The Earth Mover Distance and linear programming

François Meunier, Valentin Tréan

March 2019

1 Question 2

lp_solve is based on the revised simplex algorithm. It works in a similar fashion as the simplex algorithm but improves it computational cost by computing only the required data at each step.

The simplex problem consists of several steps :

1. Compute the canonical form of the LP problem (generally by adding slack variables) $a_{r,c}$
2. Select a column that has a positive entry in the objective row of the canonical matrix, denoted $r*$.
3. Select a row using the minimum ratio test, denoted $c*$
4. Divide column $r*$ by $a_{r*,c*}$
5. Subtract to each row $x \times r*$ with x such that $a_{r,c*} = 0$
6. Repeat at 2 until there are no more positive entries.

The final solution of the simplex algorithm is optimal. The revised simplex algorithm performs the same steps but only keep into memory the minimum information and computes additional data whenever needed.

The solver's complexity is exponential.

2 Question 3

See code. The code was tested against the default MPS example provided in lp_solve.

3 Question 4

The code is provided in python. Several datasets were computed and evaluated using lp_solve. The running times are presented in figure 1. We see that the complexity seems to be almost linear, which is way better than the expected exponential complexity of the simplex algorithm.

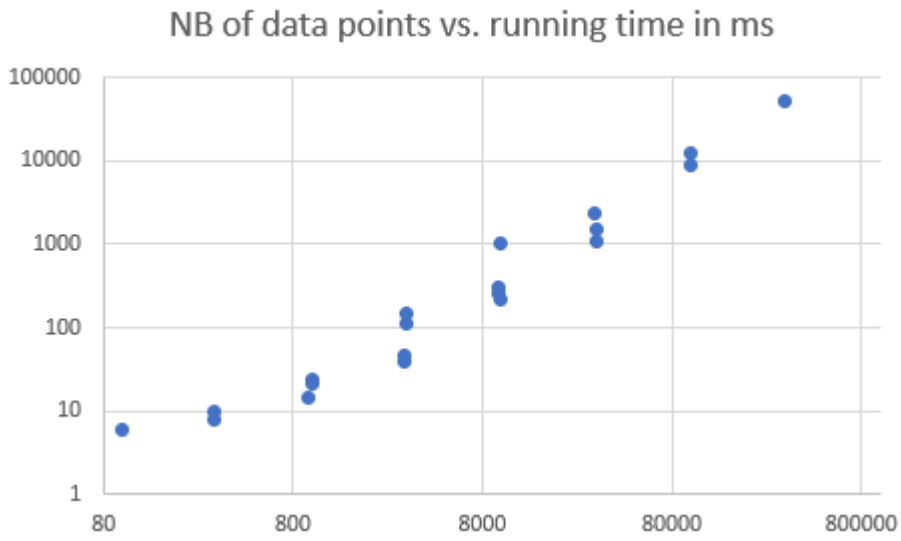

Figure 1: Time complexity of the Earth Mover Distance.

4 Question 5

To measure the distance between 2 clustering results we will use the Earth Mover Distance. "Position" will be the probability distribution that a given element belongs to a given cluster while weight is the share of data set to the most probable cluster. The distance between clusters is the sum of difference between all points in two clusters. We will then compute the distance between different clustering algorithms.

5 Question 6

The precedent algorithm failed. Indeed it ended up building impossible models according to $lp_s olve$. By relaxing the weight constraint the distance works but all final distances end up to 0. We tried also the cluster similarity sensitive distance with no success.

We tried releasing the equality constraints as they are not mentioned, but ended up again with a null distance.