README OF MULTIPLE LINEAR EQUATION SYSTEM DATABASES THAT WERE GENERATED

made by: César Miranda Meza

email: cmirandameza3@hotmail.com

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

$$y = b_0 + b_1 x_1 + b_2 x_2 \tag{1}$$

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:

$$y = 10 + (0.4)x_1 + (0.4)x_2 \tag{2}$$

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:

$$y = 10 + (0.4)x_1 + (0.4)x_2 + r \quad | \; -10 \leq r \leq 10 \tag{3}$$

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.

With the help of the Excel “text to columns” function, for the creation of the multiple linear equation system databases, the Eq. (3) was employed to generate each of the samples contained in the following .csv (comma delimited) files:

- multipleLinearEquationSystem_1systems_10samplesPerSys.csv
- multipleLinearEquationSystem_10systems_10samplesPerSys.csv
- multipleLinearEquationSystem_10systems_100samplesPerSys.csv
- multipleLinearEquationSystem_100systems_100samplesPerSys.csv
- multipleLinearEquationSystem_100systems_1000samplesPerSys.csv
- multipleLinearEquationSystem_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.

Moreover, the samples generated aimed to attempt mimicking how several real life systems behave in real life due to the bias component r. On the other hand, each listed database was generated through a separated file which was developed in Python programming language v3.7.1 in order to display a friendly and simple code:

- multipleLinearEquationSystem_1systems_10samplesPerSys.py
- multipleLinearEquationSystem_10systems_10samplesPerSys.py
- multipleLinearEquationSystem_10systems_100samplesPerSys.py
- multipleLinearEquationSystem_100systems_100samplesPerSys.py
- multipleLinearEquationSystem_100systems_1000samplesPerSys.py
- multipleLinearEquationSystem_1000systems_1000samplesPerSys.py

Created in: September 21, 2021.

Last update in: November 02, 2021.