Final Project

Logistic Regression, Suport Vector Machine & Neural Network

Charles Acevedo Díaz
System Engineering
UTB
Cartagena Bolívar
chad9591@gmail.com

***Abstract*—The purpose of this project lies in the application of the machine learning methods to a real dataset and using modules given in the programming language that has been chosen (python), we are going to design an agent that is capable of predicting results of new data based in the known information we have.**

Keywords—Logistic regression; Supervised Learning; Machine Learning; Logit; python; SVM: Support Vector Machine; Neural Network; Cross Validation; script; code; pseudo-code.

I. Introduction

Throughout the whole semester we have been learning some algorithms and methods that are used in the Machine Learning field, algorithms that help experts to classify data. This document will cover in a general form the way a learning method can be applied to solve a problem where a certain amount of data is already known, but there are some results that we would like to predict, therefore we proceed to write a python script pursuing that it would learn from the known results by itself and then design model that can be used to make those prediction, this python program is known as *intelligent agent*, and this agent will also be tested with a specific percentage of the data, in order to verify the error of the outputs that are predicted and the already known.

For this project, we have chosen a dataset that has 689 instance, and 15 features; this known instance shows whether if a person has been approved for a credit or not, but for security issues, the provider has changed the names of the feature for general ones.

The problem that is going to be solved is about classification; a group of persons have applied to a credit, and 15 features have been taken into account to make the decision; and they have been classified in two subgroups: (0) Unapproved and (1) Approved. The agent will be trained with the 70% of the record, and will be tested with the 30% left.

In this particular case, we are going to design three different agents, one will be for the implementation of Logistic Regression and the others will be for Support Vector Machine and Neural Network. Then, we will compare the metric of each, in order to analyze which one has a better performance for the dataset

II. Theory

Logistic Regression:

This is a method that helps finding discrete results, or binary outputs, which means that is intended to be used in circumstances where the results can be classified into a certain amount of groups like “Yes/No”, “To buy, To sell”, etc. The output here is a probability that takes a particular value based on combination of values taken by the predictor.

Its name is based on the function used at the core of the method, *the logistic function,* also called *sigmoid function* was developed by statisticians in order to describe properties of population growth in ecology. This is an S-shaped curve that take any real value number and gives as result a value between 0 and 1, but never exactly at those limits.

Support Vector Machine:

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Neural Network:

A neural network is a system of hardware and/or software patterned after the operation of neurons in the human brain. Neural networks -- also called artificial neural networks -- are a variety of deep learning technology, which also falls under the umbrella of artificial intelligence, or AI.

Typically, a neural network is initially trained or fed large amounts of data. Training consists of providing input and telling the network what the output should be.

III. Results

The three agents used the same dataset, and the metrics we used are: Precision, Recall, Accuracy and F1-score, each one has its own score, and the outputs we got are the following:

Logistic Regression:

Confusion Matrix

	N	**P**
N	88	19
P	8	81

F1-Score: **0.857**
Recall**: 0.912**
Precision**: 0.81**
Accuracy**: 0.862**

SVM:

Confusion Matrix

	N	**P**
N	104	3
P	29	84

F1-Score: **0.954545**
Recall**: 0.9438**
Precision**: 0.9655**
Accuracy**: 0.959**

Perceptron:

Confusion Matrix

	N	**P**
N	95	12
P	13	76

F1-Score: **0.859**
Recall**: 0.85**
Precision**: 0.86**
Accuracy**: 0.87**

IV. Conclusions

The agents has been well designed, although they have given good scores, there is no doubt that the one that gave the best overall performance is the SVM method, with 0.95 in F1-score, but more important, has fewer False positives, which means that the agent does not approve the credit to persons that are not suited for it as they may not be able to afford it.

There is a shocking info, and that's because the worst performance is given by Perceptron (Neural) agent, and it may be expected that this agent has the best overall performance; perhaps it is due to a bad implementation or, it may not be designed to sort out this kind of classification.

V. References

[1] Jason Brownlee, "Logistic Regression for Machine Learning" Phil. machinelearningmastery.com/logistic-regression-for-machine-learning/ April 2016.

[2] Medium.com/@aprendizaje.maq/regresion-lineal-con-gradiente-decendiente-c3b5ca97e27c

[3] https://searchenterpriseai.techtarget.com/definition/neural-network

[4] https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72

[5] https://archive.ics.uci.edu/ml/datasets/Credit+Approval