SEAT NO. DT-22005

# NED UNIVERSITY OF ENGINEERING & TECHNOLOGY

THIRD YEAR (Data Science)

SPRING SEMESTER EXAMINATIONS 2025

Time : 3 Hours | Batch 2022 | Dated : 04-JUN-25 | Max Marks : 60

## Data Mining - CT-356

***Note: Attempt all questions. Answer all parts of a question in a row. State your assumptions clearly. Draw diagram where necessary.***

**Q1.** a) ***Explain*** in detail the methods of measuring clustering quality. **[6]**

b) ***Explain*** in detail what is unstable gradient problem? Describe briefly any two techniques to tackle unstable gradient problem. Also outline the algorithm of Stochastic Gradient Descent (SGD). **[6]**

**[CLO 1: 12 Marks]**

**Q2.** a) Consider a Deep Neural Network that contains images of size 28 x 28 RGB of American Sign Language (ASL) dataset. Furthermore, the DNN contains three hidden layers to capture the intrinsic patterns each of size 128, 256, 512 respectively. The output layer contains the 26 possibilities. The bias term is set to 0 for the simplicity. Calculate how many number of total connections this DNN have? ***Show*** clearly each steps. **[6]**

b) The table below shows the training and validation loss of a Decision Tree model over 10 epochs. ***Interpret*** whether the model is showing signs of underfitting, overfitting, or good generalization. Justify your answer by explaining the trends in the training and validation loss curves. Also show any three remedies to resolve overfitting. Explain any three remedies to resolve overfitting. **[6]**

| Epoch | Training Loss | Validation Loss |
|---|---|---|
| 1 | 0.341 | 0.192 |
| 2 | 0.164 | 0.143 |
| 3 | 0.115 | 0.106 |
| 4 | 0.086 | 0.093 |
| 5 | 0.068 | 0.087 |
| 6 | 0.054 | 0.084 |
| 7 | 0.042 | 0.074 |
| 8 | 0.033 | 0.077 |
| 9 | 0.029 | 0.090 |
| 10 | 0.022 | 0.072 |

**[CLO 1: 12 Marks]**

**Q3.** Analyze the following hyperparameters for a neural network. ***Identify*** the challenges might this configuration present, and how could you adjust the hyperparameters to address them? **[12]**

a) Assigning too large value of Learning Rate ($\lambda$)

b) The training set contains 10 Million images and are unable to fit in memory

c) Setting epoch = 3, while training the CNN model for image classification. The size of the dataset is 10 Million three channel images.

d) Initializing weights with tiny value $w = 0.0003$

**[CLO 2: 12 Marks]**

P.T.O

**Q4.** Data have been downloaded from the Daraz 11-11 sale packages catalog by some business experts to learn the pattern of frequency of items. There are 9 different transactions (order 1 to 9) and each transactions have minimum 2 and maximum four items. Here 5 different kind of items are available. For simplicity Our Team lead used short names of items (I1- I 5) rather than full descriptive name (example I1= Hair dryer). Consider Minimum threshold support as 22.22% and confidence as 77.77%. *Apply* the most important analytics algorithm that is Apriori algorithm and find the frequent k items. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence. Show all your *steps* and *cut down transactions* with detail tabular format. **[12]**

[CLO 2: 12 Marks]

| Catalogue Item | List of Item in Package |
|---|---|
| Package 01 | I1, I2, I5 |
| Package 02 | I2, I4 |
| Package 03 | I2, I3 |
| Package 04 | I1, I2, I4 |
| Package 05 | I1, I3 |
| Package 06 | I2, I3 |
| Package 07 | I1, I3 |
| Package 08 | I1, I2, I3, I5 |
| Package 09 | I1, I2, I3 |

**Q5.** *Solve* using the DBSCAN algorithm to the given data points and create the clusters with **minPts = 4** and $\varepsilon = 1.9$. Use the Minkowski distance with $r = 2$ ($l_2$ *norm*) and calculate the distance between each points. Show the graph containing core, border and noise points. **[12]**

[CLO 2: 12 Marks]

| P1 | P2 | P3 | P4 | P5 | P6 | P7 | P8 | P9 | P10 | P11 | P12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3,7 | 4,6 | 5,5 | 6,4 | 7,3 | 6,2 | 7,2 | 8,4 | 3,3 | 2,6 | 3,5 | 2,4 |