CSC 522: Automated Learning And Data Analysis
Homework 3
Team: Srujana Rachakonda (srachak), Rajshree Jain (rjain27)
Group - G41

Q1. BN Inference

a. P(C) = P(A,C) + P(~A,C)
= P(A)* P(C|A) + P(~A)*P(C|~A)
= 0.6*0.5 + 0.4*0.4 = 0.46
b. P(D | B, ~C) = P(D = true | B = true, C = false)
Since, D is independent of C, this is equivalent to P(D = true | B = true)
= 0.2
c. P(A, B, ~C, D, E, F) = P(A = true, B = true, C = false, D = true, E = true, F= true)
= P(A) * P(B) * P(~C | A) * P(D | B) * P(E | ~C , D) * P(F | ~C)
= 0.6 * 0.4 * 0.5 * 0.2 * 0.1 * 0.6
= 0.00144
d. Given C, E and F are conditionally independent using Markov's Condition that a node is conditionally independent of it's non-descendents given the value of its parents.
e. A and B are marginally dependent since they're not connected to each other and the probability of A does not depend on B and vice versa. However, they might become dependent if the value of E is known.
f. E is dependent only on C and D. Since it's given that C = false and D = false. P(E = true | C = true, D = false) = 0.5 and P(E = false | C = true, D = false) = 1 - 0.5 = 0.5. Therefore, we can say that E is equally likely to be true or false.

Q2. Linear Regression

Linear Regression Equation: $y = w_1x^3 + w_0$

Given points are of the form (x,y)
We will consider x^3 to be x' and use x' in place of x in the computation of w_0 and w_1.

All the values are tabulated as follows:

X	**X'(X^3)**	**Y**	**X'Y**	**$(X')^2$**	**$w_1X'_i$**	**Yi - w1Xi**	**Y'**
2	8	5	40	64	4.35	0.65	3.2126
0	0	-2	0	0	0	-2	-1.1362
1	1	-3	-3	1	0.5436	-3.5436	-0.5926
-2	-8	-4	32	64	-4.35	0.3488	-5.485

SUM =	1	-4	69	129		-4.5448	

W1 = (n * sum(x'_i * y_i) - sum(x_i) * sum(y_i)) / (n * sum(x_i^2) - $(\text{sum}(x_i))^2$)
= 0.5436

W0 = sum(y_i - w_1x_i) / n = -4.5448 / 4 = -1.1362

RMSE = sqrt(sum$(y_i' - y_i)^2$ / n)
= sqrt((3.194+ 0.746 + 5.795 + 2.20) / 4) = sqrt(11.940/4) = 1.7277

Q3.

You are given the above (Figure 2) neural network with continuous input attributes X1 and X2 and continuous output variable Y . For clarity, the relationship between weights and activations is also shown in Table 1. All three activations H1, H2 and O use the linear activation function f(z) = Mz, with constant M = 1. Initial weights are as given in Figure 2 and repeated in Table 1. There is no bias (w0) added to any of the units. Answer the following:

(a) Forward Pass: If you are given one training data point: X1i = 1, X2i = −1, and Yi = 1. Compute the activations of the neurons H1, H2 and O.

w1 = 0.3,w2 = 0.6,w3 = 0.9,w4 = 0.2,w5 = 0.4,w6 = −0.1

It is given in the question that the 3 activations use the same activation function f(z) = Mz.

The value of M is 1 so the activation function will be f(z) = z.

The inputs that are given are:

X1i=1
X2i=-1

The value that will come inside the neuron would be :

$$z = \sum x_i * w_i$$

In the above formula w_i is the weight incoming the node and x_i will be the corresponding inputs in the node.

The output is Yi=1

The activation of Neuron H1:

$z = X1i * w1 + X2i * w3$

$z = 1 \; * \; 0.3 + \; -1 \; * 0.9$

$z = \; -0.6$

Therefore the activation of the node H1 is a_{H1} which will have a value of -0.6.

The activation of Neuron H2:

$z = X1i * w2 + X2i * w4$
$z = 1 * \; 0.6 \; + \; -1 * \; 0.2$

$z = 0.4$

Therefore the activation of the node H2 is a_{H2} which will have a value of 0.4

The activation of Neuron O:

$z = \; a_{H1} * w5 + \; a_{H2} * w6$

$z = \; -0.6 * \; 0.4 \; + \; 0.4 * -0.1$

$z = \; -0.28$

Therefore the activation of node O is $a_o =$ -0.28

Hence the final activation values for all the 3 neurons are:

$a_{H1} \rightarrow$ -0.6

$a_{H2} \rightarrow$ 0.4

$a_o \rightarrow$ -0.28

b) Backward Pass: At the end of forward pass, using the current training instance i: X1i = 1, X2i = −1, and Yi = 1, calculate the updated value of each of the following weights after one iteration of backpropagation:

• For CSC 522: w1, w5 and w6

• For CSC 422: w5 and w6 (w1 is optional extra credit)

Use 0.1 as your learning rate and MSE (mean squared error) as your cost function. Show your work on the following steps for each weight, w (w1, w5, w6):

i) Let aN be the activation at neuron N, X1i be the value of the attribute X1 for instance i, and Yi be the actual class of the instance i. Write equations to define the following:

A. The cost function C in terms of Yi and aO

The cost function C in terms Y_i and a_o will be the value of the Mean Squared Error.

The mean Squared error will be : C(w) = (y - f(X, w))^2

Now if we write in terms of Y_i and a_o then:

$$C = \sum(Y_i - a_o)^2$$

We can see that in the above formula Y_i is the expected output and a_o is the output that comes from the neuron O that is predicted.

B. The activation of the final layer aO in terms of second layer weights w5, w6 and the activation of the first layer aH1 and aH2 ---->

The answer: $a_0 = a_{H1} * w5 + a_{H2} * w6$

$$z = \sum x_i * w_i$$

$$a_o = f(z) = Mz = 1.(\ a_{H1} * w5 + a_{H2} * w6)$$

$$a_0 = a_{H1} * w5 + a_{H2} * w6$$

C. The activation of the node aH1 in terms of inputs X1i, X2i and weights w1 and w3

$$z = \sum x_i * w_i$$

$$a_{H1} = f(z) = Mz = 1.(\ X1i * w1 + X2i * w3\ )$$

$$a_{H1} = X_i * w1 + X2_i * w3$$

Or we can also write for simplicity:

$$a_{H1} = x_1 * w1 + x_2 * w3$$

ii. For layer-2 weights, calculate δC/δao and δao/δw . Here C is the cost function, ao is the activation at node O, and w is the weight.

$$\frac{\delta C}{\delta a_0} = \frac{\delta((Y_i - a_o)^2)}{\delta a_o} = 2(Y_i - a_o) * (-1) = 2 * (1 - (-0.28)) * (-1) = -2.56$$

$\frac{\delta a_o}{\delta w} : \frac{\delta a_o}{\delta z} * \frac{\delta z}{\delta w}$, since here $a_0 = f(z) = Mz = 1(a_{H1} * w_5 + a_{H2} * w_6)$
Hence, we can directly differentiate in terms of w.

The weights associated with this layer are w_5 and w_6

$$\frac{\delta a_o}{\delta w_5} = \frac{\delta(a_{H1} * w5 + a_{H2} * w6)}{\delta w_5} = a_{H1} = \text{-0.6}$$

$$\frac{\delta a_o}{\delta w_6} = \frac{\delta(a_{H1} * w5 + a_{H2} * w6)}{\delta w_6} = a_{H2} = 0.4$$

iii) For layer-1 weights, calculate $\frac{\delta C}{\delta a_0}$, $\frac{\delta a_o}{\delta a_{H1}}$, and $\frac{\delta a_{H1}}{\delta w}$.

$$\frac{\delta C}{\delta a_0} = \frac{\delta((Y_i - a_o)^2)}{\delta a_o} = 2(Y_i - a_o) * (-1) = 2 * (1 - (-0.28)) * (-1) = -2.56$$

$\frac{\delta a_o}{\delta a_{H1}} : \frac{\delta a_o}{\delta z} * \frac{\delta z}{\delta a_{H1}}$
$a_0 = f(z) = Mz = 1(a_{H1} * w_5 + a_{H2} * w_6)$, So we will differentiate directly wrt a_{H1}

$$\frac{\delta a_o}{\delta a_{H1}} = \frac{\delta(a_{H1} * w5 + a_{H2} * w6)}{\delta a_{H1}} = w_5 = 0.4$$

$\frac{\delta a_{H1}}{\delta w} : \frac{\delta a_{H1}}{\delta z} * \frac{\delta z}{\delta w}$
$a_{H1} = f(z) = Mz = 1.(x_1 * w1 + x_2 * w3)$, So we will differentiate directly wrt w

$$\frac{\delta a_{H1}}{\delta w} = \frac{\delta(x_1 * w1 + x_2 * w3)}{\delta w}$$

First with respect to w_1 :

$$\frac{\delta a_{H1}}{\delta w_1} = \frac{\delta(x_1 * w_1 + x_2 * w_3)}{\delta w_1} = x_1 = 1$$

Then with respect to w_3 :

$$\frac{\delta a_{H1}}{\delta w_3} = \frac{\delta(x_1 * w_1 + x_2 * w_3)}{\delta w_3} = x_2 = \text{-1}$$

iv) Calculate $\frac{\delta C}{\delta w}$ using the above values.

First calculate $\frac{\delta C}{\delta w_1}$

$\frac{\delta C}{\delta w_1} = \frac{\delta C}{\delta a_o} * \frac{\delta a_0}{\delta a_{H1}} * \frac{\delta a_{H1}}{\delta w_1}$ = -2.56*0.4*1=-1.024

Then calculating $\frac{\delta C}{\delta w_5}$

$\frac{\delta C}{\delta w_5} = \frac{\delta C}{\delta a_o} * \frac{\delta a_0}{\delta w_5}$ = -2.56*-0.6=1.536

Then calculating $\frac{\delta C}{\delta w_6}$

$\frac{\delta C}{\delta w_6} = \frac{\delta C}{\delta a_o} * \frac{\delta a_0}{\delta w_6}$ = -2.56*0.4=-1.024

v) Calculate the updated weight w′ using the $\frac{\delta C}{\delta w}$ and the learning rate

For w_1

$w_{1new} = w_1 - Rate * \frac{\delta C}{\delta w_1} =$ 0.3 - (0.1)*(-1.024) = 0.4024

For w_5

$w_{5new} = w_5 - Rate * \frac{\delta C}{\delta w_5} =$ 0.4 - (0.1)*(1.536) = 0.2464

For w_6

$w_{6new} = w_6 - Rate * \frac{\delta C}{\delta w_6} =$ -0.1 - (0.1)*(-1.024) = 0.0024

Q4. PART-1
V. Comparison:

Using linear regression we get a root mean square error value of 1300.289 and using lasso regression we get a root mean square error value of 321.3975. Since our aim is to minimize the error, lasso regression shows significant improvement over linear regression and gives a better bit.

A. All 19 variables are not useful for predicting the dependent variable. Because Lasso regression forces some of the coefficients to reduce to zero. And in this case, according to the outcome of the coef() function 16 out of the 19 attributes have zero coefficients and hence do not contribute towards the prediction of the dependent variable.
B. Attributes x6, x10 and and x12 are to be selected if lasso regression is used for feature selection.