PUNE INSTITUTE OF COMPUTER TECHNOLOGY DHANKAWADI, PUNE-43.

A PROJECT REPORT

ON

APPLICATION OF GENETIC ALGORITHM FOR FEATURE SELECTION

SUBMITTED BY

Maitraya Uttam Kakade (41427)
Pranav Kulkarni (41430)
Vaibhav Marathe (41433)

Class : BE4

Guided By:
Prof. Urmila Pawar

Problem Statement:

Apply the Genetic Algorithm for optimization on a dataset obtained from UCI ML repository.

Objective:

1) Apply genetic algorithm on dataset from UCI ML repository
2) Understand genetic algorithm and its application for optimisation

Outcome:

1) Students will have applied genetic algorithms on a dataset from UCI ML repository.

INTRODUCTION:

Feature selection is one of the most important tasks to be performed for any machine learning tasks. While feature selection can be done using some domain knowledge and data analysis, we can use genetic algorithms to help in this task. Using genetic algorithms we can find the best subset of features for a given machine learning task. This is what we have attempted to do here. We use a genetic algorithm approach to find optimal features of the iris dataset and use logistic regression for classification.

Theory:

Genetic algorithm:

Genetic Algorithms are a class of stochastic, population based optimization algorithms inspired by the biological evolution process using the concepts of **"**Natural Selection**"** and **"**Genetic Inheritance**"** (Darwin 1859) and originally developed by Holland.

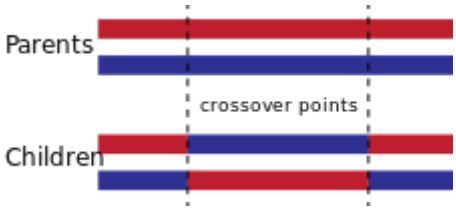
GAs are now used in engineering and business optimization applications, where the search space is large and/or too complex (non-smooth) for analytic treatment. This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation.

The process of natural selection starts with the selection of fittest individuals from a population. They produce offspring which inherit the characteristics of the parents and will be added to the next generation. If parents have better fitness, their offspring will be better than parents and have a better chance at surviving. This process keeps on iterating and at the end, a generation with the fittest individuals will be found. This notion can be applied to a search problem. We consider a set of solutions for a problem and select the set of best ones out of them.

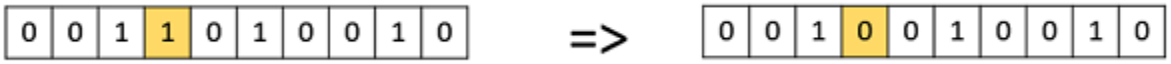
Five phases are considered in a genetic algorithm.

1. Initial population:
    a. The process begins with a set of individuals which is called a **Population**. Each individual is a solution to the problem you want to solve. An individual is characterized by a set of parameters (variables) known as **Genes**. Genes are joined into a string to form a **Chromosome** (solution).
    b. In a genetic algorithm, the set of genes of an individual is represented using a string, in terms of an alphabet. Usually, binary values are used (string of 1s and 0s). We say that we encode the genes in a chromosome.
2. Fitness function:The **fitness function** determines how fit an individual is (the ability of an individual to compete with other individuals). It gives a **fitness score** to each individual. The probability that an individual will be selected for reproduction is based on its fitness score.
3. Selection:The idea of the selection phase is to select the fittest individuals and let them pass their genes to the next generation. Two pairs of individuals (**parents**) are selected based on their fitness scores. Individuals with high fitness have more chances to be selected for reproduction.

4. Crossover:Crossover is the most significant phase in a genetic algorithm. For each pair of parents to be mated, a **crossover point** is chosen at random from within the genes. For example:



5.
6. Mutation:In certain new offspring formed, some of their random probability. This implies that some of the bits in the bit string can be flipped.



7. **Termination:** The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation). Then it is said that the genetic algorithm has provided a set of solutions to our problem.

Advantages of GAs

- Does not require any derivative information (which may not be available for many real-world problems).
- Is faster and more efficient as compared to the traditional methods.
- Has very good parallel capabilities.
- Optimizes both continuous and discrete functions and also multi-objective problems.
- Provides a list of "good" solutions and not just a single solution.
- Always gets an answer to the problem, which gets better over time.
- Useful when the search space is very large and there are a large number of parameters involved.

Limitations of GAs

- GAs are not suited for all problems, especially problems which are simple and for which derivative information is available.
- Fitness value is calculated repeatedly which might be computationally expensive for some problems.
- Being stochastic, there are no guarantees on the optimality or the quality of the solution.
- If not implemented properly, the GA may not converge to the optimal solution.

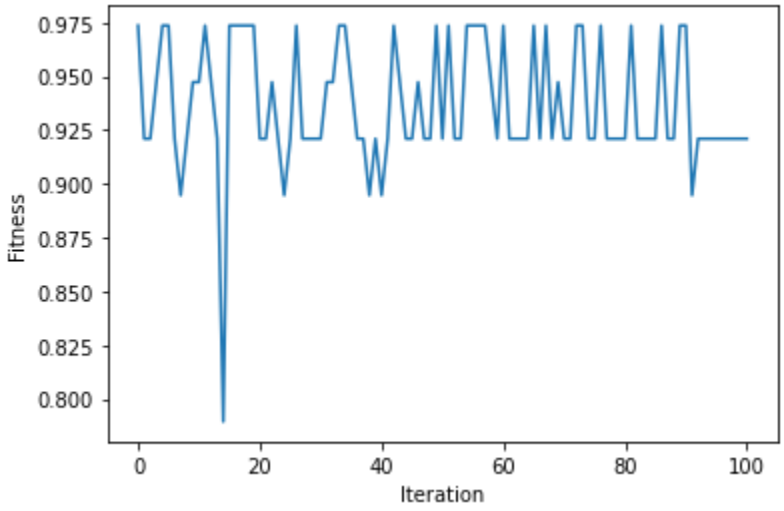
Application:

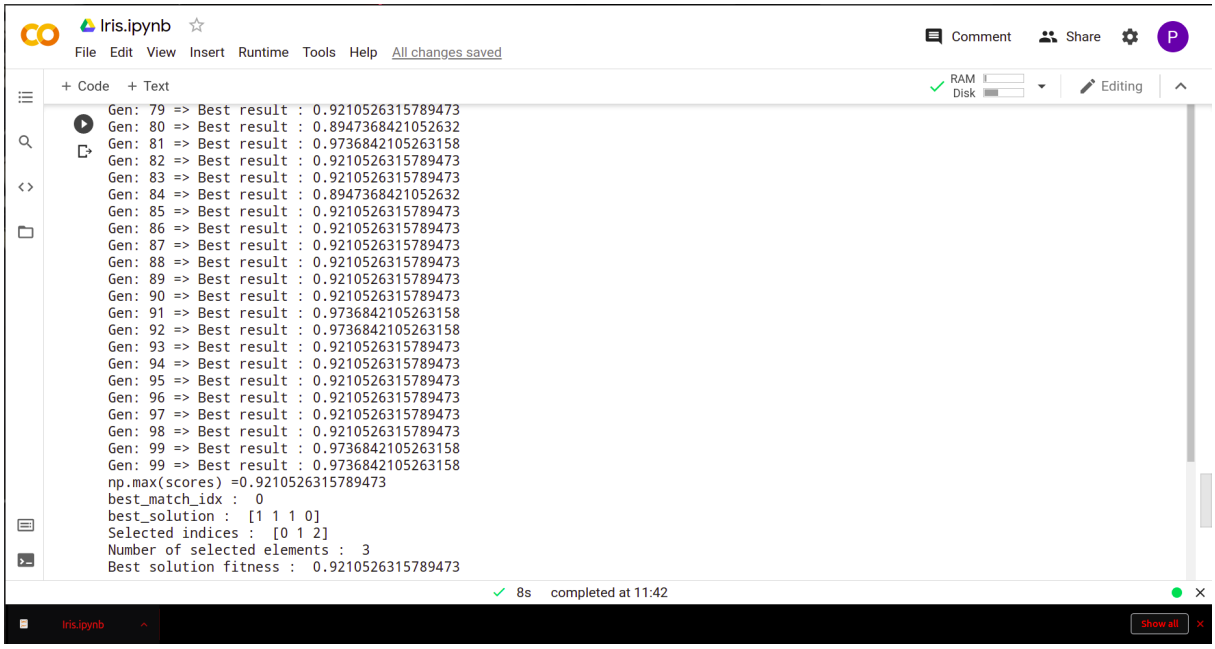
The dataset we consider is IRIS dataset.The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.The task is to predict the class attribute of the iris plant.
We use genetic algorithms to find the optimal number of features to do this task. There are a total of 4 features and we use genetic algorithms to find the optimal number of features. The classification algorithm we use is Logistic Regression. In each generation we select a subset of features as parent and calculate the classification accuracy. These parents are then used to generate the next generation through single point crossover and mutation.

Algorithm:

1. initialize population
2. find fitness of population
3. while (termination criteria is reached) do
    a. parent selection
    b. crossover with probability pc
    c. mutation with probability pm
    d. decode and fitness calculation
    e. survivor selection
    f. find best
4. return best

Results:





Conclusion:

We have successfully used genetic algorithms for optimising the number of features required for classification of Iris dataset using Logistic regression.