A(E)4M33BIA: Individual project

Task is to

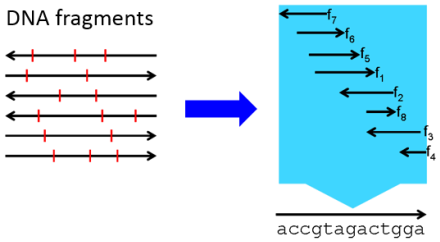
1. Design an evolutionary algorithm for chosen problem.
2. Implement and experimentally evaluate the proposed algorithm.
3. Write a report consisting of two parts
   a) Specification of chosen solution representation, selection strategy, crossover and mutation operators, evolutionary model, and definition of the fitness function.
   b) Report on the performance observed with the implementation.

      Use
      - tables to present statistics such as the absolute best solution values, the mean best-of-run values (accompanied by standard deviation), mean computation time etc. and
      - graphs showing typical, i.e. mean/median, convergence curves.

      Discuss the achieved results.

List of topics:

1. Genome Sequence Analysis



Given: A set of fragmentary gene segments. Only four letters will be used: A (adenine), C (cytosine), G (guanine), and T (thymine). The segments have the same length. For example, an input could look like:

TCGG, GCAG, ATCG, CAGC, GATC

Goal is to reconstruct the gene sequence i.e. to reassemble the fragments into the shortest possible gene that can be made from these pieces. The rules for assembling the final sequence are:

1. The sequence must use all the segments (and only the segments) provided
2. You cannot flip the segments.
3. The best answer is the shortest answer.

Example: CAGCAGATCGG and GATCGGCAGC are possible final sequences of length 11 and 10, respectively.

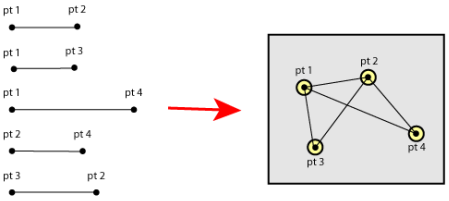
Hint: Since this is a permutation problem, you might want to try some of the crossover operators designed for the TSP, such as the Partially Mapped Crossover or Edge-Recombination Crossover.

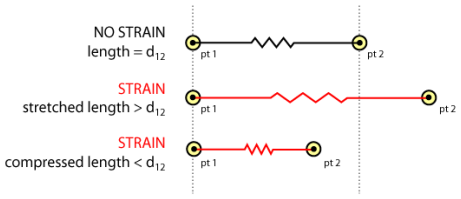
2. Design of Molecule

You are trying to model the structure of a molecule (which for the sake of simplicity will be limited to two dimensions).

Given: A collection of plastic connecting rods that represent the distance between pairs of atoms in the molecule, given in a symmetric distance matrix. A -1 in the [i,j] location signifies that there is no connection specified between the two atoms in question.

Goal is to fit the connecting rods together into a consistent two dimensional shape so that the actual lengths between atoms follow as precisely as possible the preferred ones. It is not always possible to get 100% precise solution, so some of the lengths you specify will be slightly longer or shorter than their preferred length.





This isn't always possible, so some of the lengths you specify will be slightly longer or shorter than their preferred length. **Stretching** or **compressing** the connecting rods is legal, but it causes a strain. The strain on the rod between any two atoms is defined as the absolute value of the difference between the preferred distance and the actual distance between atoms *i* and *j*. The sum of strain values over all rods should be minimized.

Hint: Try binary vs. real representation and appropriate genetic operators.

3. Gerrymandering

Gerrymandering is the process of carving up an electoral district into strange shapes in order to derive a political advantage. You are given the task of preparing for an upcoming election in the state of Rectanglia. As the director of redistricting, your job is to divide the state into N districts of equal population.

Given: A matrix A in which each element corresponds to the population in a given square mile.

Goal is to find a matrix B that indicates which voting district each square mile belongs in and voting districts must be contiguous, or connected in the four-neighbor sense (no diagonal connections).

Example. Suppose, given the following census data for Rectanglia in matrix A, you are told to divide the state into N = 3 districts of equal population.

$$A = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 12 & 11 & 2 & 12 \\ 40 & 28 & 10 & 2 \end{bmatrix}$$

Here's one way to optimally divide the state.

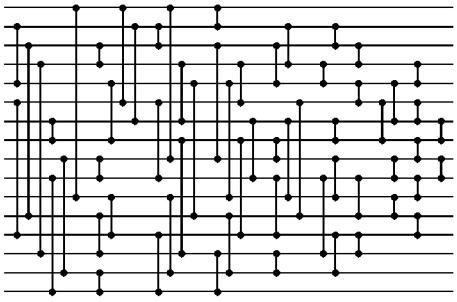
$$B = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 3 & 2 & 2 & 1 \end{bmatrix}$$

Hint: Design crossover and mutation operators that produce only valid solutions.

4. Sorting networks (Wikipedia)

Given: A sequence of *N* numbers and a set of comparators, where each comparator connects two wires and sorts the values by outputting the smaller value to one wire, and a larger value to the other.

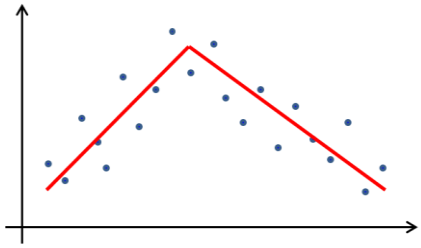
Goal is to design for the given sequence of numbers a sorting network with minimal number of comparators.



Example:

16-input sorting network. Each vertical line represents a comparator that switches the input numbers if they are not in desired order.

5. **Piecewise linear approximation of a set of points**



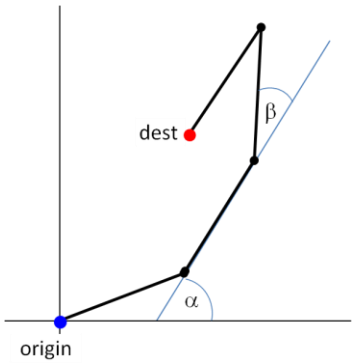
Given: A set of points [*x*, *y*].

Goal is to find a piecewise linear curve that minimizes the chosen objective (e.g. minimal sum of squared deviations or sum of absolute deviations).
The final approximation curve must be continuous.

Hints:

- Input parameter is the desired number of linear segments.
  The number of lines is not defined – then the number of linear segments needed to sufficiently approximate the data will be minimized.
- Representation
  - Set of end-points of partial linear segments.
  - Set of lines.

6. **Robotic arm with multiple joints**



Given: A robotic arm consisting of a number of segments connected by 2-degree of freedom joints. All segments are of the same length. The first joint is connected to the origin at <0,0>. The target position is at <x,y>.

Goal is to set angles of the joints so that the robotic arm reaches the target position.

Parameters of the problem are the lengths of the segments and the numbers of segments.

Hint:

- Try different representations – absolute (α) vs. relative (β) angles.

7. **Allocation of modules to processors**

Given: A set of *N* software modules; each module has assigned its computational complexity. A table of intermodule communication *C*, where position *C*[*i*,*j*] represents a communication rate between modules *i* and *j*.

Goal is to find an allocation of modules to *M* processors such that the overall computational load is evenly distributed to processors and the interprocessor communication is minimized.

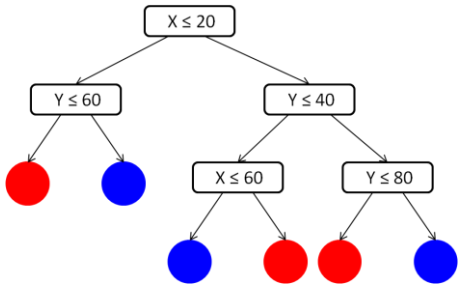
Hints:

- Multi-objective approach – NSGA-II, SPEA2.
- Single-objective with weighted individual objectives.

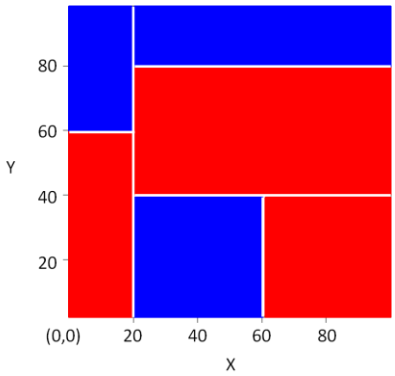
8. Classification problem

Given: A set of training points $T=\{(x_i,y_i)\text{: } i = 1..n\}$ that is split into two disjunctive sets of positive ***P*** and negative ***N*** cases.

Goal is to find a decision tree, using GP, classifying as much points as possible correctly. At each level the classified space is divided into 2 subspaces in one of the dimensions only. Sublevels continue dividing.



a) Decision tree

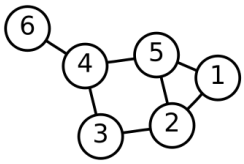


b) splitted decision space

Hints:

- Fitness takes into account only the classification accuracy.
- Fitness takes into account both the classification accuracy and the complexity of DT.

9. Large scale vertex cover problem



Given: An undirected graph with N vertices and M edges given by a symmetric binary matrix $G[N\times N]$, where $G(i, j) = 1$, iff there is an edge between i and j.

Vertex cover is a subset V' of the vertices of the graph which contains at least one of the two endpoints of each edge.

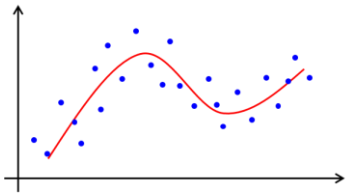
Goal is to find a vetrex cover of minimum size.

Example: vc1={1,3,5,6}, vc2={2,4,5}

10. Regression Problem

Given: Set of N data samples $[x_i,y_i]$.

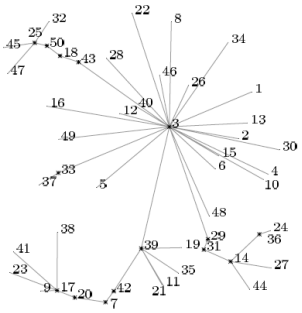
Goal: Find a function that fits to the training data as good as possible.



Hints:

- Compare GP evolving the whole function expression with GA evolving only parameters of the polynomial.
- Try GP with different function sets, high-level functions.

11. Minimum Energy Broadcast



Problem definition: The problem is defined as the problem of finding the broadcast tree T = (V,E_T) (a directed spanning tree) rooted at a source node s∈V in an ad-hoc wireless network G = (V, E, d), that minimizes the necessary total transmission power c(T) to reach all nodes of the network:

$$\min c(T) = \sum_{i \in V} \max_{(i,j) \in E} d(i,j)^{\alpha}$$

where d(i,j) refers to the Euclidean distance between nodes *i* and *j* and the constant α is the distance-power gradient which is set to 2,0.

Data: http://dag.informatik.uni-kl.de/research/meb/

Hints:

- Try hybrid approach combining EA with local search.