BigData Final Project

K. Morales Galindo , J. Escobedo Diaz

Departamento de sistemas y computación ,

Instituto Tecnológico de Tijuana

karen.morales16@tectijuana.edu.mx

jesus.escobedo16@tectijuana.edu.mx

Abstract— This investigation aims to compare the performance of the following SVM machine learning algorithms, Decision Three, Logistic Regression, Multilayer Perceptron, using a csv called bank of approximately 45,000 records

I. INTRODUCTION

Machine Learning (ML) automates the construction of data analysis models, bases its development on systems that learn from data, identify patterns and make decisions.

II. Algorithms

A. Multilayer Perceptron Classifier

Consists of multiple layers of nodes including the input layer, hidden layers (also called intermediate layers), and output layers. Each layer is fully connected to the next layer in the network.

The input layer consists of neurons that accept the input values. The output from these neurons is same as the input predictors. Nodes in the input layer represent the input data. All other nodes map inputs to outputs by a linear combination of the inputs with the node's weights w and bias b and applying an activation function. This can be written in matrix form for MLPC with K+1 layers as follows

$$\mathrm{y}(\mathbf{x}) = \mathrm{f}_K(\ldots \mathrm{f}_2(\mathbf{w}_2^T \mathrm{f}_1(\mathbf{w}_1^T \mathbf{x} + b_1) + b_2) \ldots + b_K)$$

Hidden layers are in between input and output layers. Typically, the number of hidden layers range from one to many. It is the central computation layer that has the functions that map the input to the output of a node. Nodes in the **intermediate layers** use the sigmoid (logistic) function, as follows.

$$\mathrm{f}(z_i) = \frac{1}{1 + e^{-z_i}}$$

The **output layer** is the final layer of a neural network that returns the result back to the user environment. Based on the design of a neural network, it also signals the previous layers on how they have performed in learning the information and accordingly improved their functions. Nodes in the **output layer** use softmax function [1].

$$\mathrm{f}(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}}$$

B. Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

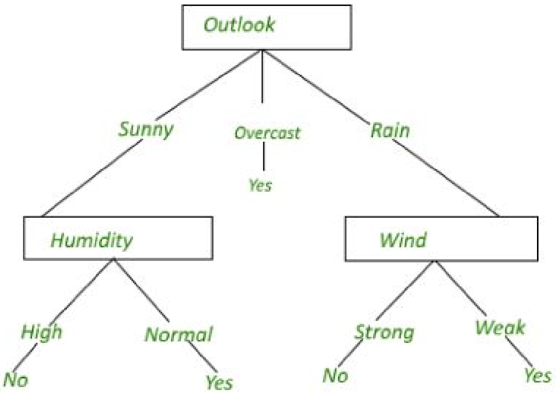

Figure 1 : A decision tree for the concept Play Tennis.

Construction of Decision Tree

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. The construction of decision tree classifier does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. In general decision tree classifier has good accuracy. Decision tree induction is a typical inductive approach to learn knowledge on classification.

Decision Tree Representation

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree,testing the attribute specified by this node,then moving down the tree branch corresponding to the value of the attribute as shown in the above figure.This process is then repeated for the subtree rooted at the new node [3].

C. Lineal Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot) [4].

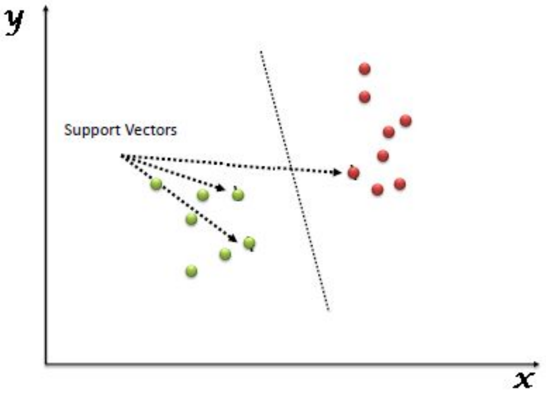

Figure 2: SVC classification example

A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier [5].

One reasonable hyperplane represents the largest separation or margin between the two categories. In other words, the distance from the chosen hyperplane to the nearest data point of each category is maximized. This hyperplane is known as the maximum-margin hyperplane and SVM aims to find this hyperplane to classify the dataset (Burges, 1998)[6].

D. *Logistic Regression*

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Sometimes logistic regressions are difficult to interpret; the Intellectus Statistics tool easily allows you to conduct the analysis, then in plain English interprets the output.

At the center of the logistic regression analysis is the task estimating the log odds of an event. Mathematically, logistic regression estimates a multiple linear regression function defined as[7]:

$$\textbf{logit(p)} = \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{12} + \beta_2 \cdot x_{12} + \_ + \beta_p \cdot x_m$$

Logistic regression is a popular method to predict a categorical response. It is a special case of Generalized Linear models that predicts the probability of the outcomes. Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical[8].

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

III. Implementation of technological tools

A. Spark

Apache Spark is an open-source, general-purpose distributed computing system used for big data analytics. Spark is able to complete jobs substantially faster than previous big data tools (i.e. Apache Hadoop) because of its in-memory caching, and optimized query execution. Spark provides development APIs in Python, Java, Scala, and R [9].

Spark enables parallelized jobs entirely in memory, greatly reducing processing times. Especially if it is an iterative process as used in Machine Learning.

B. Scala

Scales a language that runs on the Java Virtual Machine (JVM). It is a language of multiple paradigms, which allows both object-oriented and functional strategies[10][11].

We use Sacala because it is a functional and object-oriented language to demos that runs at compilation time, it allows us not to consume a lot of system memory, this helps us to get good execution time, and it allows us to work with large volumes of data.

IV. Results

Performance

	Used memory MB	**Seconds**
SVC	464.87	15.5
Logistic Regression	418.77	7.1
Decision Three	373.89	11.9
Multilayer perceptron	598.21	22.3

Accuracy

	Accuracy	**Error**
SVC	0.885	0.114
Logistic Regression	0.884	0.115
Decision Three	0.891	0.108
Multilayer perceptron	0.882	0.117

V. CONCLUSIONS

After looking at the results of the different types of classification algorithms with the banck-full dataset, we can see that on average decision trees was the one that best classified this dataset consuming less memory than the other algorithms with an average time of 11.9 seconds. and an accuracy of 89.1%. The second best algorithm that we looked at was SVC, even though it uses a little more memory, we get an effectiveness a little under decision trees with 88.5% accuracy and with an average time of 15.5 seconds, taking into account that SVC could improve its results or make them worse using different kernels.

REFERENCES

[1] Multilayer Perceptron — DeepLearning 0.1 documentation. (s. f.). Recuperado 1 de mayo de 2020, de http://deeplearning.net/tutorial/mlp.html

[2] SPARK. (s. f.). Classification and regression - Spark 2.4.5 Documentation. Recuperado 1 de mayo de 2020, de http://spark.apache.org/docs/latest/ml-classification-regression.html#multilayer-perceptron-classifier

[3]Decision Tree. (2019, April 17). Retrieved from http://www.geeksforgeeks.org/decision-tree/

[4] Sunil Ray I am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years. (2020, April 15). SVM: Support Vector Machine Algorithm in Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[5] Classification and regression. (n.d.). Retrieved from http://spark.apache.org/docs/latest/ml-classification-regression.html#linear-support-vector-machine

[6] Linear Support Vector Machine. (n.d.). Retrieved from https://www.sciencedirect.com/topics/engineering/linear-support-vector-machine

[7] What is Logistic Regression? (n.d.). Retrieved from https://www.statisticssolutions.com/what-is-logistic-regression/

[8] Swaminathan, S. (2019, January 18). Logistic Regression - Detailed Overview. Retrieved from https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[9] Spark, A. (2018). Apache spark. Retrieved January, 17, 2018.

[10] Morales, A., & Aurelio MoralesLicenciado en Geografía. Máster en Sistemas de Información Geográfica. Consultor GIS desde el año 2004. En MappingGIS desde 2012 para ayudarte a impulsar tu perfil GIS y diferenciarte de la competencia. Echa un vistazo a todos nuestros cursos de SIG online. (2019, June 28). Lenguajes de programación para realizar ciencia de datos. Retrieved from https://mappinggis.com/2019/07/lenguajes-de-programacion-para-realizar-ciencia-de-datos/

[11] The Scala Programming Language. (n.d.). Retrieved from https://www.scala-lang.org/