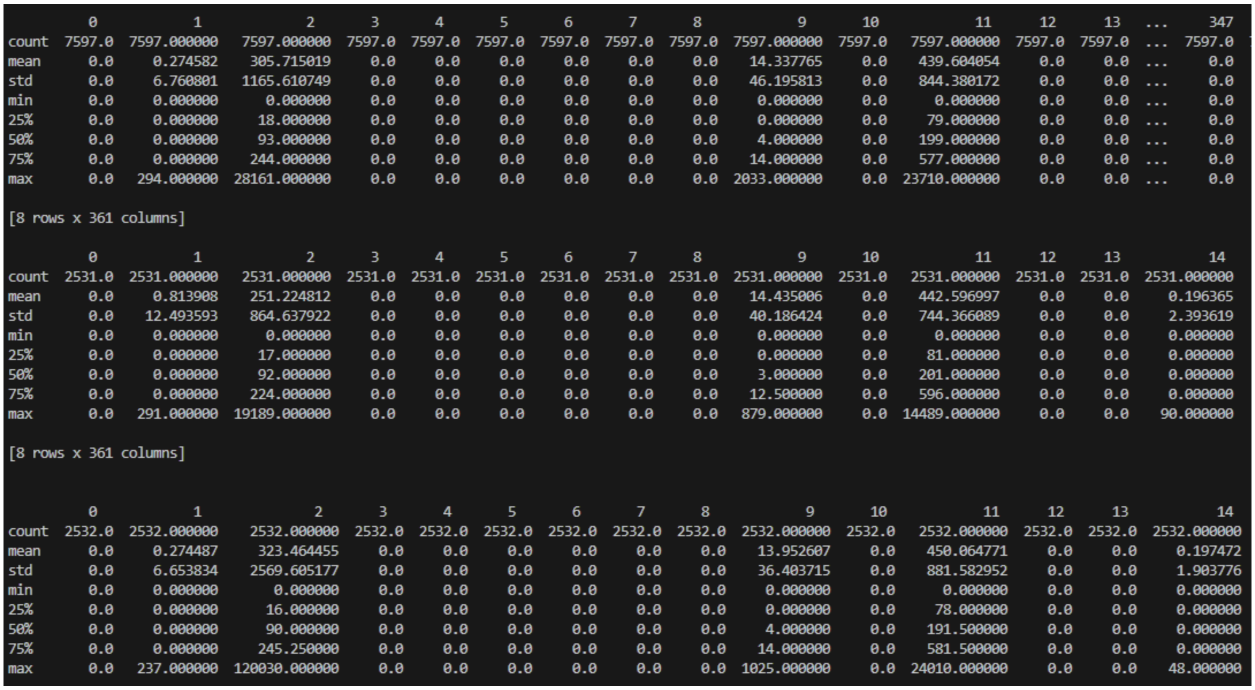
Project Milestone 1

By: Tyler Trotter

I decided to use a Perceptron to begin (since the labels were binary) and I assumed the data would be separable due to the large feature dimension. However, I found the prediction on the eval set to be perfect, but low (compared to what I am used to) for the test and train datasets. The only feature transformations that I ended up implementing was to change 0 => -1, but while that did benefit the test and train datasets, it was catastrophic for the eval dataset sending it to almost 0% accuracy. I did also play around with some scaling factors (Z-Score scaling and mapping to [0,1]). Both of these likewise ruined the accuracy on the eval set. I suspect this is because such a scaling like this wouldn't improve the linear separability of the data. Additionally, I played with different learning rate (lr) and mu values to see if I was able to improve on the test and train sets while maintaining the (at the time) 100% classification accuracy on the eval set. This didn't help much, as I found.

Having performed a .describe() on all of the datasets provided, I found that there are many features which are entirely populated with 0. Going forward, I intend to try dimension reduction by eliminating all of these features as they provided nothing for the perceptron. I wouldn't mind trying the different types of perceptron that we wrote for our second homework. Perhaps they will prove better overall at prediction. Further, I intend to implement ensembles for previously written classifiers and algorithms and look into their efficacy.

Below I've attached some of the .describe() descriptive statistics for the three data sets:



```
            0            1             2       3       4       5       6       7       8            9      10            11      12      13  ...     347
count  7597.0  7597.000000   7597.000000  7597.0  7597.0  7597.0  7597.0  7597.0  7597.0  7597.000000  7597.0   7597.000000  7597.0  7597.0  ...  7597.0
mean      0.0     0.274582    305.715019     0.0     0.0     0.0     0.0     0.0     0.0    14.337765     0.0    439.604054     0.0     0.0  ...     0.0
std       0.0     6.760801   1165.610749     0.0     0.0     0.0     0.0     0.0     0.0    46.195813     0.0    844.380172     0.0     0.0  ...     0.0
min       0.0     0.000000      0.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0      0.000000     0.0     0.0  ...     0.0
25%       0.0     0.000000     18.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0     79.000000     0.0     0.0  ...     0.0
50%       0.0     0.000000     93.000000     0.0     0.0     0.0     0.0     0.0     0.0     4.000000     0.0    199.000000     0.0     0.0  ...     0.0
75%       0.0     0.000000    244.000000     0.0     0.0     0.0     0.0     0.0     0.0    14.000000     0.0    577.000000     0.0     0.0  ...     0.0
max       0.0   294.000000  28161.000000     0.0     0.0     0.0     0.0     0.0     0.0  2033.000000     0.0  23710.000000     0.0     0.0  ...     0.0

[8 rows x 361 columns]

            0            1             2       3       4       5       6       7       8            9      10            11      12      13           14
count  2531.0  2531.000000   2531.000000  2531.0  2531.0  2531.0  2531.0  2531.0  2531.0  2531.000000  2531.0   2531.000000  2531.0  2531.0  2531.000000
mean      0.0     0.813908    251.224812     0.0     0.0     0.0     0.0     0.0     0.0    14.435006     0.0    442.596997     0.0     0.0     0.196365
std       0.0    12.493593    864.637922     0.0     0.0     0.0     0.0     0.0     0.0    40.186424     0.0    744.366089     0.0     0.0     2.393619
min       0.0     0.000000      0.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0      0.000000     0.0     0.0     0.000000
25%       0.0     0.000000     17.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0     81.000000     0.0     0.0     0.000000
50%       0.0     0.000000     92.000000     0.0     0.0     0.0     0.0     0.0     0.0     3.000000     0.0    201.000000     0.0     0.0     0.000000
75%       0.0     0.000000    224.000000     0.0     0.0     0.0     0.0     0.0     0.0    12.500000     0.0    596.000000     0.0     0.0     0.000000
max       0.0   291.000000  19189.000000     0.0     0.0     0.0     0.0     0.0     0.0   879.000000     0.0  14489.000000     0.0     0.0    90.000000

[8 rows x 361 columns]


            0            1              2       3       4       5       6       7       8            9      10            11      12      13           14
count  2532.0  2532.000000    2532.000000  2532.0  2532.0  2532.0  2532.0  2532.0  2532.0  2532.000000  2532.0   2532.000000  2532.0  2532.0  2532.000000
mean      0.0     0.274487     323.464455     0.0     0.0     0.0     0.0     0.0     0.0    13.952607     0.0    450.064771     0.0     0.0     0.197472
std       0.0     6.653834    2569.605177     0.0     0.0     0.0     0.0     0.0     0.0    36.403715     0.0    881.582952     0.0     0.0     1.903776
min       0.0     0.000000       0.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0      0.000000     0.0     0.0     0.000000
25%       0.0     0.000000      16.000000     0.0     0.0     0.0     0.0     0.0     0.0     0.000000     0.0     78.000000     0.0     0.0     0.000000
50%       0.0     0.000000      90.000000     0.0     0.0     0.0     0.0     0.0     0.0     4.000000     0.0    191.500000     0.0     0.0     0.000000
75%       0.0     0.000000     245.250000     0.0     0.0     0.0     0.0     0.0     0.0    14.000000     0.0    581.500000     0.0     0.0     0.000000
max       0.0   237.000000  120030.000000     0.0     0.0     0.0     0.0     0.0     0.0  1025.000000     0.0  24010.000000     0.0     0.0    48.000000
```