

MDL Project Report

Team Name: Sentinels

Team Number : 25

Team Members:

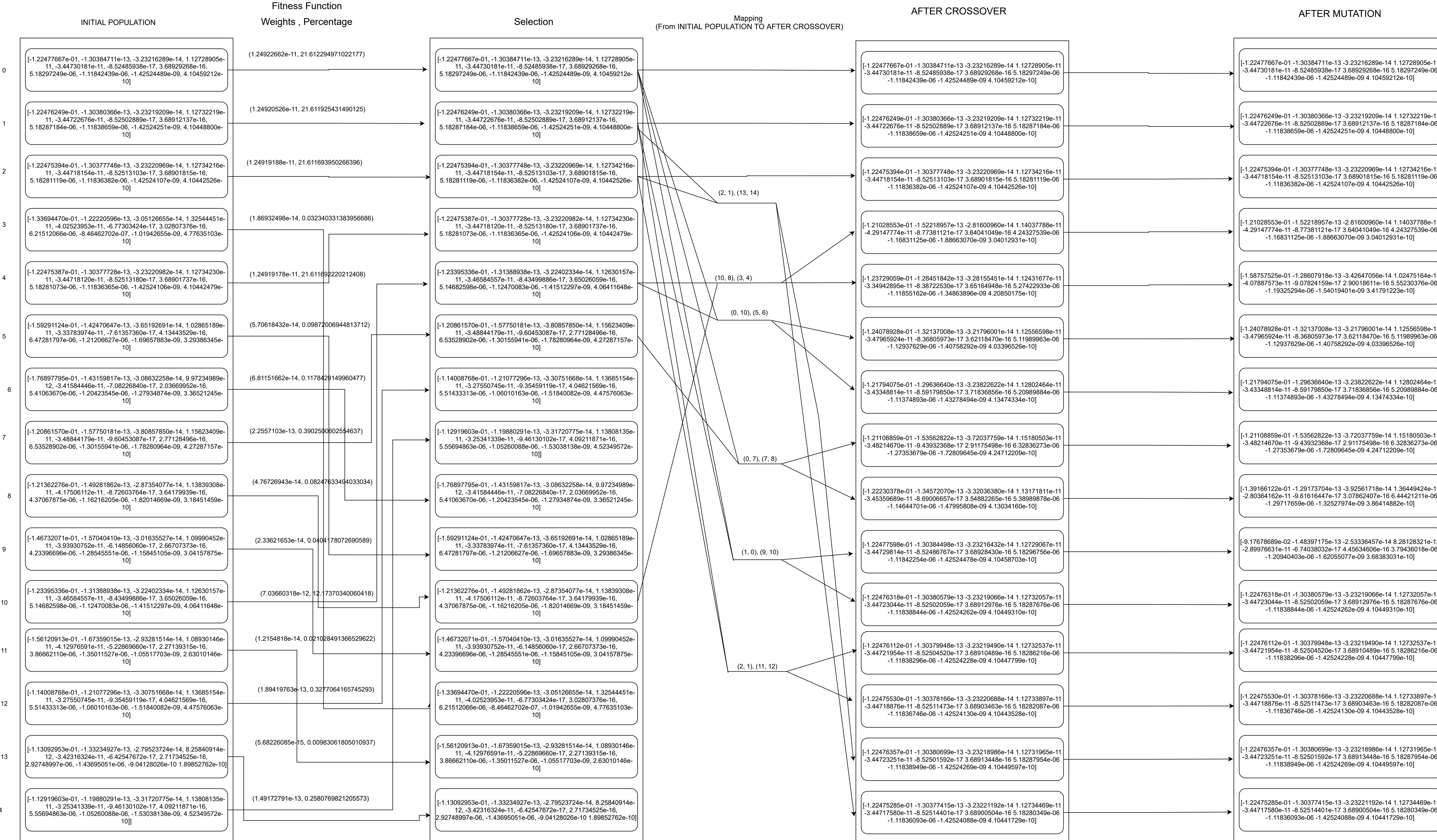
- Pratyanshu Pandey (2019101025)
- Bhaskar Joshi (2019111002)

Genetic Algorithm

1. The algorithm begins with taking the overfit vector as the base vector and creating an initial population with it. The size of the initial population was 5 then was further reduced to 15.
2. Every feature in an individual in the initial population is nothing but the corresponding feature in the overfit vector multiplied by a random number between 0 and 2 both inclusive. This is done in order to reduce the search space from $[-10, 10]$ to a smaller range of values with the overfit vector as a guideline.
3. Thus we have an initial population for a generation. Now we query the server for the train and validation errors and calculate their fitness from the fitness function.
4. Now the vectors are sorted according to fitness in decreasing order. The higher the fitness the better the vector. The top 5 vectors are transferred directly to the population for the next generation. Let's call it the next population.
5. Other vectors for the next population are created by choosing 2 parents from the population and then do crossover on them to create 2 new children. This is done in a loop till the next population has the same size as the initial population.
6. By some probability mutation is applied on individuals of the next population using mutation function.
7. Now the next population becomes the initial population for the next generation.
8. Multiple generations are run till the solution converges and no substantial difference is to be found in new populations.
9. There are multiple heuristics and tricks used in combination with the genetic algorithm to get the result. This was just an overview. See the next sections for more details.

Sample Iteration of Genetic Algorithm

Note : The generations displayed are Generation 250-252 of file generations_1.txt



Fitness Function

INITIAL POPULATION

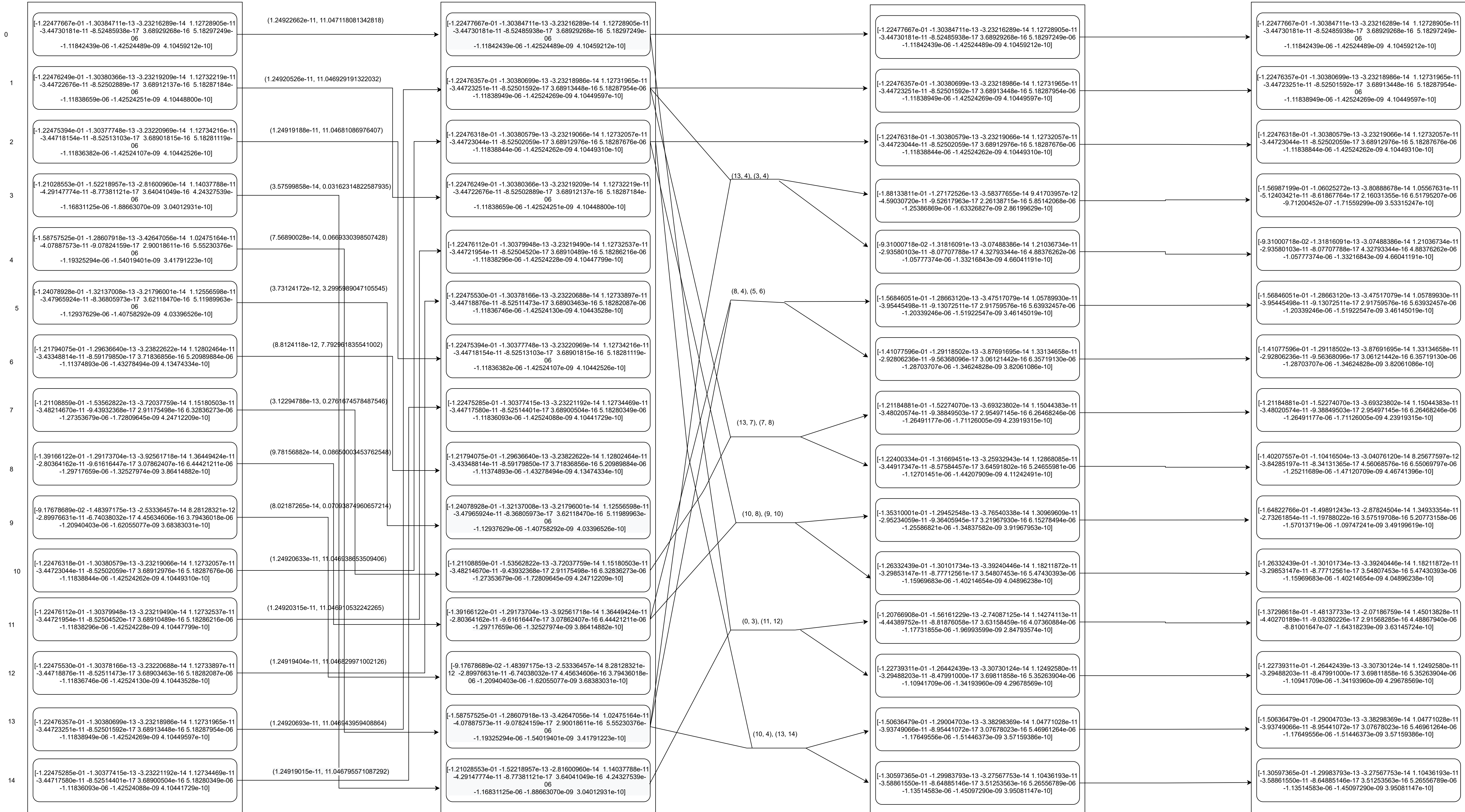
Weights , Percentage

Selection

Mapping
(From INITIAL POPULATION TO AFTER CROSSOVER)

AFTER CROSSOVER

AFTER MUTATION



Fitness Function

INITIAL POPULATION

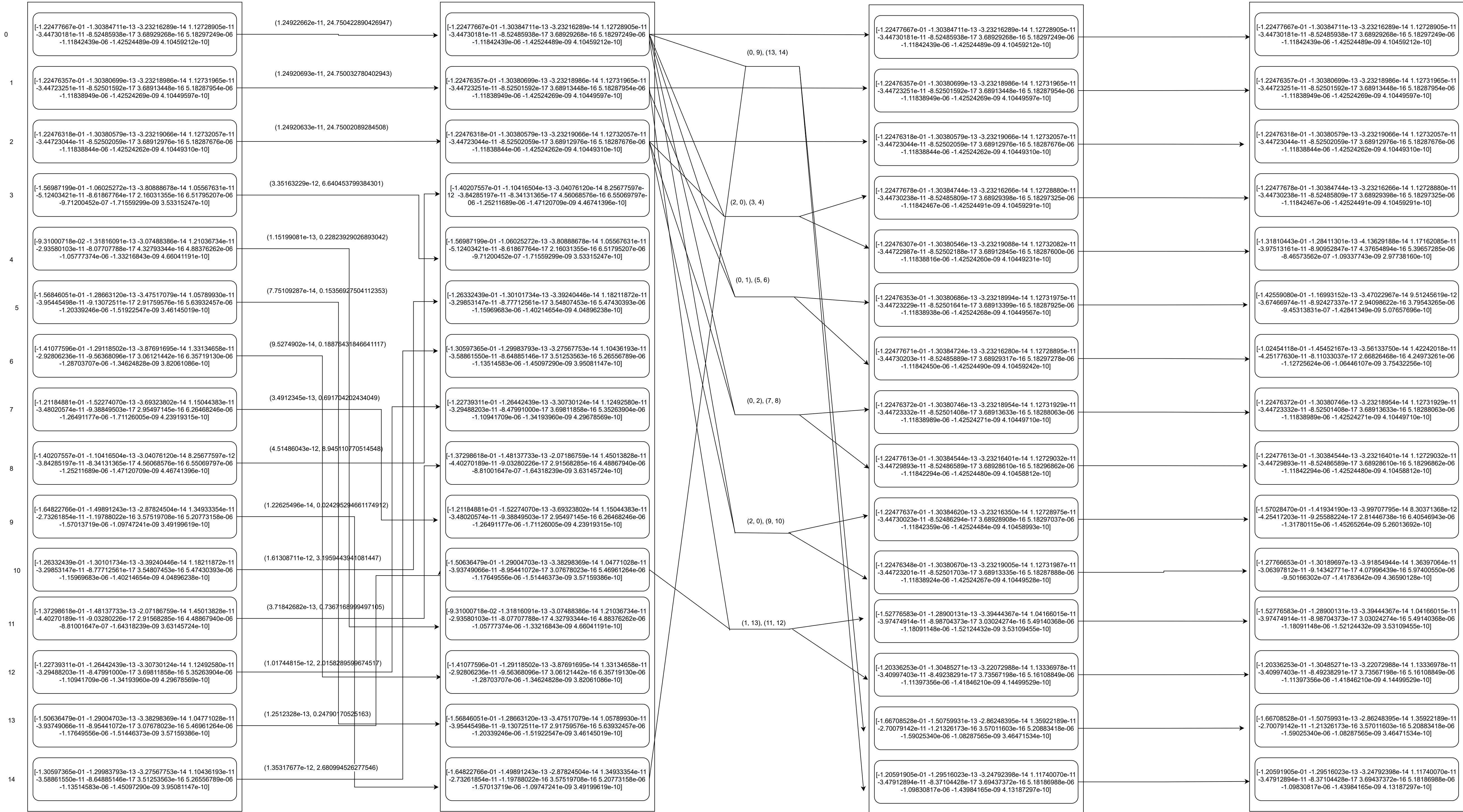
Weights , Percentage

Selection

Mapping
(From INITIAL POPULATION TO AFTER CROSSOVER)

AFTER CROSSOVER

AFTER MUTATION



Fitness Function

```
def calc_errors(individual, precalc):
    for item in precalc:
        if list(individual) == list(item[0]):
            return list(item[1][1:])
    else:
        return get_errors(ID, list(individual))

# population = (POP_SIZE * 11) matrix
def calc_weights(population, precalc):
    weights = np.zeros((len(population), 3), dtype=float)
    for i in range(len(population)):
        train_err, validation_err = calc_errors(population[i], precalc)
        weight = 1 / ((TRAIN_RATIO * train_err + VAL_RATIO *
validation_err) + DIFF_RATIO * abs(train_err - validation_err))
        weights[i] = weight, train_err, validation_err
    return weights
```

The fitness function is called `calc_weights` and it takes 2 arguments: the population for which we calculate the weights and `precalc`- the precalculated weights.

Since we are always storing 5 top individuals for the next population we need not waste a query on getting their errors as we already have them from the previous generation. Thus the top 5 individuals along with their respective train and validation errors are stored in `precalc` for saving queries.

The fitness value called `weight` is calculated using:

```
weight = 1 / ((TRAIN_RATIO * train_err + VAL_RATIO * validation_err) +
DIFF_RATIO * abs(train_err - validation_err))
```

$$weight = \frac{1}{TRAIN_RATIO * train_err + VAL_RATIO * validation_err + DIFF_RATIO * abs(train_err - validation_err)}$$

Thus high values of train and validation errors and the difference between them will mean that the vector is not good and will have low fitness value while vectors having low train and validation errors and having less difference in errors will have higher fitness values.

This fitness function is a good measure of the model as it takes everything into account.

The values of `TRAIN_RATIO`, `VAL_RATIO`, `DIFF_RATIO` are changed from time to time depending upon the errors the current generation is producing. If train error is high, `TRAIN_RATIO` is increased. Similarly for validation error, difference.

Selection Function

```
def selection(population, weights):
    length = len(population)
    od_weights = weights[:, :1].flatten()
    sum_weights = sum(od_weights, start=0)
    normal_weights = [weight / sum_weights for weight in od_weights]
    indexes = np.random.choice(a=length, size=2, replace=False,
p=normal_weights)
    parents = [population[index] for index in indexes]
    return parents, np.array(indexes)
```

Selection function is much simpler. It takes all of the individuals in the population and their corresponding weights as arguments. Then we normalize the weights by dividing them by their sum so that they add up to 1. Now these normalized weights can act as probability for selection of a vector for being a parent in crossover. We thus choose and return 2 vectors that can act as parent for crossover.

Crossover Function (Simulated Binary Crossover)

```
# p1, p2 are parents each a vector of degree MAX_DEG
def simulated_binary_crossover(p1, p2):
    u = random.random()
    b = None
    if u <= 0.5:
        b = (2 * u) ** (1 / (DISTRIBUTION_INDEX + 1))
    else:
        b = (1 / (2 * (1 - u))) ** (1 / (DISTRIBUTION_INDEX + 1))
    c1 = 0.5 * ((1 - b) * p1 + (1 + b) * p2)
    c2 = 0.5 * ((1 + b) * p1 + (1 - b) * p2)
    return c1, c2
```

Simulated binary Crossover is used for the purpose of simulating single point binary crossover for real space. The idea is to calculate childrens c1 and c2 from parents p1 and p2 such that:

$$\frac{c_1 + c_2}{2} = \frac{p_1 + p_2}{2}$$

This is done by first choosing a random number u in [0,1). Then we calculate b using:

$$b = \begin{cases} (2u)^{\frac{1}{n_c + 1}}, & \text{if } u \leq 0.5 \\ \left(\frac{1}{2(1-u)} \right)^{\frac{1}{n_c + 1}}, & \text{if } u > 0.5 \end{cases}$$

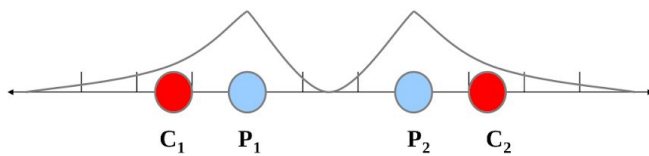
where n_c is the *DISTRIBUTION_INDEX*

Then the children are calculated with the help of p_1 , and p_2 .

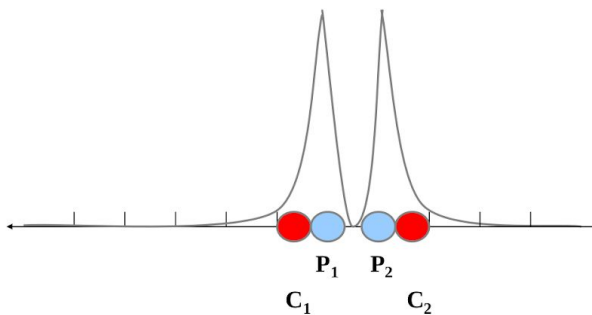
$$c_1 = 0.5 * ((1 - b) * p_1 + (1 + b) * p_2)$$

$$c_2 = 0.5 * ((1 + b) * p_1 + (1 - b) * p_2)$$

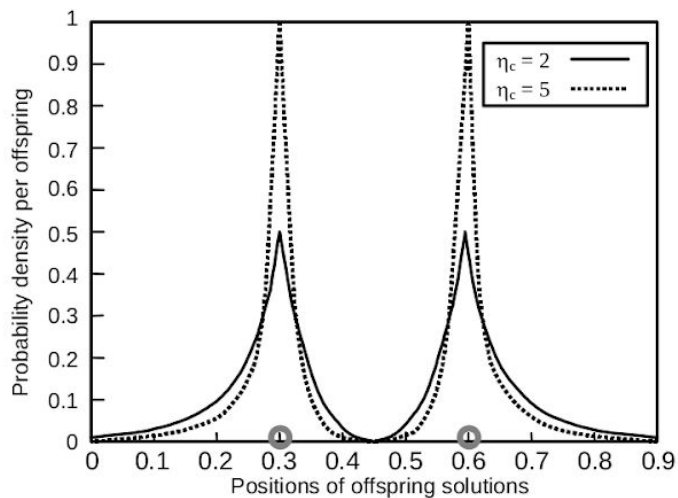
The graph below shows the probability density of offspring if the 2 parents are far enough.



The graph below shows the probability density of offspring if the 2 parents are close.



The graph below shows variation in probability density of offsprings for 2 values of distribution index 2 and 5.



Mutation Function

```
def mutation(population):
    for individual in population[ELITISM:]:
        if random.uniform(0, 1) < MUT_PROB:
            for index in range(11):
                individual[index] += individual[index] * random.uniform(-MUT_RANGE,
MUT_RANGE)

    for individual in population:
        for index in range(11):
            individual[index] = max(-10, individual[index])
            individual[index] = min(10, individual[index])
    return population
```

Mutation function takes population as input then it skips over the top elites that go directly to the next population. For the remaining individuals of the population, they get mutated with a probability that is defined by the hyperparameter `MUT_PROB`.

If an individual is to be mutated then every feature of that individual is multiplied by a random number in range $[1 - \text{MUT_RANGE}, 1 + \text{MUT_RANGE}]$. This multiplication factor is calculated independently for each feature.

Once the mutation is done a check is applied to ensure that the features of vectors remain within the acceptable bounds of $[-10, 10]$. If not they are set to the closest acceptable value.

Hyperparameters

POP_SIZE

The population size of a generation i.e. the number of individuals in the generation. It was initially set to 25 to have more vectors and more variations. But after a point it was becoming difficult to manage 25 individuals with less queries so we shifted to 15.

INITIAL_VECTOR

The initial population is formed by taking an initial vector as the first individual of the population. Then a new individual is created by taking each feature of this vector and multiplying it with a randomly chosen number in the range $[0,2]$. This is done independently for each feature. This process is done in a loop till we reach POP_SIZE number of individuals.

The INITIAL_VECTOR is in a way a very rough guidance to the initial population. It also reduces the search space from $[-10,10]$ for each feature to a very small amount. Thus the INITIAL_VECTOR was chosen to be the given OVERFIT VECTOR for the first time and was changed to other vectors when the situation demanded. (Explained in heuristics)

ELITISM

This is the number of top individuals transferred directly to the next generation. Elites have a not so subtle tendency to influence the vectors to converge in their direction. This speeds up convergence of solution and to ensure that it does not converge quickly to a local minima the MUT_RANGE is kept high.

MUT_PROB

The probability with which mutation occurs in a non-elite individual. This is kept quite stationary at 0.4 or maybe 0.5. These values ensure sufficient mutation and variation but also sufficient inheritance from parents.

MUT_RANGE

If an individual is to be mutated then every feature of that individual is multiplied by a random number in range $[1 - \text{MUT_RANGE}, 1 + \text{MUT_RANGE}]$. This multiplication factor is calculated independently for each feature. MUT_RANGE is kept high at 0.3 if large variations are needed for convergence. But once we have converged, this is systematically reduced to lower values like 0.1 and 0.05 to figure out finer variations in data.

GENERATIONS

Number of generations to run in every iteration of running the code. The lower the generation i.e. 10 to 15 the more tuning can be done at every point. But once vectors converge this can be increased to 20 to 30, as then the noticeable changes in vectors are hard to come by.

VAL_RATIO

VAL_RATIO is the coefficient of validation error in the fitness value. If VAL_RATIO is high then the algorithm prefers vectors with lower validation errors as this reduces the overall error and increases the fitness value.

Thus if validation error is very high compared to train error we increase VAL_RATIO and vice versa.

TRAIN_RATIO

TRAIN_RATIO is the coefficient of train error in the fitness value. If TRAIN_RATIO is high then the algorithm prefers vectors with lower train errors as this reduces the overall error and increases the fitness value.

Thus if train error is very high compared to validation error we increase TRAIN_RATIO and vice versa.

DIFF_RATIO

DIFF_RATIO is the coefficient of difference of train and validation errors in the fitness value.

If DIFF_RATIO is high then the algorithm prefers vectors with closer train and validation errors as this reduces the overall error and increases the fitness value.

Thus if the difference between errors is very high, we increase DIFF_RATIO and vice versa.

DISTRIBUTION_INDEX

In Simulated Binary Crossover, this is a measure of how far the children are from the parents. Large values of DISTRIBUTION_INDEX allows children to be closer to parents.

While small values mean children are farther from parents. 2 to 5 is the moderate range and the range we use. If the solution has not converged the DISTRIBUTION_INDEX is kept at 2, to get good variations. It is then systematically increased to 5 as the solution converges to have finer variations.

Heuristics and Statistical Information

Initially we chose to use the following hyperparameters:

- INITIAL_VECTOR = OVERFIT VECTOR
- POP_SIZE = 25
- ELITISM = 5
- MUT_PROB = 0.4
- MUT_RANGE = 0.3
- GENERATIONS = 10
- VAL_RATIO = TRAIN_RATIO = DIFF_RATIO = 1
- DISTRIBUTION_INDEX = 2

This is done to have high variations in vectors but also to converge it. Initially the errors are of the order $1e14$ but converge very quickly in 4 to 5 generations to order of $1e12$. In another 3 to 4 generations they converge to the order of $e11$. The algorithm then got stuck in the range $1e11$ to $3e11$. In 70 generations we finally broke through the $e10$ barrier. Throughout this we would change TRAIN_RATIO, VAL_RATIO and DIFF_RATIO depending on the errors we received.

In a failed heuristic, we continued by reducing mutation range and increasing DISTRIBUTION_INDEX and running 100 to 200 generations more with 25 as POP_SIZE. This did not work too well and even when we reached $5e10$ train and validation errors, the

leaderboard rank was not so good. We theorize that this was most probably due to overfitting.

So we changed the heuristic.

So we looked up test errors for different ranges of errors for different vectors. This was done in order to get a range of values where we are close to vectors with perfect fit but do not overfit.

We found such a vector and then took it as the INITIAL_VECTOR and restarted the genetic algorithm but this time with a reduced POP_SIZE of 15 as 25 vectors were taking a lot of queries without substantial results.

The MUT_RANGE value was 0.3 and we only changed it to 0.1 later but no further decrease. This was to ensure that overfitting does not take place and there is substantial variation in the data. The ratios were again altered dynamically according to latest data. ELITISM took values between 1,3 and 5. The elite vectors do not alter themselves but do contribute towards convergence by participating in selection and crossover. Thus the more the elite vectors the more influence they have. So we sometimes reduced ELITISM if we wanted to keep more variations in the generations.

Again the solutions converged very soon.

We repeated this process of restarting with an initial vector a few more times and then checked test errors for vectors in the range $1e11$ to $3e11$ and found very good vectors.

Final Vectors and Errors

The top vector is :

$[-0.11659813562929511, -1.2885110929990298e-13, -3.577791708504743e-14, 1.2646749212560554e-11, -3.3831773009699504e-11, -9.284248071512872e-17, 4.003342760912018e-16, 6.398320041396442e-06, -1.1245508241603073e-06, -1.4072152000100355e-09, 4.1834156007562354e-10]$

With train error = 103090039506.4589 and validation error = 312500882005.9999

While we were able to achieve train errors and validation errors as low as $4e10$ and $2.8e10$, these vectors did not give good results. This is because these vectors overfit both the train and validation datasets and thus performed badly on test dataset.

The vector we selected performs good on the test dataset because it does not overfit anything and all the 3 errors test, train and validation are close to each other. None of the errors are too high compared to one another. This is a good sign that the vector is a good fit.

Tricks

1. Since we are using ELITISM, it is bound to be the case that some vectors get directly transferred to the next population. But we already have the errors and weights for

these vectors and there is no need to calculate them again by wasting queries. So these errors along with the corresponding vectors are stored in precalc and are reused for the next generation.

2. One trick used was to initially map out the sensitivity of each feature of the data and how reliable the OVERFIT VECTOR can be for the initial population.