Problems

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 $a[z] = \psi_0 + \psi_1 z$? What kind of mapping would be created if the activation function was removed, so $a[z] = z$?

Solution:

Both are linear mappings, since linear combinations of linear functions are still linear.

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).

Solution:

there are 4 linear regions(from left to right), and 3 hidden units. for the first linear region, only h3 is active. for the second linear region, only h1, h3 are active. for the third linear region, h1, h2, h3 are all active. for the fourth linear region, only h1, h2 are active.

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.

Solution:

see official solution

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.

Solution:

pay attention to the y-intercept and slope of the third hidden unit, then it is easy to draw the figure.

Problem 3.5

Prove that the following property holds for $\alpha \in \mathbb{R}^+$:

$$\text{ReLU}[\alpha z] = \alpha \cdot \text{ReLU}[z] \tag{3.14}$$

This is known as the non-negative homogeneity property of the ReLU function.

Solution:

$$\begin{aligned}
\text{ReLU}[\alpha z] &= \max(0, \alpha z) \\
&= \begin{cases} 0 & \text{if } \alpha z \leq 0 \\ \alpha z & \text{if } \alpha z > 0 \end{cases} \\
&= \begin{cases} 0 & \text{if } z \leq 0 \text{ (since } \alpha > 0) \\ \alpha z & \text{if } z > 0 \text{ (since } \alpha > 0) \end{cases} \\
&= \alpha \cdot \begin{cases} 0 & \text{if } z \leq 0 \\ z & \text{if } z > 0 \end{cases} \\
&= \alpha \cdot \max(0, z) \\
&= \alpha \cdot \text{ReLU}[z]
\end{aligned}$$

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?

Solution:

if α is positive, then we can pick α out of the ReLU function, so the output of the network is $\alpha \cdot \frac{1}{\alpha}\text{ReLU}[z] = \text{ReLU}[z]$. if α is negative, then the ReLU function will clip the original positive part of the input, thus influencing the output of the network.

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?

Solution:

No, the loss function does not have a unique minimum.

When take deravative of the loss function,

we have to deal with the deravative of ReLU.

Consider the derivative of ReLU:

$$\frac{d}{dx}\text{ReLU}[x] = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 & \text{if } x > 0 \end{cases}$$

For any hidden unit h_i, when input x gives $h_i \leq 0$:

$$\frac{\partial y}{\partial \theta_{i0}} = 0, \frac{\partial y}{\partial \theta_{i1}} = 0$$

This means parameters θ_{i0}, θ_{i1} can take multiple values to inactive the hidden unit without affecting the output as long as $h_i \leq 0$, thus creating multiple minima.

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 function rect$[z]$, where:

$$\text{heaviside}[z] = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases} \qquad \text{rect}[z] = \begin{cases} 0 & z < 0 \\ 1 & 0 \leq z \leq 1 \\ 0 & z > 1 \end{cases} \tag{3.15}$$

Redraw a version of figure 3.3 for each of these functions. The original parameters were: $\phi = \{\phi_0, \phi_1, \phi_2, \phi_3, \theta_{10}, \theta_{11}, \theta_{20}, \theta_{21}, \theta_{30}, \theta_{31}\} =$
$\{-0.23, -1.3, 1.3, 0.66, -0.2, 0.4, -0.9, 0.9, 1.1, -0.7\}$. 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.

Solution:

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.

Solution:

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?

Solution:

Problem 3.11*

How many parameters does the model in figure 3.6 have?

Solution:

Problem 3.12

How many parameters does the model in figure 3.7 have?

Solution:

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?

Solution:

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.

Solution:

Problem 3.15*

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

Solution:

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.

Solution:

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 D_o.

Solution:

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 $\sum_{i=0}^{D_i} \binom{D}{i}$. What is the maximum number of regions if we add two more hidden units to this model, so $D = 5$?

Solution: