CTSE Assignment 2

by Vishwa Jayasekara

Submission date: 04-May-2023 07:42PM (UTC+0530)
Submission ID: 2084088127
File name: IT20217990.pdf (159.1K)
Word count: 1431
Character count: 7542

Question 1

a. Scenario:

I. Association Rule Mining (ARM) is an ideal machine learning model for organizing supermarket shelves based on customer transaction data. ARM identifies patterns and connections between frequently purchased items and establishes rules between them. These rules can then be used to decide the placement of items on the shelves. As an unsupervised learning technique, ARM can handle large datasets without the need for labeled data. By using ARM, the supermarket can gain insights into the relationships between items and create a more efficient and customer-friendly shopping experience.

II. Association Rule Mining (ARM) is an unsupervised machine learning technique that can identify connections between variables in large datasets. For the given scenario, where the dataset contains bill details of customers over the past three years, ARM can identify frequently purchased item sets and establish rules between them. This makes ARM an ideal model to decide the placement of items in the supermarket. The insights generated by ARM can be helpful to the supermarket in organizing the products on shelves, making shopping easier for the customers. Metrics can be used to evaluate the rules generated by ARM, and the most helpful rules can be used to group the items that are frequently purchased together. For example, if the ARM algorithm identifies that customers who buy bread and milk together also frequently purchase eggs, the supermarket can place bread, milk, and eggs near each other on the shelves.

III. Collaborative filtering is another machine learning strategy that can be used for organizing supermarket shelves. This technique uses customer preferences and purchase history to make product recommendations. By representing customers as rows and products as columns in a user-item matrix, collaborative filtering can identify similar customers and recommend products based on their purchase histories. This method does not require explicit features of the products or customers and can make unexpected suggestions for products that customers may not have previously considered.

b. Scenario:

I. Deep Reinforcement Learning (DRL) is a combination of deep neural networks and reinforcement learning that can be used for robotic navigation tasks. In the given scenario, DRL can be applied to a robot that is designed to navigate and collect samples on Mars. The robot's sensors gather data, which the DRL model can use to pinpoint the locations from which samples can be taken. The model can learn from its own experiences, allowing it to make predictions and decisions with greater accuracy over

time. DRL can also adapt to various circumstances and obstacles encountered while navigating.

II. Reinforcement Learning (RL) is a machine learning technique that excels at learning by doing. In the context of the Mars robot scenario, RL can be used to train the robot to maximize its performance in accordance with predefined goals, such as getting nearer to the target location and precisely extracting the samples. RL allows the robot to learn from its own experiences and adjust to changes in the environment, improving its performance over time. The robot can be trained in a simulated environment to reduce costs and time associated with real-world training. Transfer learning can also be used to speed up the learning process and improve the robot's effectiveness by using pre-trained models.

III. Convolutional Neural Networks (CNNs) are a supervised learning method that can be used for robotic navigation tasks. In the given scenario, the robot's cameras can be used to take pictures as it travels to various locations. The labeled images can be used to train a CNN to predict the best path for the robot to take while avoiding obstacles. This method simplifies the data collection process and allows for easier interpretation and comprehension of the model's decisions. Additionally, the robot is not required to explore the Martian environment, as the labeled data can be used for training.

Question 2

I.

- PassengerId: **Not Selected** , This column is a unique identifier for each passenger and does not provide any information relevant to the prediction of survival. Hence, it can be excluded from the final prediction.
- Survived: **Selected**, This is the target variable, and hence it must be included in the final prediction.
- Pclass: **Selected**, This column represents the socio-economic status of the passenger and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- Name: **Not Selected**, This column contains the names of the passengers, and it is unlikely that the name would have any bearing on their survival. Hence, it can be excluded from the final prediction.
- Sex: **Selected**, This column represents the gender of the passenger and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- Age: **Selected**, This column represents the age of the passenger and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- SibSp: **Selected**, This column represents the number of siblings/spouses the passenger had on board and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- Parch: **Selected**, This column represents the number of parents/children the passenger had on board and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- Ticket: **Not Selected**, This column contains the ticket numbers of the passengers and does not provide any information relevant to the prediction of survival. Hence, it can be excluded from the final prediction.
- Fare: **Selected**, This column represents the fare paid by the passenger and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.
- Cabin: **Not Selected**, This column contains the cabin numbers of the passengers, and it has a high percentage of missing values. Moreover, the information in this column is unlikely to have any bearing on the survival prediction. Hence, it can be excluded from the final prediction.

- Embarked: **Selected**, This column represents the port of embarkation of the passenger and can be an important factor in predicting their survival. Hence, it should be included in the final prediction.

II.

- PassengerId: This column does not contain any useful information for prediction and can be dropped.
- Survived: This is the target variable and does not require any pre-processing or feature engineering.
- Pclass: This column represents the passenger class and can be transformed into a categorical variable using one-hot encoding or dummy variable encoding. This is because there is no inherent order in the values of the passenger class.
- Name: This column contains the name of the passenger and can be used to extract useful information such as the passenger's title (Mr., Mrs., Miss, etc.). This can be done using regular expressions.
- Sex: This column represents the passenger's gender and can be transformed into a binary variable (0 or 1) to avoid any assumptions of order.
- Age: This column contains the age of the passenger and has missing values. One possible pre-processing technique is to fill in missing values with the median age. Age groups can also be created for feature engineering.
- SibSp and Parch: These columns represent the number of siblings/spouses and parents/children on board, respectively. These can be combined into a single variable representing family size.
- Ticket: This column contains the ticket number and does not contain any useful information for prediction. It can be dropped.
- Fare: This column contains the fare paid by the passenger and can have missing values imputed with the median fare. Fare categories can also be created to capture non-linear relationships with the target variable.
- Cabin: This column contains the cabin number and has many missing values. One possible pre-processing technique is to create a new variable indicating whether the passenger had a cabin or not.
- Embarked: This column represents the port of embarkation and has a few missing values. Missing values can be imputed with the most common port of embarkation. One-hot encoding

can also be used to represent different ports as binary values to learn the relationship with the survival rate.

Column	Selected/Not Selected	Preprocessing technique
PassengerId	Not selected	N/A
Survived	Selected	N/A
Pclass	Selected	Categorical/one-hot encoding
Name	Not selected	N/A
Sex	Selected	Binary encoding
Age	Selected	Impute missing values
SlbSp	Selected	Combine into one variable with Parch
Parch	Selected	Combine into one variable with SlbSp
Ticket	Not selected	N/A
Fare	Selected	Impute missing values
Cabin	Selected	Impute missing values
Embarked	Selected	Categorical/ one-hot encoding

CTSE Assignment 2

ORIGINALITY REPORT

2%	1%	1%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to De Montfort University Student Paper	1%
2	www.mdpi.com Internet Source	1%
3	Liang Zhao, Zhikui Chen, Zhennan Yang, Yueming Hu. "A Hybrid Method for Incomplete Data Imputation", 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, 2015 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		