INTRODUCTION

1 Machine Learning

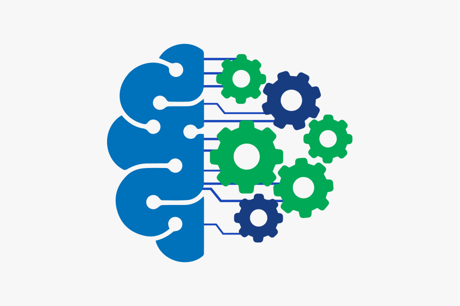

Figure 1: Machine Learning Concept, source: recorded-future.com

Machine Learning is the use of statistical techniques to make computers 'learn' with data, without being explicitly programmed. It is a subset of Artificial Intelligence in computer science. Arthur Samuel coined the term machine learning[1], and the field evolved from studies of pattern recognition in the late 1950s.

It involves the construction of algorithms that can learn from and make predictions on data – instead of following static programmed instructions. This is done by building a model from sample inputs through various learning algorithms, such as logistic regression or neural networks. It can solve problems whose solutions are difficult, if not infeasible, to explicitly program algorithmically – for example: Optical Character Recognition, Email spam filtering and computer vision.

Machine Learning is mainly concerned with 4 learning tasks –

Supervised Learning The computer is given example inputs (called data) and their known outputs (also called labels) and the objective is to learn a rule that maps each input to the given output. For example: predicting the amount of rainfall given atmospheric data as inputs, such as temperature, wind speed and humidity.

Unsupervised Learning No target labels are given to the program, instead only the input data is presented. The goal of the program is to find some structure in the given data. It is like saying to the computer 'Here, take some data and make sense out of it'. An example is the organizing of news articles on Google, where articles are automatically sorted by topic, which is inferred.

Semi-supervised Learning Incompletely labelled data is given to the program. Often, most target labels are missing and less than 10% of the data are labelled.

Reinforcement Learning The program is asked to make decisions in a dynamic environment and the training data consists of rewards or punishments given as a consequence of the action the program took. For example, teaching a computer to play a video game.

2 Classification

Classification is an application of machine learning where all the labels belong to a finite set of values that is known to the program. As such, classification falls squarely in the category of supervised learning. For eg: classifying emails as spam or not spam.

In classifications, the labels can be represented by integers starting from zero, by binary digits or by a vector of one-hot coded values.

Often in classification tasks, the purpose is to classify objects (data) into either of two categories – this is called Binary Classification.

Besides classification, two other common applications of machine learning are:

Regression where the output variable (ie the label) belongs to a range of continuous values instead of a discrete set.

Clustering where a set of inputs should be divided into groups and the set of groups is not known beforehand.

3 Data Sets Used

Two different datasets were used in this project. They are outlined below:

3.1 Iris Flower Data Set

The Iris flower data set[2] is a multivariate data set introduced by statistician and biologist Ronald Fisher in his 1936 paper *The use of multiple measurements in taxonomic problems.*[3] It was collected by Edgar Anderson in order to quantify the morphologic variation of Iris flowers of three closely related species, viz:

- *Iris setosa*
- *Iris virginica*
- *Iris versicolor*

The data set consists of 50 samples from each of the species. Each sample/record contains measurements of 4 attributes of the flower-

- Petal length
- Petal Width

- Sepal length
- Sepal Width

The data is stored and read from a CSV file where each row contains 1 record followed by the name of that species, ie, data followed by label.

It has been shown by previous analysis that out of the 3 species, *Iris setosa* is fully linearly separable from the other 2 species whereas *Iris virginica* and *Iris versicolor* are not completely linearly separable, although it is only a few examples that make the separation non-linear. Hence, with this dataset, we can investigate both the performance on linearly separable data and non-separable data for any given model.

3.2 MNIST Data Set

The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used in machine learning [4].

The MNIST database contains 60,000 training images and 10,000 testing images.

The dataset was created by Yann LeCun by combining a similar dataset written by high school students with another dataset of digits written by staff in the United States Postal Service [5].

The MNIST dataset is especially used in Computer Vision tasks involving classification and object detection.

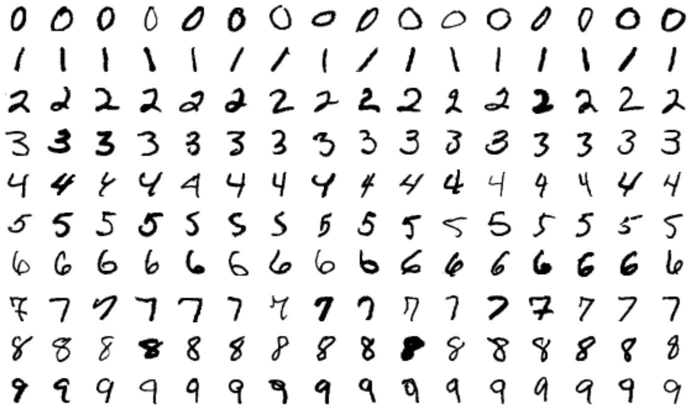

Figure 2: Some samples from the MNIST dataset, by Josef Steppan, on Wikipedia, Public Domain