CE802– Machine Learning and Data Mining

Assignment: Design and Application of a Machine Learning System for a Practical Problem

Student’s name: Monali Gothi

Registration Number: 2010397

Professor’s name: Dr. Vito De Feo

Date: 21/07/2021

Pilot-Study Proposal

(Word Count: 517)

In the days before, organizations were reluctant to carry out AI calculations in their daily exercises, either as a result of the complexity of their present cycles are unpredictable and therefore the execution would be expensive and time-consuming, or they stick to the technique that was recently performed physically or in the light of human measurements. Organizations today are developing a method of using machine learning to predict outcomes. Here, the supermarket "Tosco & Spency" would like to separate its offer and marketing strategy based on two main classes of customers. There are two types of customers: the first who only buy an expensive product and the other only buy a cheap product. So, the task is to predict new customers belong to first-class or not. From the data, we can conclude that predicting the target is binary over a set of company features. Henceforth, we can say that this scenario is a classification problem from a machine learning perspective. We try to distinguish the key features that will help us to separate customers into two classes. Below are the customer's features:

- Income
- Gender
- Occupation
- Age
- Expense
- Nationality

Now that we know the main characteristics, we will apply the appropriate machine learning classification method to the two customers. We will separate the information into train and test utilizing the sci-kit learn function train_test_split. We have performed Support Vector Machine, Decision Tree, K – Nearest Neighbor, Naïve Bayes [1]. Naïve Bayes only for classification problem. It is quicker than the linear model however less precise than the linear model. It is better for huge datasets and high-dimensional information. The KNN classification method is the easiest machine learning algorithm [1]. For, new point the method identifies the closest point from the train data. We can assign an arbitrary number, k, as n_neighbor [1] parameter of the KNN function. Here, we use sci-kit learn, NumPy, and pandas for algorithms and other methods [1]. The decision tree a lot of complicated once the dataset is immense. The Decision Trees are incredible and broadly used in supervised learning. It works recursively and goes deep till pure leave. It takes time to training data. Support vector machine works better for medium size of the dataset and requires scaling of data, sensitive to parameter [1]. SVM and KNN provide better outcomes contrasted with different calculations. We see data and target that using importing seaborn. We implement a grid search for tracking down the best params of a model. Evaluation will be finished by various metrics. There are different metrics like precision, recall, f1-score, and accuracy. The confusion metric assists us with recognizing genuine positive, genuine negative, bogus positive, bogus negative of the model. F1 model – the score determined from the accuracy and recovery rate. Below are the formula for precision, recall :

Precision = TruePositives / (TruePositives + FalsePositives)

Recall = TruePositives / (TruePositives + FalseNegatives)

For example, one model has a real value that is true and the predicted value is also true then precision is 100% and recall 0%. The above metrics and scores help us to find a better model for implementation.

Reference

1. Müller, Andreas C., and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.