Class Project 1

Saskia Reiss, Alvaro Pinedo and David Resin
CS-433 – Machine Learning
EPFL, Switzerland

***Abstract*—TODO SUMMARY**

I. Introduction

II. Methodology

First, we extract the data from the `.csv` files using the provided function `load_csv_data`. For each file it returns us the predictions `y` (which is of course empty in the case of the test data), the data `x`, and the sample identifiers `ids`. We defined the function `triage` which replaces all -999 values in the document, which are placeholders for unavailable data, with the mean of the column they are in. We consider this to be the best way to treat these values without splitting up the data into different categories. We decided not to normalize the data as it resulted in overfitting. We then use the `build_poly` method with degree 9 on both datasets, before applying `least_squares` on the training data set and feed the resulting weigths to `predict_labels` along with the testing dataset.

III. Results

- Our first attempt simply used `least_squares` on the unaltered dataset, which gave us a score of **0.74463**.
- We then improved our results by applying `build_poly` with degree 7 and feeding the result to `least_squares`, which gave us a new best score of **0.80061**.
- Our next improvement consisted of replacing all -999 values in the dataset with 0's instead for a score of **0.80596**.
- Replacing -999 values with the column average instead along with building polys of degree 9 raised our best score to **0.81553**.

IV. Discussion

V. Conclusion