EE 565 - Machine Learning I Project 0 Report

Jorge A. Garcia

***Abstract*—An introductory project that covers basic computational routines and generate data to be used in future projects for the class.**

***Index Terms*—Machine Learning, Data Science**

I. Introduction

In this report, basic routines such as generated random data from various distributions and loading data in a file are shown. The purpose is to show efficiency in routine tasks that will have to be realized throughout the semester and show the necessary programming competence.

II. Problem 1: Circular Symmetric Gaussians Data Set

A function ”circGauss” was written, which draws N data points of k dimensions, depending on the number of means μ_k provided, and with an equal variance σ^2 in all dimensions. Two different 2-dimensional distributions were generated with said function, with a combined number of samples $N = 500$. The first is centered at the origin ($\mu_1 = \mu_2 = 0$), and the second is centered at $\mu_1 = \mu_2 = 5$. Both with a variance of $\sigma^2 = 3$. Figure 1 plots the resulting distributions.

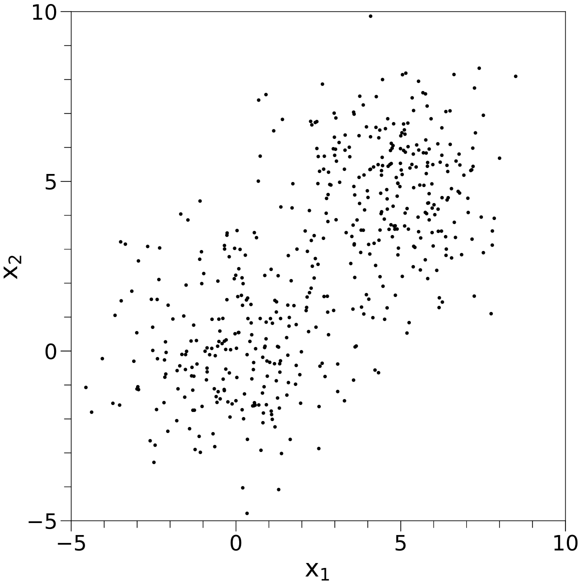

Fig. 1. Scatter plot of $N = 500$ data points drawn from the function ”circGauss”.

III. Problem 2: Double Moon Data Set

A function ”doublemoon” was written, which draws N data points from a scaled uniform distribution evenly between the two classes, that being membership of the upper (c_1) or lower (c_2) moon, of a double moon with a moon width w, moon radius r and a distance d between the lower moon and the x-axis. A total of $N = 500$ samples were drawn, with parameters $d = 0$, $r = 1$ and $w = 0.6$ used to generate the data. A plot of the resulting data distribution can be seen in Figure 2, with members of class $c_1 = +1$ plotted as blue ”+” and members of $c_2 = -1$ plotted as green ”x”.

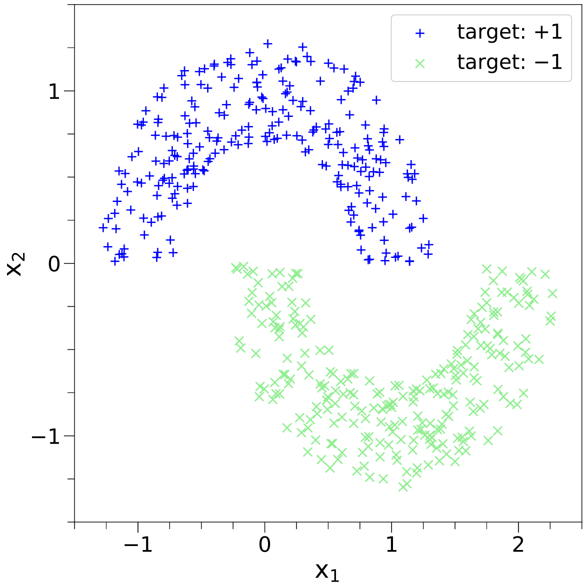

Fig. 2. Scatter plot of the two classes from the double moon data set.

IV. Problem 3: Concentric Gaussian Data Set

A function ”concentGauss” was written, which draws N from two different Gaussian distributions evenly. The inner, circular Gaussian is centered at the origin and has a variance σ^2_{in}, with members of classification c_1. The outer Gaussian annulus has a radius r and a variance σ^2_{out}, and its members having a classification c_2. It is formed by drawing normally distributed data points around the radius r and paired with angles θ drawn from a uniform distribution $[0, 2\pi]$. These pairs then are taken to be the polar coordinates of the data points, and are then projected into Cartesian coordinates. A total of $N = 500$ samples were drawn, with parameters $\sigma^2_{in} = 1$,

$r = 5$ and $\sigma^2_{out} = 1$. A plot of the resulting data distribution can be seen in Figure 3, with members of class $c_1 = +1$ plotted as blue "+" and members of $c_2 = -1$ plotted as green "x".

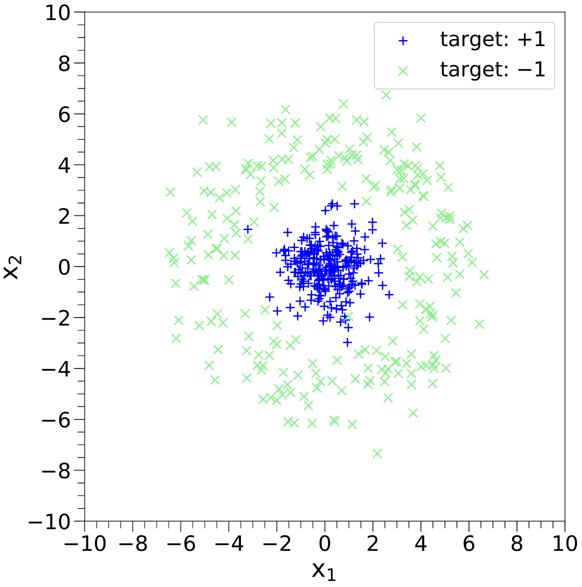

Fig. 3. Scatter plot of the two classes from the concentric Gaussian data set.

V. Problem 4: Gaussian XOR Data Set

A function "gaussX" was written, which draws N samples from a circular Gaussian distribution with means $\mu_x = \mu_y$ and a variance σ^2 and are distribtued evenly between two classes. Class c_1 corresponds to data points whose coordinate pairs (x, y) result in a product $xy = -1$, whereas members of class c_2 have a product of their coordinate pairs $xy = 1$. A total of $N = 500$ samples were drawn, with parameters $\mu_i = 0$ and $\sigma^2 = 1$. A plot of the resulting data distribution can be seen in Figure 4, with members of class $c_1 = +1$ plotted as blue "+" and members of $c_2 = -1$ plotted as green "x".

VI. Problem 5: Noisy Sinusoidal Data Set

A. Generated Data

A function "noisySin" was written, which draws N samples from the function $f(x) = sin(2\pi x) + N(0, \sigma^2)$, with N being a normal distribution with mean $\mu = 0$ and and a variance σ^2 that is added as noise. A total of $N = 50$ samples were drawn, with a variance $\sigma^2 = 0.05$ chosen for the noise. A plot of the resulting data distribution can be seen in Figure 5, with the data being plotted as blue circles. Over the noisy data, the function $g(x) = sin(2\pi x)$ is plotted to represent the ideal sinusoidal curve the data should follow.

B. Loaded Data

The data contained in the file "curvefitting.txt" is loaded and plotted in Figure 6. Alongside is plotted the ideal sinusoidal curve $g(x) = sin(2\pi x)$.

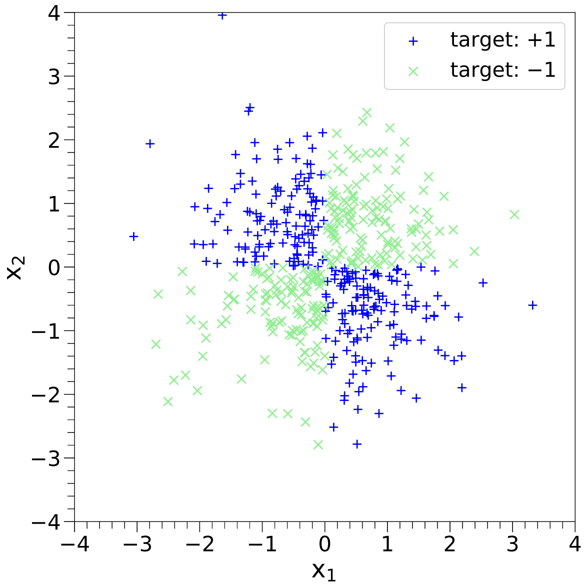

Fig. 4. Scatter plot of the two classes from the Gaussian XOR data set.

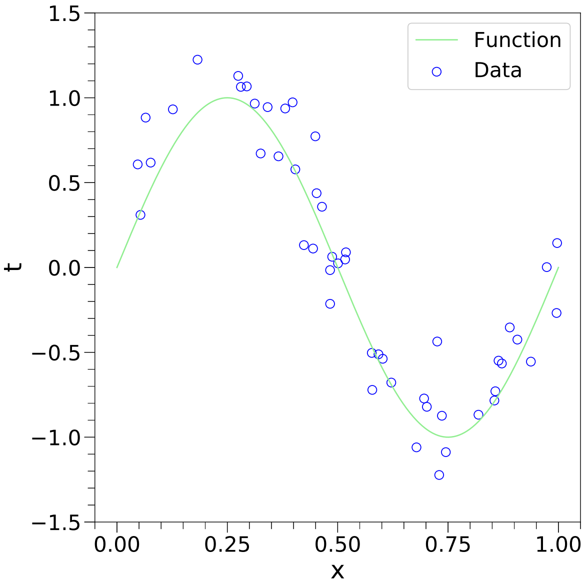

Fig. 5. Scatter plot of the generated noisy sinusoidal data, alongside the plot of the perfect sine curve.

VII. Problem 6: Old Faithful Data Set

The data contained in the file "faithful.txt" is loaded and plotted in Figure 7.

VIII. Problem 7: Neural Spike Data Set

The data contained in the file "spikes.csv" is loaded and plotted in Figure 8.

IX. Conclusion

In this project, various data sets were generated, loaded and visualized to demonstrate the necessary computational skills.

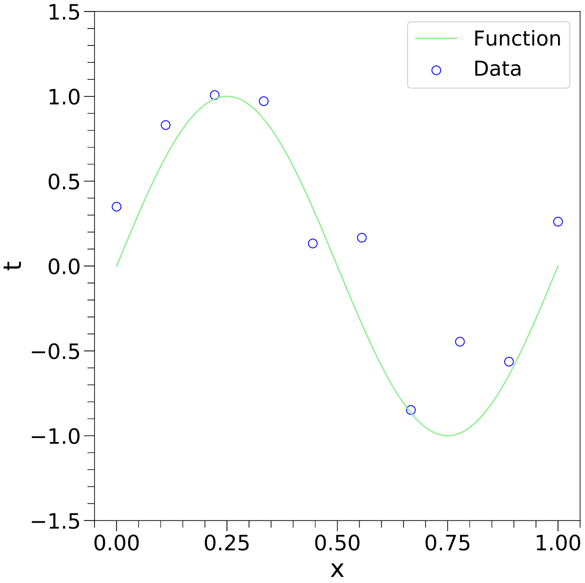

Fig. 6. Scatter plot of the loaded noisy sinusoidal data ”curvefitting.txt”, alongside the plot of the perfect sine curve.

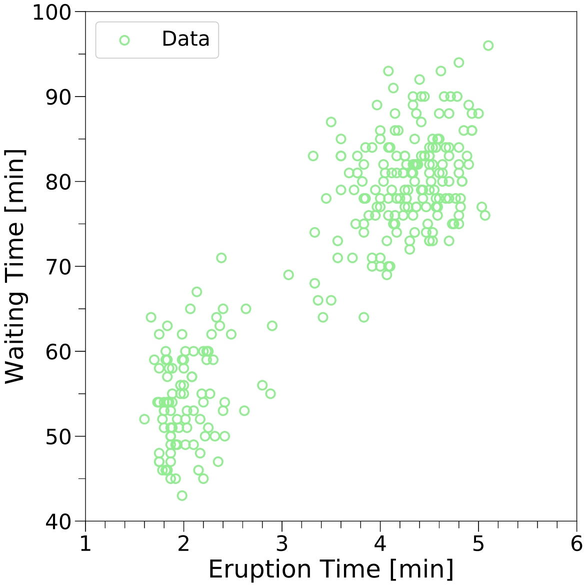

Fig. 7. Scatter plot of the data loaded from ”faithful.txt”.

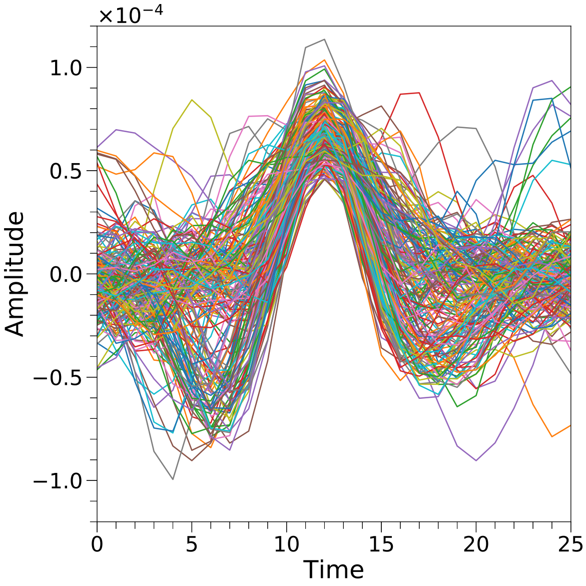

Fig. 8. Plot of the time series data loaded from ”spikes.csv”.

These data sets will be used in future projects throughout the semester.