**Electronic Assignment Cover sheet**

Please fill out and attach as the first page of Assignment.

**Student (s) Number as per your student card:** Juliana Salvadori - No: 10521647

**Course Title:** MSc Data Analytics

**Lecturer Name:** Abhishek Kaushik

**Module/Subject Title:** Machine Learning (B9DA104)

**Assignment Title:** CA1

**No of Words:** 2500

**Q1**

● Define AI, Machine learning and Deep learning (70-100 words).

AI tries to reproduce human intelligence in computer systems (i.e. make decisions and recover/adapt itself).

Machine learning is the set of techniques used to extract/learn useful information from a set of data. For instance, based on a set of data about people who have or not diabetes, the ML process will try to predict if a person has or can have diabetes.

Deep learning is the part of ML that deals with huge and non-linear amounts of data to extract information. It tries to replicate the functioning of human neurons. A common use of DL is for images.

● Define Parametric and Nonparametric models (70-100 words).

Parametric models make very strong assumptions as they have a fix set of variables and equations, and they also can have a bias effect. They do not need a huge amount of data for training, but they can nevertheless lead to a poor fit problem.

Non-parametric models do not make strong assumptions as they are free to form any function form but require more data for training. In this model can occur the overfitting problem, when it learns too much information about the model, but that does not fit well on the validation dataset.

● Define Supervised learning and its components with examples (70-100 words).

Supervised learning is a predictive model, it will try to predict the output (label) from the input data (features). In this type of learning, we train the model using the input data and the output data, that is, for each combination of the features, we say what is the desired label.

Supervised learning uses classification methods when the target variable is categorical, i.e., when trying to predict if a person has or not diabetes (yes or no), and regression methods when the target variable is numerical, i.e., when trying to predict the median sale price for a house based.

● Define Unsupervised learning and its components with examples (70-100 words).

Unsupervised learning is an explanatory model, that is, it will try to find patterns from the input data (features). In this type of learning, we do not know the output, therefore, the model receives only the features and returns which features follow the same pattern.

Unsupervised learning uses clustering methods, which divides the data into meaningful groups based on similarity, and association methods, which tries to discover meaningful relationships (association rules) between the items in the dataset, i.e., for market basket analysis, when it tries to discover which other(s) item(s) is (are) bought along with the item selected.

● What are the common types of error in Machine learning (70-100 words)?

Bias error – model does not learn enough from the dataset and then makes its own assumptions. It happens to small datasets and can lead to underfitting.

Underfitting – model was not able to correctly train because of a lack of data or the constrain of functions.

Variance error – model learns too much from the dataset and gets confused due to the small variance in the training set. It happens to large datasets and can lead to overfitting.

Overfitting – model learns too much about the training dataset, and it is very bad on the validation dataset.

Irreducible error – errors that cannot be avoid due to the structure of the dataset.

**Q2 (1000 words)**

● You are a software designer, and you want to use ML to improve the user experience.

Answer the following questions

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

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?

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.

4) Given the state of readiness you have described and the scope of the project you are 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?

ABC is a fashion retail company that was founded in 2000 and has 10 large stores across the country since then. The company offers a range of products from baby and children’s clothing, menswear, womenswear to footwear, and accessories, and it is known for the diversity and quality of the products offered and the low prices. However, just in the last 3 years, it started with the option of online sales at its new brand website, which had an impressive impact on the growth in the number of sales.

The website was developed within the company by a small specialised IT team that is currently working on improving, even more, the user experience while shopping online, as the future trend is that more and more people will be using that purchase method every day.

Machine learning can be used to help companies to improve and increase business efficiency in several areas. However, this project will primarily focus on enhancing the customer experience while shopping at the company’s website, and thereby expecting a growth in the number and amount of sales and an improvement in brand loyalty.

Therefore, this project will support the company’s aim of expanding its sales in the country, that is, trying to reach those cities where there is no physical store, and also make available the delivery option for Northern Ireland and England.

A more user-friendly and interactive website can double the chances of a user finalizing the purchase. Also, providing suggestions for similar products and products that usually are bought together with the selected item, can increase the likelihood that the customer that would buy one item will add more items to his basket.

In this way, it will be possible to evaluate whether the main desired outcome for this project is being achieved by monitoring the number of sales, the total amount of sales, and by analysing the representativeness of current and new customers in increasing sales. As one of the company’s objectives is the improvement in brand loyalty, we should be able to see an increase in the number of sales made by shoppers who were already signed up on the website and have made other purchases previously.

The main proposal is the development of an unsupervised machine learning model for market basket analysis. The idea is to make suggestion to the user every time that he selects an item, allowing him to compare similar items and suggest other items that are usually purchased together with that selected item.

There will be two main points to be developed, first, all the products need to be clustered, that is, they need to be divided into groups by similarity, for instance, one cluster could group all the running shoes for men and there may be a subgroup dividing those running shoes between the most popular and the least popular. Thus, when the customer selects one pair of running shoes for men, the website should be able to suggest other running shoes for men with similar features, showing first the most popular ones. Second, a set of association rules the model should build to link items that can often be found within the same purchase transaction, for instance, when the customer adds to his basked a pair of running shoes for men, he also might be interested in buying some appropriated socks for running, assuming that those types of socks are often bought together with those types of running shoes, the website should suggest the socks for the customer.

This primary project will focus on the first point, that is, provide a machine learning model for product clustering and, therefore, allows the website to make suggestions of similar items. Hence, the team will make use of the information available from the current company’s product catalogue to train the model based on the most relevant features, and for that, it will be necessary a prior work to lift which are the most and least significant features to be considered in this process using the domain knowledge of the internal team as the products may have several characteristics to describe them, such as title, overall description, brand, colour, size and other specifications, and not all of them will be suitable or will have the same importance when grouping the products. For example, colour would not help on this process, actually, we do not want to group products by colour, i.e., when the customer selects black running shoes for men, the website should be able to suggest other running shoes for men with similar characteristics, such as brand, material and other anatomic features, however, the shopper might be interested in check out other colours. On the other hand, characteristics like gender and some type of category that describes the shoes as appropriate for walk or running would have more relevance for clustering.

Additionally, after choosing the most appropriate algorithm to be used on this type of process, it might be necessary to run a couple of tests mixing characteristics with different relevance to identify which group of features will provide the best accuracy.

The company has been working with online sales for the last 3 years and the current website is working properly, however, there is room for improvement that will allow the business to grow. At the same time, there is a need to assess that the company’s product database has enough information to allow the model to be built, trained and validated, or it will be necessary a prior work to make that catalogue more robust and informative to allow the model to be properly developed, and thereby aggregating value to the business. Also, other costs of the project have to be measured, i.e., the need for new hardware, new and specific software, staff training and capacity and so on, to confirm that it is achievable by the company budget.

**Q3 (1000 words)**

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.

Title

Twitter Sentiment Analysis Using Machine Learning for Product Evaluation

Summary

This recent paper describes a research project which aims to inspect product reviews (in form of tweets written in English) collected from the online Twitter dataset and use sentiment analysis to classify them as positive, negative, or neutral by comparing different machine learning classification algorithms.

The paper discuss how Twitter, as a huge free online repository of people’s opinions, and sentiment analysis, as the primary method to analyse those opinions, can be used together as an important tool to source and analyse that information for the evaluation of the market for a certain product, that is, help companies to understand what people are thinking about their products. However, it also mentions that in order to obtain more structured data that can be properly used with the classifiers’ methods, an extra effort is required to pre-process Twitter data as it can contain unstructured information as emojis, hashtags, and others.

The proposed system discussed in the research consists in the following steps: data collection (extracts the relevant information from Twitter related to a certain product), pre-processing the data extracted (several tasks are carried out to make the data standardized), feature extraction, classification (different classifier algorithms are applied to get), and then the results from this last step (classification) go to an evaluation stage where they will be assessed to identify which has the best accuracy.

[…]

Reference

Yadav, N., Kudale, O., Gupta, S., Rao, A., & Shitole, A. (2020). Twitter Sentiment Analysis Using Machine Learning for Product Evaluation. *2020 International Conference on Inventive Computation Technologies (ICICT)*, 181-185.