COURSE 02457

Non-Linear Signal Processing: Quiz Exercise 5

C.M. Bishop: Pattern Recognition and Machine Learning, Sections 5.1-5.4

Questions

1. How can we make a weight-initialization range suitable for all data?

   (a) This cannot be done

   (b) By appropriately scaling data sets

   (c) By changing the optimization parameters

   (d) By re-starting the algorithm with random weight initializations several times

2. How many layers in the neural network shown in Figure 1 in Exercise 5 contain weights that are optimized during training?

   (a) 1

   (b) 2

   (c) 3

   (d) 4

3. What is an appropriate cost function for training a neural network if the response variable t is binary? Let t_n denote the response (target) for the nth observation and y_n be the prediction for the nth observation.

   (a) $\sum_{n=1}^{N}(t_n - y_n)^2$

   (b) $\frac{1}{2}\sum_{n=1}^{N}(t_n - y_n)^2$

   (c) $-\sum_{n=1}^{N}(t_n \ln(y_n) + (1 - t_n) \ln(1 - y_n))$

   (d) $-\prod_{n=1}^{N}(t_n \ln(y_n) + (1 - t_n) \ln(1 - y_n))$

4. What is the advantage of diagonal approximation of the Hessian of the error function rather than exact evaluation?

   (a) Diagonal approximation is faster and the method can be altered to give the exact diagonal values without increasing the run time.

   (b) Diagonal approximation is more accurate due to higher numerical stability.

   (c) Diagonal approximation is faster and a close approximation since the true Hessian is often nearly diagonal.

   (d) Diagonal approximation is faster and sometimes the inverse rather than the Hessian itself is needed, making a diagonal approximation desirable.

5. Neural networks can be used for

(a) Continuous response variables (output) only.

(b) Binary response variables only.

(c) Binary and continuous response variable.

(d) Continuous, binary, and multi-category response variables.

6. Consider a fully connected two-layer neural network (one hidden layer) with M hidden units, all with $\tanh(\cdot)$ as activation function. What is the minimum number of weight vectors that give rise to the same output when the same input is given? That is, what is the weight-space symmetry factor of this neural network?

(a) $M! \cdot 2^M$

(b) $M!$

(c) M^2

(d) $M! \cdot M^2$

7. Figure 1 shows a fully connected two-layer neural network with $\tanh(\cdot)$ as activation function. We consider the weight vector $\mathbf{w}^* = (w_1^{(1)}, w_2^{(1)}, w_3^{(1)}, w_4^{(1)}, w_5^{(1)}, w_6^{(1)}, w_1^{(2)}, w_2^{(2)})$. Which of the following weight vectors produce outputs of the neural network identical to $\mathbf{w}^*$?

I) $(-w_1^{(1)}, w_2^{(1)}, -w_3^{(1)}, w_4^{(1)}, -w_5^{(1)}, w_6^{(1)}, -w_1^{(2)}, w_2^{(2)})$

II) $(-w_1^{(1)}, w_2^{(1)}, -w_3^{(1)}, w_4^{(1)}, -w_5^{(1)}, w_6^{(1)}, -w_1^{(2)}, -w_2^{(2)})$

III) $(-w_2^{(1)}, w_1^{(1)}, -w_4^{(1)}, w_3^{(1)}, -w_6^{(1)}, w_5^{(1)}, -w_2^{(2)}, w_1^{(2)})$

IV) $(-w_2^{(1)}, w_1^{(1)}, -w_4^{(1)}, w_3^{(1)}, -w_6^{(1)}, w_5^{(1)}, -w_1^{(2)}, w_2^{(2)})$

(a) I), IV)

(b) II), III), IV)

(c) I), III)

(d) II), IV)

8. Which function of the inputs x_1, x_2, and x_3 does the three-layer neural network in Figure 2 represent?

(a) $\sigma\left(\sum_{k=1}^{2} w_k^{(3)} \sigma\left(\sum_{j=1}^{2} w_{kj}^{(2)} \tanh\left(\sum_{i=1}^{3} w_{ji}^{(1)} x_i\right)\right)\right)$

(b) $\sigma\left(\sum_{k=1}^{2} w_k^{(3)} \tanh\left(\sum_{j=1}^{4} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{6} w_{ji}^{(1)} x_i\right)\right)\right)$

(c) $\sigma\left(\sum_{k=1}^{2} w_k^{(3)} \tanh\left(\sum_{j=1}^{2} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{3} w_{ji}^{(1)} x_i\right)\right)\right)$

(d) $\tanh\left(\sum_{k=1}^{2} w_k^{(3)} \sigma\left(\sum_{j=1}^{2} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{3} w_{ji}^{(1)} x_i\right)\right)\right)$

9. Choose the correct ending of the following sentence: When training a neural network, the training error at the end of training is

(a) sometimes minimized as much as possible

(b) always minimized as much as possible if we have more data points than ten times the number of explanatory variables

(c) never minimized as much as possible

(d) always minimized as much as possible

10. Which of the following statements are true of backpropagation?

I) Backpropagation is only useful when the error function considered is the sum-of-squares

II) Backpropagation can be used only in neural networks

III) Backpropagation can be used to evaluate the Jacobian and Hessian matrices

(a) I)

(b) II)

(c) III)

(d) I), II)

11. Consider the two-layer neural network in Figure 3 with one hidden layer containing one unit with $\tanh(\cdot)$ as activation function. The activation function for the output unit is the identity. Considering this as a regression problem, we use the cost function $E(\mathbf{W}) = (y-t)^2$ for a prediction y with target t. Give the correct two quantities, $\delta_2 = y - t$ and $\delta_1 = \frac{\partial E}{\partial a_1}$.

(a) $\delta_2 = w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t$,
$\delta_1 = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) w_{11}^{(2)} \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right)$

(b) $\delta_2 = w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t$,
$\delta_1 = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) w_{11}^{(2)} \left(\tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right)$

(c) $\delta_2 = \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})$,
$\delta_1 = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) w_{11}^{(2)} \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right)$

(d) $\delta_2 = \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})$,
$\delta_1 = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) w_{11}^{(2)} \left(\tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right)$

12. Consider again the two-layer neural network in Figure 3 with one hidden layer containing one unit with $\tanh(\cdot)$ as activation function. The activation function for the output unit is the identity. Considering this as a regression problem, we use the cost function $E(\mathbf{W}) = (y-t)^2$ for a prediction y with true target t. Give the two derivatives $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(2)}}$ and $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(1)}}$.

(a) $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(2)}} = \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right) \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})$,
$\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(1)}} = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) x_1$

(b) $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(2)}} = \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right) \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}),$
$\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(1)}} = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right) x_1$

(c) $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(2)}} = \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}),$
$\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(1)}} = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}) - t\right) x_1$

(d) $\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(2)}} = \left(w_{11}^{(2)} \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) \tanh(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)}),$
$\frac{\partial E(\mathbf{W})}{\partial w_{11}^{(1)}} = \left(1 - \tanh^2(x_1 w_{11}^{(1)} + x_2 w_{12}^{(1)})\right) x_1$

13. **Challenge question.** Consider a fully connected two-layer neural network (one hidden layer) with M hidden units, all with the identity as activation function. What is the minimum number of weight vectors that give rise to the same output when the same input is given? That is, what is the weight-space symmetry factor of this neural network?

    (a) $M! \cdot 2^M$

    (b) $M!$

    (c) M^2

    (d) $M! \cdot M^2$

14. **Challenge question.** Consider a fully connected two-layer neural network (one hidden layer) with M hidden units, all with $\exp(\cdot)$ as activation function. What is the minimum number of weight vectors that give rise to the same output when the same input is given? That is, what is the weight-space symmetry factor of this neural network?

    (a) $M! \cdot 2^M$

    (b) $M!$

    (c) M^2

    (d) $M! \cdot M^2$

15. **Challenge question.** Consider a fully connected three-layer neural network (two hidden layers). In the first hidden layer, there are M hidden units, all with $\exp(\cdot)$ as activation function. In the second hidden layer, there are N hidden units, all with $\tanh(\cdot)$ as activation function. What is the minimum number of weight vectors that give rise to the same output when the same input is given? That is, what is the weight-space symmetry factor of this neural network?

    (a) $M! \cdot 2^M \cdot N! \cdot 2^N$

    (b) $(M+N)! \cdot 2^{M+N}$

    (c) $2^M \cdot N! \cdot 2^N$

    (d) $M! \cdot N! \cdot 2^N$

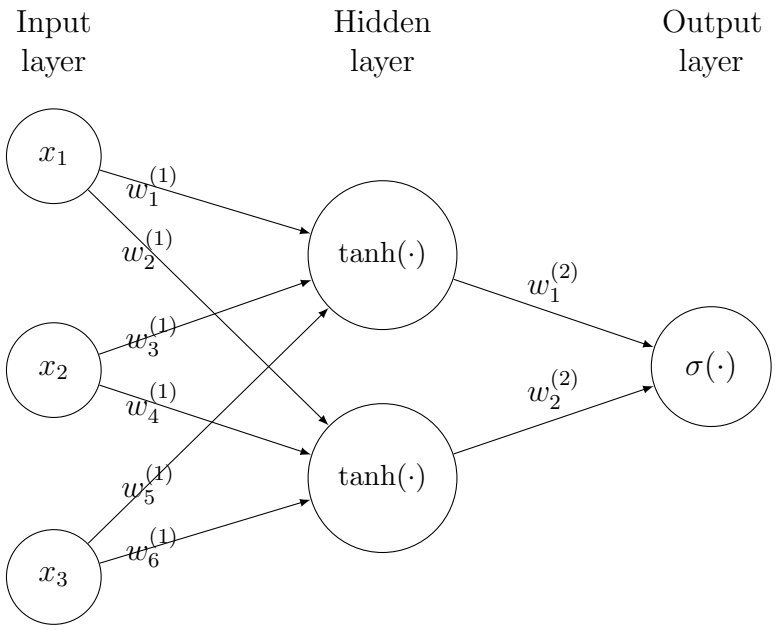

Figure 1: Two-layer neural network.

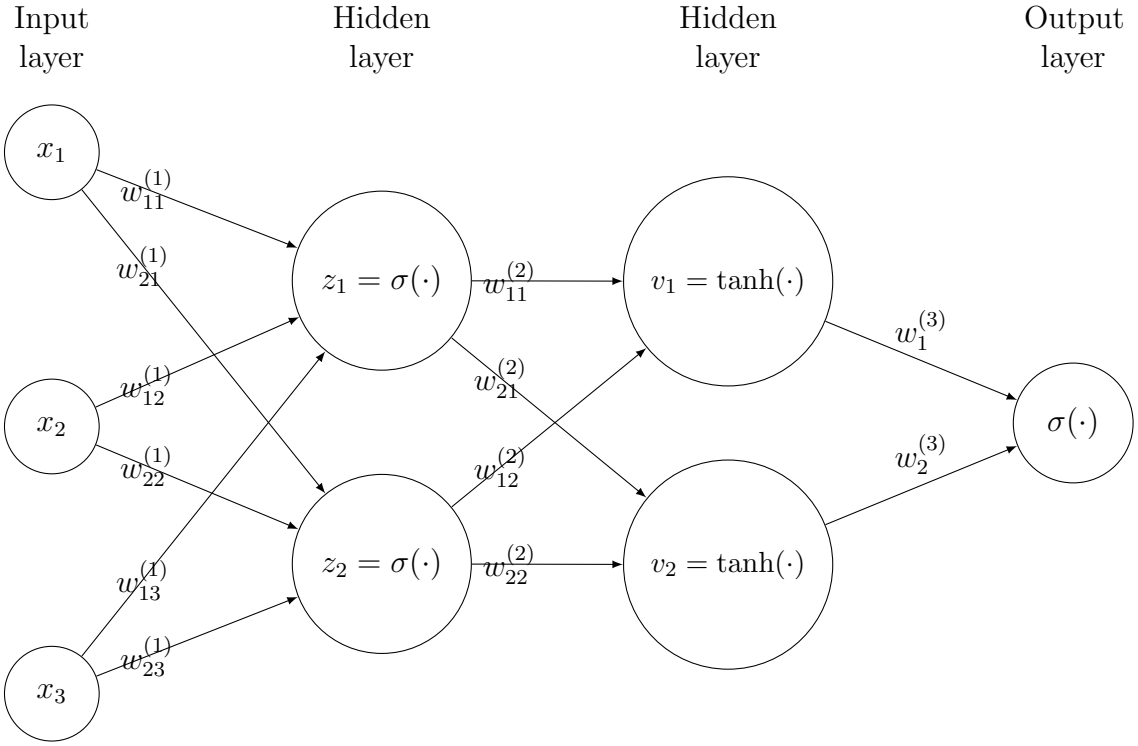

Figure 2: Three-layer neural network.

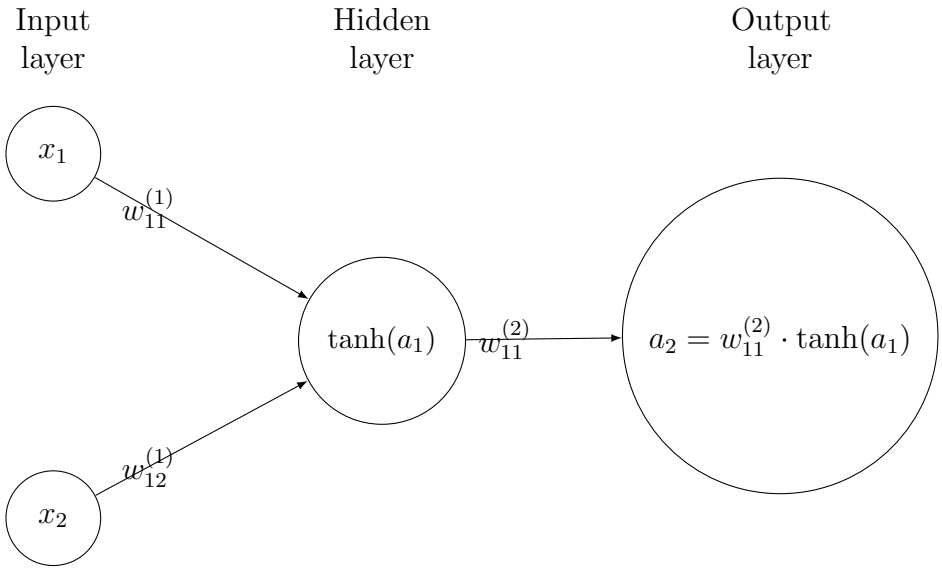

Figure 3: Two-layer neural network.

Hint - reading for each question

1. Exercise check point 5.1
2. Section 5.1 in Bishop
3. Section 5.2
4. Section 5.4
5. Section 5.2
6. Section 5.1.1
7. Section 5.1.1
8. Section 5.1
9. Section 5.2
10. Section 5.3
11. Section 5.3
12. Section 5.3
13. Section 5.1.1
14. Section 5.1.1
15. Section 5.1.1

DTU, September 2013,
Laura Frølich and Ole Winther