Genetic algorithm for feature engineering in classification tasks

Abstract

The problem of feature engineering using genetic algorithm is considered. Feature creation is one of the key elements of the contemporary machine learning which includes classification task. Two classification algorithms, random forest and k nearest neighbors, are discussed. Logistic loss metric and cross-validation technique for measuring classification performance are described. Genetic algorithm and its possible variants are presented and then implemented in Python. Usability of the implementation is experimentally verified using a sample group of medical data sets. Comparison of models fitted on raw data set, on data set composed of base features and new ones, and only on newly created features is made.

Keywords: genetic algorithm, classification, feature engineering