|  |  |
| --- | --- |
| Name | Reg No |
| Atif Bashir |  |



***Blend of Traditional Learning***

Date: 10th May, 2019

Project Report

***Abstract*-** The Paper focuses on comparing different algorithms of machine learning. In this paper two types of machine learning algorithms are implemented. i.e Supervised Learning and Unsupervised Learning. In Supervised Learning we implemented two algorithms Artificial Neural Networks and Decision Tree. In Unsupervised Learning one algorithm was implemented that was K-Means clustering. We calculated their accuracy, precision, f1\_score, recall. Then a comparison is made between different algorithms and which algorithm is better. The dataset we used was Iris Data set.

1. Introduction

In this project we are going to compare two traditional machine learning concepts i.e supervised learning and unsupervised learning. The difference between them is that in supervised learning we know the output in advance, while in unsupervised learning we don’t know the output in advance. So we are going to compare these two concepts.

We will implement different algorithms on it and experiment them and do a brief comparative analysis on them. We shall analyses their techniques and find the pros and cons for each type. The comparison will show how each type implements on the same dataset.

1. Data set

The data set we have selected for our project is Iris data set. This data set is best known data set for pattern recognition. It has a total of 150 instances.

|  |  |  |  |
| --- | --- | --- | --- |
| Data Set Characteristics | Multivariate | Number of Instances | 150 |
| Attribute Characteristics | Real | Number of Attribute | 4 |
| Associated Tasks | Classification | Missing Value | No |

Table -Iris Data Set Description

Iris data set has 4 attributes i.e Sepal Length(cm), SepalWidth(cm), PetalLength(cm), PetalWidth(cm).

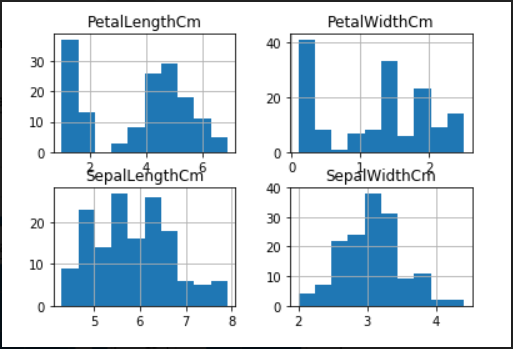


Figure 1-Data Set Visulaization

1. Supervised learning

Supervised Learning is one of the major concept of machine learning. In this method we map our inputs to the outputs where outputs are known to us. We map the input x to the output y as the function.

## Using this method we than predicate our results for next input values. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher.

Supervised learning problems can be further grouped into regression and classification problems.

**Classification:** A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

**Regression:** A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Some popular examples of supervised machine learning algorithms are:

* Linear regression for regression problems.
* Random forest for classification and regression problems.
* Support vector machines for classification problems.

## A. Artificial Neural Network

Artificial neural networks are one of the main tools used in machine learning. As the “neural” part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

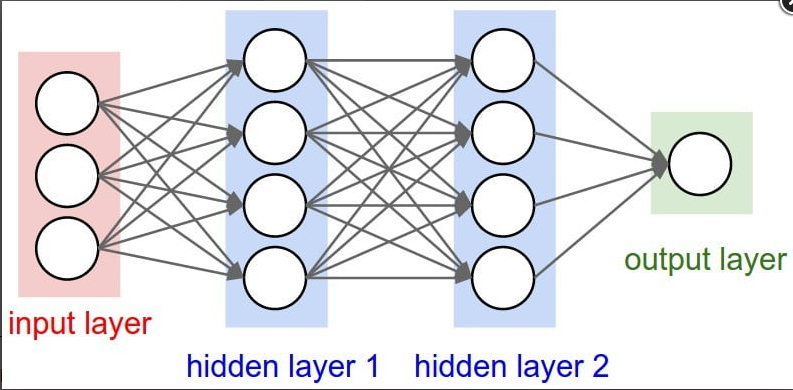


Figure 2-2 Layer ANN

## B. Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).

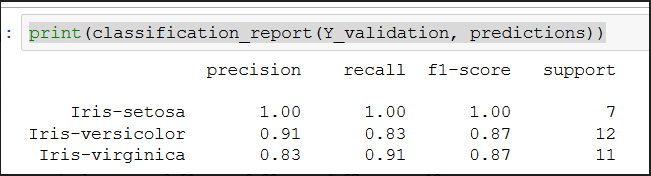


Figure 3-Decision Tree

1. Unsupervised learning

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

**Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

**Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are:

* K-means for clustering problems.
* Apriori algorithm for association rule learning problems.

*K-Means Clustering:*

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

A cluster refers to a collection of data points aggregated together because of certain similarities.

You’ll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.

Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.

In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

The ‘means’ in the K-means refers to averaging of the data; that is, finding the centroid.

1. Precision, recall, accuracy, f1\_score

We find accuracy, precision, accuracy, f1\_score to the algorithms so that they can be compared. All these terms are explained below:

**Precision:**

The ratio of correctly predicted positive observations to the total predicted positive observations is known as precision.

**Recall:**

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations is known as **recall**.

**Accuracy:**

 It is the ratio of correctly predicted observation to the total observations. It is to predict negative as negative and positive as positive in our model.

**F-measure:**

The F-measure or F-score is the harmonic mean of precision and recall.

comparison

After implementing all these algorithms, we compared these algorithms on the basis of their accuracy.

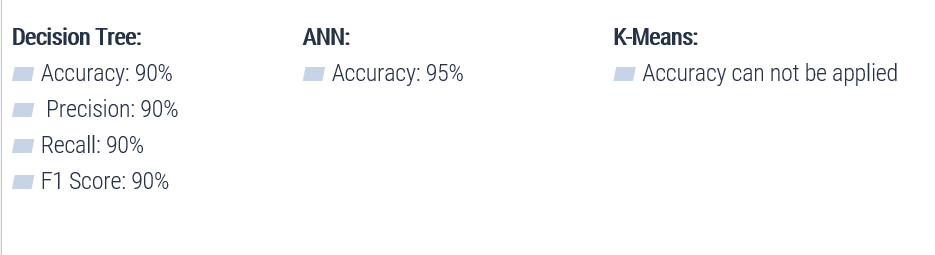


Figure 4-An Overview of Comparison

So, from Figure 4 it is clear that ANN has the highest accuracy for the iris data set.

Conclusion

All these algorithms that we have implemented deduce results and from our observations and the accuracy results we come to know that Supervised Learning method gives us better results on the selected data set. ANN algorithm was best in terms of accuracy i.e 95%, Decision Tree gave 90% accuracy on the data set.