

# Neural Networks and Machine Learning

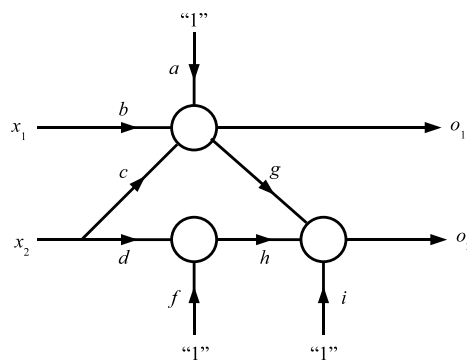
## Solutions of the Exam of 29/1/2007

Notes:

- The solutions are presented here in a brief manner, just for the students to be able to check whether the way in which they solved the problems was right or not. The “carefully justified manner” requested in the exam required some more detail than what is given here. For example, it required presenting all calculations.
- The numerical values that you have obtained may slightly differ from the ones presented here, depending on the specific way in which you performed the roundings (e.g. on how many digits you’ve kept in each step and on whether you used truncation or rounding to the nearest integer). Also note, however, that you should have retained at least three digits after the decimal point in all calculations (of course, they could be omitted if they were zero).

### Problem 1

Consider the following multilayer perceptron



The units of the first layer have as activation function (nonlinearity) the hyperbolic tangent function. The unit of the second layer (which produces output  $o_2$ ) is linear. The training set is

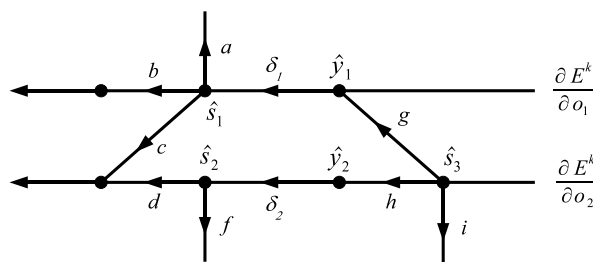
$x_1$	$x_2$	$d_1$	$d_2$
-1	-1	1	1
1	-1	2	-2

The initial values of all weights are 0.5. The cost function is the total squared error.

- 1) Draw the backpropagation network. Indicate, by means of appropriate symbols, the gains of all branches. Also indicate, by means of appropriate symbols, all the variables (node values) that you will need to compute to perform the calculation requested in item 2) below.
- 2) Compute the value of weight  $c$  after the first update, using backpropagation in **real-time** mode, assuming that the training set is repeatedly presented to the network with the training patterns in the order given above. The training is performed with a step size parameter  $\eta = 0.1$  and with no momentum term.

### Solution

- 1) The backpropagation network is



- 2) We shall number the units of the first layer as 1 (top) and 2 (bottom), and the unit of the second layer as 3. We shall designate by  $s_i$  and  $y_i$ , respectively, the input sum and the output of unit  $i$ . The corresponding variables

in the backpropagation network are designated by  $\hat{s}_i$  and  $\hat{y}_i$ , respectively, as shown in the preceding figure. The sigmoids of the units of the first layer shall be designated by  $S(\cdot)$ . The network's outputs are designated by  $o_1$  and  $o_2$  respectively. The output errors are  $e_1 = o_1 - d_1$  and  $e_2 = o_2 - d_2$  respectively.

Since training is performed in real-time mode, the weights are updated after the presentation of each pattern. The first update is performed after the presentation of the first pattern. The values obtained in the forward propagation pass, for the first pattern, are given in the following table. Note that  $o_1 = y_1$  and, since the unit 3 is linear,  $o_2 = y_3 = s_3$ . Only  $o_1$  and  $o_2$  are shown in the table.

$x_1$	$x_2$	$d_1$	$d_2$	$s_1$	$o_1$	$s_2$	$y_2$	$o_2$	$e_1$	$e_2$
-1	-1	1	-1	-0.5	-0.462	0	0	0.269	-1.462	-0.731

The contribution of the first pattern to the cost function is  $E^1 = (e_1^1)^2 + (e_2^1)^2$ , where the upper index refers to the number of the pattern. We'll omit the upper index from here on, except in  $E^1$  itself.

According to the backpropagation rule,  $\partial E^1 / \partial c = x_2 \hat{s}_1$ , where  $\hat{s}_1$  is obtained from the backpropagation network, using  $\partial E^1 / \partial o_i = 2e_i$  as inputs to that network. The values that need to be computed to find  $\partial E^1 / \partial c$  are:

$$\begin{aligned} \delta_1 &= S'(s_1) = \text{sech}^2(s_1) \\ \hat{s}_3 &= 2e_2 \\ \hat{y}_1 &= 2e_1 + g\hat{s}_3 \\ \hat{s}_1 &= \delta_1 \hat{y}_1 \\ \frac{\partial E^1}{\partial c} &= x_2 \hat{s}_1 \end{aligned}$$

These values are:

$\delta_1$	$\hat{s}_3$	$\hat{y}_1$	$\hat{s}_1$	$\partial E^1 / \partial c$
0.786	-1.462	-3.655	-2.875	2.875

Finally, the new value of  $c$  is given by

$$c^{(1)} = c^{(0)} - \eta \frac{\partial E^1}{\partial c} = 0.5 - 0.288 = 0.212$$

## Problem 2

Consider applying the k-means algorithm to cluster the following set of data points.

$$\mathcal{T} = \{ (0, 0), (4, 0), (0, 1), (4, 1) \}$$

- 1) State the truth/falsehood of the following statements:
  - a) \* The performance of the k-means algorithm depends on a learning rate parameter;  
**Solution: FALSE**
  - b) \* The convergence of the algorithm to a solution is usually asymptotic;  
**Solution: FALSE**
  - c) \* The algorithm may converge to a local minimum of its cost function;  
**Solution: TRUE**
  - d) \* Using a number of centers equal to the number of data points can lead to a trivial solution with cost equal to zero.  
**Solution: TRUE**
- 2) Find two different solutions (fixed points of the iteration), using two centers, and determine the values of the cost function for both of them;  
**Solution:** Two possible<sup>1</sup> solutions are:  $\{(0, 1/2), (4, 1/2)\}$ , and  $\{(2, 0), (2, 1)\}$ . Both solutions are fixed points of the algorithm, as can easily be checked. The costs, in terms of total squared error, are 1 and 16 respectively.
- 3) Is there any solution with three centers? If your answer is affirmative, give an example.  
**Solution:** Yes, for instance,  $\{(0, 0), (0, 1), (4, 1/2)\}$ .

<sup>1</sup>In fact they are the only solutions with two centers.

### Problem 3

Consider a set of 2-dimensional data points drawn from a multivariate Gaussian distribution with zero mean. A principal component analysis of that distribution yielded two eigenvectors,  $(\sqrt{2}/2, \sqrt{2}/2)$  and  $(\sqrt{2}/2, -\sqrt{2}/2)$ , with eigenvalues 2 and 3 respectively.

- 1) Determine the variances of the distribution ( $\sigma_1^2$  and  $\sigma_2^2$ ) along each one of the original data space dimensions.

**Solution:** The variances  $\sigma_1^2$  and  $\sigma_2^2$  are the diagonal of the correlation matrix  $R = VDV^T$ . The resulting values are  $\sigma_1^2 = \sigma_2^2 = 5/2$ .

- 2) Indicate the principal direction of greater energy (give the direction by means of a vector). What is its associated variance, in percentage of the total one?

**Solution:** That principal direction corresponds to the eigenvector with the highest eigenvalue:  $(\sqrt{2}/2, -\sqrt{2}/2)$ . The fraction of the total energy that it represents is  $3/(3+2) = 60\%$ .

- 3) Determine the first principal component of the points  $(3, 2)$  and  $(3, -2)$  (note that what is being requested are scalars, not vectors).

**Solution:** The first principal component corresponds to the coordinate of the projection of each point onto the principal direction determined above:  $(3, 2)$  projects to  $\sqrt{2}/2$ , and  $(3, -2)$  projects to  $5\sqrt{2}/2$ .

- 4) Reconstruct the points from the first principal component determined above. Then, compute the absolute value of the reconstruction error for each one of the points. Give a geometrical explanation for the disparity of errors between the two points.

**Solution:** The reconstruction of each point is computed as the product of the corresponding first principal component (found above) by the normalized principal eigenvector: The results are  $(1/2, -1/2)$  and  $(5/2, -5/2)$  respectively. The absolute reconstruction errors are  $5\sqrt{2}/2$  and  $\sqrt{2}/2$  respectively. The reason behind this disparity of values is the fact that the former vector is more aligned with the first principal direction than the latter.

### Problem 4

*Note: This problem has a somewhat long introduction, to explain the framing of what you're requested to do. That doesn't mean that what is being requested is too hard to do. Read the whole problem!*

Consider a system that is trained, in a supervised way, for classifying data into one of two classes. The desired values are 0 for one of the classes (that we shall call  $C_0$ ) and 1 for the other class (that we shall call  $C_1$ ). As you probably know, if the cost function being minimized is the expected value of the squared error, if we assume that the system is flexible enough, and if we also assume that the training reaches the absolute minimum of the cost function, then the system's output will be an estimate of the posterior probability  $P(C_1|\mathbf{x})$ , where  $\mathbf{x}$  is the input pattern.

As you may recall, in our course the proof of this fact was made in three steps: (1) considering a system with a fixed input pattern; (2) generalizing that result to a system with input patterns drawn from a countable set; (3) generalizing that result to a system with input patterns drawn from a continuous distribution. In this problem you'll only be asked to consider the first step, corresponding to a fixed input pattern.

The squared error is not the only cost function that leads the system's output to be an estimate of  $P(C_1|\mathbf{x})$ . In fact, there is an infinite number of cost functions that have that property. Here we'll consider the following one:

$$C(o) = \begin{cases} \frac{o^3}{3} - \frac{o^2}{2} + \frac{1}{6} & \text{if } d = 1 \\ \frac{o^3}{3} & \text{if } d = 0 \end{cases}$$

where  $o$  is the system's output and  $d$  is the desired value.

- 1) Sketch the graphs of the cost function, for both values of  $d$ , for  $o \in [-2, 2]$ . Indicate the exact locations of the zeros and of the local maxima and minima.

Note, to facilitate sketching the graphs: For  $d = 1$ ,  $C(o)$  has two roots, at  $o = -1/2$  and  $o = 1$ , as you can easily check. You can easily find the locations of the function's local maxima and minima yourself.

- 2) Assume that:

- The system under consideration has its output  $o$  limited to the interval  $]0, 1[$  (for example, the system could be a multilayer perceptron with a sigmoid varying between 0 and 1 in the output unit).
- The system has a fixed input pattern.

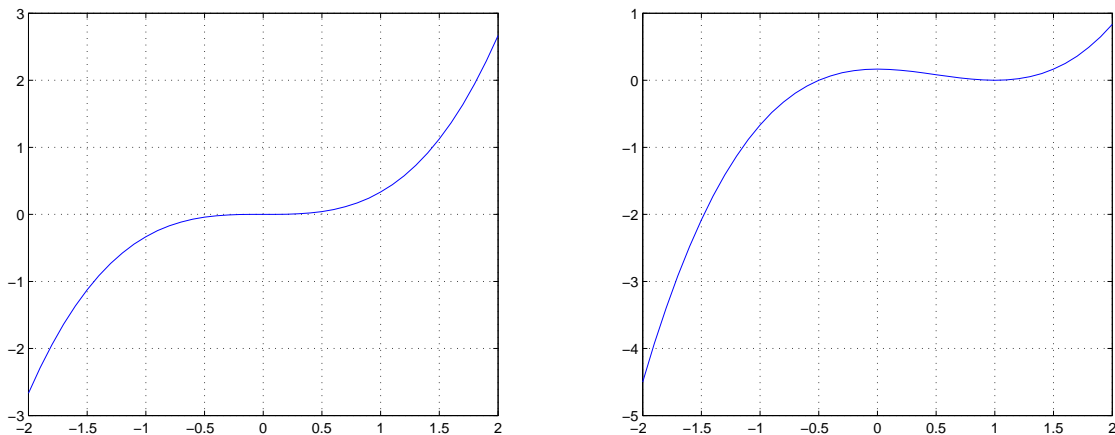
- The values of the desired output are presented randomly, with probabilities  $P(d = 1) = p$  and  $P(d = 0) = 1 - p$ , where  $p$  is some value in  $]0, 1[$ .
- The training uses as cost function the expected value of  $C(o)$  defined above.

*Assignment:* Show that, under these conditions, the absolute minimum of the cost function corresponds to  $o = p$ .

- 3) Explain what would happen if we hadn't placed any limitation on the range of permissible values of  $o$ . Which would be the largest interval to which  $o$  could be limited, to still keep the absolute minimum of the cost function at  $o = p$  for any  $p \in ]0, 1[$ ?

## Solution

- 1) For  $d = 0$ , the graph is the well known graph of  $o^3$ . For  $d = 1$ ,  $C'(o) = o^2 - o$ . Its roots are at  $o = 0$  and  $o = 1$ . Since the function has a continuous derivative, and since  $C(0) = 1/6$  and  $C(1) = 0$ ,  $o = 0$  is a local maximum and  $o = 1$  is a local minimum. The graphs are the following, for  $d = 0$  (left) and for  $d = 1$  (right):



- 2) The cost function, which is the expected value of  $C(o)$ , is

$$\begin{aligned}
 E(o) &= P(d = 0) \frac{o^3}{3} + P(d = 1) \left( \frac{o^3}{3} - \frac{o^2}{2} + \frac{1}{6} \right) \\
 &= (1 - p) \frac{o^3}{3} + p \left( \frac{o^3}{3} - \frac{o^2}{2} + \frac{1}{6} \right) \\
 &= \frac{o^3}{3} - p \frac{o^2}{2} + \frac{p}{6}.
 \end{aligned}$$

Its derivative is  $E'(o) = o^2 - po$ , which has roots at  $o = 0$  and  $o = p$ .

The function is a third degree polynomial. It has a continuous derivative. Furthermore, we have

$$\begin{aligned}
 E(0) &= \frac{1}{6}p \\
 E(p) &= \frac{1}{6}(p - p^3).
 \end{aligned}$$

Since  $E(0) > E(p)$ , there has to be a local maximum at  $o = 0$  and a local minimum at  $o = p$ . The function must decrease in  $]0, p[$  and increase in  $]p, +\infty[$ . Therefore the absolute minimum, for  $o \in ]0, 1[$  is at  $o = p$ , as we wished to prove.

- 3) From the study made in item 2) above, we also conclude that  $E(o)$  must increase in  $] -\infty, 0[$ , and that it must have a root in that interval. That root is  $o = 0$ , for  $p = 0$ , and is negative for  $p > 0$ .

If there were no limitation on the value of  $o$ , the minimum of  $E(o)$  would be at  $o = -\infty$ . The optimization would then lead  $o$  to decrease, without end, towards  $-\infty$ .

If 0 is limited, on the left, to a value  $o_l$  lower than or equal to the negative root of  $E(o)$ , there will be an absolute minimum of  $E(o)$  at  $o_l$ . Since that negative root goes to 0 as  $p \rightarrow 0$ ,  $o$  cannot be limited, on the left, to any value lower than 0, for the absolute minimum of  $E(o)$  to stay at  $o = p$  for all  $p \in ]0, 1[$ .

Since  $E(o)$  increases in  $]p, +\infty[$ , there is no need to put any limit to the value of  $o$  on the right: the absolute minimum of  $E(o)$  will still stay at  $o = p$  if  $o$  is unbounded on the right. Therefore, the largest interval to which  $o$  can be limited is  $[0, +\infty[$ .