Using Genetic Algorithm for Feature Selection

Author: Adel.Ahmadi

1. Introduction

In this chapter, we introduce the Genetic Algorithm (GA) as a feature selection technique to improve our SVM model. Feature selection is a critical step in machine learning pipelines. By using GA, a population-based optimization method inspired by natural selection, we aim to select an optimal subset of features that maximizes classification performance and reduces computations cost for life expectancy prediction. For better performance of GA feature selection, we remove country columns from GA variables.

2. Genetic Algorithm Overview

Genetic Algorithm is inspired from the process of natural evolution, where the fittest individuals are selected to produce offspring for the next generation. Over time, the algorithm converges to an optimal solution.

GA involves the following steps:

1. Initialization: A population of candidate solutions (feature subsets) is created randomly or sometimes with heuristic methods.
2. Fitness Evaluation: The fitness of each candidate is calculated based on the classification accuracy of the SVM.
3. Selection: The best-performing candidates are selected for reproduction of the next generation.
4. Crossover: Selected candidates are combined to produce offspring.
5. Mutation: A small percentage of features in the offspring are randomly altered to maintain diversity. Mutation rate is often around %5.
6. Termination: The process continues for a specified number of generations or until the algorithm converges to some answer.

- **GA Mathematics**

The optimization problem can be formalized as:

Objective:

Maximize $Fitness = Accuracy(X_{Selected})$

Constraints:

1. The number of selected features ≤ Max Features.
2. Binary encoding for feature inclusion:

   $b = [b_1, b_2, ..., b_n]$ where $b_i = 1$ if feature i is selected.

Operators:

- Crossover: Combines the binary vectors of two parents.
- Mutation: Flips (from 0 to 1 or vice versa) with a small probability.

- **Implementation Steps**

The implementation of GA for feature selection involves the following steps:

1. Data Preprocessing
2. Define Fitness Function
3. Run Genetic Algorithm
4. Evaluate Selected Features

- **Results**

After applying the Genetic Algorithm for feature selection, the SVM model was trained on a reduced feature set. The optimal subset of features was determined based on their contribution to model accuracy.

- Selected Features Recommended by GA: Adult Mortality, Alcohol, Income composition of resources, Year, thinness 1-19 years.

- Accuracy: The accuracy of the model after feature selection was 0.88, which is comparable to the original accuracy of 0.91 but with fewer predictors, improving model interpretability.

- **Conclusion**

Using the Genetic Algorithm, we successfully reduced the number of features while maintaining high classification accuracy. This demonstrates the potential of GA as a feature selection method, especially in datasets with a large number of predictors.