Course: **Data Mining** 24-Mar-2023 **(Fall 2022)**

Resource Person: **Dr. Muhammad Faheem**

ASSIGNMENT–3 (Classification and Cluster Analysis)

Total Points: 40
Submission Due: Wednesday April 05, 2023
(Google Classroom Course Page)

Instructions: Please Read Carefully!

- **This is an <u>individual</u> assignment. Everyone is expected to complete the given assignment on their own, without seeking any help from any website or any other individual.** There will be strict penalties for any work found copied from any source and the university policy on plagiarism will be strictly enforced.
- Assignment is to be submitted via Google Classroom.
    - You should already have created your account on Google Classroom as per my earlier email. If not, then follow the link in that email to create your account.
    - Submit your assignment on or before due date. **No late submissions will be possible**.
    - The viva of this assignment will be conducted.

Question 1: **[15]**

The following table consists of training data from an employee database. The data have been generalized. For example, "31 … 35" for age represents the age range of 31 to 35.

For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row.

Department	Status	Age	Salary	Count
Sales	Senior	31 … 35	46K … 50K	30
Sales	Junior	26 … 30	26K … 30K	40
Sales	Junior	31 … 35	31K … 35K	40
Systems	Junior	21 … 25	46K … 50K	20
Systems	Senior	31 … 35	66K … 70K	5
Systems	Junior	26 … 30	46K … 50K	3
Systems	Junior	41 … 45	66K … 70K	3
Marketing	Senior	36 … 40	46K … 50K	10
Marketing	Junior	31 … 35	41K … 45K	4
Secretary	senior	46 … 50	36K … 40K	4
Secretary	junior	26 … 30	26K … 30K	6

Let status be the class-label attribute.

(a) Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers.

(b) Using the multilayer feed-forward neural network obtained in (a), show the weight values after one iteration of the backpropagation algorithm, given the training instance "(sales, senior, 31 . . . 35, 46K . . . 50K)". Indicate your initial weight values and biases and the learning rate used.

(a) There is no standard answer. Every feasible solution is correct. As stated in the book, discrete valued attributes may be encoded such that there is one input unit per domain value. For hidden layer units, the number should be smaller than that of input units, but larger than that of output units.
(b) There is no standard answer. Every feasible solution is correct.

Question 2: **[5]**

How the Multilayer feed-forward neural network works? How to define a network topology of a neural network?

Refer to Section 9.2.1 and 9.2.2.

Question 3: **[10]**

Use an example to show why the k-means algorithm may not find the global optimum that is optimizing the within-cluster variation.

Consider applying the k-means algorithm on the following points, and set k = 2. A(0, 1), B(0,−1), C_i(100, 50) (i = 1, . . . , 100), and D_j(100,−50) (j = 1, . . . , 100).

If we use A and C_1 as the initial cluster centers, then the clustering process will converge to two centers, (0, 0) and (100, 0). The within-cluster variation is

$E = 1^2 + 1^2 + 100 \times 50^2 + 100 \times 502 = 1 + 1 + 100 \times 2500 + 100 \times 2500 = 500002.$

However, if we use (100, 50) and (100,−50) as the cluster centers, the within-cluster variation is

E = (1002 + 492) + (1002 + 492) + 100 × 0 + 100 × 0 = 24802, which is much smaller. This example shows k-means may be trapped by a local optimal, and cannot jump out to find the global optimum.

Question 4: **[10]**

Explain the hierarchal clustering with examples.

A hierarchical clustering method works by grouping data objects into a hierarchy or “tree” of clusters. Refer to section 10.3.