

Genetic Algorithm for Variable Selection

STAT 243 - Final Project

Fall 2023

Authors: Sean Zhou, Sun Moon, Frederik Stihler

Department of Statistics

Instructor: Christopher Paciorek

December 13th, 2023



1. Introduction

The goal of this project is to develop a genetic algorithm for variable selection, including both linear regression and GLMs. The results are consolidated in a Python package. The user can input a dataset (with covariates and corresponding responses), as well as the desired type of regression. The algorithm will perform the variable selection and tell the user which features to use.

In general, variable selection is the process of selecting a subset of relevant predictors for the model creation. Variable selection generally decreases the model complexity, which can make it easier to interpret and also reduce computational complexity.

Genetic algorithms (GA) are stochastic methods usually used for optimization or search problems. They utilize principles from biological evolution and natural selection, such as selection, crossover and mutation.

2. Programming approach

2.1 Overall structure

We identified the following main steps in the algorithm:

1. Population initialization
2. Evolution cycle for fixed number of iterations
 - Fitness assessment and ranking
 - Parent selection
 - Genetic operators
 - Crossover
 - Mutation
 - etc.
3. Output fittest individual of final population

Each (sub-)step was implemented as a modular component. We decided to use an object-oriented programming approach, which means that the individual steps are implemented as Python class methods. To keep the code organized in a logical order we applied inheritance so that each step is defined in its own class. Then, these classes and methods are inherited by the parent class “GA”, which has a primary method “select” that carries out the overall algorithm outlined above.

2.2 Default parameter values and initialization

Based on the theory of Givens and Hoeting and the examples in the next sections, we have decided on certain default parameters.

The theory suggests a population size P in the range of $C < P < 2C$, where C is the number of predictors (i.e. number of genes per chromosome). We chose the midpoint of this range as default value if no specific population size is given by the user: $P = 1.5 \times C$. As the population size should remain constant during the evolution of the algorithm and due to the crossover operation requiring an even number of chromosomes, an additional chromosome is added if the population size is odd (the user is informed about this modification). The initial population is fully randomized (random binary choice for all loci of all chromosomes).

The default genetic operators applied per generation are single split crossover and standard mutation. However, the user can also provide his own genetic operators (implemented on the full population) or modify the order of execution. In addition to the default crossover, a version with random allele selection from the parents is included in the package. The commonly used mutation probability of 1% is set as the default.

Lastly, our algorithm does not process the input data (no normalization or scaling of the data). This is something that the user needs to do in advance.

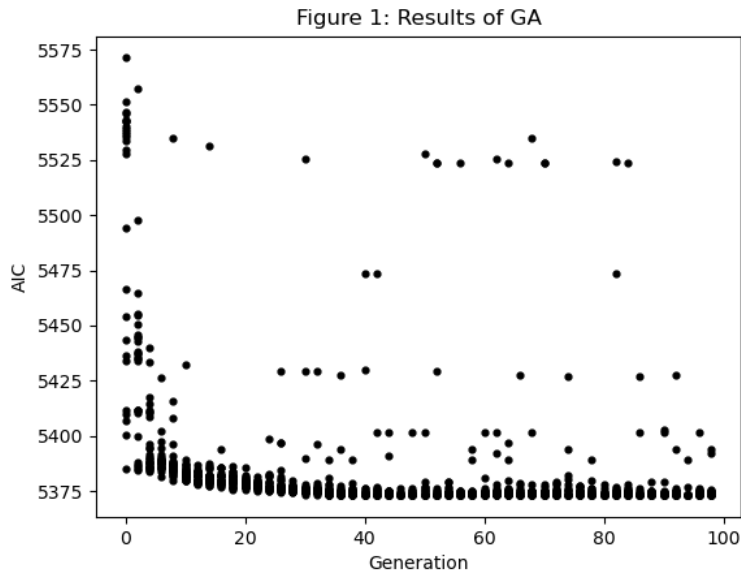
3. Testing

The overall genetic algorithm was tested using simulated and actual datasets, including running OLS and logit models from the “How Couples Meet” and “Stay Together” survey as well as the baseball data set provided by Givens and Hoeting. The individual component functions of the algorithm were also tested in unit tests (see `pytest` test suite for details).

4. Results

4.1 Example 1: Baseball data

We first run the algorithm for the baseball salaries dataset.



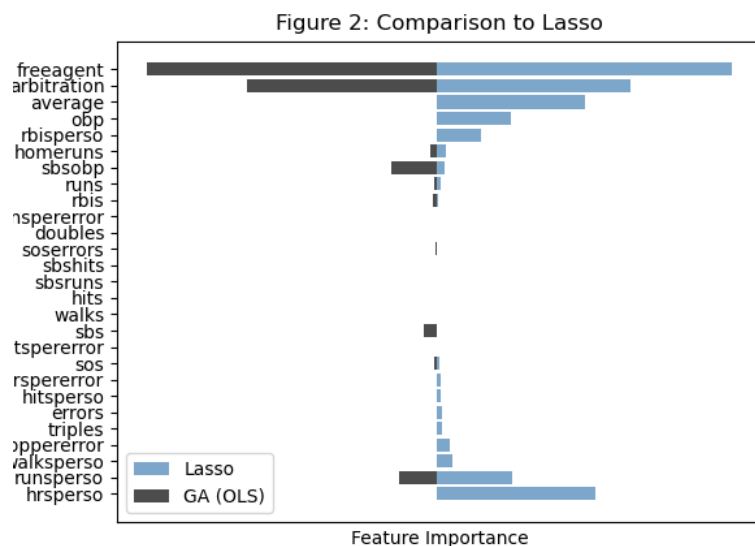
One hundred generations of size 26 were used. The data set includes 27 predictors for the response variable “salary”, so the individual chromosomes consist of 27 binary genes. The genetic operators applied were simple split crossover and a mutation rate of 1% applied independently for each chromosome and each gene position. A clear evolution or “survival of the fittest” can be observed in figure 1. The AIC of the best model was found to be around 5373.6.

4.2 Example 2: Synthetic data

We also explored the algorithms performance on a simple simulated dataset. We draw samples from five independent uniform random variables $X_1, \dots, X_5 \sim Unif(0, 1)$. We then construct the response variable as: $y = 3 \times X_1 + 2 \times X_2 + \epsilon$, where ϵ is Gaussian noise. In this straightforward example, the GA identified predictors X_1 and X_2 as the only relevant features after 7 iterations using the default parametrization outlined above.

4.3 Comparison with Lasso

Finally, the performance of the algorithm was compared to a Lasso regression (using the statsmodel package with regularization parameter $\alpha = 0.1$ on the baseball data). The Lasso method can perform variable selection by setting some of the coefficients to zero, which will “deselect” unimportant features.



The GA selected 10 predictors for the regression, meaning that 17 features were ignored. Comparing this to the coefficient sizes of the Lasso shows that similar features are considered relevant by the two models. In figure 2, the size of the horizontal bars indicates the importance of each feature (corresponding to the respective coefficient weights).

