Genetic Algorithms for Feature Selection

Goals:

- Implement a genetic algorithm to select the most impactful features in a dataset.

- Show the results of the feature selection and compare results without any feature selection.

Deadline: April 10th,2022

Assignment Background and Motivation

Feature selection is a machine learning process where redundant and irrelevant features are removed from a dataset. Evolutionary algorithms including genetic algorithms have been successfully used for the purpose of feature selection. The selected features will often provide approximately as good, or better, results, and having data with less dimensions have a few advantages when training a predictive model:

- Faster training time

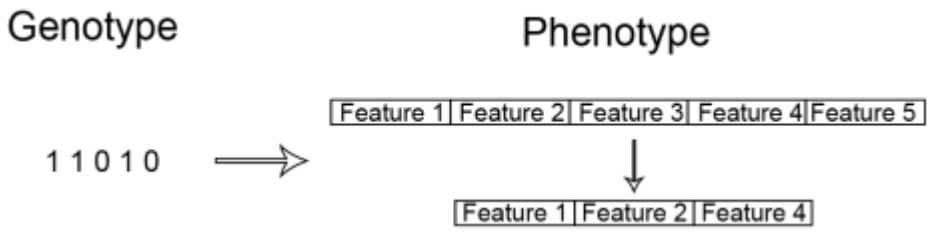
- Reducing overfitting

- Requiring less data entries

In this project you are tasked with creating a genetic algorithm for finding the best features for a given dataset. The features will be passed to a simple machine learning algorithm which returns the accuracy. This accuracy can be used as a fitness score for the implementation of the genetic algorithm for this problem. The objective is to maximize the accuracy.

Genetic Algorithm

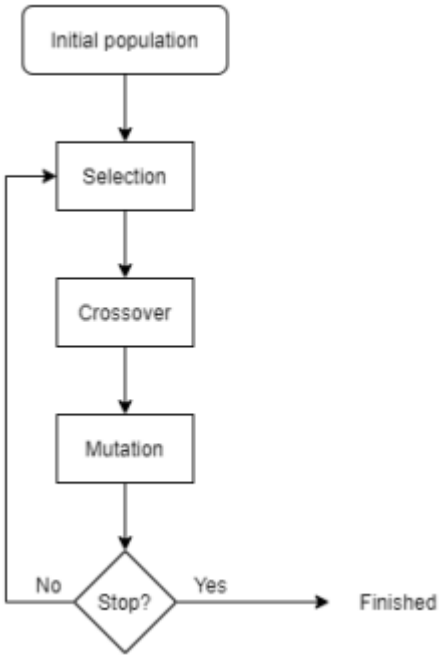
To solve this problem, you will implement a genetic algorithm (GA). It is recommended to represent the individuals as bitstrings. In the feature selection problem, each bit decides whether a feature should be included or not.



Or you can choose a subset of features as discussed in class. Note that GA parameter values (population size, generation number, crossover rate, mutation rate, etc.) are important to optimize and are typically correlated. Your GA may successfully find optimal or near-optimal solutions if you use the right parameter values. However, there is no definite rule as to how to find such parameter values. Therefore, you should test different sets of parameter values to decide on appropriate values.

Genetic Algorithm:

The simple genetic algorithm consists of generating an initial population, selecting the best parents of the offspring, doing crossover and mutation, and replacing the population. This process is repeated until an end condition is reached.



Fitness Function:

You can use any machine learning algorithm for the provided dataset and return accuracy. This accuracy can be used as a fitness function for a genetic algorithm. It is important to keep in mind that your genetic algorithm should attempt to maximize the accuracy.

Dataset:

The dataset to be used for machine learning and feature selection consists of 8482 rows each having 58 columns. The first 57 columns represent the data, while the last column represents the value of the row.

There will be a demo for this assignment.

Here are the tasks you need to implement and be ready to demonstrate and answer questions about during the demo day.

- Implement a function to generate an initial population for your genetic algorithm.
- Implement a parent selection function for your genetic algorithm. You have to implement selection method as discussed in the class.
- Implement a function that creates two offspring from two parents through crossover. The offspring should also have a chance of getting a random mutation.

Run the genetic algorithm on the provided dataset. Show the results and compare them to the results of not using any feature selection (given by running the machine learning algorithm with all features selected).

Deadline:

You should deliver a zip file with your code on GCR. The submission system will be closed on April 10th.

Functions that must be included in your code (you can make other functions too but these are compulsory):

- Preprocessing
- Create_population
- Fitness_function
- Crossover
- Mutation

Note:

You are not allowed to use any library other than NumPy, Pandas and Scikit-learn.

You cannot drop any column manually in preprocessing

No Deep learning algorithm

Plagiarism and collusion are taken extremely seriously.