# Machine Learning Project 1 - Fall 2020

Julien Hu, Matthieu Masouye, Sebastien Ollquist
*Department of Computer Science, EPFL, Switzerland*

***Abstract*—Implement basic Machine Learning methods on a given data set and analyze the predictions from it.**

## I. Introduction

The goal of this first mini project is to implement all basic Machine Learning methods on a given data set and analyze the results we obtained from running these algorithms. Essentially, the demanded algorithms were:

1) Linear regression using Gradient Descent and Stochastic Gradient descent
2) Least squares regression and ridge regression using normal equations
3) Logistic regression using Gradient Descent
4) Regularized logistic regression using Gradient Descent

## II. Data normalization

Before performing any algorithm, we essentially have to do two things in order to clean the data:

1) We first have to normalize the training data. For each data sample $X$, we want to compute the value $Y = (X - \mu)/\sigma$ where $\mu$ is the computed mean and $\sigma$ the standard deviation.
2) Then, we want to add a one in front of the $X^T$ matrix. This represents the bias which allows the linear function not to pass from the origin. In a function $y = ax + b$, $b$ is the bias.

## III. Algorithms implementation details

### *A. Linear regression*

This first algorithm is essential to Machine Learning. It consists of taking a data set that often contains two different data point types and split them using a line described by a linear function in order to divide the points the best way possible. We have performed two different implementations of it: one using Gradient Descent (GD) and the other one using Stochastic Gradient Descent (SGD).

Note that the GD implementation does not work due to the fact that we are treating a big amount of data, so the SGD will help us resolve that problem by only taking a batch of for example 50 randomly selected data samples.

After data standardization, we can run the SGD algorithm on our data set. We get an F1-score of 0.732 on the first run which is pretty good given that there are much better algorithms than SGD for predictions.

### *B. Least squares regression*

### *C. Ridge regression*

### *D. Logistic regression*

## IV. Results obtained

## V. Conclusion