CS7.301 **2^{nd} April, 2020**

MDL Assignment-3

Vikrant Dewangan, Roll No.- 2018111024
Mohammad Nomaan Qureshi, Roll No.- 2018111027

Contents

1 File Structure

Upon running the file *solution.py*, we get

```
team71
├── report.pdf:  contains report
├── client.py:  contains necessary functions for server
├── set_cur_state.py:  contains necessary functions for vectors
├── solution.py:  main solution
└── trace_71.txt:  trace obtained
```

2 Problem Statement

Given coefficients of vectors corresponding to over-fit model, apply Genetic Algorithm (G.A) to reduce over-fitting.
Input:

$$\begin{aligned}\bar{v}_{overfit} = [&-0.00016927573251173823, 0.0010953590656607808,\\ &0.003731869524518327, 0.08922889556431182,\\ &0.03587507175384199, -0.0015634754169704097,\\ &-7.439827367266828e-05, 3.7168210026033343e-06,\\ &1.555252501348866e-08, -2.2215895929103804e-09,\\ &2.306783174308054e-11]\end{aligned}$$

Given a query server to allow for submission of weights (w_i), the following is returned for each query -

$$M.S.E = \frac{\sum_{x \in data} \big(y - f\big(x\big)\big)^2}{N}$$

where $f\big(x\big)$ is calculated using the weights submitted $[w_1, w_2 \dots w_{10}]$ -

$$f\big(x\big) = \sum_{i=1}^{10} w_i \cdot x_i$$

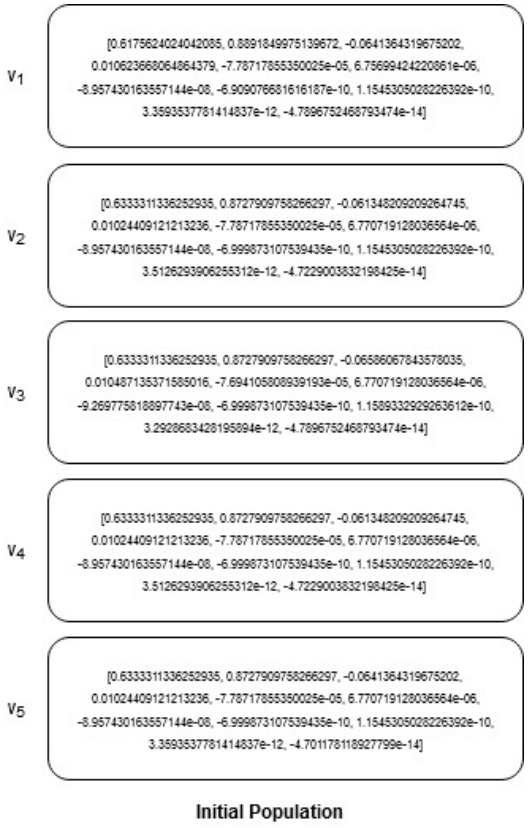
Output: A set of weights $[w_1, w_2 \dots w_{10}]$ corresponding to best-fit under the conditions-

$$|w_i| \leq 10$$

3 Algorithm Definition

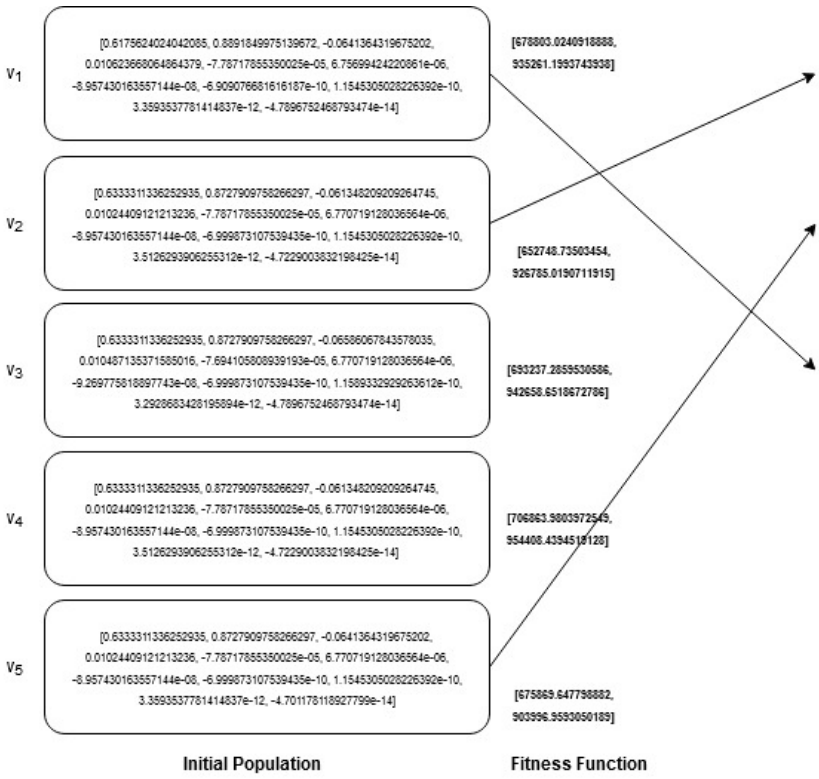
Our Genetic Algorithm consists of 4 steps -

1. **Initialisation of Population**:
   Each generation contains 5 vectors. This is chosen from previous set of population or derived initially randomly.After the algorithm runs, this population is instead chosen from the already saved state in *current_state.pkl*.



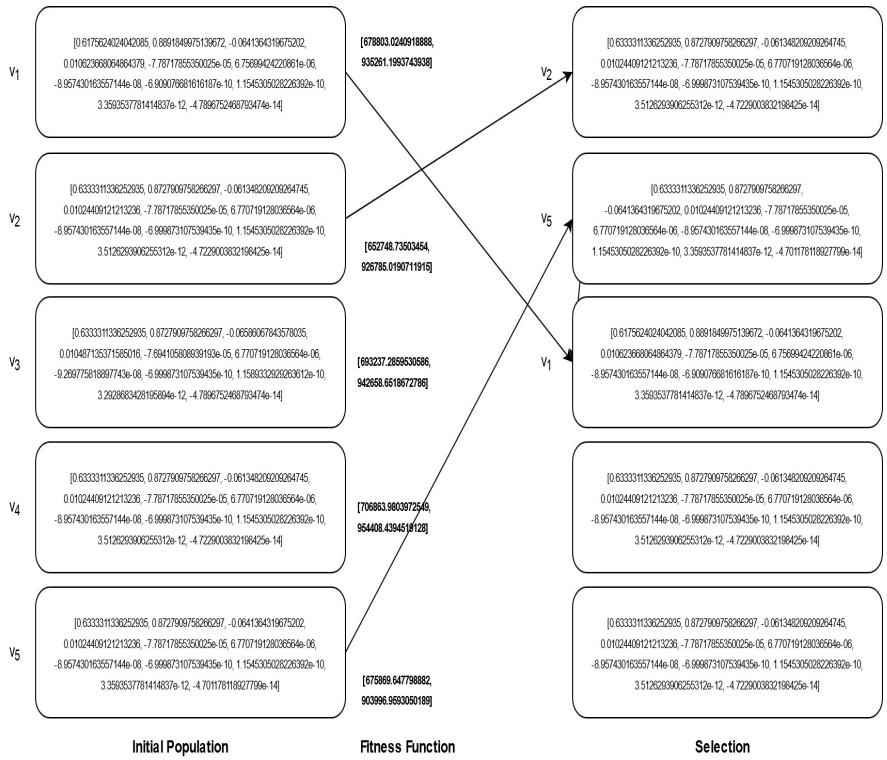
population.jpg

2. **Fitness Function**:
   The best three of each iteration survive and give birth to next generation. We chose training error to be out fitness function.



fitness.jpg

3. **Selection**
   Since we have used training error as our fitness function , the best 3 vectors are **deterministically** chosen from the population. This is where we differ from the normal Genetic Algorithm that the selection is not random.



selection.jpg

So if A, B, C, D, E are in sorted order of their fitness function then A, B, C survive.The best three of each iteration survive and give birth to next generation.

4. **Crossover with Mutation**
   Crossover
   We did a probabilistic cross over which is explained below.

   A is best vector and B is second best vector, then let Crossover vector be C.

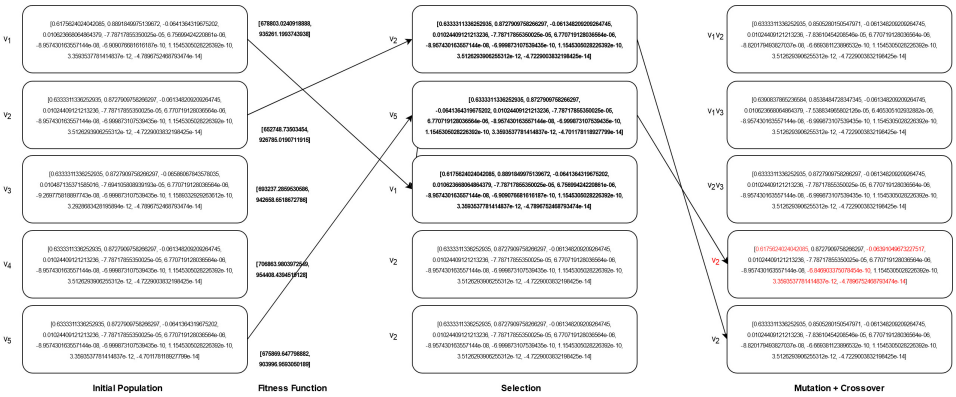
$$C_i = \begin{cases} A_i, & \text{with } p_i = 0.6 \\ B_i, & \text{with } q_i = 0.4 \end{cases}$$

   Mutation The mutation was also probabilistic. Let M be the new mutation vector.

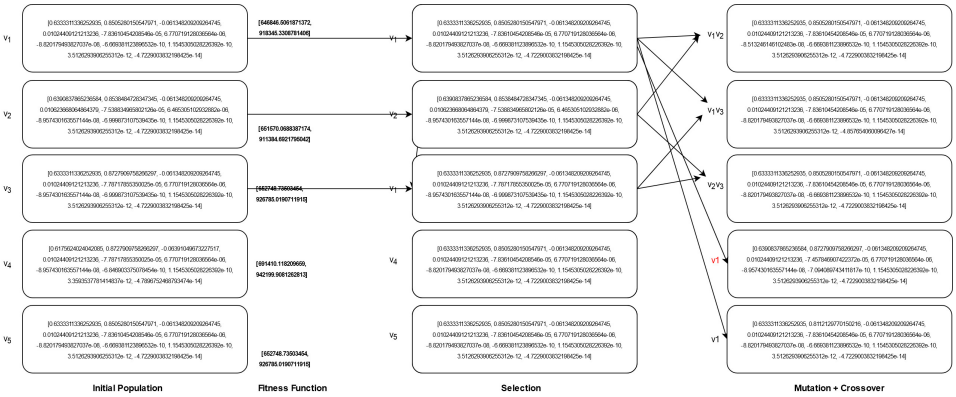
$$M_i = \begin{cases} A_i, & \text{with } p_i = 0.6 \\ \in (0.95 * A_i, 1.05 * A_i), & \text{with } q_i = 0.4 \end{cases}$$

Suppose the chosen population was A,B and C (with A giving the best value in fitness function, B the next best and then C). The new vectors are

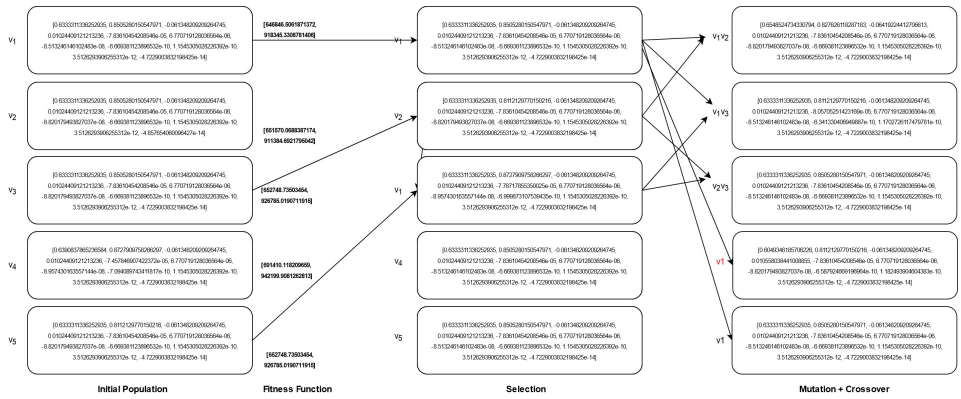
- A-B (Crossover with Mutation).
- A-C (Crossover with Mutation).
- B-C (Crossover with Mutation).
- A with mutations: This is A with the following mutation.
- A without any mutation: This is just plain A without any mutation.



iteration1.jpg



iteration2.jpg



iteration3.jpg

The heuristics We did a probabilistic cross over which is explained above. We introduced the concept of survivability The best vector always survive. The best vector also produces a copy of itself with some mutations. This helps in converging towards a minimum. Also as the error becomes less we did a trick. (A-B)[0] survives only if it is better than vector of previous generation.