Machine Learning

Module Code:B9DA104

Submitted By:

Prasad Tambe

(1051513)

Table of Contents

Q1 Define the following.

1.Define AI, Machine learning and Deep learning.?

Artificial Intelligence is characterized as the technique which makes a machine to learn and understand to solve issues or make judgements. [Negnevitsky, M., 2005.]

Machine learning is the use of artificial intelligence (AI) which enables systems to learn and improve their experience automatically without being specifically programmed. (Expert System, 2017)

Deep learning is an AI method that recreates the operation in the processing of data for decision-making activities of the human brain. (Investopedia, 2020)

2.Define parametric and Nonparametric models.?

A parametric model is a term for characterizing a variable with all its data used in statistics. In simple terms,
Parametric model represents a model used in statistics to represent all the details in its parameters.

Nonparametric models are mathematical models that often do not agree with no rmal distribution since they depend rather than on discrete value on continuous data. (DeepAI, 2019)

3.Define Supervised learning and its components with examples?

The well-classified and labelled study data is known as supervised learning.
It is just like the kid learns to understand fruit or colour under the supervision of a teacher.

1)Classification

Example: if it is a lion or tiger? The separation of data into classes is done by the Classification algorithm.

2)Regression

Example: It is used to predict the cost of a house positioned by parameters such as the size of the house etc. (Softwaretestinghelp.com, 2019)

4.Define Unsupervised learning and its components with examples?

Unsupervised learning is which is unlabelled sample data and a process which is independently learned. As the bird flies by itself which doesn't need supervision.

1)Clustering

Cluster analysis is the method of identifying and aggregating the correlations be tween data objects as the same type, height, colour, weight, etc.

2)Association

It is a type of mining which discovers the most common articles or connections between elements. (Softwaretestinghelp.com, 2019)

5.What are the common types of error in Machine learning?

The common types of error in Machine learning are as follows few listed below

1) Data Collection Error
   When the data is incorrectly collected. For example, for insurance the age, gender, can cause an error at different levels.
2) Data Storage error
   Information can be stored in error because some records can be saved wrongly or a portion of the information may be lost in storage.
3) Bias error
   This occurs when the model is being conditioned with too few functions. The pattern is too basic or unsuitable in this situation. (Obi, 2020)

Q2 Please solve the case study.

Scenario: You are the head of the Corporate clients division of a major Insurance provider, and you want corporate insurance applications decided faster and better.

1) Provide a brief description of your company, industry, or business?

I am CTO of an organization of a leading insurance provider that helps the corporate insurance if the insurance is lapsed and provides them predicting which insurance would be best suited and what claims would an insurer would get depending on the factors such as Designation, Salary. I receive clients from all sectors whether it is related to IT or construction. The applications would also anticipate an insurance generator that recommends the best-suited insurance i.e. which Discovers where there can be too much or too little protection from an individual insured and then help them get their current situation proactively covered in their insurance.

2.What business outcome are you supporting with your machine learning project? How is this outcome relevant and important for the company? How will you evaluate whether the desired outcome is being achieved?

One of the challenges faced by our organization was targeting new customers because in the last few quarters new customer's insurance rate is decreased by 60 percent. In the last few months due to an increase in the number of clients, our current process is efficient but very slow because of this reason we are losing many clients. So, we have come up with a solution in such a way that we wanted to create a model that would extract companies' key points along with rivalry company's positive points and pitch in a way such that the new companies would be convinced to take our insurance and would retain for a longer time.

The insurance industry is considered one of the most dynamic and volatile areas of business. It has to do with threats and risks immediately. Consequently, it was always statistically based.

Insurance undertakings have a wider range of sources of knowledge to determine the applicable risk. To predict, track and evaluate threats and claims. Machine learning techniques are implemented to establish productive consumer acquisition and retention approaches. Within the domains of their great interest, the insurance companies certainly benefit from Machine Learning use.

3.What machine learning project will you propose to support this business outcome? At a high level, what will your machine learning model be doing? Make the case that this is a viable project (at least in theory) and relates to your overall business goals.?

The factor which affects the insurer would be the insurance lapsed which would be a crucial decision for an insurer to decide which insurance would be best for him/her. The model would predict which factors such as Designation, Salary would be suitable for the customer concerning the claims would be claimed. There are few cases where the customer is expecting to less and getting to more and the price hike would result in loss of the client. Therefore, Clients are classified into algorithms that will be used in our Machine learning prediction by their maturity, age, place, etc. We would be using All consumers are therefore grouped by chance into their own identities, desires and preferences and personal information. This distinction allows us to establish habits and strategies that are especially relevant for each person. This can lead to the

development of sales goals and the implementation of individual segments of personal services.

The following measurement parameters were used: maturity, age, place, Insurance areas to test the outputs of different algorithms for machine learning (Naive Bays, Multi-Layer Perceptron, Random Forest, Logistic model forest). In contrast with the Naive Bayes and the Random Tree Algorithms, Logistic Model Tree (LMT), Random Forest algorithms have provided better performance. Concerning the accuracy, the insurance would be decided faster as well as the better for the insurer for the current situation as well as in the future.

The whole process has been dived into 4 stages.

1) Data Gathering
   We have provided the digital forms to every cop rate organisation we ask them to fill out the digital forms.
   We also request them for background check data for every employee.
2) Data Filtering
   Here it is important that the data is completely cleaned and usable for stage 3 and stage 4.
   We extract only the important and usable attributes for stage 2 and stage 3.
   We create subsets for the whole data set according to their designation and salary.
3) Finding the best fit claim
   We would be finding the best fit claims for a particular subset.i.e. designation and salary using Naive Bays, Multi-Layer Perceptron, Random Forest, Logistic model forest.

4) Predicting future claims
   We would be finding the range of expenses for our organization that is when would the next time range pattern would arrive for every subset and would merge that data to create future claims overdue.

4.Given the state of readiness you have described and the scope of the project you're proposing, is this a risky project, broadly speaking? That is, is it appropriate to the stage your company is at or will it provide particular challenges?

Yes, the same question of human behaviour and the probability of gains and losses is discussed by insurance firms as it's a risky project and can have a financial loss as well as risk. The aim is to determine based on certain assumptions like Designation,Salary,zip code, marital status, age, research, occupation, hobbies, payment history, and a lot of other creative people to predict when a person will be dead, shot, promoted, injured, burnt, have babies, terminated the contract, etc. The consistency of the models determines the company's success. If an insurance company takes too little care, it will lose business by having high premiums and coverage. If you're too reckless, you get many clients, but you have to spend as much as you can. The best competitor has the best model. Machine learning have many ways to refine a similar model for a particular organization based on historical data. It recognizes these trends as "the most likely changes in policy after two years are individuals from that area," and retains policies. So, to gain many clients there can be a risk or a financial loss if there is less accuracy and we mind loss up losing many clients.

Q3 Please write the summary of the journal paper you read and explain about the machine learning techniques you learn from the journal and what are the advantage and disadvantage of the application discussed in the research project.

3.1 Summary:

The article which I am going to summarize is about Cyber Physical System which consists of structures composed of physical (hardware), software and possibly other kinds of systems (for example, human) systems. Both the physical as well as psychological trends are merged together to give a local mindset behaviour of a human emotion. Thus, CPS is often assisted by hardware that communicates with the actual world and with complex software components, such as sensors, actuators and related embedded systems. The article provides as an overview of how the external stimulus can affect the human mood/emotion. This research explores the effect of sensations and feelings in decision-making around organizational management activities in a data centre. The research was made is two ways which is a general and a specific one. A set of questionnaires was examined how the human reacts on such situations.

The general way was an entire data was refined to predict the behaviour of human with 85 percent of accuracy. The specific way was to examine the interaction of the external stimulus and the final decision of the subject. The random forest was used in this scenario to extract the behaviour. I have experienced how specific interactions within the psychological component and how feelin gs can be diverted through an external stimulation cause. Weka model is used in this research article for gathering the results obtained using Weka models. (Ncl.ac.uk, 2020)

3.2 Machine Learning Technique:

The Random forest Algorithm was used for this article which was sub grouped as general random forest algorithm and specific random forest algorithm. If we want a high performance with very less clarification Random forest is being the best model in such case.

1) **General Random forest algorithm**

The first model has been able to predict the possibility of acting / reporting to the command line, with the whole data flow. We perform two data evaluation procedures, the first one to examine the relationship between the variables and the rejection which is higher than 95 percent. The second one is to study the p-value which excludes the factors which has no predictive value p=0.05.

2) **Specific random forest algorithm.**

In comparison to a general model which can simulate the behaviour of a subject in a data centre, the role of triggers and emotions in decision-making was very important. A second model was therefore produced. For this reason. Stimulus, and the responses after stimulus were the chosen variables for the study. The main objective is to evaluate the likelihood of behaviour by intention and stimulus.

3.3 Random Forest:

Random forest is an algorithm of analysis. The "forest" it creates is a series of decision-making trees normally equipped with the process of "bagging." A mixture of research models increases the overall outcome, as a general idea of the bagging methodology. In simpler words random forest frames multiple decision trees and combines them together to give an accurate prediction. Random forests are bagged models of the decision tree which divide on each subdivision. A decision tree will be tested, bagged decision-making trees addressed and a random subset of the functionality separated. From this article we acknowledged how the external stimulus can be more

biased to the actual predictions so the random forest algorithm groups the positive stimulus and the negative stimulus to obtain an accurate prediction using the bagging methodology. Instead of search the nearest answer to the questionnaire it searches for the best match answer for the questionnaire. (Built In, 2019)

3.4 Advantages of Random Forest

1. Random forests can overcome classification and regression problems and make a fair prediction on both sides.
2.The ability of managing large data sets with greater dimensionality is one of the advantages of the Random Forest. It can manage thousands of input and identity variables most importantly so it is considered as one of the methods of reducing dimensionality. In addition, the model generates variable value, which can be very useful.
3. It has an effective method to measure loss of data and preserves consistency if much of the data is lacking.
4. It has approaches in data sets where groups are unbalanced for consistency errors as positive stimulus is balanced by 50 and the negative stimulus is balances by 50 in the total 100 sample data.

3.5 Disadvantages of Random Forest:

1. It is certainly good at classifying but not for a regression problem, since it does not provide an accurate prediction of a continuous nature. In the case of a regression, the training data cannot estimate beyond the range which is the questionnaire before stimulus and may be particularly disturbing over suitable data sets and provide in accurate decision after stimulus.

2. If the interaction between dependent (before stimulus) which is the mood and independent (after stimulus) which is the emotion variables is not continuous, it might not fit well.

Reference

1] Built In. (2019). *A complete guide to the random forest algorithm*. [online] Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 10 Mar. 2020].

2] Ncl.ac.uk. (2020). *Cyber-Physical Lab - Newcastle University*. [online] Available at: https://research.ncl.ac.uk/cplab/aboutthelab/whatarecyber-physicalsystems/ [Accessed 10 Mar. 2020].

3] Obi, B. (2020). *Sources of Error in Machine Learning*. [online] Medium. Available at: https://medium.com/towards-artificial-intelligence/sources-of-error-in-machine-learning-33271143b1ab [Accessed 10 Mar. 2020].

4] Softwaretestinghelp.com. (2019). *Types Of Machine Learning: Supervised Vs Unsupervised Learning*. [online] Available at: https://www.softwaretestinghelp.com/types-of-machine-learning-supervised-unsupervised/ [Accessed 10 Mar. 2020].

5] DeepAI. (2019). *Parametric Model*. [online] Available at: https://deepai.org/machine-learning-glossary-and-terms/parametric-model [Accessed 10 Mar. 2020].

6] Investopedia. (2020). *How Deep Learning Can Help Prevent Financial Fraud*. [online] Available at: https://www.investopedia.com/terms/d/deep-learning.asp [Accessed 10 Mar. 2020].

7] Expert System. (2017). *What is Machine Learning? A definition - Expert System*. [online] Available at: https://expertsystem.com/machine-learning-definition/ [Accessed 10 Mar. 2020].

8] Negnevitsky, M., 2005. *Artificial intelligence: a guide to intelligent systems*. Pearson education.