AI/ML assignment

# Description of data

Given are unknown objects (and we are not telling you what these are later either, try guessing from data, that’s a part of challenge) exhibiting 17 features called “CA” (2 letter abbreviation – one does not get more telecom & scientific than that). These objects were examined by a set of models producing assessment vectors – think about x-ray under different angles or CAT scan (definitely, data is not related to either). At the same time, it was known if given object exhibits given CA feature or not. Data related to these is present in dataset.zip. Rows in different CA “x” files present assessment vectors. Features are encoded in “y” files by supposedly binary labels denoting if an object exhibited the feature or not (1 means it does, captain obvious saves the day and flies away). Supposedly binary as there is some noise in these labels what’s interesting on its own and will require denoising. Rows in different x and y files correspond to same objects.

Challenge accepted!

# Serious description of data

For each CA 2 csv files are created: x\_train and y\_train. x\_train file has a vector of rational numbers per each row. These numbers depict assessment of an object via given assessment model. Values in single column (i.e. column 1) of one file (i.e. x\_train\_CA1) were produced with same model. Corresponding y\_train has an expected binary label (with a noise) related to a row in x\_train file. In that sense .

Having x\_train\_CA1 as defined in table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x\_train\_CA1 | | | | |
| 0.5 | 0.2 | 0.3 | 0.1 | 0.2 |
| 0.8 | 0.2 | 0.7 | 0.9 | 0.4 |

And y\_train\_CA1 as defined in a table below:

|  |
| --- |
| y\_train\_CA1 |
| 0 |
| 1 |

One can produce joint data array that contains both input and corresponding output by appending y as a column.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x\_train\_CA1 | | | | | y\_train\_CA1 |
| 0.5 | 0.2 | 0.3 | 0.1 | 0.2 | 0 |
| 0.8 | 0.2 | 0.7 | 0.9 | 0.4 | 1 |

Data from different CA files can be merged to produce metadata. It must be possible to map response produced back to specific CA. Consider example below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x\_train\_ | | | | | | | | y\_train\_ | | |
| CA1 | | | | | CA2 | | | CA1 | CA2 | |
| 0.5 | 0.2 | 0.3 | 0.1 | 0.2 | 0.9 | 0.7 | 1 | 0 | | 1 |
| 0.8 | 0.2 | 0.7 | 0.9 | 0.4 | 0.3 | 0.8 | 0.9 | 1 | | 1 |

# Goal

Using provided data and performing its analysis build prediction model that based on x vector will produce y value without knowing function (frankly, we do not know it either!).

# Expected outcome

A Jupyter (not to be mixed with Jupiter, we are not having any premises on that planet) notebook containing description of data, its analysis and information about reason and purpose of each step, its input, output and result. Notebook should produce a final model or models that based on x vector predicts y value for each CA with analysis of results.