**Subject Code:** MDS6304
**Subject Name:** Deep Learning Principles & Applications
**Sample Final Exam Questions:** Covers topics in Segments 1-6

---

1. The MAHE registrar has the complete list of courses taken by each graduating student in a program. This data is represented as a matrix $\mathbf{X}$ with $m$ rows (samples) and $n$ columns (features) as follows:

| Student \ Course | **1** | **2** | $\cdots$ | **n** |
|---|---|---|---|---|
| **1** | 1 | 1 | $\cdots$ | 0 |
| **2** | 0 | 1 | $\cdots$ | 0 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| **m** | 1 | 0 | $\cdots$ | 1 |

The entries of the data matrix are 1s and 0s representing whether a particular student has taken a particular course. For example, the red-highlighted entry 1 means that the 1st student has taken the 1st course and the blue-highlighted entry 0 means that the $m$th student has not taken the 2nd course. Recall that the $i$th student vector is represented as $\mathbf{x}^{(i)}$ and the $j$th course vector is represented as $\mathbf{x}_j$.

   (a) The total number of courses the 5th student has taken is $\boxed{?}^{\mathrm{T}}\boxed{?}$.

   (b) In English, explain what the quantity $\mathbf{x}^{(3)^{\mathrm{T}}}\mathbf{1}$, where $\mathbf{1}$ is the vector full of ones of compatible shape, represents w.r.t. the data?

   (c) The total number of students who have taken both class 1 and class 6 is $\boxed{?}^{\mathrm{T}}\boxed{?}$.

   (d) In English, explain what the quantity $\left\|\mathbf{x}^{(1)} - \mathbf{x}^{(6)}\right\|^2$ represents w.r.t. the data?

   (e) In English, explain what the quantity $X^{\mathrm{T}}\mathbf{1}$ represents w.r.t. the data?

2. Consider the logistic regression classifier applied to a sample $\mathbf{x}$ (bias feature added) with correct label $y$ using weights vector $\mathbf{w}$ (bias corrected). The forward propagation for calculating the sample's loss is:

$$z = \mathbf{w}^{\mathrm{T}}\mathbf{x} \longrightarrow a = \sigma(z) \longrightarrow L = -\log\left(a^y \times (1-a)^{1-y}\right).$$

Calculate the gradient of the loss using the chain rule $\nabla_{\mathbf{w}}(L) = \nabla_{\mathbf{w}}(z) \times \nabla_z(a) \times \nabla_a(L)$ and express it in terms of $\mathbf{x}, a$, and $y$.

3. For a particular binary classification task, your friend modifies the logistic regression loss function as follows:

$$L = -\log\left(a^{\boldsymbol{\alpha} y} \times (1-a)^{\boldsymbol{\beta}(1-y)}\right),$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are coefficients set by your friend for a dataset in hand. What kind of a binary classification task might your friend be solving? Keep your answer short to explain how your friend would have set the coefficients based on the dataset.

4. onsider the $5 \times 5$-matrix

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

   (a) For a 5-vector $x$, how are $Bx$ and $x$ related?

   (b) Compute $B^2$ defined as the matrix-matrix product $B \times B$. Use the result to identify what $B^5$ would be without any further calculations.

5. For the following vector $\mathbf{z}$, without any calculation, write down what is softmax($\mathbf{z}$) approximately, and answer to which category (among 1, 2, and 3) would a sample be predicted to belong to if it had this raw score:

$$\mathbf{z} = \begin{bmatrix} 10^{-10} \\ 10^{6} \\ 10^{-12} \end{bmatrix}.$$

6. You want to predict if a patient survived (label 1) or not (label 0) by training a logistic regression model on a dataset with 4 samples $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}$ corresponding to correct labels $1, 1, 0, 0$, respectively, using full-batch gradient descent. The following quantities are observed in consecutive epochs:

| **epoch** $k$ | **epoch** $k+1$ |
|---|---|
| $\sigma\left(w \cdot x^{(1)}\right) = 0.97$ | $1 - \sigma\left(w \cdot x^{(1)}\right) = 0.01$ |
| $1 - \sigma\left(w \cdot x^{(2)}\right) = 0.09$ | $1 - \sigma\left(w \cdot x^{(2)}\right) = 0.07$ |
| $\sigma\left(w \cdot x^{(3)}\right) = 0.51$ | $\sigma\left(w \cdot x^{(3)}\right) = 0.49$ |
| $1 - \sigma\left(w \cdot x^{(4)}\right) = 0.04$ | $\sigma\left(w \cdot x^{(4)}\right) = 0.98$ |

For each epoch, calculate the training loss and the proportion of samples that are correctly classified and compare them across the epochs. Are you surprised by the answer? Write your observation in one sentence.

7. Suppose $\mathbf{X}$ represents the data matrix (samples along columns) containing information about 100 individuals

$$\text{with features} \begin{cases} \text{gender (female or male),} \\ \text{age,} \\ \text{weight,} \\ \text{BMI,} \\ \text{blood glucose,} \\ \text{total cholesterol,} \end{cases} \qquad \text{and categories} \begin{cases} \text{non-diabetic,} \\ \text{pre-diabetic,} \\ \text{diabetic,} \\ \text{severely diabetic.} \end{cases}$$

   (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?

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

$$w_{:,2},\ w_{4,:},\ w_{18},\ w_{25}.$$

   (c) Suppose the gradient of the loss with respect to some weight parameter evaluated at its current value is 4. Justify what will happen to the loss if we increase that weight parameter a little bit while keeping the other parameters fixed? What if we decrease it a little bit?

8. Calculate the sensitivity of $L(w) = 4w + w^3$ w.r.t. the input $w$ at $w = 1$. What happens to $L$ if $w$ increases slightly from its current value.

9. Calculate the sensitivity of $L(w) = 4w - w^3$ w.r.t. the input $w$ at $w = 1.5$. What happens to $L$ if $w$ increases slightly from its current value.

10. Calculate the sensitivity of $L(\mathbf{w}) = w_1 + w_2^2$ w.r.t. the input $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ at $w_1 = 1, w_2 = 2$.

11. You want to train a neural network to classify a sample with 19 continuous features into one of 3 possible categories.

   (a) How many parameters will have to be trained if you use an 8-layer deep neural network with 8 nodes in each of the hidden layers?

   (b) Now you want to train using a 2-layer shallow neural network such that the number of parameters will not exceed the number identified in the previous part for the deep neural network. How many nodes will this shallow neural network have in the hidden layer?

   (c) Which architecture would be preferable, the deep or the shallow one? Justify using one or two sentences at maximum.

12. Consider the following forward propagation through a fully-connected deep neural network architecture (256 nodes in hidden layer) for a $32 \times 32$-image sample represented as vector $\mathbf{x}$ with one-hot encoded correct label 3-vector $\mathbf{y}$ and predicted probability vector $\hat{\mathbf{y}}$ (indexing starts from 0):

$$\boxed{\underbrace{\mathbf{x}_B}_{?} = \begin{bmatrix}\mathbf{x}\\ 1\end{bmatrix}} \rightarrow \boxed{\underbrace{\mathbf{z}^{[1]}}_{?} = \underbrace{\mathbf{W}^{[1]}}_{?\times ?}\mathbf{x}_B} \rightarrow \boxed{\underbrace{\mathbf{a}^{[1]}}_{?} = \text{ReLU}\left(\mathbf{z}^{[1]}\right)} \rightarrow \boxed{\mathbf{a}_B^{[1]} = \begin{bmatrix}\mathbf{a}^{[1]}\\ 1\end{bmatrix}} \rightarrow \boxed{\underbrace{\mathbf{z}^{[2]}}_{?} = \underbrace{\mathbf{W}^{[2]}}_{?\times ?}\underbrace{\mathbf{a}_B^{[1]}}_{?}} \rightarrow \boxed{\mathbf{a}^{[2]} = \text{softmax}\left(\mathbf{z}^{[2]}\right)} \rightarrow \boxed{L = \sum_{k=0}^{?} -y_k \log\left(\hat{y}_k\right)}.$$

Identify the missing shapes corresponding to the question marks.

13. Suppose we have the following raw scores vector corresponding to dense layer 4 of a deep neural network:

$$\mathbf{z}^{[4]} = \begin{bmatrix}-10\\ 1\\ 10\\ -1\end{bmatrix}.$$

(a) Calculate the local gradient of activation layer 4 which is tanh activated. Round your calculations to 2 decimal places.

(b) Clearly state for which nodes of dense layer 4, learning of the corresponding parameters will be very minimal? Justify your answer briefly. What could you do to improve the learning of the parameters of all nodes in dense layer 4?

14. Suppose we want to implement a dropout layer for a dense hidden layer $l$ of a deep neural network with a dropout probability of 0.2. Consider the following dropout-matrix:

$$\mathbf{D} = \begin{bmatrix}0.49 & 0.47 & 0.7 & 0.99\\ 0.86 & 0.49 & 0.76 & 0.96\\ 0.13 & 0.98 & 0.07 & 0.54\\ 0.48 & 0.96 & 0.76 & 0.32\end{bmatrix}$$

(a) What is the number of nodes in dense layer $l$?

(b) What is the batch size?

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

(d) Write down the forward propagation equation through the dropout layer for each batch sample.

15. Consider the following sample:

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

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.

16. CNN question.

Consider the convolutional neural network defined by the layers in the left column below. Fill in the shape of the output volume and the number of parameters at each layer.

Notation:

- CONV5-N denotes a convolutional layer with N filters with height and width equal to 5. Padding is 2, and stride is 1.
- POOL2 denotes a 2x2 max-pooling layer with stride of 2 and 0 padding.
- FC-N denotes a fully-connected layer with N neurons

| Layer | Output Volume Dimensions | Number of parameters |
|---|---|---|
| Input | $32 \times 32 \times 1$ | 0 |
| CONV5-10 | | |
| POOL2 | | |
| CONV5-10 | | |
| POOL2 | | |
| FC10 | | |