A Study on Regression Algorithm in Machine Learning

M.Ameerunnisa Begam
Guest Lecturer, Blossom College of Distance Education,
Manonmaniam Sundaranar University, Dindigul-3

M.Guhapriya
Guest Lecturer, Blossom College of Distance Education,
Manonmaniam Sundaranar University, Dindigul-3

Abstract:- In this fast-moving world, millions of data and information exist and accessible to all. But from those collections, gathering exactly required data leads to predict accurate results. ML plays a vital role in converting the data into knowledge. Obliviously people are interacting with ML every day. From each and every interaction it constantly learns and improves the interaction. Regression is an important factor in ML. It determines the relationship among variables. This paper provides a study about regression algorithms such as Linear regression, Support Vector Machine, Random Forest along with their strengths and weaknesses.

Keywords:- *AI, ML, Regression, Support Vector Machine, Linear regression, Random Forest.*

I. INTRODUCTION

A. Machine Learning

Machine Learning is irrefutably one of the most influential and powerful technologies in today's world. It is a great tool that turns information into knowledge. In traditional programming, data is given as input and set of rules are drafted to get accurate output. Now, ML discovers the rules based on the data and output. Multiple forms of machine learning are available. They are supervised, unsupervised, semi-supervised and reinforcement learning.

Process of Machine learning consists of Data collection, Data Preparation, Model Fitting, Model Evaluation and Hyperparameter Tuning. The three basic capabilities of ML are

- *Classification* — divides objects into multiple classes.
- *Regression* — discover relationships between variables.
- *Clustering* — objects with similar characteristics are grouped.

Machine learning techniques are used in Natural Language Processing, Image recognition and computer vision, Cyber Security, Predictive analysis, Marketing and chatbots.

ML is used across a range of industries such as Financial and banking service, Medical and healthcare, Education, Manufacturing, etc.

B. Regression

Regression is one of the basic capabilities of machine learning. It is a form of supervised learning. It discovers relationship among variables and provide output in terms of numbers rather than class. Hence, it is useful in predicting number--based problems like stock market prices, student test performance, temperature for a given day.

There are various regression algorithms, that performs the task efficiently and produce accurate results. They are: *Linear Regression, Logistic Regression, K nearest neighbors and Decision Trees, Support Vector machine, Random Forest and Naive Bayes.*

Each regression algorithm has its own features. Based on the data preparation regression algorithm will be selected and trained and then machine learning model will be generated.

II. LINEAR REGRESSION

Linear Regression is one of the most common regression techniques in machine language. It attempts to fit a straight hyperplane to the dataset, takes the features and predict a continuous output. For example, if the dataset consists of only two variables LR attempts to fit a straight line. When the relationship between the variables in the dataset are linear, this LR works well and provide accurate results. It finds a linear curve solution to every problem.

$$h_\theta=\theta_0+\theta_1x_1+\theta_2x_2+\ldots \qquad \text{- Eqn. 1}$$

Weight parameter is allocated and each training features are hold in theta. At the initial stage of training, theta is initialized randomly. Later, based on the changes in the expected and predicted output, the values of theta will be corrected. To align the θ values in right direction, gradient descend algorithm is used.

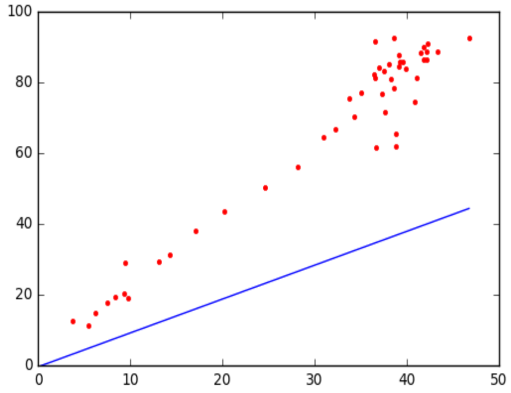

Fig 1:- Linear Regression

The above diagram shows the derived solution in blue line and training data in red dots.

➢ *Strengths:*
- Straight forward to understand and implement.
- Space complex solution.
- Easy to update with new data.

➢ *Weaknesses:*
- Poor in handling non-linear relationships.
- Not flexible to capture more complex pattern.

III. SUPPORT VECTOR MACHINE

Support Vector Machine is a type of machine learning technique which is used for both classification and regression. It is a supervised learning technique that is used to perform classification and regression analysis. There are two major variants available to support linear and non-linear problems.

Linear SVM separates the problem space by deriving a hyperplane and maximizes the classification margin. It has no kernel. An optimal hyperplane is drawn between the maximum margin at the midpoint. The nodes in the feature space that are in the boundary of the maximal margin are said to be support vectors. (Fig. 2)

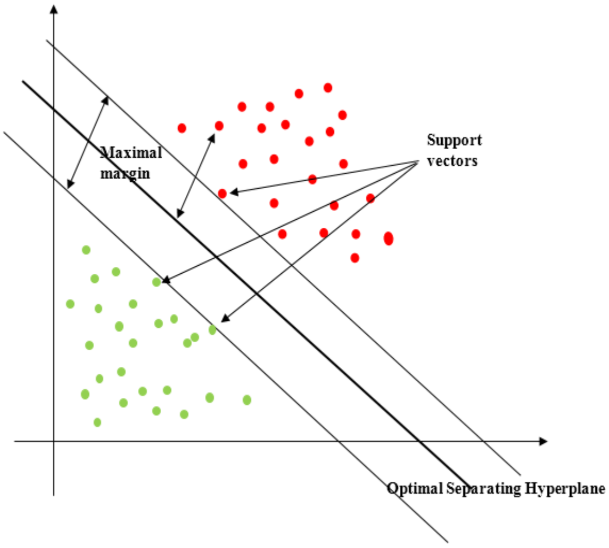

Fig 2:- Linear SVM

Value of margin(m) will be inversely proportional to ||w||. For maximizing margin, we need to minimize||w||. The optimization problem is shown below.

$$Minimize\ \frac{\|\underset{w}{\rightarrow}\|^2}{2}\ where\ y_i(\underset{w}{\rightarrow}.\underset{x}{\rightarrow}+b)\ \geq 1$$
$$for\ any\ i = 1, \ldots., n$$ -- Eqn. 2

where, w – set of weight matrices

When the datasets are fully linear, Eqn. 2 works well and produce optimal solution. But to handle outliers, hinge loss has to be used to get slack variable.

$$Minimize\ \frac{1}{2}\|w\|^2 + c\sum_i \max(0,1 - y_i\,(w^T x_i + b))$$

--Eqn. 3

where, w - tune with maximum margin between the classes.

C - decides the level of margin.

By using the above equation, the cost function is minimized.

When the dataset is linearly separable using Eqn. 2 and Eqn. 3 problem space can be separated. If the dataset is not linearly separable, then Non-linear SVM has to be used. For deriving a new hyperplane, kernel function is used in Non-linear SVM. New hyperplane forms linearly separable curve to classify the dataset. (Fig. 3)

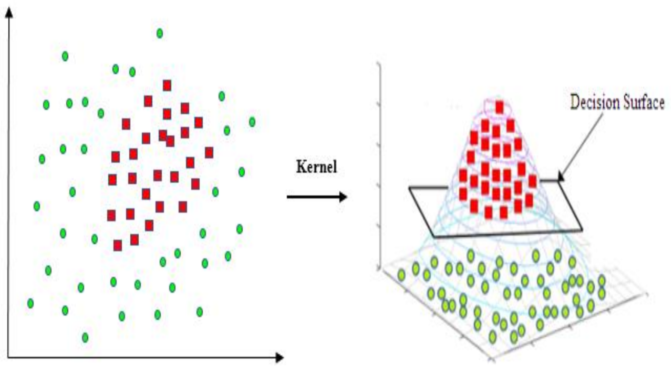

Fig 3:- Non-Linear SVM

Non-linear SVM includes new kernel function to Eqn. 2 which forms the new equation as shown below

$$Minimize\ \frac{\left\|\underset{w}{\rightarrow}\right\|^2}{2} + \mathrm{C}\sum_i \boldsymbol{\zeta_i}$$
$$where\ y_i(\underset{w}{\rightarrow}.\phi(x_i) + b)\ \geq\ -\boldsymbol{\zeta_i}$$
$$for\ all\ 1\ \leq i\ \leq n, \boldsymbol{\zeta_i} > \mathbf{0}$$ --Eqn. 4

Kernels such as Gaussian kernel, polynomial kernel, Sigmoid kernel, Laplace RBF kernel etc. can be used in non-linear kernels.

➢ *Strengths:*
- Complex problems are solved using kernel tricks.
- Effective when number of dimensions is greater than number of samples.
- SVM works well in high dimensional spaces.
- Hinge loss provides higher accuracy.

➢ *Weaknesses:*
- Difficult to choose kernel trick.
- Memory requirement is high.
- Hinge loss leads to sparsity.
- Longer training time for larger datasets.

IV. RANDOM FOREST

Random Forest is a collection of models. It consists of multiple decision trees that are combined (as shown in Fig. 4) and form a strong model to carryout classification and regression. The newly derived model will be more robust, accurate and handles overfitting better than basic models. Based on the majority voting, it calculates the output for classification. For regression, mean is calculated.

Random forest model is good in handling tabular data or categorical features. It captures non-linear interaction between the features and the target. Tree-based models are not designed to work with very sparse features. While dealing with sparse input data, sparse features can be pre-processed to generate numerical statistics, or switch to a linear model. This suits better for such scenarios.

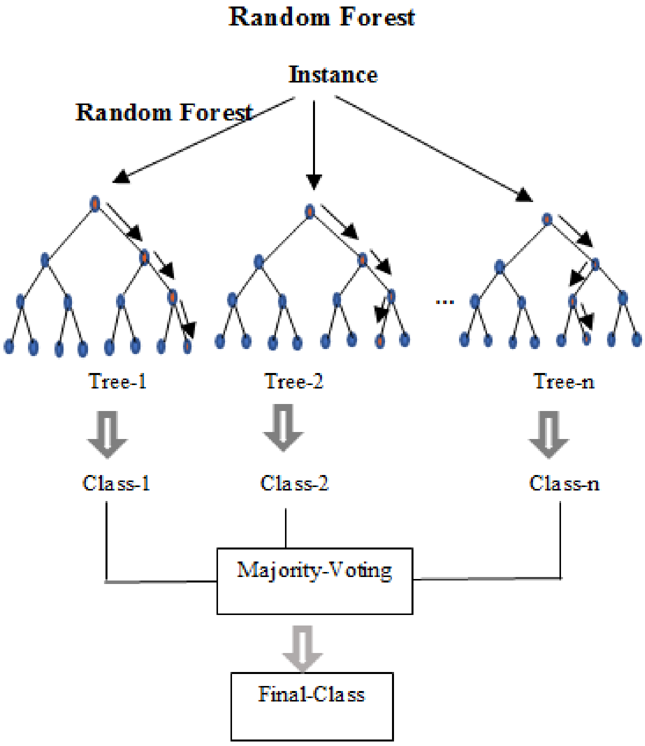

Fig 4:- Random Forest

➢ *Strengths:*

- Robust, Accurate and powerful model.
- Reduce overfitting, variance.
- Works well with categorical and continuous variables.
- Supports implicit feature selection and derives feature importance.

➢ *Weaknesses:*

- High computational cost when forest becomes large.
- Slow in prediction.

V. CONCLUSION

Machine learning techniques are used to find the underlying patterns within complex data automatically, else it is hard to discover. Future event prediction and complex decision making can be done with the hidden patterns and knowledge about a problem. Machine learning converts the data into knowledge. Classification and Regression are most important part of ML. Regression discovers models based on the given dataset. In this paper, some regression techniques with their strengths and weaknesses are described. Each type of models has different special features. Optimal result depends on the selection of regression algorithm that most suits to the data. In future, with the collaboration of various regression techniques, we can derive new regression technique that works well in short time with high accuracy even for large datasets.

REFERENCES

[1]. Kotsiantis, S.B., Zaharakis, I.D., Pintelas, P.E.: “Machine learning: a review of classification and combining techniques.” Artif. Intell. Rev. 26(3), 159–190 (2006)

[2]. Cristianini N, Shawe-Taylor J (2000) “An introduction to support vector machines and other Kernel-based learning methods.” Cambridge University Press, Cambridge.

[3]. Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood: “Random Forests and Decision Trees” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012 ISSN (Online): 1694-0814

[4]. Breiman, L.: “Random forests. Machine. Learning.” 45, 5–32 (2001). DOI 10.1023/A:1010933404324