Neural networks 2018 Fall
paper-based closed room test
REPLACEMENT

PPCU-ITK

3rd December, 2018

Name: Ekart CYBA

NEPTUN ID: ZWPMKP

Question	Points	Student's points
1	12	12
2	12	8
3	12	12
4	30	30
Total	66	62

94 %

Instructions:

1. This examination contains 6 pages, including this page.
2. You have **ninety (90) minutes** to complete the examination. As a courtesy to your classmates, we ask that you not leave during the last fifteen minutes.
3. You may not use any external resources, including lecture notes, books, other students or other engineers.
4. You may use a calculator. You may not share a calculator with anyone.
5. Please sign the below statement.

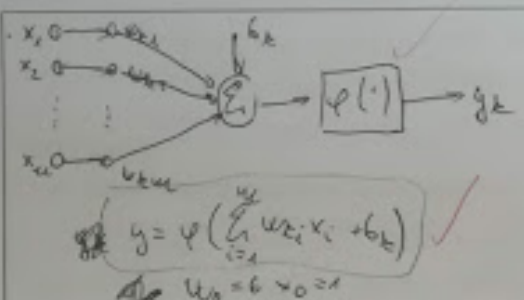
I hereby certify that I will neither give nor receive unpermitted aid on this examination.

Signature: Ekart Cyba

Question 1: The artificial neuron

Overall: [12 pts]

(a) [5 pts] Please write down the working steps of an artificial neuron! (What happens when the perceptron receives input(s) through its synapse?)



$$y = \varphi\left(\sum_{i=1}^{n} w_{ki} x_i + b_k\right)$$

$w_0 = b \quad x_0 = 1$

or

$$y = \varphi\left(\sum_{i=0}^{n} w_{ki} x_i\right) = w^T x$$

0. receives input vector on the synapse
1. multiply it with the corresponding weights
2. sum the values
3. add the bias (for non-linearity)
4. apply activation function on the whole (also for non-linearity)
5. output

5

(b) [3 pts] State the perceptron convergence theorem! (No proof required.)

The perceptron convergence theorem states that the Perceptron Learning Algorithm a.k.a. Rosenblatt algorithm converges in finite number of steps for a linearly seperable dataset.

3

(c) [4 pts] Write down the output equation of a multi-layer perceptron (a.k.a. fully connected neural network)!

$$NN(x, w) = \varphi^{(L)}\left(w^{(L)} \varphi^{(L-1)}\left(w^{(L-1)}\left(\varphi^{(L-2)}\left(\dots \varphi^{(2)}\left(w^{(2)}\left(\varphi^{(1)}\left(w^{(1)} x\right)\right)\right)\right)\right)\right)$$

4

12

Question 2: Regularization techniques

Overall: [12 pts]

(a) [4 pts] Please describe the dropout normalization technique in details!

Uses mini-batch approach, and for each minibatch select a random subset of neurons which will be deactivated in the training
Selection & deactivation probability : p ✓
When testing multiply the outputs with p to account the deactivated activation

4

(b) [5 pts] What is local response normalization? Please give your answer in detail.

~~[illegible]~~
~~(batch-norm, input vector normalization, [illegible] ([illegible]))~~
We want to normalize to avoid overfitting & make the network less sensitive to errors.
~~[illegible]~~ We learned about input vector norm, batch norm. and weight norm.
In normalization processes we want to squeeze the range, where these values exist.

Input vector norm:
x : input vector
$\bar{x}$: mean
σ : deviation

$$x_{norm} = \frac{x - \bar{x}}{\sigma}$$

L_1 & L_2 norm

$$C_0 = \frac{1}{2n} \sum_x \|y - a^L\|^2 \qquad C_2 = \frac{1}{2n} \sum_x \|y - a^L\|^2 + \frac{\lambda}{2n} \sum_w w^2$$

$$C_1 = \frac{1}{2n} \sum_x \|y - a^L\|^2 + \frac{\lambda}{n} \sum_w |w|$$

Batch norm

$$\hat{x} = \frac{x - E[x]}{\sqrt{Var[x]}}$$

E: expected
Var: variability

YES THIS IS WHY WE USE NORMALIZATION FE

1

$$b^i_{x,y} = a^i_{x,y} / \left(k + \alpha \cdot \sum a^j_{x,y}\right)$$

(c) [3 pts] Write 3 common loss functions!

Quadratic-loss function: ✓

$$R(w) = \frac{1}{K} \sum_{k=1}^{K} (d_k - Net(x_k, w))^2$$

Negative-log likelihood: ✓

$$L(w) = \frac{1}{K} \sum_{k=1}^{K} -\log(P(y_k | x_k, w))$$

Cross entropy: ✓

$$C = -\frac{1}{n} \sum_x [y \ln a + (1-y) \ln(1-a)]$$

$$C(w) = -\frac{1}{K} \sum_{k=1}^{K} [d_k \log P(y_k | x_k, w) + (1 - d_k) \cdot \log(1 - P(y_k | x_k, w))]$$

3

(3)

Question 3: Optimization techniques

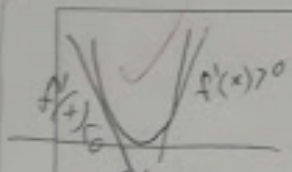
Overall: [12 pts]

(a) [3 pts] What is the mathematical definition of the Newton method (first order gradient based optimization method)?

$$\Delta x = -H(f(x[u]))^{-1} \nabla f(x(u)) \quad x(u+1) = x(u) - \eta H(f(x(u)))^{-1} \nabla f(x(u))$$

3

(b) [5 pts] Write down the general idea of gradient descent in optimization problems!



→ Goal: find local minima
→ Start from a random point and calculate the gradient of the function.
↳ always follow the descending gradient.
→ follows the delta-rule

5

(c) [4 pts] Explain the minibatch approach in gradient-based optimization algorithms!

3 main approaches: instant update, batch, minibatch
If we use the minibatch approach, it mean, that we select a random set from the input vectors, calculate the updates, then take their average and apply.

↳ update weights!

4

if $f'(x) < 0$ we should increase x to get closer to the min.
if $f'(x) > 0$ we should decrease x to get closer to the min.

m_b: size of minibatch usually ≈ 100

$$e = \frac{1}{m_b} \sum_{k=1}^{m_b} [d_k - y_k(x,w)]^2$$

12

Question 4: Mathematical methods in practice

Overall: [30pts]

(a) [12pts] Apply the Rosenblatt perceptron training algorithm to learn the implication logical function ($f(x,y) = \neg x \vee y$ (NOT x or Y)) with a single perceptron! Validate your results graphically! Please show your work.

x_1	x_2	$f(x_1,x_2)$
-1	-1	1
-1	1	1
1	-1	-1
1	1	1

$\varphi(h) = sign(\cdot)$

$\eta = \frac{1}{2}$

$w(0) = 0$

$y(k) = \varphi(w^T x(k))$

$k = 1$

① $w(k+1) = w(k) + \eta\, x(k)$

② $w(k+1) = w(k) - \eta\, x(k)$

① $y(1) = sign\left(\begin{bmatrix}0 & 0 & 0\end{bmatrix}\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right) = 0$

$\varepsilon(1) = d_1 - y(1) = 1 - 0 = 1 \rightarrow$ ① $\rightarrow w(1) = \begin{bmatrix}0\\0\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\-1\\-1\end{bmatrix} = \begin{bmatrix}1/2\\-1/2\\-1/2\end{bmatrix}$

② $y(2) = sign\left(\begin{bmatrix}1/2 & -1/2 & -1/2\end{bmatrix}\begin{bmatrix}1\\-1\\1\end{bmatrix}\right) = 1$

$\varepsilon(2) = d_2 - y(2) = 1 - 1 = 0 \Rightarrow$ OK

③ $y(3) = sign\left(\begin{bmatrix}1/2 & -1/2 & -1/2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right) = -1$

$\varepsilon(3) = d_3 - y(3) = -1 - (-1) = 0 \Rightarrow$ ②

$\hookrightarrow w(4) = \begin{bmatrix}1/2\\-1/2\\-1/2\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\-1\\-1\end{bmatrix}$

④ $y(4) = sign\left(\begin{bmatrix}0 & -1 & -1\end{bmatrix}\begin{bmatrix}1\\-1\\1\end{bmatrix}\right) = 1$

$\varepsilon(4) = d_4 - y(4) = -1 - 1 = -2 \Rightarrow$ ②

$w(5) = \begin{bmatrix}0\\-1\\-1\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\-1\\1\end{bmatrix} = \begin{bmatrix}-1/2\\-1/2\\-3/2\end{bmatrix}$

⑤ $y(5) = sign\left(\begin{bmatrix}-1/2 & -1/2 & -3/2\end{bmatrix}\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right) = 1$

$\varepsilon(5) = d_5 - y(5) = 1 - 1 = 0 \Rightarrow$ OK

⑥ $y(6) = sign\left(\begin{bmatrix}-1/2 & -1/2 & -3/2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right) = -1$

$\varepsilon(6) = d_6 - y(6) = 1 - (-1) = 2 \Rightarrow$ ① $\rightarrow w(6) = \begin{bmatrix}-1/2\\-1/2\\-3/2\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\0\\-1\end{bmatrix}$

⑦ $y(7) = sign\left(\begin{bmatrix}0 & 0 & -1\end{bmatrix}\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right) = 1$

$\varepsilon(7) = d_7 - y(7) = -1 - 1 = -2 \Rightarrow$ ②

$w(7) = \begin{bmatrix}0\\0\\-1\end{bmatrix} - \frac{1}{2}\begin{bmatrix}1\\-1\\-1\end{bmatrix} = \begin{bmatrix}-1/2\\1/2\\-1/2\end{bmatrix}$

⑧ $y(8) = sign\left(\begin{bmatrix}-1/2 & 1/2 & -1/2\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right) = -1$

$\varepsilon(8) = d_8 - y(8) = 1 - (-1) = 2 \Rightarrow$ ① $\rightarrow w(8) = \begin{bmatrix}-1/2\\1/2\\-1/2\end{bmatrix} + \frac{1}{2}\begin{bmatrix}1\\1\\1\end{bmatrix} = \begin{bmatrix}0\\1\\0\end{bmatrix}$

⑨ $y(9) = sign\left(\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right) = 1$

$\varepsilon(9) = d_9 - y(9) = 1 - 1 = 0 \Rightarrow$ OK

⑩ $y(10) = sign\left(\begin{bmatrix}0 & 1 & 0\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right) = 1$

$\varepsilon(10) = d_{10} - y(10) = 1 - 1 = 0 \Rightarrow$ OK

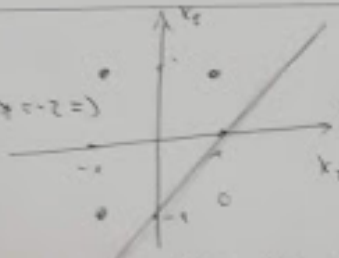
$\varepsilon(11) = 0 \Rightarrow$ OK

12

(b) [2pts] If we are able to learn this logical function using a single perceptron, what property of the learned function does this imply?

linear separability

2



$x_2 = x_1 - 1$

$w^T x = 0$

$0 = \begin{bmatrix}1/2\\-1/2\\1/2\end{bmatrix}^T \begin{bmatrix}1\\x_1\\x_2\end{bmatrix} = \frac{1}{2} - \frac{1}{2}x_1 + \frac{1}{2}x_2$

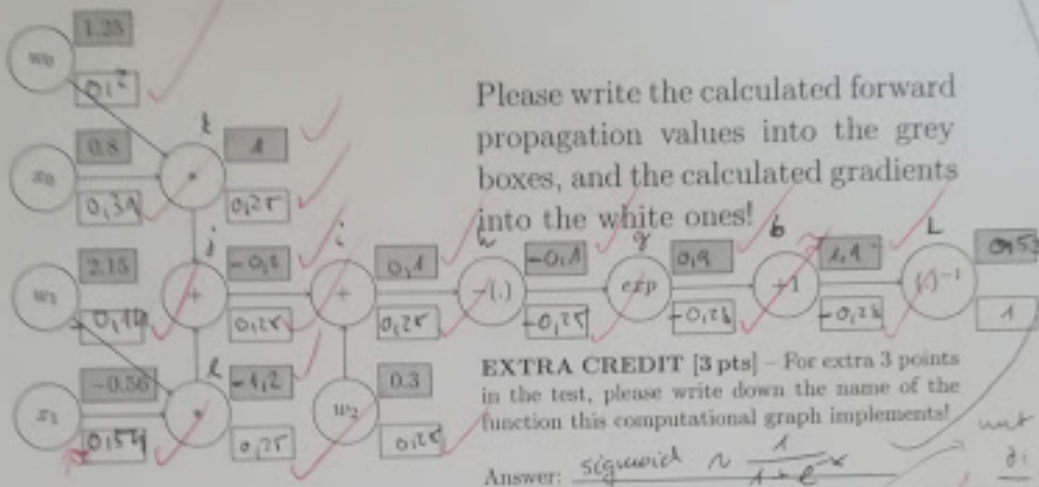
$-\frac{1}{2}x_1 + \frac{1}{2}x_2 + \frac{1}{2} = 0$

14

$$= (w_0 x_0 + w_1 x_1 + w_2) \qquad + 1$$

(e) [16 (13+3) pts] Calculate the gradient for each node of the computational graph below! Please show your work! Also, here is some help with derivation :)

$$f(x) = e^x \to \frac{df}{dx} = e^x \qquad f(x) = ax \to \frac{df}{dx} = a$$
$$f(x) = \frac{1}{x} \to \frac{df}{dx} = \frac{-1}{x^2} \qquad f(x) = c + x \to \frac{df}{dx} = 1$$



Please write the calculated forward propagation values into the grey boxes, and the calculated gradients into the white ones!

EXTRA CREDIT [3 pts] – For extra 3 points in the test, please write down the name of the function this computational graph implements!

Answer: sigmoid $\sim \frac{1}{1+e^{-x}}$

$\frac{\partial L}{\partial L} = 1$

$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial L} \cdot \frac{\partial L}{\partial b} = 1 \cdot \left(-\frac{1}{(1,9)^2}\right) = -0,28$

$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial L} \cdot \frac{\partial L}{\partial b} \cdot \frac{\partial b}{\partial y} = (-0,28) \cdot 1 = -0,28$

$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial L} \cdot \frac{\partial L}{\partial b} \cdot \frac{\partial b}{\partial y} \cdot \frac{\partial y}{\partial w} = (-0,28) \cdot e^{-0,1} = -0,25$

$\frac{\partial L}{\partial i} = \frac{\partial L}{\partial L} \cdot \frac{\partial L}{\partial b} \cdots \frac{\partial y}{\partial w} \cdot \frac{\partial w}{\partial i} = (-0,28) \cdot (-1) = 0,25$

$\frac{\partial L}{\partial j} = \frac{\partial L}{\partial L} \cdots \frac{\partial k}{\partial i} \cdot \frac{\partial i}{\partial j} = 0,25 \cdot 1 = 0,25$

$\frac{\partial L}{\partial w_c} = \frac{\partial L}{\partial L} \cdots \frac{\partial i}{\partial w_c} = 0,25 \cdot 1 = 0,25$

$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial L} \cdots \frac{\partial j}{\partial e} = 0,25 \cdot 1 = 0,25$

$\frac{\partial L}{\partial k} = \frac{\partial L}{\partial L} \cdots \frac{\partial l}{\partial k} = 0,25 \cdot 1 = 0,25$

$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial L} \cdots \frac{\partial k}{\partial w_0} = 0,25 \cdot 0,8 = 0,2$

$\frac{\partial L}{\partial w_0} = \frac{\partial L}{\partial L} \cdots \frac{\partial e}{\partial w_0} = 0,25 \cdot (-0,56) = -0,14$

$\frac{\partial L}{\partial x_0} = \frac{\partial L}{\partial L} \cdots \frac{\partial k}{\partial x_0} = 0,25 \cdot 1,25 = 0,31$

$\frac{\partial L}{\partial x_1} = \frac{\partial L}{\partial L} \cdots \frac{\partial j}{\partial e} \cdot \frac{\partial e}{\partial x_1} = 0,25 \cdot 2,15 = 0,54$

16