NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 2 EXAMINATION 2021-2022

EE6227 – GENETIC ALGORITHMS AND MACHINE LEARNING

April / May 2022 Time Allowed: 3 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 5 pages.
2. Answer all 4 questions.
3. All questions carry equal marks.
4. This is a closed book examination.
5. Unless specifically stated, all symbols have their usual meanings.

---

1. Imagine yourself to be in the business of individualized customization of protective headgear. You have a system that configures the basic cupping structure of the helmet. For our purpose, we can visualize in Figure 1 on page 2 a simplified 2-dimensional version of the fitting process. To be considered a good fit, let us assume that 3 points P1, P2 and P3 are identified as vital points which are required to be in physical contact with the wearer's head. The characteristic equation of the cupping structure is defined by the following equation:

$$px^2 + qy^2 = h$$

(a) The customization process involves finding suitable values of the parameters p, q and h such that all 3 points should be in physical contact with the headgear to achieve a good fit. Write a suitable objective function $f(\lambda)$ in terms of p, q, h, x_1, y_1, x_2, y_2, x_3 and y_3, where λ denotes a triple (p, q, h) for an optimization algorithm that seeks to minimize the objective function.

(7 Marks)

Note: Question No. 1 continues on page 2.

(b) To solve the problem, a genetic algorithm is configured where the genotype is represented as a coding of 12-bit binary string. Let the first 4 bits represent a positive integer a, the 5th to 8th bits for integer value b while the remaining four bits represent integer value h. The phenotypic traits of the string can be written as $p = a/(a+b)$ and $q = b/(a+b)$ while h is the value as decoded. Denoting the binary string S as $<\alpha_1\alpha_2 \ldots \alpha_{12}>$, write the values of p, q and h in terms of α_i for $i = 1, 2, \ldots 12$.

(7 Marks)

(c) Suppose the values of P1, P2 and P3 are (12,5.6), (7,3.4) and (-10,4.5) respectively. Calculate the fitness values of the following strings:

100001001011, 010110011100, 010011011110

(6 Marks)

(d) An alternative is to use real values encoding to code the solution strings and solve it as a continuous optimization problem. Compare and comment on the advantages and disadvantages between the two approaches.

(5 Marks)

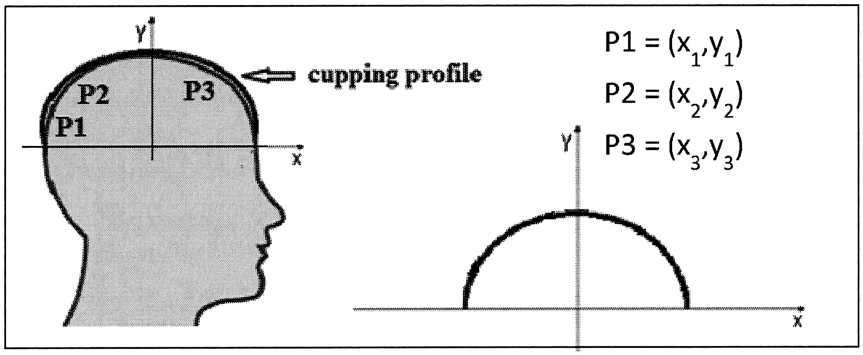

Figure 1

2. This question consists of 6 parts.

(a) The travelling salesman problem is considered an NP-hard problem. Describe in your own words what this means from a problem-solving perspective.

(5 Marks)

(b) The (x, y) coordinates of 10 points for a 10-city travelling salesman problem are given in Table 1. Calculate the inter-city 10x10 distance matrix $\mathbf{D} = [d_{i,j}]$ for the 10 cities. Present your answer with up to 2-decimal-point accuracy.

(4 Marks)

Note: Question No. 2 continues on page 3.

(c) Let $\pi = <\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6\alpha_7\alpha_8\alpha_9\alpha_{10}>$ represents a permutation string that decodes into a visitation sequence of the 10 cities. The fitness function for π can be written as follow:

$$f(\pi) = K - (\sum_{i=1}^{n-1} d_{\pi(i,i+1)} + d_{\pi(1,10)}),$$

where K is an appropriately chosen constant, $d_{\pi(i,j)}$ is the distance between nodes α_i and α_j of the sequence π and n is the number of cities.

A valid tour sequence is one where each city is visited once with the salesman returning to the original starting point. Consider the following permutation strings $\pi_1 = <8\ 4\ 3\ 10\ 7\ 6\ 2\ 5\ 9\ 1>$ and $\pi_2 = <5\ 2\ 10\ 3\ 4\ 1\ 9\ 8\ 7\ 6>$ representing two closed circuit tour sequences. Calculate the travel distance for the two tour sequences $d(\pi_1)$ and $d(\pi_2)$.

Table 1

Nodes	x	y
0	10	25
1	3	27
2	14	22
3	1	13
4	20	3
5	20	16
6	28	12
7	30	31
8	11	19
9	7	3

(4 Marks)

(d) Apply the cycle crossover operation based on the longest cycle and show the two offspring π_1* and π_2* derived from crossover of π_1 and π_2. Determine if the fitness of the offspring improves?

(4 Marks)

Note: Question No. 2 continues on page 4.

(e) In the above scenario, we considered all the nodes to be connected, which can be depicted as a complete graph with values of 1 for the whole matrix. Assuming that not all nodes are connected and the adjacency matrix A[i,j] for i, j = 1 to 10 is as follow:

$$A = [a_{i,j}] = \begin{pmatrix} 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \\ 0\,1\,0\,1\,0\,1\,0\,1\,0\,1 \\ 1\,0\,1\,0\,1\,0\,1\,0\,1\,0 \end{pmatrix}$$

Define C(π) as the connectivity fitness for the tour sequence π. Write the connectivity fitness equation for C(π) in terms of $a_{i,j}$. Compute C(π_1), C(π_2), C(π_1*), and C(π_2*).

(4 Marks)

(f) The original fitness function for the complete graph is f(π) = K – d(π), where K is an appropriately chosen constant. Write an appropriate equation of a fitness function v(π) for a GA that searches for an optimal valid tour. The fitness function applies a penalty according to the connectivity of the given sequence.

(4 Marks)

3. A dataset contains 2000 samples, where 1200 samples are from class 1, and the remaining 800 samples are from class 2. Assume the samples in the two classes follow two independent normal distributions, and the estimation of the mean vectors and covariance matrices of the two classes are given as follows:

$$\hat{\boldsymbol{\mu}}_1 = \begin{bmatrix} -0.51 \\ -0.39 \end{bmatrix}, \qquad \hat{\boldsymbol{\mu}}_2 = \begin{bmatrix} 1.12 \\ 2.09 \end{bmatrix},$$

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 1.02 & -0.02 \\ -0.02 & 0.88 \end{bmatrix}, \qquad \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 1.18 & -0.12 \\ -0.12 & 1.16 \end{bmatrix}.$$

(a) Design a Bayes decision rule for the 2-class classification problem, and use the rule to classify the following sample:

$$\mathbf{x}_1 = \begin{bmatrix} 0.2576 \\ 0.2544 \end{bmatrix}.$$

(10 Marks)

Note: Question No. 3 continues on page 4.

(b) Design a Fisher linear discriminant classifier for the 2-class classification problem, and use the classifier to classify the following sample:

$$\mathbf{x_2} = \begin{bmatrix} 2.244 \\ 1.491 \end{bmatrix}.$$

(10 Marks)

(c) Discuss the procedures and metrics that can be used for evaluating pattern classifiers.

(5 Marks)

4. (a) Supervised feature selection algorithms can be categorized into filter and wrapper methods. Draw the block diagrams of the two feature selection methods and discuss the main difference and respective advantages and disadvantages.

(10 Marks)

(b) List three types of clustering methods and give one example algorithm for each type. Discuss the application scenario for each of the three types of clustering methods.

(9 Marks)

(c) Describe the procedure of the DBSCAN clustering algorithm.

(6 Marks)

END OF PAPER

EE6227 GENETIC ALGORITHMS & MACHINE LEARNING

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.