README OF MULTIPLE LINEAR EQUATION SYSTEM DATABASES THAT WERE GENERATED

made by: César Miranda Meza

email: cmirandameza3@hotmail.com

The multiple linear classification equation that was used as a reference to create the databases labeled as “the multiple linear classification equation systems” is the following:

$$y \in \begin{cases} 1 & if \quad b_0 + b_1 x_1 + b_2 x_2 > 50 \\ 0 & if \quad b_0 + b_1 x_1 + b_2 x_2 \leq 50 \end{cases} \tag{1}$$

Where y is the dependent variable (output of the current sample); x_1, x_2 represents the independent variables (inputs of the current sample); and b_0, b_1, b_2 stand for the coefficient values of the equation. Furthermore, the values that were selected for b_0, b_1, b_2 are the following:

- $b_0 = 10$
- $b_1 = 0.4$
- $b_2 = 0.4$

such that the Eq. (1) will turn into the following:

$$y \in \begin{cases} 1 & if \quad 10 + (0.4)x_1 + (0.4)x_2 > 50 \\ 0 & if \quad 10 + (0.4)x_1 + (0.4)x_2 \leq 50 \end{cases} \tag{2}$$

However, the Eq. (2) was modified by adding to it a bias component r, that would represent a random value and should be generated each time a new sample is calculated:

$$y \in \begin{cases} 1 & if \quad 10 + (0.4)x_1 + (0.4)x_2 + r > 50 \\ 0 & if \quad 10 + (0.4)x_1 + (0.4)x_2 + r \leq 50 \end{cases} \quad | \; -10 \leq r \leq 10 \tag{3}$$

Where the independent variable was restricted to be sampled with values according to the following way $0 < x \leq 100$ and where if no random bias value is needed, then it should be negated by setting $r = 0$ or, Ec. (2) should be used instead.

Nevertheless, several regression databases governed by the term $10 + (0.4)x_1 + (0.4)x_2 + r$ from the Eq. (3) have already been created (see databases in the directory databases/regressionDBs/randMultipleLinearEquationSystem). Therefore, it was decided to recycle them and use a copy of those files in Excel to apply to them the threshold defined in the Eq. (3), which is 50. As a consequence, the following .csv (comma delimited) files were generated for the creation of the multiple linear classification equation systems:

- randLinearEquationSystem/1systems_10samplesPerSys.csv

- randLinearEquationSystem/10systems_10samplesPerSys.csv
- randLinearEquationSystem/10systems_100samplesPerSys.csv
- randLinearEquationSystem/100systems_100samplesPerSys.csv
- randLinearEquationSystem/100systems_1000samplesPerSys.csv
- randLinearEquationSystem/1000systems_1000samplesPerSys.csv

And for the ones made from the Eq. (2), which were created with the same strategy (see databases in the directory databases/regression/multipleLinearEquationSystem), the following .csv (comma delimited) files were generated:

- linearEquationSystem/1systems_10samplesPerSys.csv
- linearEquationSystem/10systems_10samplesPerSys.csv
- linearEquationSystem/10systems_100samplesPerSys.csv
- linearEquationSystem/100systems_100samplesPerSys.csv
- linearEquationSystem/100systems_1000samplesPerSys.csv
- linearEquationSystem/1000systems_1000samplesPerSys.csv

For all these files, note that they try to mimic how a real database would normally be organized by a professional and in which you will encounter four columns, whose headers and purpose are the following:

1. **id:** Represents the unique identifier for the current row of the database.
2. **system_id:** Represents the unique identifier for the current system sampled. This is because the databases will contemplate having several samples for several systems that manifest the same phenomenon.
3. **dependent_variable:** Represents the output value of the current sample.
4. **independent_variable_1:** Represents the input value 1 that generated the current sample.
5. **independent_variable_2:** Represents the input value 2 that generated the current sample.

Created in: September 21, 2021.

Last update in: November 26, 2021.