Bio AI

Project 3 - Step 6

Step 6 is about finding the optima in specific landscapes. In this document you will find the details about implementation of the ML model you need to use and which parameters to use, how to map the bitstring to the dataset, as well as the specific tasks and datasets you should use.

Task 1 is an *easy* task, with plenty of local optima with respect to search space size.

Task 2 is a *somewhat difficult* task, with a lower proportion of local optima.

Task 3 is a *hard* instance, with plenty of plateaus and very few local optima.

Tasks 1 & 2 are in this document, and Task 3 will be provided later in the week.

# The ML model

* Using this model: **Random Forest**
* Set these specific hyper-parameters:
  + **30 trees** (n\_trees = 30)
  + No limits on maximum depth tree (max\_depth=0)
  + No minimum information gain to allow for a node's partition (min\_gain=0)
  + Minimum number of records a node must have to consider partitioning: 2 (min\_records=2)
  + Maximum number of features to consider at each partitioning: 0 (max\_features=0)
  + Splitting criterion should be according to **Gini index** (or *impurity*)
  + Beta parameter should be 0, i.e., **every tree should have the same weight**.
* Train **30 random forests** using the **selected columns**, record their accuracy, and calculate **the average**. This is the fitness of **1 individual**.

**Default Parameters**

If the implementation of Random Forest you are using does not have some of these parameters, **just use the default**. Just make sure you use 30 trees and the Gini index.

# 

# Training Considerations

**Reverse order of columns when training**

We consider a *reverse* order of columns while training! This means that binary strings grow from right to left. And so, individual 1 will get the rightmost column. For example, assume you have a 5 column dataset:

* An individual [0, 0, 0, 0, 1] will return the last column (column 5). Its encoding in decimal is 1.
* An individual [0, 0, 1, 1, 1] will return the last 3 columns (columns 3, 4 and 5). Its encoding in decimal is 7.
* An individual [1, 0, 0, 0, 0] will return the first column (column 1). Its encoding in decimal is 16.

**Keep this in mind** while comparing your individuals with our solutions!

# What to include in your report

You must include the following items in your report, per task:

1. **The individual** that achieved the solution, and **its function value**
2. **A screenshot** showing your algorithm **reaching the solution** and its function value

We highly **recommend** that you also include the **Hamming distance** between **your found solution and the solution we provide**. That will also make it easier for you to debug.

As per usual, **your code** needs to accompany your report in your Blackboard submission as a separate file(s).

# 

# Task 1

* Using this dataset: [**Heart Disease (Cleveland)**](https://archive.ics.uci.edu/dataset/45/heart+disease)
  + Use the processed.cleveland.data file (303 rows, 13 columns)
* For the fitness function, minimize
  + is the classification error, obtained as
  + Set , that is, no penalty.

## The solution

* **1 global optimum**
* **270 local optima** (including the global optimum)

Get **any** of the local optima with an error value **less than 0.425**, i.e.,

* 2 points if the Hamming distance from the global optimum is .
* 1 point if the Hamming distance from the global optimum is
* 0 points otherwise.

### Check yourself

We provide a matrix in HDF5 format that you can use to check your results:

.

└── data (8191 x 2)

├── accuracy

└── time

The matrix is ordered, from individual 1 to individual 8191.

You have to calculate the error yourself.

## 

# Task 2

* Using this dataset: [**Zoo**](https://archive.ics.uci.edu/dataset/111/zoo)
  + Use the zoo.data file (101 rows, 16 columns)
* For the fitness function, minimize:
  + Set
  + Set

## The solution

* **1 global optimum**
* **958 local optima** (including the global optimum)

Get **any** of the local optima with an error value **less than 0.65**, i.e.,

* 2 points if the Hamming distance from the global optimum is .
* 1 point if the Hamming distance from the global optimum is
* 0 points otherwise.

### Check yourself

We provide a matrix in HDF5 format that you can use to check your results:

.

└── data (65535 x 2)

├── accuracy

└── time

The matrix is ordered, from individual 1 to individual 65535.

You have to calculate the error and regularization using the provided penalty yourself.