International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Predictive Analytics: The Core of Data Mining

Neeta Dhane[1], Rajendra Gurao[2]

[1]Research Scholar, JJTU, Vidyanagari, Jhunjuhunu, Rajasthan, India 333001

[2]Brihan Maharashtra College, Pune, Maharashtra, India

Abstract: ***Data Mining has been defined in a number of ways by researchers. The two major types of analytics in Data Mining are descriptive and predictive. The goal of descriptive analytics is to discover patterns in data. Predictive Analytics, by contrast, goes beyond finding patterns in data and attempts to predict the outcome for unobserved (future) data. This paper discusses the role of predictive analytics in Data Mining. Some algorithms of predictive analytics are described in detail.***

Keywords: Data Mining, Predictive Analytics, Loss Function, SVMs, Descriptive Model

1. Introduction

The literature on Data Mining contains many definitions of Data Mining. However the most common part of all definitions is "Knowledge Discovery in Data" or KDD, as it is popularly known. This aspect of Data Mining is more about descriptive than predictive analytics because it does not include methods of deriving benefits from the discovered patterns in data. Predictive analytics is about developing methods for predicting future outcomes on the basis of past or historical data. For example a customer data base may contain information about customers who have failed to pay their bills. The goal of predictive analytics will then be to predict which other customers may fail to pay their bills in future. In short, predictive analytics has a focus on making predictions.

Finding common patterns in data is often fascinating and can also be very useful, but generally it is predictive analytics that can obtain direct benefit of this discovery. For example suppose that customers of AMUL Dairy Products tend to be urban higher middle class families. This pattern may be interesting, but what should AMUL dairy do with this finding? It is possible to think of two possible future actions due to the same finding. Moreover, these two possible actions are contradictory, but both are reasonable in some sense. One would suggest that AMUL Dairy should direct its marketing towards more urban middle class families. The other would suggest that the urban middle class has already been covered and hence AMUL Dairy should target a less tapped segment of the population in its marketing. This shows the limitation of descriptive analytics. By contrast the predictive analytics is about helping in making decisions that maximize the benefit to the decision-maker. For example, the credit limit of those customers can be reduced who are more likely to avoid paying in the future. The difference between decision and prediction is very important to understand. Data

Mining helps us in making predictions, but predictions are useful only if they allow us to make decisions that lead to better outcomes.

In the context of business or industry, it is sometimes felt that maximizing profit may be at the expense of customers. It should be kept in mind that maximizing profit is generally maximizing efficiency. Also, increased efficiency usually arises from improved accuracy in targeting and benefits the people it targets. A business has no motive, for example to send advertisements to people who will only be annoyed and will not respond.

The other side of the coin is that sometimes Data Mining incurs a profit, but it is small in comparison to the cost of Data Mining. In such cases some additional social benefit might be derived by directing the same effort towards a different objective.

1.1 Limitations of Predictive Analytics

It must be kept in mind that predictive analytics is a supervised learning method and hence requires training and testing data sets. The limitations of predictive analytics must be understood with this background. First, the training data set must have adequate size and quality. Second, the concept to be predicted must be clearly defined and there must be historical examples of the concept. Consider the following extract of an article that appeared in the London Financial Times on May 13, 2009:

"FICO, the company behind the credit score, recently launched a service that pre-qualifies borrowers for modification programs using their in-house scoring data. Lenders pay a small fee for FICO to refer potential candidates for modifications that have already been vetted for inclusion in the program. FICO can also help lenders find borrowers that will best respond to modifications and learn how to get in touch with them."

It is questionable to consider this as an application of Data Miningbecause it is not possible to think that a useful training data set would exist. The target is to find "borrowers that will best respond to modifications" such a borrower is a person who would not pay under the current contract, but would pay under a modified contract. In 2009, there was no long historical experience of offering modifications to borrowers and hence FICO had no relevant data. Also, the target is defined on the basis of what cannot be observed. In other words, FICO claimed to be reading the minds of borrowers. However, Data Mining cannot and does not read minds.

For successfully applying Data Mining the action that would be taken on the basis of prediction must be defined clearly

and must have reliable beneficial consequences. These actions should not have unintended consequences. In the example above, modifications may change behavior in undesired ways. An individual requesting a modification is already thinking of not paying. If modification is offered, this individual may be motivated to request more concessions.

In addition to these considerations, it is also necessary that the training data is representative of test data. It is common to have training data to come from the past, while the test data arise in the future. If the phenomenon to be predicted changes over time, then predictions are not likely to be useful. Here, again, changes in economy are likely to change the behavior of borrowers in future. Finally, it helps if the consequences of actions are independent of the current data. This may not be the case in the example. Rational borrowers who come to know about modifications offered to others may try to appear to be candidates for a modifications. As a result, every modification generates a cost that is not limited to the loss caused by the person getting the modification.

We can think of an even more clear example of a situation where predictive analytics is not likely to work is a model to predict which individuals will commit a terrorist act. This is so because statistically reliable patterns cannot be learned since there are so few positive training examples. Also, intelligent terrorist will take care that they do not fit in with patters exhibited by earlier terrorist. Another issue is that Data Mining can tend to put an ever increasing focus on optimizing existing processes, at the expense of understanding a broader situation. Big data and Data Mining can give a false sense of security.

1.2 Opportunities of Predictive Analytics

Certain criteria need to be developed for judging the potential success of Data Mining applications. For this purpose, a sample application is considered where the objective is to predict the success rate of calls made from a mobile phone. Every call can terminate due to one of the following reasons: normal termination, call dropped by the network, call dropped by the calling phone, call dropped by the receiving phones, and perhaps some more reasons. A sequence of questions is given here in reasonable orders. According to an explanation of every question, there is a discussion of the answers with reference to the sample application.

1.2.1 Does the domain involve many individual cases?
Data Mining and predictive analytics are not about making one-off decisions for a company or organization. In the domain of example, a case is one telephone call, and there are many of these.

1.2.2 Is there a clear objective to be optimized?
There is not definite problem to solve if it is not clear as to what the goal is. The objective is usually from someone's point of view. In case of commercial transactions, for example, the seller and the buyer may have some conflicting objectives. Data Mining is applied to achieve the goals of whoever is doing the Data Mining. In the domain of the example, the telephone company is mining data and hence the goal is to make every call terminate successfully. Even though customers have the same general goal, objectives are not perfectly matching. For example, every customer is interested in his/her call ending successfully while the company may be motivated to prioritize the calls made by its profitable customers.

1.2.3 Is it possible to take actions that can influence the objective?
This is very crucial. There is nothing to do if the action cannot change the outcome. In case of the example, changing the transmission power of the phone or the base station is a possible action a higher level of transmission power will generally increase the chance of a successful call.

1.2.4 Is it possible to predict an unknown target value that is relevant to the objective?
Predictive analytics involves predicting some relevant characteristic of individual cases that cannot be observed at the time when it would be useful to know it. In the example, the target is whether the call will fail.

1.2.5 Is the target value known for many historical cases?
Yes, at the end of every call it is known whether or not the call was successful.

1.2.6 Is it possible to observe, for every individual case, features that are correlated with the target value?
Yes. These features will include the weather, locations of the base station and the phone, the phone model, the relevant power levels, and derived features like the distance between the base station and the phone.

1.2.7 Are the individual cases independent of one another?
In terms of the example, does failure of one call influence the success of another call? There is no apparent reason to think so.

This paper discusses more recent methods that are still not well understood in the statistical community. In particular, support vector machines as a classifier has still not being used as much as logistics regression and linear discriminant function. Random forests are another recent method that can handle non linearity better than support vector machines. Random forests are also easy to understand and implement.

2. Predictive Analytics in General

Supervised learning algorithms have the goal of learning from examples in training data. Once developed, a classifier can be used to make predictions for cases in test data. This type of learning is called "supervised" because known outcomes in training data have guided the classifier to optimality.

Every case in the training and testing data is represented as a vector of fixed length p. Every element in the vector is called a feature value. It may or may not be a real number. The training set has vectors with known labels. It is important to distinguish between a feature and a feature value. The training data set is usually organized in the form of a matrix, where rows are formed by cases and column are formed by features. The outcome is often denoted by y, which is known for every case in training data set and unknown for cases in test data. The classifier produces a

predicted y-value. When y is a real number, supervised learning is called a regression and the classifier is called a "Regression model." The word "classifier" is used when y values are discrete or non-numerical. The simplest case is when there are only two label values. They are then denoted by 0 and 1, -1 and +1 or negative and positive.

The training data consist of a matrix having n rows corresponding to the n cases and p columns corresponding to the p features. It has an additional column of the y values, n is called the size and p is called the dimensionality of training data. The feature values corresponding to case i and features j is denoted by x_{ij} and the label of case i isy_i. Label values are known for all cases in the training data but not for the test data.

2.1 The Problem of Overfitting.

In any particular application, the training data set contains examples with labels, while examples in the test data set have no labels. The purpose is to predict the labels for examples in the test data set. However, in research it is required to measure the performance achieved by a learning algorithm. For this, the test data set contains examples with known labels. The only difference between training dataset and test dataset is that the algorithm is allowed to learn from the training dataset, where labels are available to the learning algorithm. The labels of examples in the test dataset are not made available to the learning algorithm when it predicts them. Once unknown labels are predicted the predicted labels are compared with the already known labels to assess the accuracy of prediction. Here it is important to emphasize that the performance of a classifier must be tested on an independent data set. The learning algorithm looks for patterns in the training dataset, so that unknown labels can be predicted if these patterns occur in the test cases. It can happen that some of the discovered patters are spurious. That is, these patterns may appear in the training data set but do not appear in the test dataset. The algorithm will have high accuracy in the training dataset if it relies on these spurious patterns. However, when labels in the test data set are predicted, the accuracy will be low because the spurious patterns may not appear in test data. The phenomenon of relying on patterns that are strong only in the training dataset is called overfitting. In practice, every learning algorithm faces the risk of overfitting.

It is important to avoid overfitting as much as possible. The method suggested in the literature toward this end was to randomly divide the available dataset into training and test datasets. The recent methodology goes beyond this and divides the entire dataset into three mutually exclusive subsets. These are training dataset, validation set and test set. A set of labeled examples is called a validation set when it is used to measure accuracy of the learning algorithm. The final test set must be used only once, while training and validation sets may be used multiple times in order to achieve better learning.

Division of the available data into training, validation, and test sets must be done randomly in order to guarantee that every set is a random sample from the same distribution. In practice, however, it is possible that test cases may be observed in future and hence may not be a random sample from the same distribution.

3. Support Vector Machines.

This section explains support vector machines (SVMs). The SVMs described here are most common and are called soft margin SVMs. We discuss linear as well as nonlinear kernels. We also discuss the use of regularization for preventing overfitting.

Regression is a well-known predictive model, but it is hardly known as a learning algorithm. In the simplest form, SVMs are used when it is required to predict a binary label. For convenience, it is assumed that the binary label y takes the values +1 and -1.

3.1 Loss Function

Let $\underline{x}$ be an instance, that is, a numerical vector of dimension p, let y be its true label, and let f(x) be the prediction function. It is assumed that the prediction function is real-valued. In order to convert it to a binary prediction, a threshold value of zero is used to define.

$$\hat{y} = 2.\mathrm{I}[f(x) \geq 0] - 1$$

Where I(.) is an indicator function taking values +1 if its argument is true and 0 if its argument is false.

The loss function $L(.,.)$ measures how good the prediction is. The loss functions usually do not depend on $\underline{x}$ and a loss of zero corresponds, to a prefect prediction. Otherwise, the loss is positive. The most obvious loss function is $L(f(x), y) = I(\hat{y} \neq y)$ and it is called the zero-one loss function. It is common to describe accuracy in terms of this function. However, it has some undesirable properties. First and foremost, it loses information in the sense that it cannot distinguish between predictions that are nearly right and predictions that are extremely wrong. Second, regarding mathematical properties, its derivative with respect to the value of f(x) is either zero or is undefined. This makes it difficult to use the 0/1 loss function in training algorithms that try to minimize the loss by modifying the parameter values using derivatives.

A more preferable loss function is the squared error loss defined as

$$L(f(x), y) = [f(x) - y]^2.$$

This function is differentiable everywhere, and does not incur any loss of information while making real valued prediction. However, when the true label is 1, this loss function says that the prediction $f(x) = 1.5$ is as bad as the prediction$f(x) = 0.5$. Intuitively, it is natural to think that a predicted value greater than 1 should not be considered incorrect if the true label is 1.

Another loss function, known as hinge loss function, fits the intuitive reasoning stated above. This function is defined as follows.

$$L(f(x), -1) = \max\{0, 1 + f(x)\}$$
$$L(f(x), +1) = \max\{0, 1 - f(x)\}$$

To understand this loss function, consider a situation where the true label is +1. There is no error as long asf(x) ≥ 1. Similarly when the true label is -1, there is no error as long as f(x) ≤ −1.By contrast, when the true label is +1, the error increases monotonically as the function goes away from +1 in the wrong direction. This loss function can be defined mathematically as follows.

$$L(f(x), y) = \max\{0, 1 - yf(x)\}.$$

The hinge loss function is the first insight behind SVMs. In order to minimize the hinge loss function, the training aims to achieve predictions $f(x) \geq 1$ for all training instances having true label $y = +1$, and to achieve predictions $f(x) \leq -1$ for all training instances having $y = -1$. Let us now consider the range of the function f, because nothing has been said about it so far. Training aims to classify instances correctly, but it does not restrict the predicted values to be +1 or -1. In this sense, the training process does not impose any additional restrictions on the classifier. As a result, the training process attempts to distinguish between the two classes that are inside the interval (-1,1).

3.2 Regularization

When the training dataset contains n training examples$\{(\underline{x}_i, y_i), i = 1,2 \ldots, n\}$, where y_i is the actual label of the example $\underline{x}_i$, the total training loss is given by the sum of individual losses.
$\sum_{i=1}^{n} L(f(\underline{x}_i), y_i)$.

This sum is also called the empirical loss because it is computed from the available data. Suppose F is the space of possible functionsf(.), and f is obtained by minimizing the empirical loss.
$f = \underset{f \in F}{\operatorname{argmin}} \sum_{i=1}^{n} L(f(\underline{x}_i), y_i)$.

We run the risk of overfitting if F is too flexible or the training dataset is too small. On the other hand, if F is too restricted, we run the risk of under fitting. Since the best space F is not known in advance, pertaining to a particular training data set, it is suggested to allow F to be flexible while imposing a penalty on the complexity of f. Suppose c(f) is a real-valued measure of complexity. The learning process then attempts to solve.
$f = \underset{f \in F}{\operatorname{argmin}} \lambda c(f) + \sum_{i=1}^{n} L(f(\underline{x}_i), y_i)$.

The parameter λ controls the relative strength of the two objectives, namely minimizing the complexity of and minimizing training error.

The space of candidate functions can be defined by a vector $\underline{w} \in \mathcal{R}^p$ of parameters, so that we can write

$$f(\underline{x}) = g(\underline{x}, \underline{w})$$

Where g is some fixed function. The complexity of every candidate function can then defined to be the norm of the corresponding vector $\underline{w}$. The square of the L_2 norm is most commonly used, and is given by

$$c(f) = \|\underline{w}\|_2^2 = \sum_{j=1}^{p} w_j^2.$$

It is however, possible to use other norms like the L_0 norm. Defined by

$$c(f) = \|\underline{w}\|_0 = \sum_{j=1}^{p} I(w_j \neq 0),$$

Or the L_1, norm, defined by

$$c(f) = \|\underline{w}\|_1 = \sum_{j=1}^{p} |w_j|.$$

The L_0 norm directly identifies the vectors $\underline{w}$ that are sparse, but is NP- hard to optimize. The L_1 norm can be optimized with gradient method and tends to obtain sparse vectors $\underline{w}$. Nevertheless, the squared L_2 norm is mathematically most convenient and works reasonably well.

3.3 Linear Soft-Margin SVMs

A linear classifies is specified by$f(\underline{x}) = g(\underline{x}, \underline{w}) = \underline{x}.\underline{w}$, the dot product function. The objective is then to find.

$$\underline{w} = \underset{\underline{w} \in \mathcal{R}^p}{\operatorname{argmin}} \lambda\|\underline{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(\underline{w}.\underline{x}_i)\}.$$

where the average in the second term makes λ freefrom the data size n. The solution to this problem is called a liner soft-margin SVM classifier. Uniqueness of the solution can be proved by virtue of convexity of the objective function. The optimization problem can alternatively be written as

$$\underline{w} = \underset{\underline{w} \in \mathcal{R}^p}{\operatorname{argmin}} \|\underline{w}\|^2 + C\frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(\underline{w}.\underline{x}_i)\}.$$

where $C = \frac{1}{n\lambda}$ .Sometimes it is left to the user to specify C instead of λ. The smaller the value of C, the stronger is the regularization. For a given dimensionality p, a smaller training data set should require a smaller value of C. Even then, there are no guidelines regarding the best value of C for a given dataset. It is necessary to try several values of C to find the best value experimentally.

In mathematical terminology, the above optimization problem is primal formulation. It is an easy description of SVMs and most of the fast algorithms for training liner SVMs are based on it.

3.4 Dual Formulation

The dual of the primal formulation of the previous section is as follows.

$$\max_{\underline{\alpha} \in \mathcal{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\underline{x}_i, \underline{x}_j)$$

subject to constraints
$0 \leq \alpha_i \leq C, i = 1, \ldots, n.$

The primal and dual are different problems but have a common unique solution. The solution of the dual problem gives a coefficient α_i for every training example. Notice that the optimization is over$\mathcal{R}^n$, whereas the primal had optimization over $\mathcal{R}^p$. The trained classification is
$f(\underline{x}) = \underline{w}.\underline{x}$where
$\underline{w} = \sum_{j=1}^{n} \alpha_i y_i \underline{x}_i$This equation shows that $\underline{w}$ is a weighted linear combination of the training examples $\underline{x}_i$and the weights are between 0 and C, while the labely_i is the sign of every example.

The solution of the dual problem contains many $\alpha_i's$ having the value zero. The training examples $\underline{x}_i$ corresponding to

$\alpha_i > 0$ are called support vectors. These are the only examples that contribute to the classifier.

4. Conclusions

Data Mining has two major goals as sate below

- To generate descriptive models to solve problems.
- To generate predictive model to solve problems.

Descriptive models have already been developed in statistical methodology. It is predictive model building that caught attention because this is the activity that involves both supervised and unsupervised learning. Supervised learning is a generalization of many statistical methods. However, classical statistical methods do not have a separate test data set. The concept of dividing the available data into training, validation and test sets came about only due to machines learning and Data Mining. As consequences, predictive model is are being used more and more in marketing, insurance, banking, manufacturing, supply chain management, customer relation management, and so on. This paper has avoided a discussion. On linear and logistic regressions because they are classical statistical models. SVMs are discussed in the detail because they are the gift of Data Mining to statistics.

References

[1] Abe, N., Pednault, E.,Wang, H., Zadrozny, B., Fan,W., and Apte, C. (2002). Empirical comparison of various reinforcement learning strategies for sequential targeted marketing. In Proceedings of the IEEE International Conference on Data Mining, pages 3–10. IEEE.

[2] Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. Journal of Machine Learning Research, 6(1):503–556.

[3] Huang, J., Smola, A., Gretton, A., Borgwardt, K. M., and Sch¨olkopf, B. (2006). Correcting sample selection bias by unlabeled data. In Proceedings of the Neural Information Processing Systems Conference (NIPS 2006).

[4] Jonas, J. and Harper, J. (2006). Effective counterterrorism and the limited role of predictive data mining. Technical report, Cato Institute.

[5] Michie, D., Spiegelhalter, D. J., and Taylor, C. C. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood.

[6] Murphy, S. A. (2005). A generalization error for Q-learning. Journal of Machine Learning Research, 6:1073–1097.

[7] Neumann, G. (2008). Batch-mode reinforcement learning for continuous state spaces: A survey. O¨GAI Journal, 27(1):15–23.

[8] Powell, W. B. (2007). Approximate Dynamic Programming. John Wiley & Sons, Inc.

[9] Riedmiller, M. (2005). Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In Proceedings of the 16th European Conference on Machine Learning (ECML), pages 317–328.

[10] Simester, D. I., Sun, P., and Tsitsiklis, J. N. (2006). Dynamic catalog mailing policies. Management Science, 52(5):683–696.

[11] Tang, L. and Liu, H. (2009a). Relational learning via latent social dimensions. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 817–826. ACM.

[12] Tang, L. and Liu, H. (2009b). Scalable learning of collective behavior based on sparse social dimensions. In Proceeding of the 18th ACM Conference on Information and Knowledge Management (CIKM), pages 1107–1116. ACM.

[13] Vert, J.-P. and Jacob, L. (2008). Machine learning for in silico virtual screening and chemical genomics: New strategies. Combinatorial Chemistry & High Throughput Screening, 11(8):677–685(9).

[14] Viviani, P. and Flash, T. (1995). Minimum-jerk, two thirds power law, and isochrony: converging approaches to movement planning. Journal of Experimental Psychology, 21:32–53.

[15] Yu, V. (2007). Approximate dynamic programming for blood inventory management. Honors thesis, Princeton University.

[16] Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1):1–130