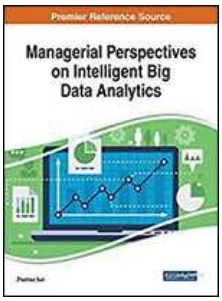

Managerial Perspectives on Intelligent Big Data Analytics

by Zhaohao Sun

Chapter 11: Analysis of Cutting-Edge Regression Algorithms Used for Data Analysis

Indivar Mishra,
KIIT University,
India

Ritwik Bandyopadhyay,
KIIT University,
India

Sourish Ghosh,
KIIT University,
India

Aleena Swetapadma,
KIIT University,
India

ABSTRACT

Considering the growing applications of big data analytics in the various fields such as healthcare, finance, e-commerce, and web services, it is essential to continuously develop techniques useful for big data. Among various techniques used for big data analytics, regression analysis is very important. In this chapter, an attempt is made to take a detailed look into some of the main regression algorithms and their origin that are used for big data analytics. In this study, some of very famous works related to regression along with some latest research are analyzed. Regression is the process of deducing a predictive model for real-world information based on verified information that is already received. It is used for making predictions, optimizing solutions to complex problems, and understands trends in large and big data analytics. The goal of this study is to promote and facilitate a better understanding of regression algorithms that are in use in the real world for big data analytics.

INTRODUCTION

With advancement of technology the amount of data is increasing exponentially which makes it difficult for data scientists to retrieve much of useful information out of it. Classical data analysis techniques focus on data collection followed by the imposition of a model, analysis, estimation and testing that depends on the parameters of that model. This type of analysis is not suitable for all types of data as it is a hectic job to explicitly program the system with respect to new datasets. This is where big data analytics comes in to the picture. Big data analytics uses various techniques for analysis such as rule-based systems, machine learning, multi-agent systems techniques, neural networks systems, fuzzy logic systems, cased-based reasoning techniques, genetic algorithms techniques, data mining algorithms, cognitive computing, natural computing, intelligent agents etc. Among these techniques for big data analytics, regression is an important method which is focused in this study.

In this study various types regression algorithms has been described in brief and its importance in big data analytics has been explored. Various types of regression algorithm that has been used for big data analytics are linear regression, polynomial regression, support vector machine regression, decision tree regression, ensemble learning regression, neural network regression, pattern aided regression etc. Above described regression algorithms and its relation with respect to big data analytics will be described in the next section.

BACKGROUND

Before diving into regression and methods to implement it, first it must be mentioned clearly basics and importance of a dataset in big data analytics. Data as the formal definition goes are unprocessed facts, in predictive modeling data has two parts dependent and independent data/variable/feature, goal is to build relation for dependent variable with respect to one or more independent variables. Dependent variables are those variables whose scalar/vector magnitude value depend on independent variables and can be found through some relation among independent variables, on the other hand independent variables are the variables which are not dependent or have any relation with other variables in the dataset. Datasets can be discrete data or continuous data (Pasta, 2009).

Regression algorithms are used to predict continuous variables/data on the basis of one or more independent variables (Rawlings, Pantula, & Dickey, 1998). Attempts at regression date back to the days of the legendary Isaac Newton and Joseph-Louis Lagrange. The crudest attempt at this was interpolation- where, given (n+1) points on a 2-D plane, a polynomial of degree n satisfying all those points can be derived. Then the values at all other points by using this polynomial can be

predicted. Unfortunately, such a simple solution cannot be applied in the fabrication of an intelligent device. This method is used for mathematical purposes with purely academic ambition.

The solution relies overly on the training data supplied to it. This is called over-fitting. In Figure 1, the red line shows a curve that uses a method like interpolation and exhibits the problem of over-fitting. The green line does not touch each point given, but it is a better predictive model because the total error is less. It shows the difference between a useful regression model and a curve obtained by interpolation which goes through each given data point. At the point labeled 0.9, the approximate error for the red line is much higher than that for the green line. This means, that the green line is a much more accurate predictive model- as mentioned earlier. This is achieved through some of the methods discussed in this study.

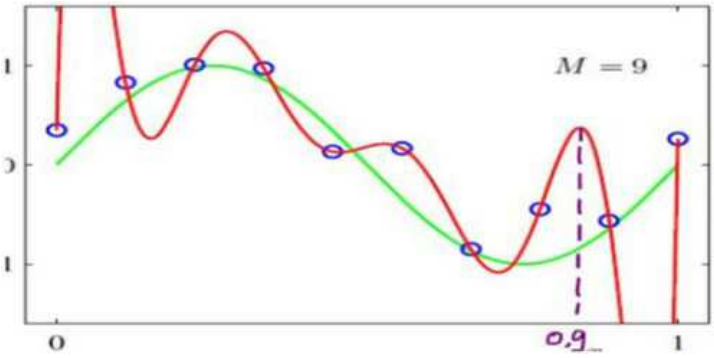

Figure 1: A simple example demonstrating over-fitting

REGRESSION TECHNIQUES

Above described problems paved the way for mathematicians, who eventually evolved into computer scientists, to chalk up algorithms for regression. The ultimate objective of these attempts was to find a way to fashion a predictive model which is neither too reliant on the data provided, nor overly generalized (Umam, n.d.). This optimization between bias and variance is the main challenge of regression algorithms. This has many approaches giving weight-age to certain points, using error functions to determine the usefulness of a data point or even using simple conditional statements. There are various types of regression algorithm which are being used in big data analytics such as linear regression, polynomial regression, support vector machine regression, decision tree regression, ensemble learning regression, neural network regression, pattern aided regression etc. Some of the regression methods have been discussed below.

Linear Regression

It is a very basic and popularly used algorithm to find a linear relation between dependent and independent variables. If there is only one independent variable then it is called simple linear regression as shown in (1),

$$(1)\quad Y = A_0 + A_1 X_1 + E$$

Similarly there is one more type of regression called as multiple linear regressions. Multiple regressions are the relationship between several independent or predictor variables and a dependent or criterion variable (Su, Gao, Li, & Tao, 2012). If a relation is to be built from more than one independent variables then it's called multiple linear regressions as shown in (2).

$$(2)\quad Y = A_0 + A_1 X_1 + A_2 X_2 + \ldots + A_n X_n + E$$

Here the variables $X_1, X_2, \ldots, X_n$ are the independent variables and Y is the dependent variable and E is the error related to prediction. The coefficients $A_0, A_1, \ldots, A_n$ are chosen such that the line drawn through these equations have minimum distance from each point causing minimum error.

Figure 2 shows the example graph of simple linear regression prediction line with corresponding errors. In Figure 2, a case of simple linear regression has been considered in which the red dots are dependent value on y-axis which is dependent on an independent value on x-axis. When some dataset to the algorithm is provided it will find coefficients A0 and A1 such that the line using these coefficients will be closest to each point of the training dataset. The blue line represents equation shown above. The magnitude of green line shows the error related to each prediction. If all the possible error with every possible coefficient from the equation will plotted it will always form out to be a convex contour plot as show in Figure 3.

In Figure 3 the two horizontal axes are the two coefficients (in case for simple linear regression) and corresponding vertical axis gives error associated with those coefficients. Coefficients of linear equation mentioned above are calculated by using cost function. The very purpose of the cost function is to find the coefficients of equation such that the magnitude of error is minimal.

This algorithm is known as gradient descent algorithm (Hall'en, 2017) as shown in (3).

$$J(a) = 1/2m \sum_{i=1}^{m} (h(x^i) - y^i)^2 \quad (3)$$

The above formula is repeated until convergence with arbitrary coefficients usually (0, 0). For every iteration, coefficients are chosen such that the resulting value of the cost function is minimum, and the coefficients are updated after everyone complete execution. In Cervellera & Macciò (2014) a method has been suggested for linear regression which concludes that the estimation improves as the discrepancy of the observation points becomes smaller. It allows to treat indifferently the cases in which the samples come from a random external source and the one in which the input space can be freely explored. Lasso is a popular technique for joint estimation and continuous variable selection for sparse linear regression problems. In Mateos, Bazerque, & Giannakis (2010), an algorithm had been developed to estimate the regression coefficients via Lasso when the training data are distributed across different agents, and their communication to a central processing unit is prohibited. Other than linear regression there is one more type of linear regression called polynomial linear regression which will be explored in the subsection below.

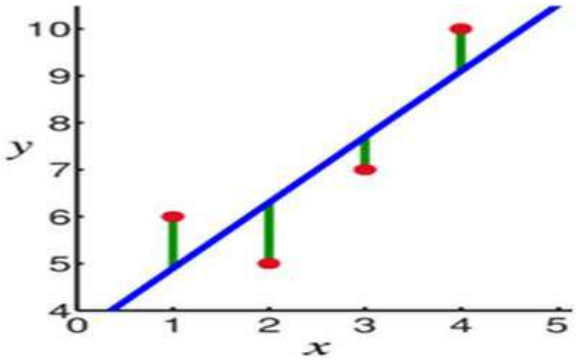

Figure 2: Graph of simple linear regression prediction line with corresponding errors

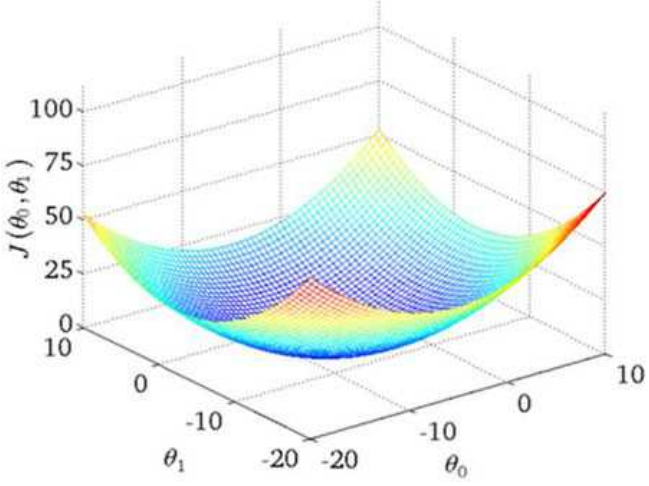

Figure 3: Contour plot

Polynomial Linear Regression

Linear regression as discussed above is a good method for data fitting when there is a linear relation among the variables. But when dataset is concerned about growth of population, tissues etc., where there is exponential increase in dependent variable over small change in independent variable, then the linear models fail to show a good fit/prediction. In these cases polynomial linear regression come in handy. Polynomial linear regression is capable of fitting non linear data. Generalized equation for linear model with one independent variable is shown in (4) and generalized equation for polynomial model with one independent variable is shown in (5),

$$Y = A_0 + A_1.X_1 \quad (4)$$

$$Y = A_0 + A_1.X_1 + A_2.X_2^2 \quad (5)$$

Figure 4 shows two plots, first line is fitted using linear regression and latter is fitted using polynomial regression. It is clear from

Figure 4 that polynomial model is far better fitted than linear model. It's the X_2^2 factor that gives the curve parabolic effect and hence covering most of the points in the plot with minimum error. Like linear regression, polynomial linear regression also works on the concept of ordinary least square method. Generalized polynomial equation for one independent variable is shown in (6),

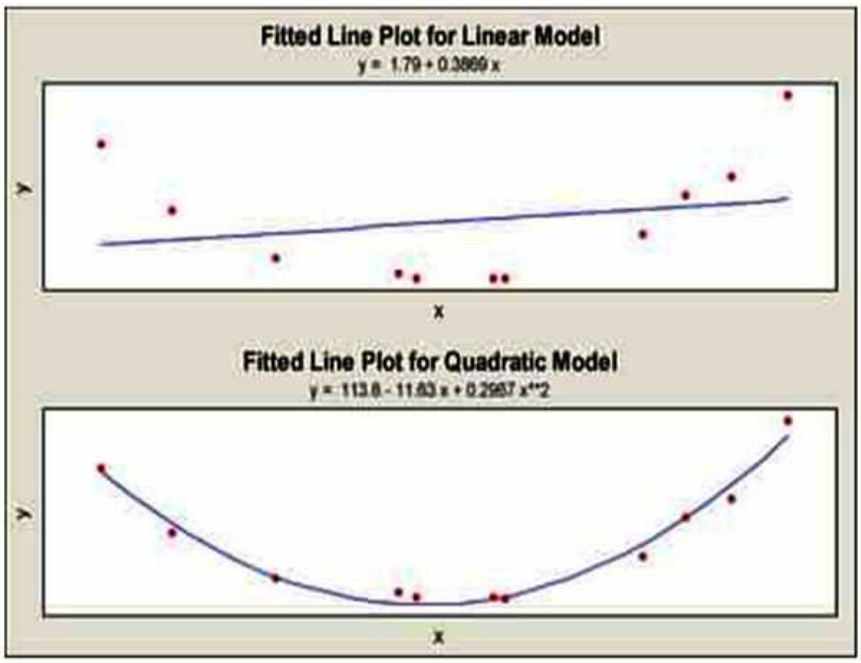

Figure 4: Comparison of fit between simple linear and polynomial fit

(6) $$Y = A_0 + A_1.X_1 + A_2.X_2^2 + A_3.X_3^3 + \ldots + A_n.X_n^2$$

Depending on the highest power used in an equation they are named as quadratic, cubic, quadratic and so on. Polynomial models are still called linear models as there coefficients which are known are have linear dependency i.e. a0, a1, a2… an, hence polynomial linear is a special case of multiple linear regression.

Support Vector Machines Regression (SVR)

Now that there is a general idea about regression and its importance in classifying dataset by generating decision boundaries. But the problem arises when a dataset is classified where the two or more classes of data are not linearly separable (Osuna, Freund, & Girosi, 1997). This is where the idea of SVM, championed by Vladimir Vapnik in Statistical Learning Theory (1998), comes in which can generate a non linear decision boundary by drawing hyper-planes with the help of kernels as shown in Figure 5 (Lin, Huang, & Chiueh, 1998).

SVM, a supervised learning algorithm, is a large margin learner that is it chooses such a decision boundary where there is maximum distance between the hyper-plane and the corresponding classes of data or the best line that can fit the training data keeping a maximum margin. Classification is the process by which an algorithm analyzes the decision boundary taking help from the training data containing known observations. Likewise in regression, it estimates the relationship among one or more independent variables. SVM works on empirical risk minimization which leads us to an optimization function as given in (7).

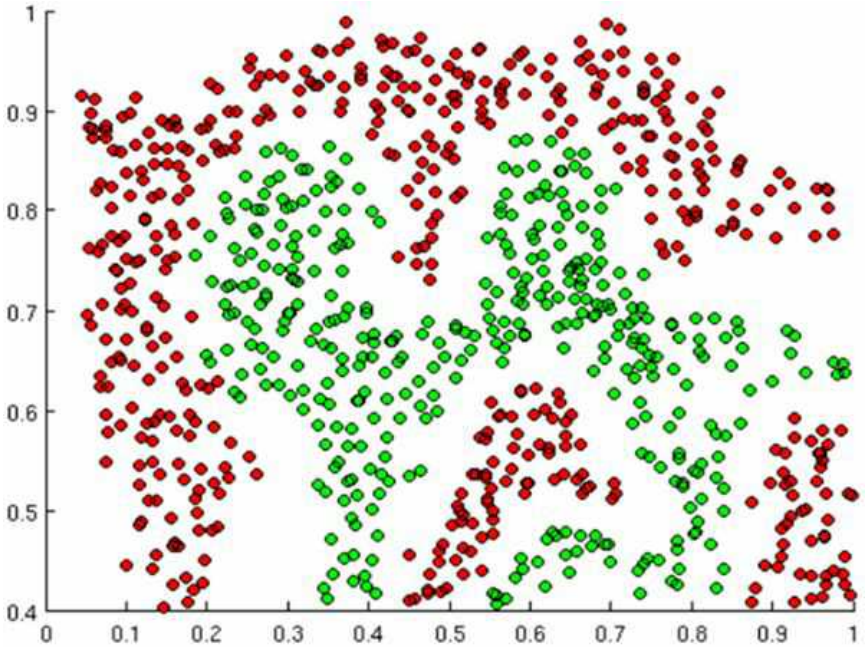

Figure 5: A typical case of support vector machine

(7) $$minimumof\omega \sum \ell\left(x_i, y_i, \omega\right) + \lambda \cdot r\left(\omega\right)$$

where *l* is the hinge loss function and r is the regularization function. In case of support vector regression (SVR), where the output takes on continuous values, it becomes very difficult to predict the best fit line. A loss function is defined that ignores the errors, which are situated at certain distance of the true value. This function is known as Epsilon (ξ) intensive loss function (Cherkassky & Ma, 2002), which is a margin of tolerance. However, the main idea is to minimize error while individualizing the hyperplane which maintains a large-margin regressor, keeping in mind that a part of the error has to be tolerated. The SVR model has been shown in Figure 6.

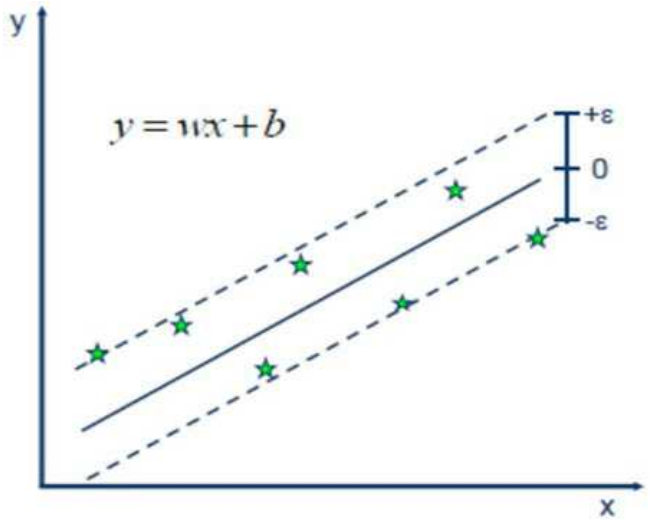

Figure 6: A support vector regressor mode forming the best line between the 2 dotted lines

Thus, the general solution would be as shown in (8),

(8) $$min\,\frac{1}{2}\left|\left|\omega\right|\right|^2 \text{ with constraint } \mathrm{y}_i - \omega x_i - b \leq \varepsilon$$

where $\left\|\omega\right\|$ is the length of the vector $\vec{\omega}$. Another important part is the error minimization as shown in (9),

(9) $$\frac{1}{2}\left|\left|\omega\right|\right|^2 + c\sum_{i=1}^{N}\left(\xi i + \xi' i\right)$$

When moving from non-linear SVR where no best line is possible to fit the data it is needed to transform that data into a much higher dimensional feature vector. This is achieved with the help of kernel functions (Smola & Schölkopf, 1998). There are mainly 2 types of kernel functions that are used mostly. Those are polynomial function and Gaussian radial basis function which are given in (10) and (11) respectively.

(10) $$k\left(xi, xj\right) = \left(xi \cdot xj\right)^d$$

(11) $$k\left(xi, xj\right) = exp\left(-\frac{\left|\left|xi - xj\right|\right|^2}{2\sigma^2}\right)$$

SVR have been very successful in pattern recognition and function estimation problems for crisp data. In (Hong & Hwang, 2005), a new method to evaluate interval linear and nonlinear regression models combining the possibility and necessity estimation formulation with the principle of quadratic loss SVR has been proposed. It utilizes quadratic loss function, unlike the traditional SVM. When function approximation in SVR is non-stationary the single kernel approach may be ineffective. In (Bellocchio, Ferrari, Piuri, & Borghese, 2012), a hierarchical SVR model has been presented which aims to provide a good solution. It consists of a set of hierarchical layers, each containing a standard SVR with Gaussian kernel at a given scale. However with all these advantages, SVR suffers from a basic disadvantage. When number of training examples is much more than number of features, SVR can take up a lot of time to train. In such circumstances, logistic regression is preferable over SVR.

Decision Tree Regression

Decision tree as name suggests works on the principles of traditional data structures tree (Mohamed & Robert, 2010). Decision tree, also known as CART (Breiman, Friedman, Olshen, & Stone, 1987), is a powerful tool to predict dependent variables compared to other algorithms, the basic/generalised idea behind working of decision trees are to keep splitting the dataset to smaller homogeneous groups, and then check if a data belongs to that particular group. Consider the plot shown in Figure 7 (a). Suppose it is required to predict a dependent variable with two independent variables x1 and x2, so the dependent variable will be perpendicular to the plane (x1, x2). Now if a point in this plane say (x01, x02) is placed then and its corresponding value which is the average of the data points in the plane, which will give a very poor prediction if we apply this idea to our machine learning algorithm. Ironically decision tree uses somewhat similar approach the algorithm divides the plane into sub planes (nodes) until each plane has considerable homogeneity, or maximum features/depth is reached (leaf nodes). For decision tree regression the splits in the regions are made taking those independent variables first which have minimum value of cost function i.e. error, and it keeps dividing the recursively until a criterion like maximum splits/maximum depth is reached.

Figure 7 (c) shows decision making tree model to predict the value given dependent variable in our case x1 and x2, the small triangles are decision-making nodes and rectangular boxes are the predicted values of particular x1,x2 model also called leaf node of tree which basically contain average value of the split region in Figure 7 (b). Suppose x1 = 0.6 and x2 = 0.6, now from Figure 7 (c) it can be observed that x2 is not smaller than 0.55 therefore move to the right child (decision node) of tree which gives another condition clearly x1 is not smaller than 0.23 therefore again move to right child of the decisive node where next condition is satisfied so move to left child node of the decisive node and apparently reach leaf node which will contain average value of data points in the region 4 in Figure 7 (b).

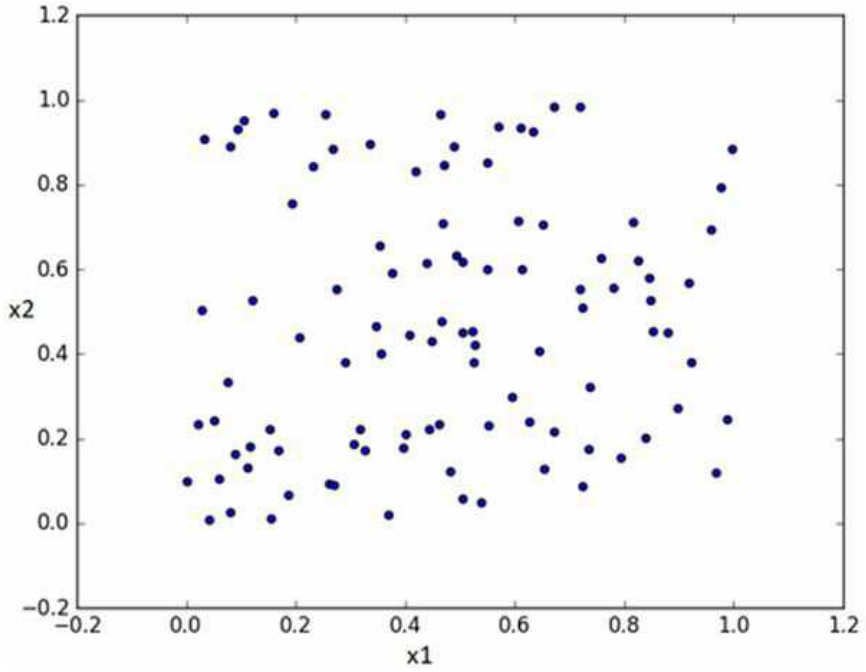

Figure 7a: A random data set scattered in 2-D plane

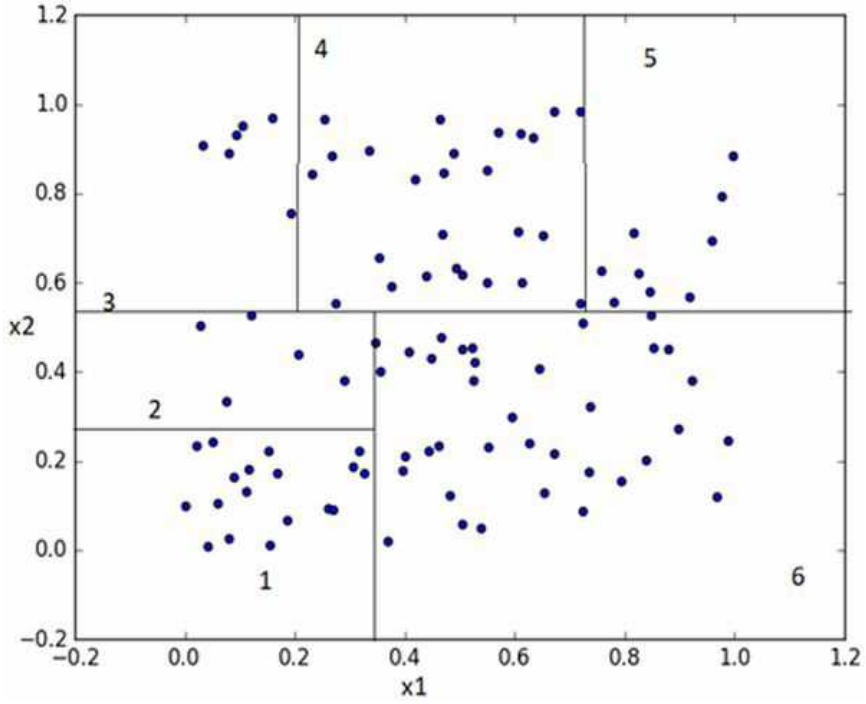

Figure 7b: After applying Decision tree algorithm 2-D plane with virtual/imaginary lines

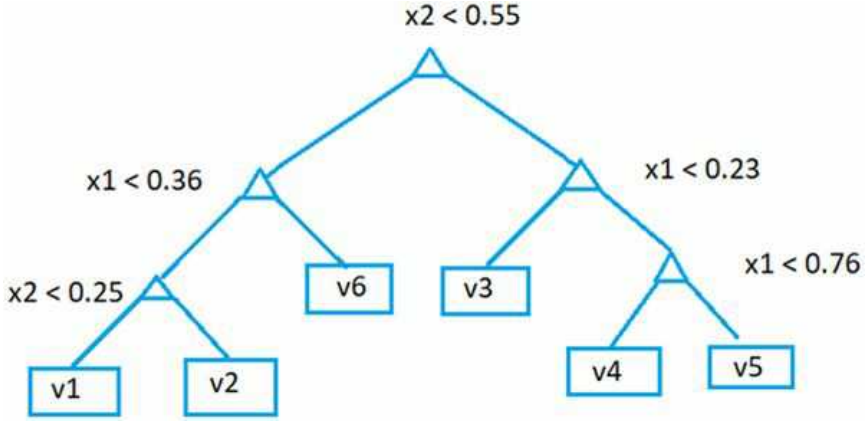

Figure 7c: Decision tree graph of said dataset

Decision tree are much better than other classifiers as they give an intuitive tree from which one can understand how the classification is done and which feature has most contribution in the problem. They are one of the powerful tool for predictive modelling in machine learning arsenal but still it has some drawbacks like it easily gets over fitted, it works on a greedy approach etc. These problems are handled by a method called pruning which simply means removing the decisive nodes which have less contribution to the classification/regression (Patel & Upadhyay, 2012). In Gey & Nedelec (2005), performance of the regression trees pruning algorithm and the final discrete selection by test sample as a functional estimation procedure is considered. The validation of the pruning procedure is applied to Gaussian and bounded regression. In Malerba, Esposito, Ceci, & Appice (2004) a method for the data-driven construction of model trees is presented called as stepwise model tree induction method. Main feature of the method is two types of nodes: regression nodes, which perform only straight-line regression, and splitting nodes, which partition the feature space. The multiple linear models associated with each leaf are then built stepwise by combining straight-line regressions reported along the path from the root to the leaf. Internal regression nodes contribute to the definition of multiple models which have global effect. Straight-line regressions at leaves have local effects. These trees are simple and easily interpretable and their analysis often reveals interesting patterns.

Ensemble Learning Regression

Sometimes, these predictive models by themselves are not very effective. That is where ensemble learning comes in. In Freund & Schapire (1999), a bottom-up approach where a number of weak learners are given and a better predictive model is obtained by combining them. A weak learner is defined as a learning algorithm that has an error rate less than a half. In this paper, we come across the concept of AdaBoost. In this algorithm, a weak learner is trained on the training dataset for T times. Every time, the training data has a distribution of weight over itself and after every round of training, the weight is increased for samples with higher error rate. Which means, as the number of rounds of training increases, the error keeps going down. The maximum training error and the generalisation error are also calculated.

Similar algorithms include Gradient tree boosting and XGBoost. These are basically ways to improve the basic weak learner used and then incorporate them to the simple AdaBoost processing that is discussed in Freund & Schapire (1999). In simpler words, the difference between them is similar, in a manner, to the difference between simple and compound interest on a fixed sum of money. Another widely used type is bootstrap aggregating or bagging. In this, the training dataset is divided into subsets. Then the learner is trained on each of these training data subsets. Then average all the learning models derived and do not score. Figure 8 shows the plot of error and round for training and testing.

This gives a much better predictive model than others. This is better than the other regression models because the models that are biased due to over-fitting have very little effect. In ensemble learning regression, due to repetitive use of the algorithms, the generalization error is very hard to remove.

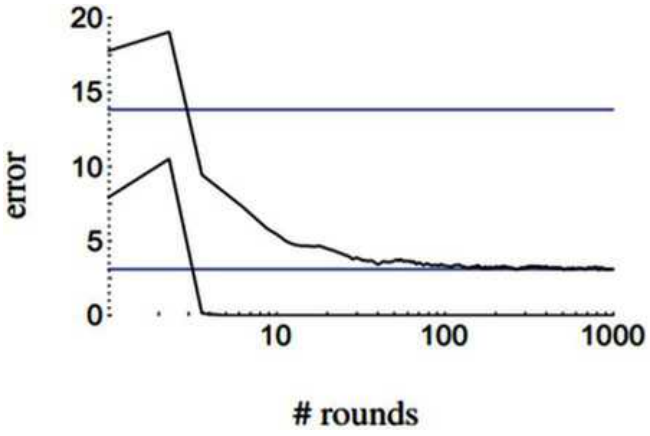

Figure 8: Lower curve for training and upper curve for testing

APPLICATIONS OF REGRESSION IN BIG DATA ANALYTICS FRAMEWORK

The above described regression algorithms are used most frequently in big data analytics framework. Some of the applications of regression algorithm in big data analytics will be discussed here. In Berberidis, Kekatos, & Giannakis (2016), a method was suggested for identifying and omitting less informative observations in an online and data-adaptive fashion. Here the maximum-likelihood estimator is sequentially found using first- and second-order stochastic approximation algorithms. These schemes are well suited when data are inherently censored or when the aim is to save communication overhead in decentralized learning setups.

In Lin, Chen, & Tsai (2016), a novel robust non-contact technique has been suggested for the evaluation of heart rate variation.

Here ensemble empirical mode decomposition of the Hilbert-Huang transform has been used to acquire the primary heart rate signal. The instantaneous frequencies from intrinsic mode functions are implemented by the multiple-linear regression model to evaluate heart rates. Predicting the bug number of a software system is important for project managers and software end-users. In Zhang, Du, Yoshida, Wang, & Li (2018), a method called SamEn-SVR has been proposed to combine sample entropy and support vector regression (SVR) to predict software bug number. Template vectors are used with the smallest complexity as input vectors for SVR classifiers to ensure predictability.

Transmission lines are one of the major components of the electrical power system. With the growing demand of the power the number of transmission lines is increasing day by day. With increase in transmission lines the type and volume of the data are also increasing. To efficiently manage this type big data in electrical power system various data analytics techniques has been used. In Swetapadma & Yadav (2017), a decision tree regression based fault distance estimation scheme for double-circuit transmission lines has been presented. Here decision tree regression was chosen because it requires less training time, offers greater accuracy with a large data set, and robustness than all other techniques like artificial neural networks, support vector machines, adaptive neuro-fuzzy inference systems, etc. The scheme is relatively simple and easy in comparison with complex equation-based fault-location estimation methods.

CONCLUSION AND FUTURE WORK

Big data analytics is the most interesting topic of the data science due to the increasing and large data in today's world in almost all fields. Regression is one of the most important techniques used in big data among various other techniques. In this study, concept of various regression algorithms has been presented. It was observed that there is no best algorithm for regression. The ones mentioned are the primary types of regression algorithms. These are used based on the application intended and the constraints imposed by the nature of training data. Algorithms implementing other learners to keep working on a problem are called auto machine learning which the next stage of everything in this study is. However, there is no algorithm to determine the optimal method needed for a particular problem. That would need a algorithm coupled with natural language processing to come to a credible decision for big data analytics.

REFERENCES

Bellocchio, F., Ferrari, S., Piuri, V., & Borghese, N. (2012). *Hierarchical Approach for Multi-scale Support Vector Regression*. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9), 1448–1460. doi:10.1109/TNNLS.2012.2205018 PMID:24807928

Berberidis, D., Kekatos, V., & Giannakis, G. (2016). *Online Censoring for Large-Scale Regressions with Application to Streaming Big Data*. *IEEE Transactions on Signal Processing*, 64(15), 3854–3867. doi:10.1109/TSP.2016.2546225 PMID:28042229

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1987). *Classification and Regression Trees*. *Cytometry*, 8, 534–535.

Cervellera, C., & Macciò, D. (2014). *Local Linear Regression for Function Learning: An Analysis Based on Sample Discrepancy*. *IEEE Transactions on Neural Networks and Learning Systems*, 25(11), 2086–2098. doi:10.1109/TNNLS.2014.2305193 PMID:25330431

Cherkassky, V., & Ma, Y. (2002). *Selecting of the Loss Function for Robust Linear Regression*. Academic Press.

Freund, Y., & Schapire, R. (1999). *A Short Introduction to Boosting. Proceedings of the 16th international joint conference on Artificial intelligence*, 2, 1401-1406.

Gey, S., & Nedelec, E. (2005). *Model selection for CART regression trees*. *IEEE Transactions on Information Theory*, 51(2), 658–670. doi:10.1109/TIT.2004.840903

Hall'en, R. (2017). *A Study of Gradient-Based Algorithms*. Mathematical Statistics, Lund University Libraries.

Hong, D., & Hwang, C. (2005). *Interval regression analysis using quadratic loss support vector machine*. *IEEE Transactions on Fuzzy Systems*, 13(2), 229–237. doi:10.1109/TFUZZ.2004.840133

Lin, K., Chen, D., & Tsai, W. (2016). *Face-Based Heart Rate Signal Decomposition and Evaluation Using Multiple Linear Regression*. *IEEE Sensors Journal*, 16(5), 1351–1360. doi:10.1109/JSEN.2015.2500032

Lin, S., Huang, R., & Chiueh, T. (1998). *A tunable Gaussian/square function computation circuit for analog neural networks*. *IEEE Transactions on Circuits and Systems. 2, Analog and Digital Signal Processing*, 45(3), 441–446. doi:10.1109/82.664259

Malerba, D., Esposito, F., Ceci, M., & Appice, A. (2004). *Top-down induction of model trees with regression and splitting nodes*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5), 612–625. doi:10.1109/TPAMI.2004.1273937 PMID:15460282

Mateos, G., Bazerque, J., & Giannakis, G. (2010). *Distributed Sparse Linear Regression*. *IEEE Transactions on Signal Processing*, 58(10), 5262–5276. doi:10.1109/TSP.2010.2055862

Mohamed, H., & Robert, P. (2010). *Dynamic Tree Algorithms*. arXiv:0809.3577 [math.PR]

Osuna, E., Freund, R., & Girosi, F. (1997). *An Improved Training Algorithm for Support Vector Machines*. *Proceedings of the IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing VII*, 276-285. 10.1109/NNSP.1997.622408

Pasta, D. (2009). *Learning When to Be Discrete: Continuous vs. Categorical Predictors*. San Francisco, CA: ICON Clinical Research.

Patel, N., & Upadhyay, S. (2012). *Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA*. *International Journal of Computers and Applications*, 60(12), 20–25. doi:10.5120/9744-4304

Rawlings, J., Pantula, S., & Dickey, D. (1998). *Applied Regression Analysis: A Research Tool*. Springer. doi:10.1007/b98890

Smola, A., & Schölkopf, B. (1998). *A Tutorial on Support Vector Regression NeuroCOLT Technical Report*. NC-TR-98-030, Royal Holloway College, University of London, UK.

Su, Y., Gao, X., Li, X., & Tao, D. (2012). *Multivariate Multilinear Regression*. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, 42(6), 1560–1573. doi:10.1109/TSMCB.2012.2195171 PMID:22677310

Swetapadma, A., & Yadav, A. (2017). *A Novel Decision Tree Regression-Based Fault Distance Estimation Scheme for Transmission Lines*. *IEEE Transactions on Power Delivery*, 32(1), 234–245. doi:10.1109/TPWRD.2016.2598553

Umam, A. (n.d.). *Ardian Umam Blog*. Retrieved from *https://ardianumam.wordpress.com/tag/regression/ArdianUmam.* Blog

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley-Interscience Publication.

Zhang, W., Du, Y., Yoshida, T., Wang, Q., & Li, X. (2018). *SamEn-SVR: Using sample entropy and support vector regression for bug number prediction*. *IET Software*, 12(3), 183–189. doi:10.1049/iet-sen.2017.0168

ADDITIONAL READING

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. North Carolina, USA: Wiley.

Cady, F. (2017). *The Data Science Handbook*. New Jersy, USA: Wiley. doi:10.1002/9781119092919

Dasgupta, N. (2018). *Practical Big Data Analytics*. Birmingham, United Kingdom: Packt Publishing.

Dhiraj, A., Minelli, M., & Chambers, M. (2012). *Big Data, Big Analytics*. New Jersy, USA: Wiley CIO Series.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. USA: Wiley.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

Zikopoulos, P., & Eaton, C. (2011). *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. USA: McGraw-Hill.

KEY TERMS AND DEFINITIONS

Big Data:
It is a massive volume of both structured and unstructured data that is difficult to process using traditional techniques.
CART:

It is the classification and regression trees which work on decision tree algorithms that can be used for classification or regression predictive modelling.

Data Analytics:

It is the process of examining data to draw conclusions about the information they contain.

Ensemble Learning:

It uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone.

Linear Regression:

It is an approach to modelling the relationship between a scalar response and explanatory variables.

Regression:

It is a statistical measure used to determine the strength of the relationship between one dependent variable and a series of other changing variables.

SVM Kernels:

SVM algorithms use a set of mathematical functions called kernel which take data as input and transform it into the required form.