Data Science

Problem sheet 3

Ex 1

Let (0, 0, –2); (0, 1, 1); (1, 0, 2) be three records on the x_1–x_2 plane, where the third coordinate is the y target variable. Determine the coefficients of the following linear regression that minimizes the squared error: $y = w_1x_1 + w_2x_2 + w_0$.

a) Determine the optimal coefficients analytically!

b) Approximate the optimal coefficients using the gradient descent method (few steps enough).

c) Approximate the optimal coefficients using the stochastic gradient descent (few steps enough).

For gradient methods use the following initialization of the weights: $w_0 = w_1 = w_2 = 1$. Let the learning rate be 0.25.

Solution

Let's write the records in a table:

	x_1	x_2	y
(1)	0	0	-2
(2)	0	1	1
(3)	1	0	2

$y = w_1x_1 + w_2x_2 + w_0$

a) Analytical solution

We have to minimize the following Err^2 error function that is the sum of squared errors:

$$
\begin{aligned}
Err^2 &= \sum_{i=1}^{3} Err_i^2 = \sum_{i=1}^{3} \left(w_1x_1^{(i)} + w_2x_2^{(i)} + w_0 - y^{(i)}\right)^2 \\
&= \left(w_1x_1^{(1)} + w_2x_2^{(1)} + w_0 - y^{(1)}\right)^2 + \left(w_1x_1^{(2)} + w_2x_2^{(2)} + w_0 - y^{(2)}\right)^2 + \left(w_1x_1^{(3)} + w_2x_2^{(3)} + w_0 - y^{(3)}\right)^2 \\
&= (w_1 \cdot 0 + w_2 \cdot 0 + w_0 - (-2))^2 + (w_1 \cdot 0 + w_2 \cdot 1 + w_0 - 1)^2 + (w_1 \cdot 1 + w_2 \cdot 0 + w_0 - 2)^2 \\
Err^2 &= (w_0 + 2)^2 + (w_2 + w_0 - 1)^2 + (w_1 + w_0 - 2)^2
\end{aligned}
$$

The goal is to find the w_0, w_1 és w_2 coefficients, which minimize the Err^2 function, hence first we have to calculate the partial derivatives:

$$
\begin{aligned}
\frac{\partial Err^2}{\partial w_0} &= 2(w_0 + 2) + 2(w_2 + w_0 - 1) + 2(w_1 + w_0 - 2) = 6w_0 + 2w_1 + 2w_2 - 2 \\
\frac{\partial Err^2}{\partial w_1} &= 0 + 0 + 2(w_1 + w_0 - 2) = 2w_0 + 2w_1 - 4 \\
\frac{\partial Err^2}{\partial w_2} &= 0 + 2(w_2 + w_0 1) + 0 = 2w_0 + 2w_2 - 2
\end{aligned}
$$

Thus, we have to solve the following system of linear equations:

$$
\begin{array}{rrcrcrcl}
\text{I.} & 6w_0 & + & 2w_1 & + & 2w_2 & - \; 2 & = 0 \\
\text{II.} & 2w_0 & + & 2w_1 & & & - \; 4 & = 0 \\
\text{III.} & 2w_0 & + & & & 2w_2 & - \; 2 & = 0
\end{array}
$$

$$
\begin{array}{rcrcl}
\text{II.} \implies & 2w_1 = 4 - 2w_0 & \implies & w_1 = 2 - w_0 \\
\text{III.} \implies & 2w_2 = 2 - 2w_0 & \implies & w_2 = 1 - w_0
\end{array}
$$

By substituting $w_1 = 2 - w_0$ and $w_2 = 1 - w_0$ in equation I., we have:

$$\begin{aligned} 6w_0 + 2(2 - w_0) + 2(1 - w_0) - 2 &= 0 \\ 6w_0 + 4 - 2w_0 + 2 - 2w_0 - 2 &= 0 \\ 2w_0 + 4 &= 0 \\ 2w_0 &= -4 \\ w_0 &= -2 \end{aligned}$$

Moreover,

$$w_1 = 2 - w_0 = 2 - (-2) = 4 \qquad\qquad w_2 = 1 - w_0 = 1 - (-2) = 3$$

Hence, the analytical solution is:

$$\begin{aligned} w_0 &= -2 \\ w_1 &= 4 \\ w_2 &= 3 \end{aligned}$$

b) Gradient descent

For the sake of simplicity, let w be a column vector, containing the w_0, w_1 and w_2 coefficients.

$$w^{(0)} = \begin{pmatrix} w_0^{(0)} \\ w_1^{(0)} \\ w_2^{(0)} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

The gradient descent updating rule, where Err^2 is the same error function as in part a), namely the sum of squared erros

$$w^{(i+1)} = w^{(i)} - lrate \begin{pmatrix} \frac{\partial Err^2}{\partial w_0} \\ \frac{\partial Err^2}{\partial w_1} \\ \frac{\partial Err^2}{\partial w_2} \end{pmatrix} \Bigg|_{w^{(i)}}$$

First iteration:

$$w^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.25 \begin{pmatrix} 6w_0 + 2w_1 + 2w_2 - 2 \\ 2w_0 + 2w_1 - 4 \\ 2w_0 + 2w_2 - 2 \end{pmatrix} \Bigg|_{\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}$$

$$w^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.25 \begin{pmatrix} 6 + 2 + 2 - 2 \\ 2 + 2 - 4 \\ 2 + 2 - 2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - 0.25 \begin{pmatrix} 8 \\ 0 \\ 2 \end{pmatrix}$$

$$w^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ \frac{1}{2} \end{pmatrix}$$

Second iteration:

$$w^{(2)} = \begin{pmatrix} -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} - 0.25 \begin{pmatrix} 6 \cdot (-1) + 2 \cdot 1 + 2 \cdot \frac{1}{2} - 2 \\ 2 \cdot (-1) + 2 \cdot 1 - 4 \\ 2 \cdot (-1) + 2 \cdot \frac{1}{2} - 2 \end{pmatrix} =$$

$$= \begin{pmatrix} -1 \\ 1 \\ \frac{1}{2} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -5 \\ -4 \\ -3 \end{pmatrix} = \begin{pmatrix} -1 + \frac{5}{4} \\ 1 + 1 \\ \frac{1}{2} + \frac{3}{4} \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ 2 \\ \frac{5}{4} \end{pmatrix}$$

Third iteration:

$$w^{(3)} = \begin{pmatrix} \frac{1}{4} \\ 2 \\ \frac{5}{4} \end{pmatrix} - 0.25 \begin{pmatrix} 6 \cdot \frac{1}{4} + 2 \cdot 2 + 2 \cdot \frac{5}{4} - 2 \\ 2 \cdot \frac{1}{4} + 2 \cdot 2 - 4 \\ 2 \cdot \frac{1}{4} + 2 \cdot \frac{5}{4} - 2 \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{4} \\ 2 \\ \frac{5}{4} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} \frac{6}{4} + \frac{10}{4} + 2 \\ \frac{1}{2} \\ \frac{12}{4} - 2 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ 2 \\ \frac{5}{4} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 6 \\ \frac{1}{2} \\ 1 \end{pmatrix}$$

$$w^{(3)} = \begin{pmatrix} -\frac{5}{4} \\ \frac{15}{8} \\ 1 \end{pmatrix}$$

c) Stochastic gradient descent

Here the error function is calculated using only one data point in each iteration. Let's assume, that the stochastic gradient descent method first selects the first, then the second, and finally the third record.

First iteration (using the first record) First record: (0, 0, –2). Now the error function and its partial derivatives are the following:

$$Err^2 = (w_1 \cdot 0 + w_2 \cdot 0 + w_0 - (-2))^2 = (w_0 + 2)^2$$

$$\frac{\partial Err^2}{\partial w_0} = 2(w_0 + 2) \qquad \frac{\partial Err^2}{\partial w_1} = \frac{\partial Err^2}{\partial w_2} = 0$$

$$w^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - lrate \begin{pmatrix} 2(w_0 + 2) \\ 0 \\ 0 \end{pmatrix} \Bigg|_{w^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$$

$$w^{(1)} = \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 1 \end{pmatrix}$$

Second iteration (using the second record) Second record: (0, 1, 1). Now the error function and its partial derivatives are the following:

$$Err^2 = (w_1 \cdot 0 + w_2 \cdot 1 + w_0 - 1)^2 = (w_2 + w_0 - 1)^2$$

$$\frac{\partial Err^2}{\partial w_0} = 2(w_2 + w_0 - 1) \qquad \frac{\partial Err^2}{\partial w_1} = 0 \qquad \frac{\partial Err^2}{\partial w_2} = 2(w_2 + w_0 - 1)$$

$$w^{(2)} = \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 2\left(1 + \left(-\frac{1}{2}\right) - 1\right) \\ 0 \\ 2\left(1 + \left(-\frac{1}{2}\right) - 1\right) \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -\frac{1}{2} + \frac{1}{4} \\ 1 \\ 1 + \frac{1}{4} \end{pmatrix}$$

$$w^{(2)} = \begin{pmatrix} -\frac{1}{4} \\ 1 \\ \frac{5}{4} \end{pmatrix}$$

Third iteration (using the third record) Third record: (1, 0, 2). Now the error function and its partial derivatives are the following:

$$Err^2 = (w_1 \cdot 1 + w_2 \cdot 0 + w_0 - 2)^2 = (w_1 + w_0 - 2)^2$$

$$\frac{\partial Err^2}{\partial w_0} = 2(w_1 + w_0 - 2) \qquad \frac{\partial Err^2}{\partial w_1} = 2(w_1 + w_0 - 2) \qquad \frac{\partial Err^2}{\partial w_2} = 0$$

$$w^{(3)} = \begin{pmatrix} -\frac{1}{4} \\ 1 \\ \frac{5}{4} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 2\left(1 + \left(-\frac{1}{4}\right) - 2\right) \\ 2\left(1 + \left(-\frac{1}{4}\right) - 2\right) \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} \\ 1 \\ \frac{5}{4} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -2 - \frac{1}{2} \\ -2 - \frac{1}{2} \\ 0 \end{pmatrix} = \begin{pmatrix} -\frac{1}{4} \\ 1 \\ \frac{5}{4} \end{pmatrix} - \frac{1}{4} \begin{pmatrix} -\frac{5}{2} \\ -\frac{5}{2} \\ 0 \end{pmatrix} =$$

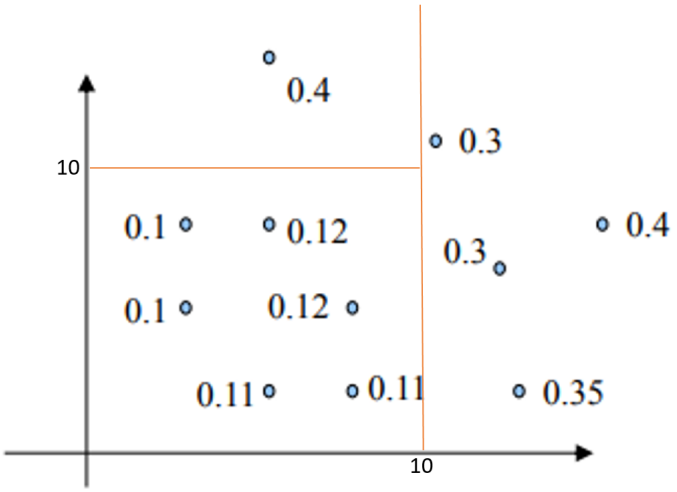
$$w^{(3)} = \begin{pmatrix} \frac{3}{8} \\ \frac{13}{8} \\ \frac{5}{4} \end{pmatrix}$$

Ex 2

We would like to solve a regression problem using a decision tree. How would it split the data given in the coordinate system below? The maximum number of leaves is set to 3. Sketch the splits on the figure. (No precise calculation is required.)

Solution

The goal of a decision tree regressor (the splitting criterion) is to minimize the variance in the emerging child notes. A possible solution is the following: first, it splits at $x = 10$. Then, it splits the $x < 10$ child node further at $y = 10$. At this point there are three leaves, so the algorithm terminates. The areas corresponding to the three leaves are bounded by the orange lines.



Ex 3

Consider the following data, where the first and the second coordinates are binary attributes, and the third is the class label: (1, 1, –); (1, –1, +); (–1, 1, +); (–1, –1, –). Which Boolean function do you recognize in the data? Is it linearly separable? If so, give the equation of the separating line with maximal margin. If the function is not linearly separable, then transform the data into the following three-dimensional feature space: $(x_1, x_2, x_1 \cdot x_2)$. Furthermore, find the equation of the separating plane with the maximum margin in the transformed space!

Solution

Note that these data points represent the XOR function, which is not linearly separable. After the data transformation we get the following points: (1, 1, 1, –); (1,–1, –1, +); (–1, 1, –1, +); (–1, –1, 1, –) in the xyz three-dimensional space. Note that the third (z) coordinate is –1 of all the records with '+' label, moreover the z coordinate is +1 of all the instances with '–' class label. Thus, in the three-dimensional space, these points (namely the XOR function) are linearly separable. The separating plane is the xy plane, which is described by the $z = 0$ equation.

Ex 4

Represent the following logical functions with a perceptron or show that it is not possible to do so. In the latter case, construct a neural network with one hidden layer.

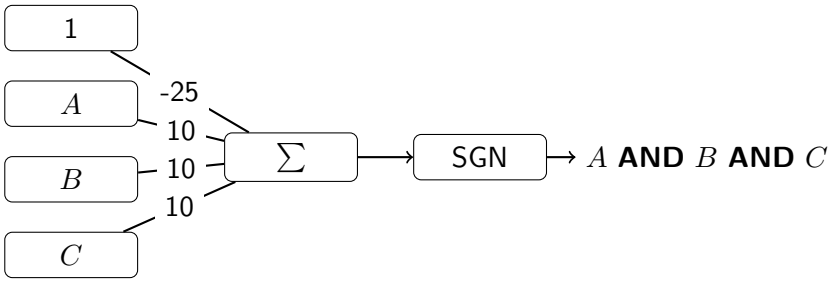
a) A **AND** B **AND** C

b) (A **XOR** B) **AND** (A **OR** B)

Solution

b)

$y = 10A + 10B + 10C - 25 \cdot 1$. The activation function takes 1, if the input is positive, and 0 otherwise.

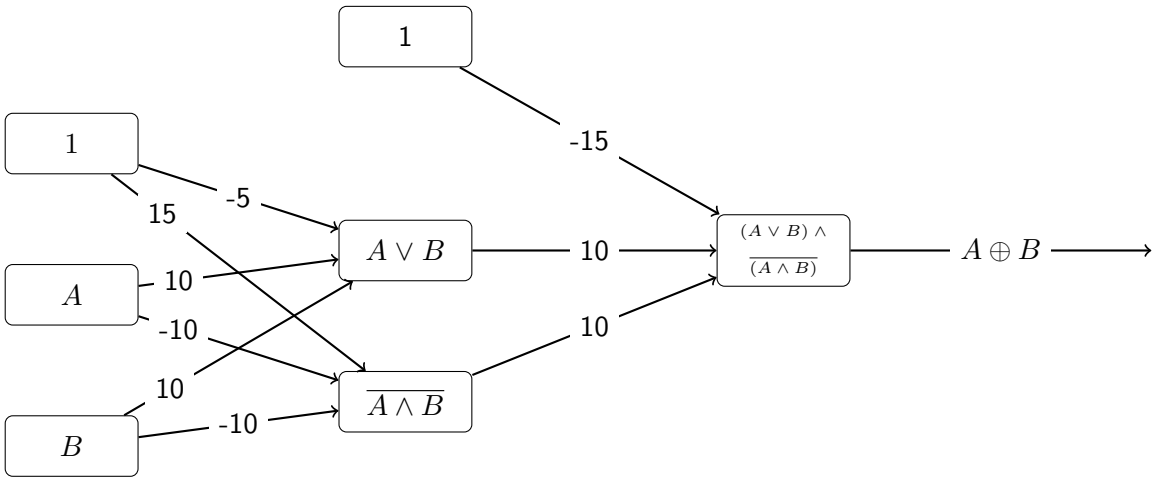


b)

Note that (A **XOR** B) **AND** (A **OR** B) = A **XOR** B.
The definition of XOR is:

$$A \oplus B = (A \vee B) \wedge \overline{(A \wedge B)}.$$

Hence, in the first layer we calculate $(A \vee B)$ and $\overline{(A \wedge B)}$, then in the second layer we calculate the calculate the conjunction of these:

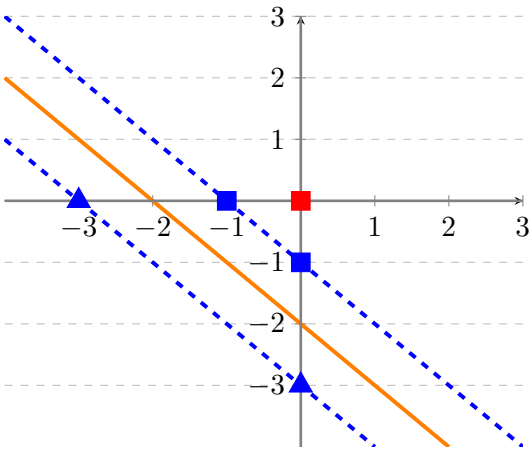


Ex 5

How does the $2 + x + y = 0$ line separates the 2-dimensional plane? Draw the line in a coordinate system. Which of the following records are support vectors of the given line: (–3, 0); (0, –3); (–1, 0); (0, –1); (0, 0)?

Solution

$2 + x + y = 0 \implies y = -x - 2$
In the figure the orange line is the separating line. The blue dashed lines are the margins, and the blue points are the support vectors (the two different classes are denoted by the squares and triangles).



Ex 6

Consider a chain of two neurons. The input of the first neuron is x_1 and the output is $y_1 = ax_1 + b$. The input of the other neuron is x_2 and the output is $y_2 = cx_2 + d$ (the activation function is the identity function in both cases). Connect the two neurons so that the second neuron's input is y_1, i.e. $x_2{=}y_1$.

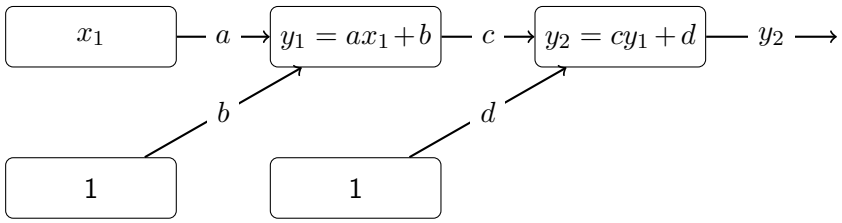
a) Draw this neural network!

b) Give the final output y_2 as a function of x_1!

c) Let the input of the ANN be x and the output be y. Using the gradient descent method, show how the weights (a, b, c, d) are updated after one training step if the squared error is minimized.

Solution

a)



b)

$$y_2 = cy_1 + d = c(ax_1 + b) + d = cax_1 + (cb + d)$$

c)

We have to minimize the following F error function:

$$F = (y_2 - y)^2$$

Gradient descent updating rule:

$$\frac{\partial F}{\partial d} = 2(y_2 - y) \Longrightarrow d = d - lrate \cdot 2(y_2 - y)$$
$$\frac{\partial F}{\partial c} = 2(y_2 - y)y_1 \implies c = c - lrate \cdot 2(y_2 - y)y_1$$
$$\frac{\partial F}{\partial b} = 2(y_2 - y)c \implies b = b - lrate \cdot 2(y_2 - y)c$$
$$\frac{\partial F}{\partial a} = 2(y_2 - y)cx_1 \implies a = a - lrate \cdot 2(y_2 - y)cx_1$$