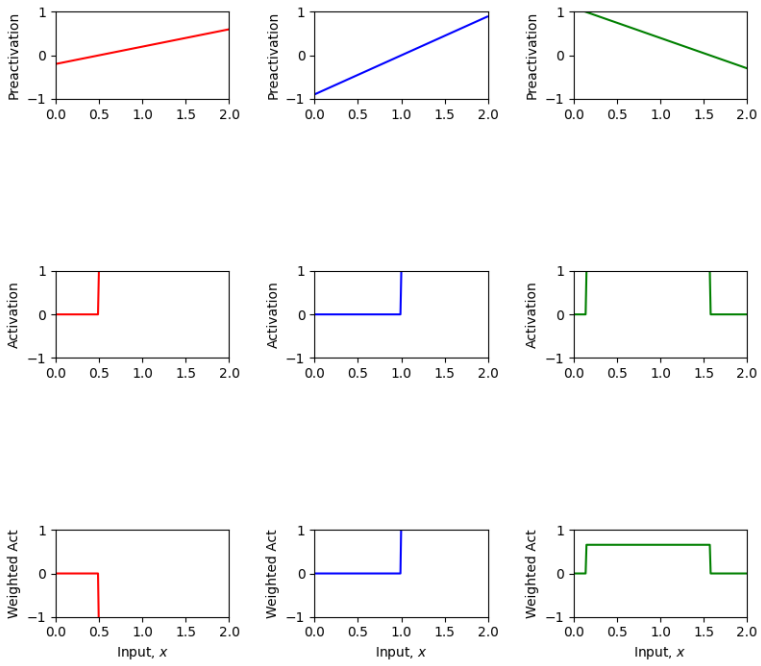


Result:

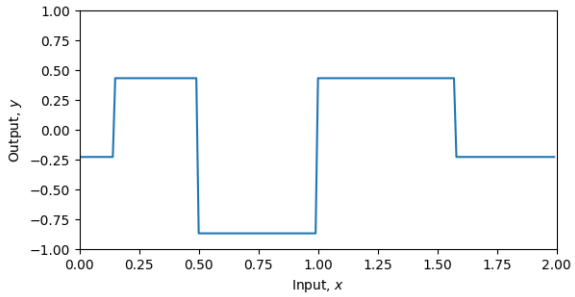


- Using $rect[z]$

  Pre-activation, activation and weighted-activation:

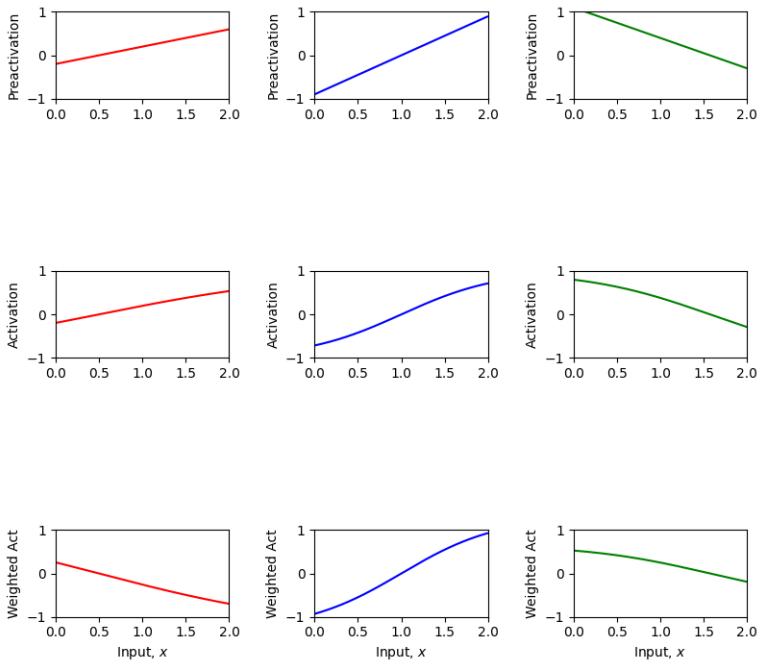


Result:

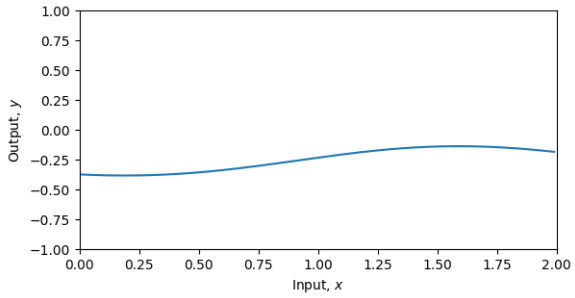


- Using $tanh[z]$

  Pre-activation, activation and weighted-activation:



Result:



Description:

- Using $heavside[z]$: Using the activation function $heavside[z]$ creates a step-like function with as many steps as joints there are if using a $ReLU[z]$. The value of each step is given by the ϕ_i due to the fact that the activation function only returns values of 1 or 0.
- Using $rect[z]$: Using the activation function $rect[z]$ creates a step-like function with more steps as joints there are if using a $ReLU[z]$ due to the fact that the function only returns 1 if $0 \leq z \leq 1$ and 0 otherwise. Just as with the $heavside[z]$ the value of each step is given by the ϕ_i due to the fact that the activation function only returns values of 0 and 1.
- Using $tanh[z]$: Using the activation function $tanh[z]$ creates a curve (not a straight line) that does not show where the joints are (if any). This is due to the fact that the function only returns the value 0 when $z = 0$, so the hidden units are really never

inactive. Also, since $tanh[z]$ is a sigmoid-like function, it curves the lines that one can see with the pre-activation values, hence, the result is a curve and not a straight line.

9. Problem 3.9 Show that the third linear region in figure 3.3 has a slope that is the sum of the slopes of the first and fourth linear regions.

   Using problem 3.2, one may note that the region 3 has all hidden units active, and that region 1 has only hidden unit 3 active while region 4 has hidden units 1 and 2 active.

   We can consider the resulting line (in 3.3j) in each region as the weighted sum of each of the corresponding lines in the post-activation lines for each hidden unit (3.3 g-i). Taking only the graphs shown in 3.3, we can suppose that the weight of each of the post-activation lines is equal to zero. Therefore, region 1, since only the hidden unit 3 is active, the slope would be just the sum of the slope of post-activation hidden unit line 3; and region 4, where hidden units 1 and 2 are active, the slope would be the sum of the corresponding slopes of each of the post-activation hidden unit lines 1 and 2. The region 3, has all hidden units active. Therefore, the slope would be the sum of all slopes in the post-activation hidden units lines.

   Since this shallow network has only 3 hidden units, we have showed that the slope in region 3 is the sum of the slopes of the first and third regions.

10. Problem 3.10 Consider a neural network with one input, one output, and three hidden units. The construction in figure 3.3 shows how this creates four linear regions. Under what circumstances could this network produce a function with fewer than four linear regions?

    A shallow network with one input, one output and three hidden units would create less than four linear regions if the lines of two or more pre-activation hidden units have the same x-intercept. That is,

$$
\begin{aligned}
y &= \theta_{i0} + \theta_{i1}x & \qquad y &= \theta_{j0} + \theta_{j1}x \qquad \text{para } i, j = 1, 2, 3 \;\; i \neq j \\
0 &= \theta_{i0} + \theta_{i1}x_0 & 0 &= \theta_{j0} + \theta_{j1}x_0 \\
x_0 &= 0 - \frac{\theta_{i0}}{\theta_{i1}} & x_0 &= 0 - \frac{\theta_{j0}}{\theta_{j1}} \\
& & -\frac{\theta_{i0}}{\theta_{i1}} &= -\frac{\theta_{j0}}{\theta_{j1}} \\
& & \frac{\theta_{i0}}{\theta_{i1}} &= \frac{\theta_{j0}}{\theta_{j1}}
\end{aligned}
$$

    This means, the shallow network would have less than for regions if for any two or three hidden units we have $\frac{\theta_{i0}}{\theta_{i1}} = \frac{\theta_{j0}}{\theta_{j1}}$.

11. Problem 3.11 How many parameters does the model in figure 3.6 have?

We can count the number of parameters the following way (in this case, we are going to include the bias):

(a) We have 1 input, therefore each hidden unit would have only 2 parameters (θ_{i0} and θ_{i1}).

(b) We have 4 hidden units, therefore we have 4 parameters (ϕ_i) and the bias (ϕ_0) per output.

(c) Based on the above, since the hidden units are "shared" for both outputs, we have, $4 * 2$ parameters from the hidden units and $5 * 2$ corresponding to the "independent" parameters per output.

(d) Therefore we have $4 * 2 + 5 * 2 = 18$ parameters in total.

If we do not want to consider the bias, we substract one parameter for each hidden unit and one for each output, that is, $18 - 6 = 12$.

12. Problem 3.12 How many parameters does the model in figure 3.7 have?

    We can count the number of parameters using the method described above (in this case, we are going to include the bias):

    (a) We have 2 inputs, therefore each hidden unit would have 3 parameters (θ_{i0}, θ_{i1} and θ_{i2}).

    (b) We have 3 hidden units, therefore we have 3 parameters (ϕ_i) and the bias (ϕ_0) per output.

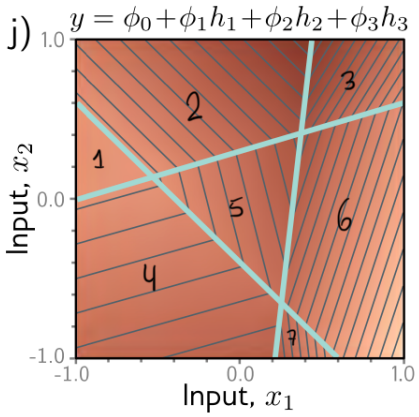
    (c) Based on the above, since the hidden units are "shared" for both outputs, we have, $3 * 3$ parameters from the hidden units and $3 * 1$ corresponding to the "independent" parameters per output.

    (d) Therefore we have $3 * 3 + 4 = 13$ parameters in total.

    If we do not want to consider the bias, we substract one parameter for each hidden unit and one for each output, that is, $13 - 4 = 9$.

13. Problem 3.13 What is the activation pattern for each of the seven regions in figure 3.8? In other words, which hidden units are active (pass the input) and which are inactive (clip the input) for each region?

    We consider the numbering of the regions the following way:



We now can describe the activation pattern using the following table, active hidden units are marked with an 'X':

Region	HU1	HU2	HU3
Region 1			
Region 2	X		
Region 3	X	X	
Region 4			X
Region 5	X		X
Region 6	X	X	X
Region 7		X	X

14. Problem 3.14 Write out the equations that define the network in figure 3.11. There should be three equations to compute the three hidden units from the inputs and two equations to compute the outputs from the hidden units.

    Hidden Units:

    - $h_1 = \theta_{10} + \theta_{11}x_1 + \theta_{12}x_2 + \theta_{13}x_3$
    - $h_2 = \theta_{20} + \theta_{21}x_1 + \theta_{22}x_2 + \theta_{23}x_3$
    - $h_3 = \theta_{30} + \theta_{31}x_1 + \theta_{32}x_2 + \theta_{33}x_3$

    Outputs:

    - $y_1 = \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3$
    - $y_2 = \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3$

15. Problem 3.15 What is the maximum possible number of 3D linear regions that can be created by the network in figure 3.11?

In figure 3.11 we have: $D_i = 3$, $D = 3$, $D_o = 2$. Let p be the maximum number of regions. Using Zaslavsky (1975) we have:

$$
\begin{aligned}
p &= \sum_{j=0}^{3} \binom{3}{j} \\
p &= \binom{3}{0} + \binom{3}{1} + \binom{3}{2} + \binom{3}{3} \\
p &= 1 + 3 + 3 + 1 \\
p &= 8
\end{aligned}
$$

In this case, we have at most 8 regions per output. Combining the maximum number of regions for the $D_o = 2$ outputs, we have that this shallow network may have at most 16 regions.

16. Problem 3.16 Write out the equations for a network with two inputs, four hidden units, and three outputs. Draw this model in the style of figure 3.11.

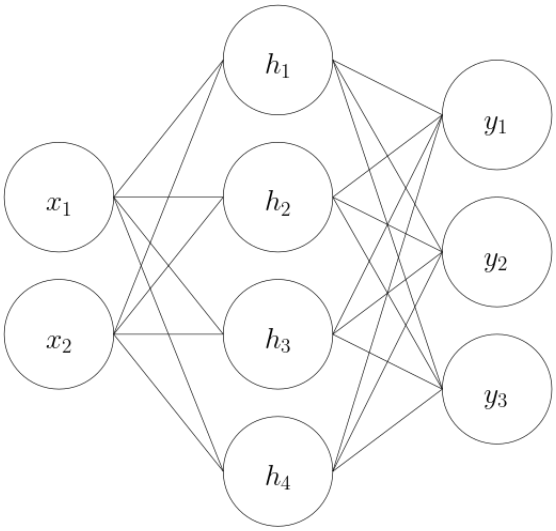
    Hidden Units:

    - $h_1 = \theta_{10} + \theta_{11}x_1 + \theta_{12}x_2$
    - $h_2 = \theta_{20} + \theta_{21}x_1 + \theta_{22}x_2$
    - $h_3 = \theta_{30} + \theta_{31}x_1 + \theta_{32}x_2$
    - $h_4 = \theta_{40} + \theta_{41}x_1 + \theta_{42}x_2$

    Outputs:

    - $y_1 = \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 + \phi_{14}h_4$
    - $y_2 = \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3 + \phi_{24}h_4$
    - $y_3 = \phi_{30} + \phi_{31}h_1 + \phi_{32}h_2 + \phi_{33}h_3 + \phi_{34}h_4$

    Image of the shallow network (arrows were omitted, but flow is from left to right):



17. Problem 3.17 Equations 3.11 and 3.12 define a general neural network with D_i inputs, one hidden layer containing D hidden units, and D_o outputs. Find an expression for the number of parameters in the model in terms of D_i, D, and Do.

    There are some things to take into consideration:

    - For each hidden unit, there are the number of inputs plus one (the bias) parameters.
    - For each output, there are the number of hidden units plus one (the bias) parameters.

    Based on that, we can get two equations (whether or not we want to count the biases). Let p be the number of parameters:

    - Counting the biases: $p = D * (D_i + 1) + D_o * (D + 1)$
    - Not counting the biases: $p = D * D_i + D_o * D$

18. Problem 3.18 Show that the maximum number of regions created by a shallow network with $D_i = 2$-dimensional input, $D_o = 1$-dimensional output, and $D = 3$ hidden units is seven, as in figure 3.8j. Use the result of Zaslavsky (1975) that the maximum number of regions created by partitioning a D_i-dimensional space with D hyperplanes is $\sum_{j=0}^{D_i} \binom{D}{j}$. What is the maximum number of regions if we add two more hidden units to this model, so $D = 5$?

    (a) $D_i = 2$, $D = 3$, $D_o = 1$. Let p be the maximum number of regions. Using Zaslavsky

(1975), we have:

$$
\begin{aligned}
p &= \sum_{j=0}^{2} \binom{3}{j} \\
p &= \binom{3}{0} + \binom{3}{1} + \binom{3}{2} \\
p &= 1 + 3 + 3 \\
p &= 7
\end{aligned}
$$

In this case, we have at most 7 regions.

(b) $D_i = 2$, $D = 5$, $D_o = 1$ Let p be the maximum number of regions. Using Zaslavsky (1975), we have:

$$
\begin{aligned}
p &= \sum_{j=0}^{2} \binom{5}{j} \\
p &= \binom{5}{0} + \binom{5}{1} + \binom{5}{2} \\
p &= 1 + 5 + 10 \\
p &= 16
\end{aligned}
$$

In this case, we have at most 16 regions.