DBS
Dublin
Business
School

# Electronic Assignment Cover sheet

Student (s) Number as per your student card:
ASHWIN RAMDAS
10534162

Course Title: MSc DATA ANALYTICS

Lecturer Name: Abhishek Kaushik

Module/Subject Title: B9DA104 MACHINE LEARNING

Assignment Title: CA1

No of Words: 1804

**Q1)**

A= 1)

AI

AI is also known as machine intelligence, because it is the intelligence demonstrated by a machine, as opposed to Natural Intelligence shown by humans. Any device is said to be Artificially Intelligent, when it can make reasonable predictions based on limited parameters.

Machine Learning

Machine Learning is the part of AI which deals with algos and statistics used to make predictions without specific instructions instead relying on pattens and observations from the data.

Deep Learning

Deep Learning is a broader method in the machine learning paradigm which is based on artificial neural networks and feature engineering.

2)

Parametric models are a particular class of statistical models in statistics. It is also known as parametric family or finite-dimensional models. It, basically means models with finite parameters in probability distribution

Whereas, non-parametric models are another class of statistical models in statistics. They are also known as distribution free models. These models are either distribution free or have specific distribution where the distribution parameters are unspecified.

3)

On the basis of how they make their prediction, there various types of machine learning algorithm, one of which is Supervised learning. The scientist trying to work on the data "supervises" the algorithm, deciding what conclusions and predictions, it should make.

In supervised learning, the data contains both input and output variables and the training data is used to find a link between them, while the test data is used to test our evaluations and make test predictions to test the accuracy of the model.

Types of supervised learning algorithms are Support Vector Machine(SVM), Linear Regression, Logistic Regression, Random Forest, Classification Trees etc.

An example of this would be- if you find out if a person wants to buy a car, so the outcome in that will be either yes or no. Therefore, on the basis of the previous car owner data, in supervised learning, we try to predict if a person will buy a car or not.

4)

Unsupervised Learning is a type of machine learning algorithm that make their predictions without the “supervision” of the scientist performing the algorithm, deciding what conclusions to get and predictions to make itself.

In Unsupervised Learning, the data given to the system contains just input variables and the algorithms are left on their own to find and show any interesting hidden pattern in the data.

Types of Unsupervised Learning algorithms are Clustering Algorithm, K-means, Hierarchical clustering, Anomaly Detection etc.

Unsupervised learning is used to classify mails as Spam or not Spam, specifically in clustering algorithm where we form two different clusters, based on previous data and try to determine which cluster the new mail will go in.

5)

There are many types of statistical errors which are related to machine learning and predictive analysis. These are mostly of the types in-sample errors and out-sample errors. In-sample uses the training data to find error rate while generalisation errors (or out-sample errors) are found on the new dataset

Some of the most common types of error in machine learning are-:

1. Root Mean Squared Error- It is the measure of how spread out the data is. It is found out by taking the standard deviation of the distances of various data points from the regression line.
2. Mean Squared error- It is the average of the square of the distances of each data point from its predicted value. Like RMSE, it is also used to find the spread of the data.
3. Mean Absolute Error- It is the summation of absolute vertical distance between the data point and its predicted value. It is calculated to know how far the predicted values are from the actual.
4. Median Error- it is the measure of how does a univariate sample of quantitative data varies. For this, the median of the absolute deviation from the median of the data.
5. Confusion Matrix- It is a table layout from which visualization of the performance of an algorithm, typically a supervised learning one, can be done. It is also known as an error matrix,
6. Precision-It is used in pattern recognition. It refers to the fraction of relevant Instances from the retrieved instances
7. Recall- It is also used in pattern recognition and refers to the fraction of all the relevant instances that were retrieved.

**Q2)**

A= 1)

Our company “Insura” is a leader in providing insurance cover to various corporate as well as independent clients for their businesses in almost every part of the developed world. Our insurance policy is based on checking the past financial track record of the client, and current state of business and providing insurance if our analyst deem if the product or company is worth insuring and are they likely to pay on time.

2)

The business outcome that we are trying to implement and support with our machine learning project is reduced costs, reduced time on a single client and improvement in overall efficiency of the company “Insura”.

This outcome is relevant and important for our company as if we can make a machine learning model which can predict if the client is worth insuring and is likely to pay on time, based on the past financial record, we can have reduce cost as our analysts don't have to check it manually and it also leads to reduced cost on a single client for the same reason. Moreover, if it works correctly in assisting our analysts in making the correct decision, then it also leads to improvement in overall efficiency.

To check if the desired outcome is being achieved, we will first beta test this model, with only some of our selected senior analysts, for a year, before rolling out fully for all the analysts.

3)

To achieve this business outcome, we will implement a machine learning project which will take the past financial history of the company to decide if the client is likely to pay on time and if they are worth insuring i.e. if the current track record doesn't suggest bankruptcy.

The Machine learning model will check if the current financial status and trend is above a certain threshold, depending on whether the company is a small, medium or a large-scale business. Then, according to the past data of various clients of our company and the past financial record of the company, the model shall decide in yes or no, if the client is viable to be invested in. In general, the algorithm used by it will logistic regression, a supervised learning algorithm, which according various past data will classify the viability of an investment in the company as yes or no

4)

Although our company is among the best in the world and have state-of-art equipment, the scope of this project is such that even in the beta phase, it needs to be carefully handled by the best data scientists, or it could lead to hefty losses. However, we employ a team of the best data scientists in the world, who are equally up for the job. So, the risk involved in the project would be medium, broadly speaking.

It is appropriate for the stage the company is at as we are looking to perform better and we believe that such a project is a step in the right direction for our company and whatever challenges we may encounter, we are well equipped to deal with them.

**Q3)**

A= Studies have shown us how we can determine someone's creativity based on the making of their semantic memory. But it is only one of the many aspects of the whole lexical memory. In this paper, we make a neural network of symbols and phonetics and determine how it is correlated to any man's creativity and influence if he would be a creative person or not. this repository of words is used by our network to parse words and is known as the viable cluster. In this study, we turn our focus on constructing animal names by the low as well as the highly creative individuals which we use in a semantic fluency exercise. When we go through the entire viable cluster, the outcome of navigation is what the exercise is modelled on, in the lexical network. Low and high creative people are very different in their approach to the viable cluster in this task. The higher creative individuals go through this cluster very rarely and usually are with low uncertainty. in comparison with low creative individuals, they get to more peripheral words and they go through larger multiplex network distances in midst of concepts. A machine learning classifier of creativity levels is constructed based on these differences. This has an accuracy of 65.0 _0.9% and the area under the curve is 68.0 _0.8%. The two groups of low and high creativity individuals had highly dissimilar features. A cross-validation was performed assigning each list randomly, to training or validation uniformly. In this way, 1000 random splits were made. A logistic regression classifier was trained for every split. As it provided to have a higher classification precision, we used logistic regression while not using other algorithms like random forests, support vector machines etc. A "low" and "high" label was given based on the predictions from the cv set. The potential use of combining psychological measures with neural network models for modelling our mental navigation and, also, we were able to classify people as creative or not as we can see from the outcome of the study. There are a few limitations too though.

Firstly, creativity was assessed based on a questionnaire known as CAQ, used to assess creativity as subjective accolades through many varying fields. Further studies might be able to achieve the unthinkable of even telling us the different tasks we are creative at. Also, creativity in a unique quality of a human being, while we try to judge creativity based on the people who are creative and people who are not, but mental navigation and creativeness can vary person-to-person. Such approaches are being developed for every individual as we mull over it here. Therefore, future studies may have a similar study on an individual level. Even for noisy datasets, deep learning can be used instead of logistic regression for pattern detection. While the regressor can catch small dissimilarities, higher accuracy and AUC may be possible in further research. The network used in this study contains limited representations of the similar words present in the brain. If we include more similar words using attribute sharing in the form of layers, may increase the precision of the model made in this study. However, That may require us to make larger datasets.