SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

1. (10pts) True or false questions (+2pts for correct answer, -2pt penalty for incorrect answer)

(a) Stochastic gradient descent is always guaranteed to converge to the global optimum of a loss function. (True / False)

F no guarantee.

(b) You train a model and you observe that the validation accuracy is much lower than the training accuracy. If you use batch normalization, the new model is likely to have a smaller gap between the train and validation accuracies (True / False)

regularization

T

(c) Deep neural networks typically have many hyperparameters that can be determined based on performance on the training dataset. (True / False)

F

(d) The momentum optimizer better avoids local minima by keeping running gradient statistics. (True / False)

T

기존 기울기 통계 유지
⇒ momentum.

(e) Dropout is a technique used to prevent overfitting by randomly dropping out some nodes during training. (True / False)

T

SEOULTECH
Seoul National University of Science & Technology

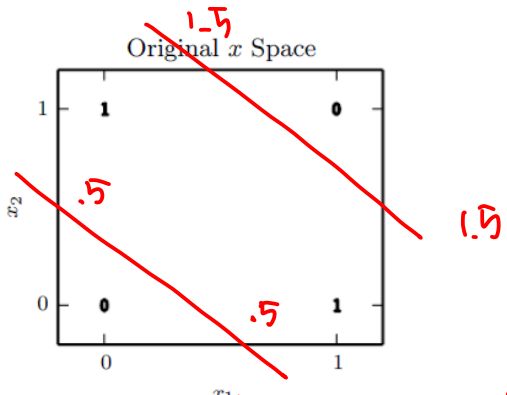
Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

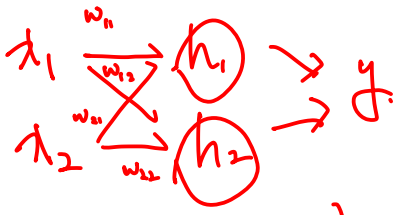
diff ⇒ true

2. (10pts) Given the following XOR problem, design a neural network that can classify these points. Consider a neural network that has a single hidden layer with two nodes and ReLU as an activation function for hidden nodes.



$x_1 + x_2 - 0.5 = 0$
$x_1 + x_2 - 1.5 = 0$

x_1 & x_2



x_1	x_2	h_1, h_2	y
0	0	0, 0	0
0	1	0.5, 0	1
1	0	0.5, 0	1
1	1	1.5 0.5	0

$$\text{ReLU}\left(\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}+\begin{bmatrix}-0.5\\ -1.5\end{bmatrix}\right)$$

$h_1 - 3h_2 + (-0.25)$

$(1, -3)$ -0.25

$$W = \begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} \quad b_h = \begin{pmatrix}-0.5\\ -1.5\end{pmatrix}$$

$$V = (1 \ -3) \quad b_y = (-0.25)$$

$\text{ReLU}(WX + b_h) = h$

$\text{ReLU}(Vh + b_y) = y$

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

3. (10pts) After training a neural network, you observe a large gap between the training accuracy (100%) and the test accuracy (50%). Provide at least three methods that can be used to reduce this gap, and explain why you choose them.

overfitting occured. regularization is required.

L2 : add squared sum of parameters to minimize loss function
avoid weights being too large. make weights small & even smooth model

L1 : add norm value of weights to loss function.
not important features param ⇒ 0.

dropout : randomly deactivate neurons for each iteration
prevent dependence on certain neurons.

early stopping : stopping model learning at optimal point
{ training error still decrease
{ validation error increase again at certain point
which is starting point of overfitting.
early stopping stops learning this point

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

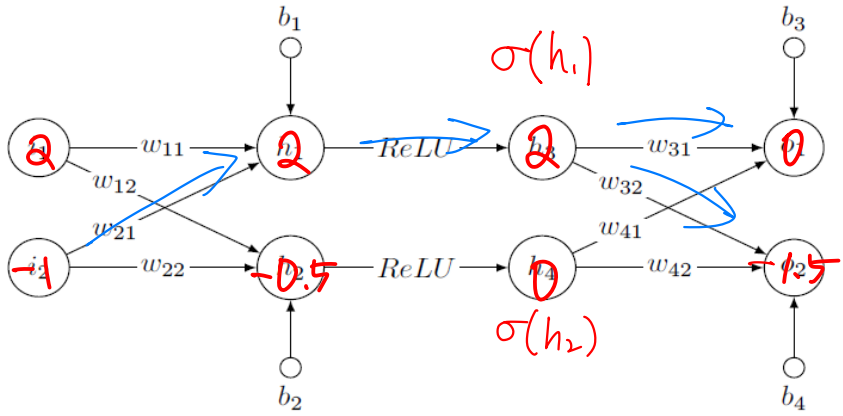
Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

4. (20pts) Given the following neural network with fully connected layers and ReLU activations, including two input units (i_1, i_2), four hidden units (h_1, h_2) and (h_3, h_4). The output units are indicated as (o_1, o_2) and their targets are indicated as (t_1, t_2). The weights and bias of fully connected layer are called w and b with specific sub-descriptors.

The values of variables are given in the following table.

Var.	i_1	i_2	w_{11}	w_{12}	w_{21}	w_{22}	w_{31}	w_{32}	w_{41}	w_{42}	b_1	b_2	b_3	b_4	t_1	t_2
Val.	2.0	-1.0	1.0	-0.5	0.5	-1.0	0.5	-1.0	-0.5	1.0	0.5	-0.5	-1.0	0.5	1.0	0.5



(a) (5pts) Compute the output (o_1, o_2) with the input (i_1, i_2) and network parameters as specified above. Write down all calculations including intermediate layer results.

$$\begin{bmatrix} 1 & 0.5 \\ -0.5 & -1 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix}+\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}=\begin{bmatrix} 2 \\ -0.5 \end{bmatrix}$$

$$\begin{bmatrix} 0.5 & -0.5 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ 0 \end{bmatrix}+\begin{bmatrix} -1 \\ 0.5 \end{bmatrix}=\begin{bmatrix} 0 \\ -1.5 \end{bmatrix}$$

$(0, -1.5)$

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

(b) (5pts) Compute the mean squared error of the output (o_1, o_2) calculated above and the target (t_1, t_2). $MSE = \frac{1}{n}\sum_i (t_i - o_i)^2 = \frac{1}{2}\left((1-0)^2 + (-1.5-0.5)^2\right) = \frac{5}{2}$

$E_i = \frac{1}{2}(t_i - o_i)^2$

(c) (5pts) Compute the gradient of the mean squared error with respect to w_{ij}.

$\frac{dE}{dw_{11}} = \frac{dE_1}{dO_1}\frac{dO_1}{dh_1}\frac{dh_1}{dw_{11}} + \frac{dE_2}{dO_2}\frac{dO_2}{dh_2}\frac{dh_2}{dw_{11}} = -(1-0)\times 0.5 \times 1 \times 2 - (-1.5-0.5)\times(-1)\times 1\times 2$

$= -1 - 4 = -5$

$\frac{dE}{dw_{12}} = 0$

$\frac{dE}{dw_{21}} = -(1-0)\times 1\times 1\times -1 \; -(-1.5-0.5)\times 0.5 \times 1 \times (-1)$

$= -2?$

$\frac{dE}{dw_{22}} = 0$

$\frac{dE}{dw_{31}} = \frac{dE_1}{dO_1}\cdot\frac{dO_1}{dw_{31}} = -(1-0)\cdot 2 = -2$

$\frac{dE}{dw_{32}} = \frac{dE_2}{dO_2}\cdot\frac{dO_2}{dw_{32}} = -(-1.5-0.5)\cdot 2 = 4$

$\frac{dE}{dw_{41}} = \frac{dE_1}{dO_1}\frac{dO_1}{dw_{41}} = -(1-0)\cdot 0 = 0$

$\frac{dE}{dw_{42}} = 0$

(d) (5pts) Update the weights using gradient descent with learning rate 0.1.

$w_{11} = w_{11} - 0.1\times(-5) = 1 + 0.5 = 1.5$

$w_{12} = w_{12} - 0.1\times 0 = -0.5$

$w_{21} = w_{21} - 0.1\times(-2) = 0.5 + 0.2 = 0.7$

$w_{22} = w_{22} - 0.1\times 0 = -1$

$w_{31} = w_{31} - 0.1\times(-2) = 0.5 + 0.2 = 0.7$

$w_{32} = w_{32} - 0.1\times 4 = -1 - 0.4 = -1.4$

$w_{41} = w_{41} - 0.1\times 0 = -0.5$

$w_{42} = w_{42} - 0.1\times 0 = 1$

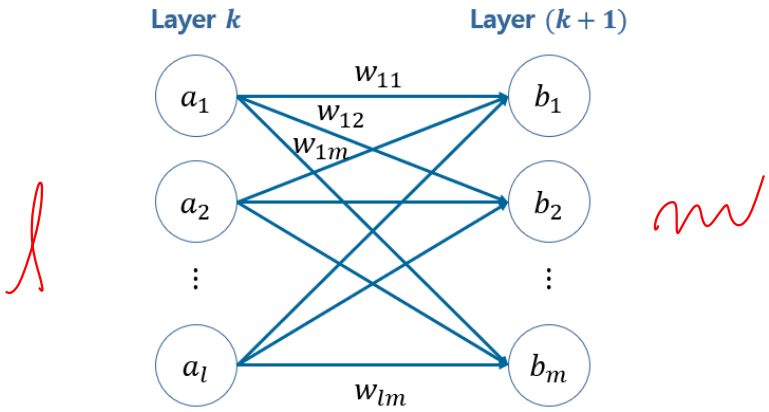
SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	**2 hours**	**Questions**	**6**	**Weighting**	**30%**
Student's Number			**Student's Name**		

5. (20pts; each 5pts) We want to apply the dropout method to **Layer k** in the following fully-connected layers. For dropout, we set the dropout rate (i.e., the probability of an element to be zeroed) to p. Here, we omit an activation function for simplicity.



(a) Describe how the dropout works during training in detail.

lxp units in layer k will be deactivated (zeroed-out). randomly.
network cannot depend on certain hidden node

(b) For a specific node i in **Layer** $(k+1)$, what is the expected activation value $\mathbb{E}[b_i]$?

$b_i = \sum_{j=1}^{l} w_{ij} a_j$ dropout $\Rightarrow$ $b_i = \sum_{j=1}^{l} w_{ij} a'_j$, $a'_j = \begin{cases} 0 & \text{(deactivation)} \\ \frac{a_j}{1-p} & \text{(activation.)} \end{cases}$

$\mathbb{E}a'_j = 0 \times P(a'_j=0) + \frac{a_j}{1-p} \times P(a'_j = \frac{a_j}{1-p}) = 0 + a_j/1-p \times 1-p = a_j$

$\mathbb{E}b_i = \sum_{j=1}^{l} w_{ij} a_j$

(c) If the dropout is also applied to the test phase, we cannot expect consistent predictions due to the randomness of the dropout. Explain how we can resolve this issue with your answer in (b).

in test phase, turn off the dropout (dropout rate = 0) because $\mathbb{E}[b_i]$ is already reflected scaling up.
just use $\mathbb{E}[b_i]$ for test, not randomly selected activation value for stable testing with average output.

(d) Explain why dropout in a neural network provides a regularization effect.

prevent co-adaptation between features
network depends on certain hidden layers. force to learn in-out relationship w/o certain hidden units
=> robustness improves.

virtual ensembling (test without dropout = averaging dropouted networks predictions)

weights smaller. smoother output function. decrease overfitting.

SEOULTECH
Seoul National University of Science & Technology

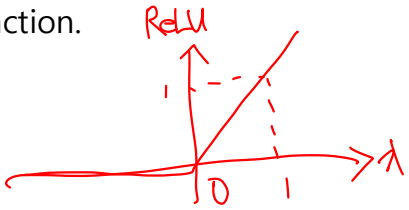
Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

6. (30pts; each 5pts) Write answers with precise explanations for the following questions.

(a) Draw a plot that represents a ReLU activation function and its advantages over a sigmoid activation function.



avoid vanishing gradients occurs update suspension
efficient backpropagation calculation.

$ReLU = \max(0, x)$

$Sigmoid = \frac{1}{1+e^{-x}}$

(b) You are training a neural network and notice that the validation error is significantly lower than the training error. Provide (at least) two possible reasons for this to happen.

increasing model capacity makes more flexibility. variance.
additional parameters are used for modeling noise of training set.

data set itself is noisy & limited, then the network learns noise.
Cannot distinguish noise & data features. variance
how sensitive to training set

(c) Explain why we need nonlinear activation functions.

using nonlinear activation function, we can model complex nonlinear relationship.
if we use linear function as activation function, it just became linear regression problem.

ReLU

(d) You want to solve a classification task. You first train your network on 20 samples. Training converges, but the training loss is very high. You then decide to train this network on 10,000 samples. Is your approach to fixing the problem correct? If yes, explain the most likely results of training with 10,000 samples. If not, give a solution to this problem.

problem : underfitting.

Sol : bigger data set is not solution.
current model is underfitted status. Training already converges. it means the model is too simple to represent input data features.
Adding more layers. more units per layer can be helpful. / adjusting learing rate, or degree of regularization reduction is needed.

SEOULTECH
Seoul National University of Science & Technology

Mid-term Exam

Date : 2023.11.01

Code	ITM 626		Title	Artificial Intelligence	
Time for Exam	2 hours	Questions	6	Weighting	30%
Student's Number			Student's Name		

(e) Given a convolutional layer with 8 filters, a filter size of 6, a stride of 2, and a padding of 1. For an input feature map of 32x32x32, what is the output shape after applying the convolutional layer to the input?

$\left(\frac{32+2\times1-6}{2}+1\right)\left(\frac{32+2\times1-6}{2}+1\right)\times 8.$

output = 15x15x8

(f) Why softmax function is often used for classification problems? Explain with the formular of softmax function.

convert output to probability dist (logit)

$\hat{y}_i = e^{h_i} / \left(\sum_i e^{h_i}\right) > 0$

Differentiable

softmax.

$\sum \hat{y}_i = 1$