EE 565 Machine Learning
Project 0 Report

Anthony M. DeRocchis

***Abstract*—In this report I explore the creation and visualization of randomly generated synthetic data sets. These data sets are to be used in subsequent projects for test and evaluation of algorithm performance. In addition to the randomly generated data sets, I also visualize some predefined data sets by loading the relevant data from files and plotting.**

***Index Terms*—Data Sets**

I. Introduction

This document is an example of a project report. The document shall be formatted in accordance with the IEEE Transactions journals templates. Each report will consist of an introduction, a section (or subsection) for each problem part, a conclusion, and references. Each figure will be numbered with a caption, well-labeled (axes, color, symbols, etc.), and shall be called out at least once in the report text. I will attempt to be as concise as possible, whilst still fully addressing all problem parts.

II. Problem 1: Circular Symmetric Gaussian

A function named "circGauss" was written which returns N samples drawn from a circular symmetric multi-variate Gaussian distribution of dimension D with variance σ^2

$$X \sim \begin{bmatrix} \mathcal{N}(\mu_1,\, \sigma^2) \\ \mathcal{N}(\mu_2,\, \sigma^2) \\ \vdots \\ \mathcal{N}(\mu_D,\, \sigma^2) \end{bmatrix} \tag{1}$$

where μ_k is the mean for the kth dimension [1]. The function was used to generate a data set (shown in figure 1) consisting of samples drawn from two circular symmetric multi-variate Gaussian distributions in two dimensions with one distribution at the origin and one at the point $(5, 5)$, where both distributions have $\sigma^2 = 3$.

III. Problem 2: Double Moon Data Set

A function named "doublemoon" was written which returns N samples drawn from a double moon distribution [2]. Figure 2 shows an example of data drawn from the double moon distribution with parameters $N = 500$ samples, $d = 0$, $r = 1$, and $w = 0.6$. In figure 2 the members of class $\mathscr{C}_1$ are plotted as "blue +" and $\mathscr{C}_2$ are plotted as "green x".

The author is with the Klipsch School of Electrical and Computer Engineering, New Mexico State University (NMSU), Las Cruces NM 88003 USA. e-mail: tonydero@nmsu.edu

IV. Problem 3: Concentric Gaussian Data Set

A function named "concentGauss" was written which returns N samples drawn from a data set consisting of a circular symmetric Gaussian centered at the origin with variance σ^2_{center} and a Gaussian annulus centered at the origin with radius mean r and variance σ^2_{outer}. Figure 3 shows an example of data drawn from the concentric Gaussian data set with $N = 500$ samples, $\sigma^2_{center} = 1$, $r = 5$, and $\sigma^2_{outer} = 1$ with members of class $\mathscr{C}_1$ plotted as "blue +" and class $\mathscr{C}_2$ plotted as "green x".

V. Problem 4: Gaussian XOR Data Set

A function named "gaussX" was written which returns N samples drawn from a Gaussian XOR distribution. Figure 4 shows an example of data drawn from a Gaussian XOR data set with $N = 500$ samples and $\sigma^2 = 1$, where members of class $\mathscr{C}_1$ are plotted as "blue +" and members of $\mathscr{C}_2$ as "green x".

VI. Problem 5: Noisy Sinusoidal Data Set

A. Randomly Generated

A function named "noisySin" was written which returns N samples drawn from a noisy sine distribution with a sinusoidal amplitude of 1 and period 2π, where the added noise has a variance σ^2. Figure 5 shows an example of data drawn from a noisy sine data set with $N = 50$ samples and $\sigma^2 = 0.05$, where the noise-added sinusoidal data is plotted as "blue o" and the clean sinusoidal waveform is plotted as a green curve.

B. Loading from File

To demonstrate the instance of needing to use a pre-existing, given, data set, the file "curvefitting.txt" was loaded. The data set consists of input x and target t values. These values were plotted, as shown in figure 6, using the same method as section VI-A, with the noise-added sinusoidal data plotted as "blue o", and the clean sinusoid plotted as a green curve.

VII. Problem 6: Old Faithful Data Set

The Old Faithful data set was loaded from the file "faithful.txt" and plotted. The results are shown in figure 7.

VIII. Problem 7: Neural Spike Data Set

The Neural Spike data set was loaded from the file "spikes.csv" and plotted. The results are shown in figure 8.

IX. Conclusion

In this project I loaded, generated, and visualized several data sets which will be useful for test and evaluation of machine learning algorithms. In addition, this document serves as a simple example of what is expected in a project write-up.

References

[1] C. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.

[2] S. Haykin, Neural Networks and Learning Machines. Prentice Hall, 2009.

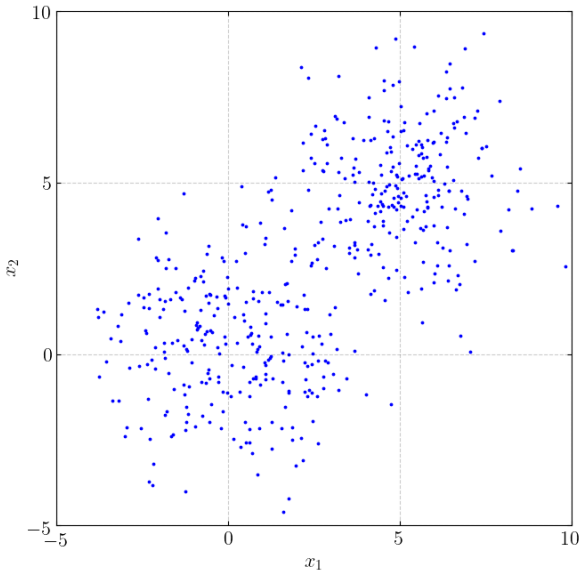

Fig. 1. A data set consisting of samples drawn from two circular symmetric multi-variate Gaussian distributions with paramters $\sigma^2 = 3$ and $\mu = (0, 0)$ for one and $\mu = (5, 5)$ for the other.

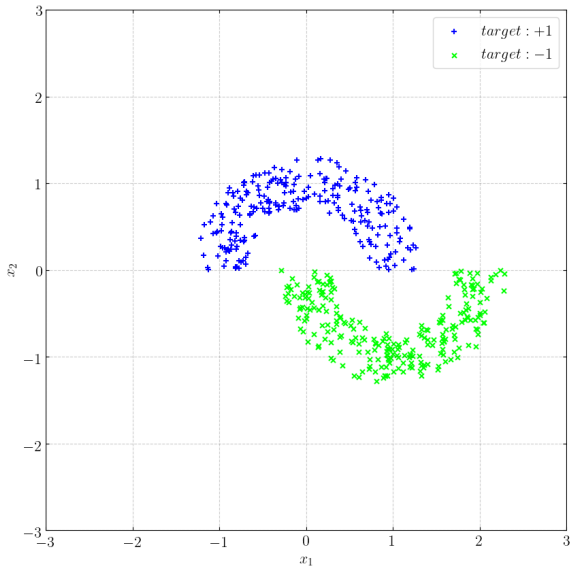

Fig. 2. A data set consisting of samples drawn from a double moon distribution with parameters $N = 500$ samples, $d = 0$, $r = 1$, and $w = 0.6$. The members of class $\mathscr{C}_1$ are plotted as "blue +" and $\mathscr{C}_2$ as "green x".

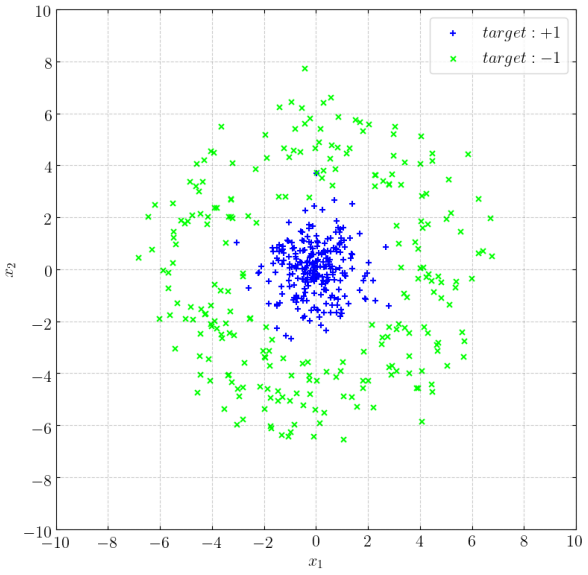

Fig. 3. A data set consisting of samples drawn from a circular symmetric Gaussian centered at the origin with variance σ^2_{center} and a Gaussian annulus centered at the origin with radius mean r and variance σ^2_{outer} with parameters $N = 500$ samples, $\sigma^2_{center} = 1$, $r = 5$, and $\sigma^2_{outer} = 1$. Members of class $\mathscr{C}_1$ are plotted as "blue +" and class $\mathscr{C}_2$ plotted as "green x".

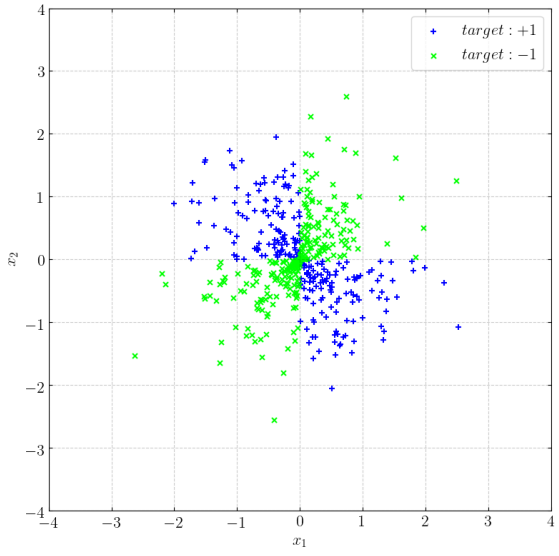

Fig. 4. A data set consisting of samples drawn from a Gaussian XOR distribution with parameters $N = 500$ samples and $\sigma^2 = 1$. Members of class $\mathscr{C}_1$ are plotted as "blue +" and members of $\mathscr{C}_2$ as "green x".

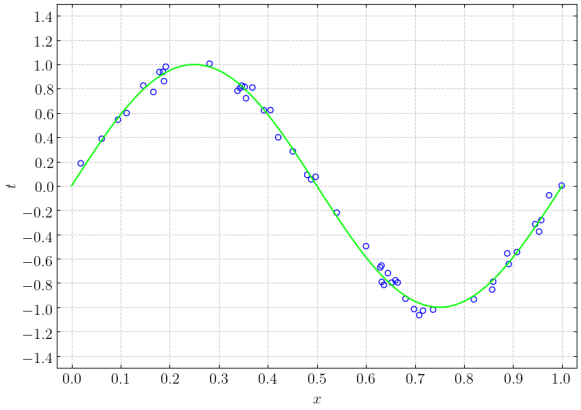

Fig. 5. A data set consisting of samples drawn from a noisy sine distribution with a sinusoidal amplitude of 1 and period 2π, where the added noise has a variance σ^2 with parameters $N = 50$ samples and $\sigma^2 = 0.05$. The noise-added sinusoidal data is plotted as "blue o" and the clean sinusoidal waveform is plotted as a green curve.

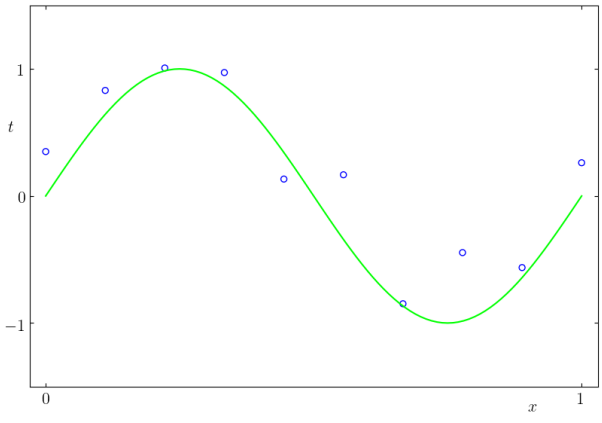

Fig. 6. The data set read from the file "curvefitting.txt" plotted as "blue o" on top of a clean sinusoidal waveform plotted as a green curve.

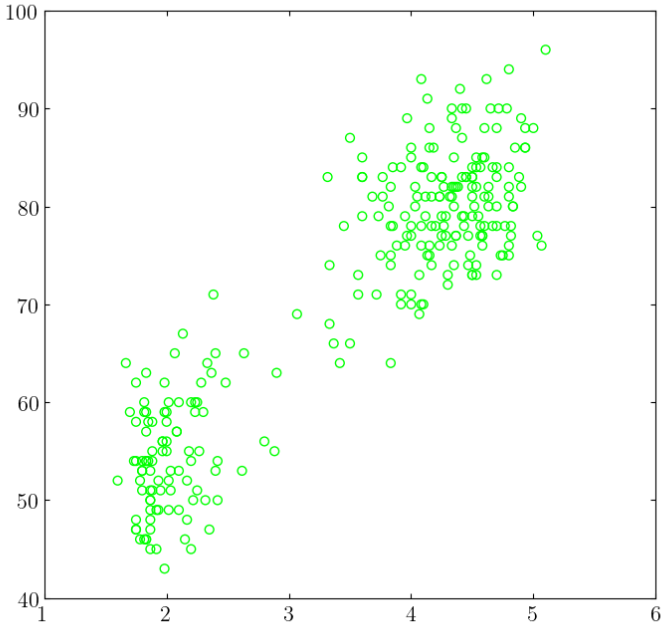

Fig. 7. The data set read from the file ”faithful.txt” and plotted.

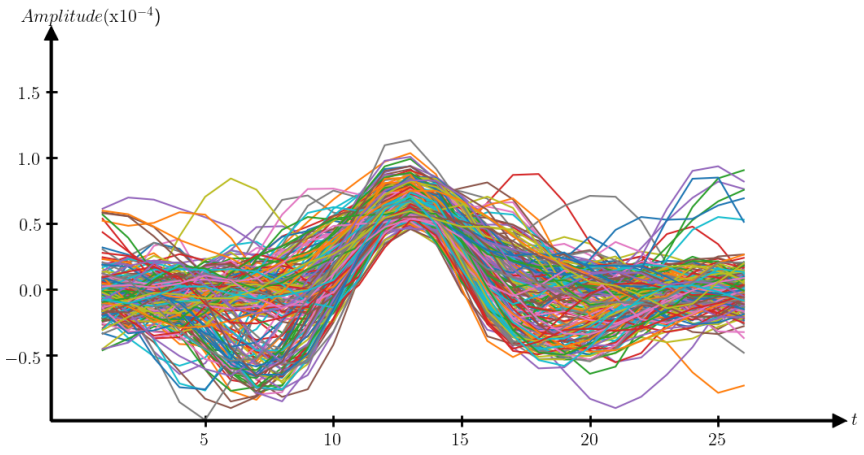

Fig. 8. The data set read from the file ”spikes.csv”, with the amplitudes scaled up by a factor of 10^4 and plotted.