MATH0462 - Discrete Optimization
Project

Box search for data mining

Data mining consists in finding relevant information in a large set of data that is usually automatically generated. Companies try to automatically collect as much data as possible. An important question is therefore to understand how to handle this new information correctly. In this project, we are interested in a particular problem that arises in this context and that may be of interest for very complex industrial processes. In particular, we consider a process for which we automatically retrieve a number of variables that might influence a given output variable. For example, we may consider a line creating some products and the output variable could be an estimation of the quality of the product. It is clear that a large number of variables may influence the quality of the production and we assume that most of them can be retrieved along the way for each produced item. The question that we ask is to find a set of rules, i.e. intervals on the variables, for which the output is maximized.

More specifically, we assume that each produced item i is represented by a D-dimensional vector $x^{(i)}$ of the input variables. The output value of the item i is given by $c_i \in \mathbb{R}$. We set up a threshold for the output value C. All data points are split in two sets $X = \{x^{(i)} \mid c_i > C\}$ and $Y = \{x^{(i)} \mid c_i < C\}$. On the other hand, we define the concept of a box, defined by a set of rules $B^{(u_1,\ldots,u_D)}_{(l_1,\ldots,l_D)} = \big\{x \in \mathbb{R}^D \mid l_i \leq x_i \leq u_i \text{ for all } i \in \{1,\ldots,D\}\big\}$.

x_i = vector of input variables for item i

A data point is x_i, a row.

c_i = quality/output of item i, last column

C is given

Each data point is split in 2 based on a threshold C.

X = all data points for which c_i > C
Y = all data points for which c_i < C

Box is a set of rules that restraints x_i

Goal is to find Box that is maximize

The goal of this project is to find the values $(l_1,\ldots,l_D)$ and $(u_1,\ldots,u_D)$ that define a box B such that $B \cap X = \emptyset$. You can download a file on the `dox` repository that includes a database coming from an industry. Each row of the file corresponds to a datapoint. The last column is the output. The threshold C is considered as a parameter of the problem.

Questions

1. Write a MIP model that finds a box B that has the largest value of $\sum_{i=1}^{D}(u_i - l_i)$. For this problem to make sense, you must normalize all variables in such a way that they all belong to $[0,1]$.
2. Write a heuristic that solves the same problem.
3. Write a MIP model that finds a box B that maximizes the cardinality of $B \cap Y$.
4. Write a heuristic that solves the same problem.

Instructions

All projects will be done by groups of 2. You must write a very short report (3 pages maximum) including a human-friendly short version of your models and your heuristics. Everything (code+report) should be sent by e-mail to `q.louveaux@uliege.be`. The deadline for submitting your project is May 16. The presentation of the project should be done on Wednesday May 18. No formal presentation is needed but you should possibly be able to discuss the various tests that you have performed.