AML 5202 | Deep Learning | Sessional-2 Solutions

1. [10 points] [L5, CO 1] Suppose $\mathbf{X}$ represents the data matrix (samples along columns) containing information about 100 individuals

$$\text{with features} \begin{cases} \text{age}, \\ \text{annual income}, \\ \text{credit limit}, \\ \text{gender (female or male)}, \\ \text{education level (graduate, high school, post graduate)}, \end{cases} \quad \text{and categories} \begin{cases} \text{does not default payment}, \\ \text{defaults payment}. \end{cases}$$

   (a) Suppose we want to apply softmax classifier to the dataset. What will be the shape of the weights matrix $\mathbf{W}$ assuming that the bias trick has been done?

   Solution: Note that "gender" is a categorical feature which will be one-hot encoded resulting in the 2 new features "genderfemale" and "gendermale" and that "education level" is also a categorical feature which will be one-hot encoded resulting in the 3 new features "education levelgraduate", "education levelhigh school", and "education level post graduate". This results in 8 features:

   (1) age (2) annual income (3) credit limit (4) genderfemale (5) gendermale (6) education levelgraduate (7) education levelhigh school (8) education level post graduate

   As there are 2 output categories, the weights matrix W will have shape $2 \times (8+1) = 2 \times 9$, taking into account the bias feature.

   (b) What will be the shape of the raw scores matrix comprising the raw scores of all 100 samples?

   Solution: 2×100.

(c) In plain English and using the data as context, explain what each of the following represents assuming indexing starts from 1:

$$w_{:,2},\ w_{1,:},\ w_{1,8},\ w_{2,5}.$$

Solution:

$w_{:,2}$: weights applied to the 2nd feature "annual income" to get all output category raw scores.

$w_{1,:}$: weights applied to all features to get the raw score for the 1st output category "does not default payment".

$w_{1,8}$: weight applied to the 8th feature "education level post graduate" to get the 1st output category "does not default payment" raw score.

$w_{2,5}$: weight applied to the 5th feature "gendermale" to get the 2nd output category "defaults payment" raw score.

2. [10 points] [L5, CO3] Consider the following initial weights matrix (assuming that the bias trick has been done) for dense layer l in a deep neural network:

$$\mathbf{W} = \begin{bmatrix} 0.01 & -0.01 & 0.08 & 0.1 \\ 0.01 & -0.01 & 0.08 & 0.1 \end{bmatrix}.$$

(a) How many nodes are there in layer $l-1$ and layer l?

Solution: 2 nodes in layer l and 3 nodes in layer $l-1$.

(b) Is there any issue with the initial values of the weights given here? Justify your answer in 1-2 lines.

Solution: The weights and bias values are identical for all nodes in layer l which will lead to symmetric learning.. To break the symmetry in learning, and therefore to enhance the learning of the nodes, w=the weights must be randomly assigned across all nodes.

(c) Calculate the $L2$-regularization loss for dense layer l.

Solution: $L2$-regularization loss $= \lambda \times \Big[\underbrace{(0.01)^2 + (-0.01)^2 + (0.08)^2}_{1^{\text{st}}\text{ node}} + \underbrace{(0.01)^2 + (-0.01)^2 + (0.08)^2}_{2^{\text{nd}}\text{ node}}\Big]$, where λ is the regularization strength. Note that the bias is not included in regularization.

3. [10 points] [L5, CO3] Suppose we want to implement a dropout layer after dense layer l of a deep neural network with a dropout probability of 0.2. Consider the following dropout-matrix:

$$\mathbf{D} = \begin{bmatrix} 0.49 & 0.47 & 0.7 & 0.99 \\ 0.86 & 0.49 & 0.76 & 0.96 \\ 0.13 & 0.98 & 0.87 & 0.54 \\ 0.48 & 0.96 & 0.76 & 0.32 \\ 0.62 & 0.15 & 0.23 & 0.58 \end{bmatrix}$$

Solution: 5.

(a) What is the batch size?

Solution: 4.

(b) Each batch sample contributes to the learning of specific nodes of dense layer l. Identify those nodes for each batch sample.

Solution:

Batch sample	Neurons
0	0, 2, 3, 4
1	0, 1, 4
2	0, 1, 3, 4
3	2, 3, 4

(c) How does the activations vector for dense layer l for the 0th sample denoted as $\mathbf{a}^{[l](0)}$ gets forward propagated through this dropout layer? Your answer should be a vector whose elements involve the elements of the vector $\mathbf{a}^{[l](0)}$.

Solution:

$$\mathbf{a}_D^{[l](0)} = \frac{1}{1-0.2}\begin{bmatrix} a_0^{[l](0)} \\ 0 \\ a_2^{[l](0)} \\ a_3^{[l](0)} \\ a_4^{[l](0)} \end{bmatrix} = \frac{1}{0.8}\begin{bmatrix} 0.49 \\ 0 \\ 0.13 \\ 0.48 \\ 0.62 \end{bmatrix}.$$

(d) Compare and contrast dropout vs. loss-based regularization in not more than 2-3 lines.

Solution: Both dropout and loss-based regularization approaches shrink the weights to zeros. However, dropout-based approach does not require additional calculations because it is randomized dropping of nodes across activation layers for samples in the batches. Loss-based regularization approaches like L1 or L2 require calculation of regularization loss and its gradient. L1 regularization can be used to shrink most of the weights to zeros which will result in a compact model.

(e) Is dropout applied to test data? In one line, justify your answer.

Solution: Dropout is applied only during training to prevent overfitting; during testing, all features are taken into account across all layers of the neural network.

4. [10 points] [L5, CO4] Consider the following sample matrix:

$$\mathbf{X} = \begin{bmatrix} -6 & -5 & 6 & 4 \\ 7 & -10 & -2 & 5 \\ 1 & 1 & 7 & -8 \\ -2 & 2 & 1 & 1 \end{bmatrix}.$$

(a) Suppose you are told that this sample represents a grayscale image. Justify why you see negative numbers in the sample matrix.

Solution: The image is possibly mean-cenetered resulting in negative intensities.

(b) Convolve this sample with the kernel $K = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ using no zero padding and unit stride.

Solution: The first convolution operation with the center of the filter resulting in the $(1, 1)$ element of the output is highlighted below:

$$\begin{bmatrix} -6 & -5 & 6 & 4 \\ 7 & -10 & -2 & 5 \\ 1 & 1 & 7 & -8 \\ -2 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} -9 & -15 \\ 0 & -5 \end{bmatrix}$$

(c) In one line, intuitively interpret the effect of convolving the image with the given kernel.

Solution: Convolving with this kernel helps in detecting vertical edges in the image.

5. [10 points] [L3, CO4] Consider the convolutional neural network defined by the layers in the left column in the table below. Fill in the shape of the output volume and the number of parameters corresponding to each layer using the notations below:

- CONVx-N denotes a convolutional layer with N filters with kernel height and width both equal to x. Padding is 2, and stride is 1.
- POOL-N denotes an $N \times N$ max-pooling layer with stride of N and no zero padding.
- FLATTEN flattens its inputs.
- FC-N denotes a fully-connected layer with N neurons.

Layer	Output Volume Shape	Number of Parameters
Input	$32 \times 32 \times 3$	0
CONV3-16	**?**	**?**
Leaky ReLU	**?**	**?**
POOL-2	**?**	**?**
FLATTEN	**?**	**?**
FC-10	**?**	**?**

Solution:

Layer	Output Volume Shape	Number of Parameters
Input	$32 \times 32 \times 3$	0
CONV3-16	$34 \times 34 \times 16$	$16 \times (3 \times 3 \times 3 + 1)$
Leaky ReLU	$34 \times 34 \times 16$	0
POOL-2	$17 \times 17 \times 16$	0
FLATTEN	$17 * 17 * 16 = 4624$	0
FC-10	10	$10 \times (4624 + 1)$